UiO : University of Oslo

Håvard Tveit Ihle

Bayesian Data Analysis for Intensity Mapping and CMB Experiments

Thesis submitted for the degree of Philosophiae Doctor

Institute of Theoretical Astrophysics
Faculty of Mathematics and Natural Sciences

2021

© Håvard Tveit Ihle, 2021

Series of dissertations submitted to the
Faculty of Mathematics and Natural Sciences, University of Oslo

ISSN ISSN

All rights reserved. No part of this publication may be
reproduced or transmitted, in any form or by any means, without permission.

Cover: Hanne Baadsgaard Utigard.
Print production: Reprosentralen, University of Oslo.

To Ludvig, for being a good boy

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of *Philosophiae Doctor* at the University of Oslo. The research presented here was conducted at the University of Oslo and, under the supervision of Ingunn K. Wehus and Hans Kristian K. Eriksen. This work was supported by the Norwegian Research Council through grant 251328.

Most of my work has been on the COMAP data analysis pipeline, including calibration, low level data processing, systematics, simulation, signal validation, data selection and power spectrum methods. Much of this work is presented in Paper I (Foss et al., in prep) which describes the data analysis pipeline all the way from raw data to final maps.

For the last year or so I have also worked on developing power spectrum methods that have a high sensitivity, yet are robust to instrumental systematics. This work is presented in Paper II (Ihle et al., in prep).

Papers I and II are both meant to be part of a set of papers to be published together making up the first release of COMAP results. As they are both in a fairly complete state we present them here as drafts, but we note that both may receive significant modifications before final submission.

I have also worked more abstractly on inference from line intensity maps: How do you get as much information possible out of an observed map? Paper III (Ihle et al., 2019) demonstrates how, for a COMAP-like experiment, using a combination of the power spectrum and the voxel intensity distribution gives significantly stronger constraints on the CO luminosity function than using either observable separately. Lately my focus has been more on how to bridge the gap between forecasts like this, with a bunch of simplifying assumptions, and real world data. As I discuss in the thesis, this gap needs to be closed for the work we do on inference methods to be useful.

For the last two years or so I have also spent some of my time working on the BeyondPlanck project (Paper IV, BeyondPlanck Collaboration, 2020). My main focus starting out was on implementing the methods for sampling the correlated noise and the corresponding noise parameters. As the work evolved we encountered some interesting systematics which seem to involve both the noise estimation and calibration, which makes sense since they are so closely related. This work involved using housekeeping data from the Planck satellite to interpret and understand the results. Our main findings and more discussion can be found in Papers V and VI (Ihle et al., 2020, Gjerløw et al., 2020).

Acknowledgements

I want to thank my supervisors **Ingunn** and **Hans Kristian** for believing in me, for giving me the chance to develop as a researcher, and for all the help you have provided. Thanks for suggesting these projects for me, I have learned so much during these years.

Thank you to **Marie**, it's been a pleasure working with you on COMAP for the last four and a half years.

Thanks to my master's students, **Maren**, **Jonas**, **Jowita** and **Nils**, you are all brilliant, and it has been really fun supervising and working with you!

Thanks to everyone I have worked with in COMAP, **Kieran**, **Tim**, **David**, **James**, **Rick**, **Stuart**, **Bade**, **Junhan**, **Liju**, **Dongwoo**, **Marco**, **George**, **Patrick**, **Marta** and **Hamsa**, it has been so interesting to meet and work with great people like you from all over the world!

Thanks to all the people I have met here at ITA over these many years, both in the basement, at the first floor and at the institute in general. I really enjoy it at ITA and you people are what makes it what it is.

Thanks to **Frode**, for for giving me the chance to lecture the Cosmology II course. That was a really cool and inspiring experience, and I learned so much!

To **Caroline**, my love, you are kind, beautiful and wise! Thank you for putting up with me, and for all your love and support.

Thank you also to my family, **Mor**, **Far**, **Aslak**, **Maria** and **Bjarte**, for the unconditional love that you provide!

Håvard Tveit Ihle
Oslo, March 2021

List of Papers

Paper I

M. Foss, H. T. Ihle and the COMAP collaboration. “First Season COMAP Results: CO Data Processing” (draft).

Paper II

H. T. Ihle and the COMAP collaboration. “First Season COMAP Results: Power spectrum methodology and preliminary data quality assessment” (draft).

Paper III

H. T. Ihle, D. Chung, G. Stein and the COMAP collaboration. “Joint Power Spectrum and Voxel Intensity Distribution Forecast on the CO Luminosity Function with COMAP”. In: ApJ, Jan 2019, volume 871, no. 1 p. 75. DOI: 10.3847/1538-4357/aaf4bc.

Paper IV

The BeyondPlanck collaboration. “BeyondPlanck I. Global Bayesian analysis of the Planck Low Frequency Instrument data”. In: A&A, to be submitted, arXiv: 2011.05609.

Paper V

H. T. Ihle, M. Bersanelli, C. Franceschet, E. Gjerløw and the BeyondPlanck collaboration. “BeyondPlanck VI. Noise characterization and modelling”. In: A&A, to be submitted, arXiv: 2011.06650.

Paper VI

E. Gjerløw, H. T. Ihle, S. Galeotta and the BeyondPlanck collaboration. “BeyondPlanck VII. Bayesian estimation of gain and absolute calibration for CMB experiments”. In: A&A, to be submitted, arXiv: 2011.08082.

Contents

Preface iii

List of Papers v

Contents vii

1 Cosmology 1
- 1.1 Standard Cosmology . . . 1
- 1.2 Observational Cosmology . . . 5
- 1.3 Unresolved Questions and Future Prospects . . . 9

2 COMAP 11
- 2.1 COMAP Experiment . . . 11
- 2.2 COMAP Instrument . . . 11
- 2.3 Observations . . . 13
- 2.4 Data and Systematics . . . 15
- 2.5 Calibration . . . 28
- 2.6 Data Analysis Pipeline . . . 31
- 2.7 Data selection . . . 40
- 2.8 Inference from line intensity maps . . . 45

3 BeyondPlanck 55
- 3.1 The BeyondPlanck approach . . . 55
- 3.2 Noise Modelling and Characterization . . . 56
- 3.3 Calibration . . . 58
- 3.4 Future prospects . . . 61

Summary and Outlook 63

Bibliography 65

A Appendix 73
- A.1 Binomial likelihood and mean estimation . . . 73

Papers 76

I First Season COMAP Results: CO Data Processing (draft) 77

II First Season COMAP Results: Power spectrum methodology and preliminary data quality assessment (draft) 103

III Joint Power Spectrum and Voxel Intensity Distribution Forecast on the CO Luminosity Function with COMAP **119**

IV BeyondPlanck I. Global Bayesian analysis of the Planck Low Frequency Instrument data **133**

V BeyondPlanck VI. Noise characterization and modelling **213**

VI BeyondPlanck VII. Bayesian estimation of gain and absolute calibration for CMB experiments **241**

Chapter 1

Cosmology

Cosmology is the study of the universe as a whole. What are the physical properties of the universe? What is the history and future of the universe? What does the universe consist of?

Although we have had a fairly good qualitative understanding of some parts of these questions for about a century, it is really first in the last few decades that we have been able answer any part of these questions with any kind of certainty or precision.

This development has mostly been led by experimental advances in the study of the cosmic microwave background (CMB) and the large scale structure of the universe.

1.1 Standard Cosmology

Here we will discuss some of the main theoretical components underlying modern cosmology, including general relativity (GR), dark matter (DM), dark energy (DE) and inflation, culminating in the ΛCDM model.

1.1.1 General Relativity

GR is an incredibly successful theory of gravity describing everything from gravitational phenomena measured in laboratories here on earth to the expansion and evolution of the universe as a whole. The fundamental equation of GR is the *Einstein equation*

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}, \tag{1.1}$$

where $G_{\mu\nu}$ is the *Einstein tensor* and $T_{\mu\nu}$ is the *energy-momentum tensor*. The Einstein equation relates the curvature of the universe, represented by $G_{\mu\nu}$, to the matter and energy content of the universe, represented by $T_{\mu\nu}$. This means that any distribution of matter and energy will create curvature in the spacetime around it, much like any distribution of electric charge will create an electromagnetic field.

The ΛCDM model assumes that the spatial part of the universe is homogeneous, isotropic and flat, and that the energy content of the universe is given by radiation, baryonic matter, cold dark matter (CDM) and DE. In this case the Einstein equation can be reduced to

$$H^2(t) = H_0^2\left(\frac{\Omega_\mathrm{r}}{a^4} + \frac{\Omega_\mathrm{b} + \Omega_\mathrm{c}}{a^3} + \Omega_\Lambda\right), \tag{1.2}$$

where $a(t)$ is the scale factor, which keeps track of the size of the universe at any given time, $H(t) \equiv \frac{\dot{a}(t)}{a(t)}$ is the Hubble parameter denoting the expansion rate of the universe, $\dot{a}(t)$ denotes the time derivative of $a(t)$, H_0 is the current value of the Hubble rate, often called the *Hubble constant* and $\Omega_{\mathrm{r,b,c},\Lambda}$ denotes the fractional energy density today in radiation, baryonic matter, CDM and DE respectively.

1.1.2 Dark Matter

Dark matter (DM) refers to a matter component additional to the visible matter we can see and interact with. The term was first used by Fritz Zwicky (Zwicky, 1933) in 1933 when he found larger velocity dispersion in the Coma cluster than could be explained by the visible matter, indicating another "dark" matter component.

Today DM is a cornerstone of cosmology with multiple independent lines of evidence in its favor, including rotation curves of spinning galaxies (Borriello and Salucci, 2001), velocity dispersion in galaxy clusters (Saro et al., 2013), weak (van Uitert et al., 2012) and strong (Moustakas and Metcalf, 2003) gravitational lensing, the CMB and non-linear structure formation (Planck Collaboration VI, 2020).

These lines of evidence, as well as null results in direct- and indirect detection experiments, put strong constraints on the properties of dark matter:

- It must be *dark* (i.e. non-luminous): In practice this means no coupling (or extremely weak) to the photon. In general DM can not have any large coupling to any light standard model particle.
- It must be non-interacting: The self-interaction of dark matter is severely constrained, especially on large scales.
- It must be cold: DM has to be non-relativistic during structure formation.
- It must be stable: It must have a lifetime significantly longer than the age of the universe.

It is remarkable that simply positing a new heavy non-SM particle with small or no interactions with the visible sector can explain such a breadth of independent observations.

It is interesting to note, however, that in all the stated cases, DM is inferred from its gravitational effects. This has led some to suggest a modification to gravity to explain these phenomena (see e.g. Milgrom, 2010). It has, however, not been possible to devise a modification of gravity that can explain more than a few of the above lines of evidence for DM at a time.

1.1.3 Dark Energy

In 1998 supernova measurements suggested, for the first time, that the expansion of the universe was accelerating (Riess et al., 1998). This acceleration cannot be achieved by ordinary matter, but requires a cosmological component with positive energy and negative pressure. This new component is what is called *Dark Energy*. Since then multiple other lines of evidence have confirmed this accelerated expansion (Huterer and Shafer, 2018) and we now know that DE is actually the largest contribution to the energy budget of the universe at around 70%.

We do not know the nature of DE, but the simplest explanation is that empty space has a certain energy associated with it, a *vacuum energy*. This constant energy density has exactly the properties we need to explain the accelerated expansion we observe. However, the big open question is: Why does the vacuum energy take this

specific value, and not another? Or, more specifically: Why is the energy density of the vacuum of comparable magnitude as the energy density of matter? This is called the *Coincidence Problem*, and we do not have any satisfactory answers to it. Perhaps even more disconcertingly, if we try to estimate the contributions to the vacuum energy from quantum field theory we naively expect contributions of the order of the Planck scale, which is about 120 orders of magnitude off from the value we measure in cosmology. Clearly there is a significant piece missing in our understanding of DE, and it remains one of the major challenges for the future.

1.1.4 Inflation

Inflation is the theory that the universe went through a period of accelerated expansion in the very early universe. During inflation, the universe expanded by at least a factor of 10^{26}. This accelerated expansion solves two problems (and other related problems) with the traditional big bang model, *the flatness problem* and *the horizon problem* (Guth, 1981).

If the universe is dominated by matter or radiation, which it was during most of its history, any initial spatial curvature would tend to increase as the universe expands. This means that since we observe the universe to be very close to flat today, it must have been extremely close to flat in the very early universe, which would require a large degree of fine-tuning, unless we have some mechanism for making space flat to begin with. This is called the flatness problem, and inflation solves it since during inflation, the universe becomes more and more flat as the expansion goes on.

The horizon problem can be understood by observing that the CMB photons reaching us from opposite points of the sky have the same temperature, even though these points in space have never, in the traditional big bang model, been in causal contact and so have no reason to be at the same temperature. This is called the horizon problem, and inflation solves this by ensuring that regions that were in causal contact before inflation got spread out to super-horizon scales. The effect of this is that even regions which (if we did not know about inflation) we would think never had been in causal contact, actually could have had time to reach thermal equilibrium.

The simplest model of inflation posits a single scalar field, $\phi(\boldsymbol{x}, t)$, rolling slowly down towards the minimum of its potential. In such a *slow-roll* scenario, the energy of the scalar field is dominated by the potential term that acts almost like a vacuum energy and leads to exponential expansion.

Another crucial effect of the accelerated expansion is that small scale quantum fluctuations in the scalar field get blown up to large scales and then frozen in as these scales move outside the cosmological horizon. After inflation, when these modes eventually move into the horizon again, these initial perturbations act as seeds for the subsequent growth of structure in the universe.

Single field slow-roll inflation makes three predictions for the generated perturbations, they will be:

- Gaussian: Since the quantum state of the perturbation field is basically in the vacuum state of an harmonic oscillator.

- (Nearly) Scale invariant: Since the conditions under which the fluctuations were generated changes only very slowly, the properties of the generated perturbations will only depend very weakly on scale.
- Adiabatic: Since the perturbation in all the different cosmological components are generated from the same scalar perturbations, they will be in phase with each other.

Observations of the CMB confirms all these predictions, which is why inflation is generally accepted, even though we do not have very direct evidence for it. On the other hand, one can argue that in each of these three cases, if you had asked the question "What is the simplest thing you would expect?" the answer would be adiabatic Gaussian fields with a scale invariant spectrum. In this sense we can say that these are very generic predictions, and thus less impressive. Nevertheless inflation is still the simplest theory on the market to explain these observations (as well as solving the horizon problem etc.).

Another important prediction of inflation is the presence of tensor perturbations in addition to the scalar perturbations we have been discussing. Tensor perturbations are primordial gravitational waves induced by the inflation field. Tensor modes are harder to detect than scalar modes, but they can be observed because they induce so-called B-modes in the polarization spectrum of the CMB.

The scalar and tensor power spectra are conventionally defined as follows (Baumann, 2018)

$$P_{\mathcal{R}}(k) = \frac{k^3}{2\pi^2} A_{\mathrm{s}} \left(\frac{k}{k_*}\right)^{n_{\mathrm{s}}-1}, \tag{1.3}$$

$$P_h(k) = \frac{k^3}{2\pi^2} A_{\mathrm{t}} \left(\frac{k}{k_*}\right)^{n_{\mathrm{t}}}, \tag{1.4}$$

where $\mathcal{R}$ is the scalar curvature perturbation, h is the tensor perturbation, A_{s} and A_{t} are the scalar and tensor perturbation amplitudes, respectively. n_{s} and n_{t} are the scalar and tensor spectral indices, respectively. The amplitudes are conventionally defined at $k_* = 0.05\ \mathrm{Mpc}^{-1}$ (Planck Collaboration VI, 2020).

We then define the tensor-to-scalar ratio, r, as

$$r \equiv \frac{A_{\mathrm{t}}}{A_{\mathrm{s}}}. \tag{1.5}$$

Inflation does not predict the overall level of the tensor perturbations, but it predicts the following relation $n_{\mathrm{t}} = -r/8$. We have not observed any cosmological tensor modes, and the current observational constraints on the tensor-to-scalar ratio is given by $r < 0.044$ (95 % CL) (Tristram, M. et al., 2021). Therefore, since we already know that r is fairly small, this means that in practice, since it will be very hard to measure, inflation predicts $n_{\mathrm{t}} \approx 0$.

Measuring the tensor-to-scalar ratio, r, is one of the main goals of future CMB experiments. This would not only be more evidence for inflation (especially if we could confirm that the tensor spectral index is consistent with inflation), but it would also tell us (at least in the simplest inflation models) about the energy scale of inflation, and about the shape of the inflationary potential.

1.1.5 The ΛCDM Model

The ΛCDM model combines the theories of GR and inflation[1] with DE and (Cold) Dark Matter (CDM), as well as known particle physics and thermodynamics, into a single 6 parameter model for explaining the expansion and evolution of the universe, in addition to the origin and growth of structures in the universe.

The model assumes that the universe is spatially flat, that the energy density of dark energy is constant over time and that there are no primordial tensor perturbations. Table 1.1 shows the constraints on the six parameter model from the latest Planck release. We see that all these parameters, except for the optical depth to reionization, τ, are constrained to more than one percent accuracy, which is quite remarkable.

Table 1.1: Constraints on the 6 ΛCDM parameters based on Planck data, in addition to lensing and baryon acoustic oscillation data (Planck Collaboration VI, 2020). Here $h \equiv H_0 \,/\, 100\ \mathrm{km\ s^{-1}\ Mpc^{-1}}$, θ_* is the angular acoustic scale and τ is the optical depth to reionization.

Parameter	68% limits
$\Omega_\mathrm{b}h^2$	0.02242 ± 0.00014
$\Omega_\mathrm{c}h^2$	0.11933 ± 0.00091
$100\theta_*$	1.04119 ± 0.00029
τ	0.0561 ± 0.0071
$\ln\left(10^{10}A_\mathrm{s}\right)$	3.047 ± 0.014
n_s	0.9665 ± 0.0038

1.2 Observational Cosmology

Observational cosmology is a wide ranging field with many different sources of data, including galaxy surveys, observations of supernovae, gravitational lensing, Lyman-α forest, CMB, line intensity mapping (LIM) and gravitational waves. Here we will focus on CMB and LIM, as those are the topics of this thesis.

1.2.1 CMB observations

The cosmic microwave background (CMB) is the radiation left over from the hot big bang. It essentially captures a snapshot of the distribution of structure in the very early universe, only about 400 thousand years after the big bang. This radiation has been a treasure trove of information about our universe and has brought us into the era of precision cosmology.

Planck was the fourth satellite mission to study the CMB, following the *RELIKT* (Klypin et al., 1987), *COBE* (Mather et al., 1994) and *WMAP* (Bennett et al., 2013) satellites. Planck produced the ultimate map of the temperature anisotropies of the CMB,

[1] The ΛCDM does not explicitly include inflation, but assumes that the initial power spectrum is of the form given in Eq. 1.3 and that $r = 0$.

limited only by cosmic variance and fundamental physical processes. This has made Planck one of the most important datasets for constraining cosmological parameters, further cementing the ΛCDM model as the standard model for cosmology. With the temperature map all but settled by Planck, the future of CMB lies in the polarization signal.

The success of the CMB field comes from the combination of precise observations and precise theoretical predictions. The early universe was incredibly homogeneous, allowing the use of linear perturbation theory to study the growth of perturbations, which means that the coupled Einstein-Boltzmann equations can be solved very precisely (see Dodelson, 2003 for more details). Figure 1.1 shows the close agreement between theoretical predictions and CMB observations.

1.2.2 Line intensity mapping

Line intensity mapping (Madau, Meiksin, and Rees, 1997) (LIM) aims to map out the 3D structure of the early universe in large volumes using various different spectral lines, like 21 cm, Lyα, CII, or CO. Compared to galaxy surveys, these surveys do not rely on resolving and detecting individual sources of radiation, but instead collect the aggregate emission from the full population of sources from each cross-sectional area of the sky (see Fig. 1.2). By extracting the information present in the 3D maps LIM aims to help us understand both galaxy evolution, reionization and fundamental cosmology.

Intensity mapping is now a growing field with multiple efforts both experimental and theoretical. For a nice review of the motivation for LIM see (E. D. Kovetz et al., 2017; E. Kovetz et al., 2019). The number of running and planned LIM experiments is growing, including SKA (Koopmans et al., 2015; Square Kilometre Array Cosmology Science Working Group et al., 2020), GBT (Masui et al., 2013), BINGO (Battye et al., 2013), HIRAX (Newburgh et al., 2016), LOFAR (van Haarlem et al., 2013), CHIME (Bandura et al., 2014), HERA (DeBoer et al., 2017), GMRT (Pen et al., 2009), MWA (Tingay et al., 2013), PAPER (Parsons et al., 2014) and MeerKAT (Santos et al., 2017) targeting 21 cm, COMAP (Cleary et al., 2016), COPSS (Keating, Bower, et al., 2015; Keating, Marrone, Bower, Leitch, et al., 2016), mmIME (Keating, Marrone, Bower, and Keenan, 2020) and AIM-CO (Bower et al., 2016) targeting CO, CCART-prime (Stacey et al., 2018), TIME (Crites et al., 2014) and CONCERTO (Lagache, Cousin, and Chatzikos, 2018) targeting CII and CDIM Cooray et al., 2016, SPHEREx (Doré et al., 2014), Origins (The OST mission concept study team, 2018) and HETDEX (Hill et al., 2008) targeting other lines.

A simple analytic model for the signal power spectrum, $P(k)$, of a line intensity map is given by

$$P(k) = \langle T_{\text{line}} \rangle^2 b_{\text{line}}^2 P_m(k) + P_{\text{shot}}, \tag{1.6}$$

where $\langle T_{\text{line}} \rangle$ is the mean brightness temperature of the line, b_{line} is the luminosity weighted bias of the line, $P_m(k)$ is the matter power spectrum and P_{shot} is the scale invariant shot noise power spectrum. Together the three factors of the first term are referred to as the *clustering* power spectrum as it follows the underlying clustering of the matter distribution. The clustering term dominates at large scales, while the shot noise term dominates at small scales (see Fig. 1.3).

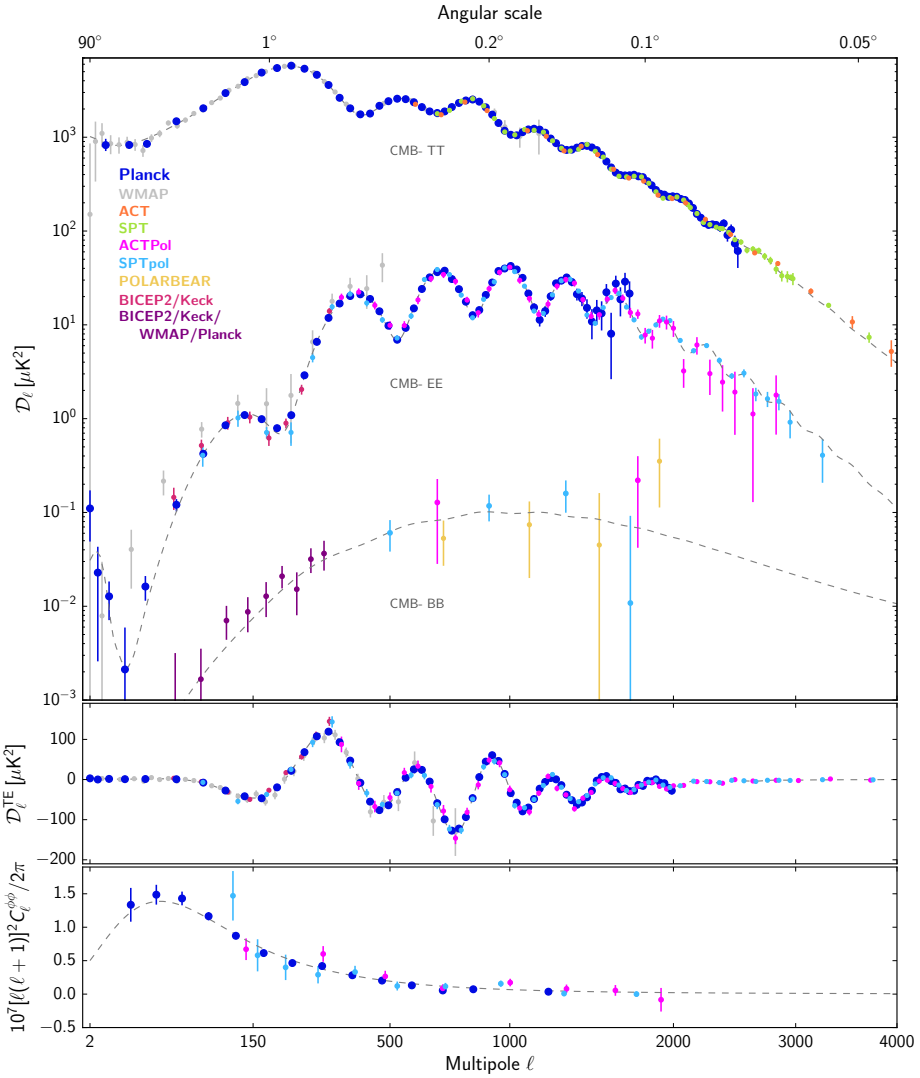

Figure 1.1: Comparison of CMB power spectrum measurements and theoretical predictions from the ΛCDM model. Top panel: Temperature (TT), E-mode (EE) and B-mode (BB) polarization auto spectra measured by several different CMB experiments. Middle panel: Cross-spectrum between temperature and E-mode polarization maps. Bottom panel: Lensing deflection power spectrum. Note that the B-mode spectrum measured here is consistent with that induced by lensing, and shows no sign of cosmological tensor modes. For more details see Planck Collaboration et al., 2020, from where this figure was taken.

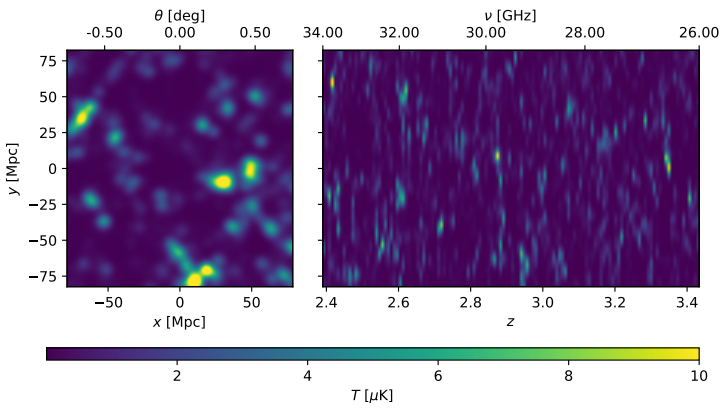

Figure 1.2: Simulated 3D cube of CO line emission brightness temperature. This corresponds roughly to the survey volume of one of the COMAP fields.

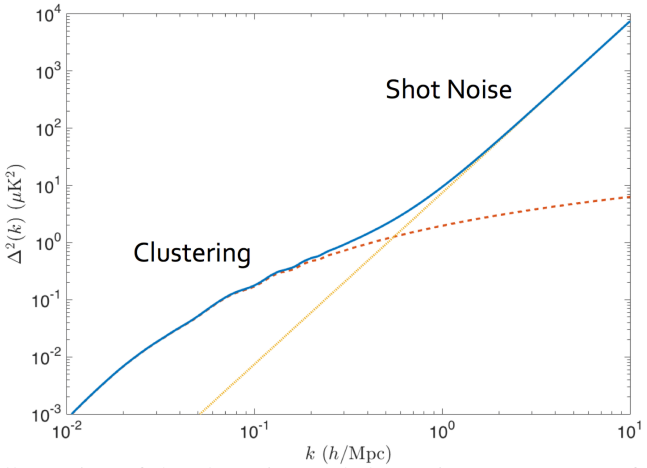

Figure 1.3: Illustration of the clustering and shot noise components of the intensity mapping power spectrum. Figure taken from E. D. Kovetz et al. (2017).

By tracing spectral lines originating in star forming regions, the interstellar medium and the intergalactic medium, LIM can map out galaxy formation and evolution over cosmic time, and, crucially, follow the process of reionization by tracing both the sources of ionizing radiation, and the neutral hydrogen gas being gradually ionized. These data will, among other things, help constrain the optical depth to reionization, τ, which is the least constrained parameter in the ΛCDM model.

In addition to these unique insights into astrophysics, LIM also provides a great opportunity to test fundamental cosmology. As we can see from Eq. 1.6, LIM traces the underlying matter distribution, and as such, it gives us access to new cosmological volumes largely inaccessible to other probes. Where CMB observations are limited to the number of 2D angular modes from the surface of last scattering, LIM can give us access to the full 3D modes up to high redshifts.

This could allow us, as an example, to follow the baryon acoustic oscillation (BAO) scale through cosmic time, which would give us a robust, model independent, measurement of the cosmic expansion history (Bernal, P. C. Breysse, and E. D. Kovetz, 2019). This would be a powerful tool to test things like modified gravity and dark energy properties. We could also get much tighter constraints on primordial non-Gaussianity (e.g. Furlanetto et al., 2019), which could help us distinguish different models of inflation. Even dark matter decay or annihilation could be detected using LIM (Creque-Sarbinowski and Kamionkowski, 2018). These are just some of the use cases of LIM in cosmology currently being discussed. As the field develops, many more will surely emerge.

1.3 Unresolved Questions and Future Prospects

The standard model of particle physics has been incredibly successful, explaining all particle interactions that we have observed, with the detection of the Higgs boson at the large hadron collider (LHC) in 2012 putting the last puzzle piece into place (Chatrchyan et al., 2012; Aad et al., 2012).

Motivated by the WIMP miracle and the hierarchy problem (Jungman, Kamionkowski, and Griest, 1996; Susskind, 1979), there was a significant expectation that the LHC would detect supersymmetry (SUSY), or other new beyond standard model (BSM) physics, at the weak scale. No such signs have yet been seen. While low energy SUSY is still not ruled out, the fact that we have not seen any signs of it at least lets us consider the possibility that we will not make much progress towards BSM physics using particle colliders in the near future.

From cosmology, however, we know that 95 % of the energy content of the universe (DE and DM) comes from sources that are not part of the standard model of particle physics. Inflation is also an important process in cosmology that cannot be explained within the standard model.

This suggests that at this point cosmology might be the most likely way to move fundamental physics forwards. What is the nature and properties of DM and DE? Are either of these explained by modifications of gravity? Did inflation happen? If so, what is the source of inflation? If not, then what explains the flatness and horizon

problem? And what gives us the spectrum of initial perturbations? How does any of this fit together in a coherent picture with the standard model?

With new probes like line intensity mapping coming online, in addition to well established ones like CMB analysis and galaxy surveys, observational cosmology and astrophysics is now in a position to shed light on these questions. Detecting primordial gravitational waves, non-Gaussianities or large scale anisotropies would shed light on inflation or any other source of primordial perturbations. Mapping out the whole expansion history of the universe, and the growth rate of structure, could help us nail down the properties of DE, possibly modified gravity, or even new early universe physics. Detecting the small scale structure of galaxies can tell us about the properties and interactions of DM. All this makes observational cosmology an exciting area to work in.

Chapter 2

COMAP

2.1 COMAP Experiment

The CO molecule is the second most abundant molecule in the interstellar medium, after H_2, and is a great tracer of cold molecular clouds, and thus star formation. CO has a ladder of equally spaced rotational transitions at integer multiples 115.27 GHz, which opens the possibility of studying different transitions of the same CO sources by observing at different frequencies.

The CO Mapping Array Pathfinder (COMAP, Cleary et al., 2016) is a line intensity mapping experiment targeting the rotational transitions of the CO molecule. It is an international collaboration between California Institute of Technology, Canadian Institute of Theoretical Astrophysics, Jet Propulsion Laboratory (NASA), University of Manchester, Maryland University, Miami University, Stanford University, University of California Berkeley, Princeton University and University of Oslo. The Oslo team is responsible for the end-to-end data analysis of the main science observations from COMAP.

Phase One of COMAP consists of a single telescope observing in the K_a-band at 26-34 GHz, which for CO $1 \rightarrow 0$, emitted at 115 GHz, corresponds roughly to the redshift range 2.4 – 3.4, the Epoch of Galaxy Assembly. The K_a-band will also pick up a, presumably weaker, signal from the CO $2 \rightarrow 1$ transition, emitted at 230 GHz, or the redshift range 6 – 8, during the Epoch of Reionization. See Fig. 2.1 for a group photo of the COMAP collaboration in front of the COMAP telescope.

COMAP Phase One is a pathfinder experiment meant to demonstrate the feasibility of large scale line intensity mapping using CO. Future planned phases of COMAP will involve more telescopes observing at K_a, as well as the addition of one or several telescopes observing the K_u-band. The K_u-band, at 12-20 GHz, would pick up the $1 \rightarrow 0$ emission from the same redshifts as the $2 \rightarrow 1$ emission from the K_a-band. This will allow us to separate out the $2 \rightarrow 1$ from the $1 \rightarrow 0$ signal in the K_a data. The long term goal is to do CO line intensity mapping from space (Bowman et al., 2012; M. B. Silva et al., 2019).

We started the main science observations for Phase One in June of 2019, and are currently analyzing the data taken during the first year of observations.

2.2 COMAP Instrument

For COMAP Phase One we utilize a 10.4 m telescope located at the Owens Valley Radio Observatory (OVRO) in California. The telescope was part of an array of dishes previously used for the Combined Array for Research in Millimeter-wave Astronomy (CARMA) experiment, but has now been moved back to OVRO to be used for other experiments. The angular resolution obtained is roughly 4.5 arcmin FWHM at 30 GHz.

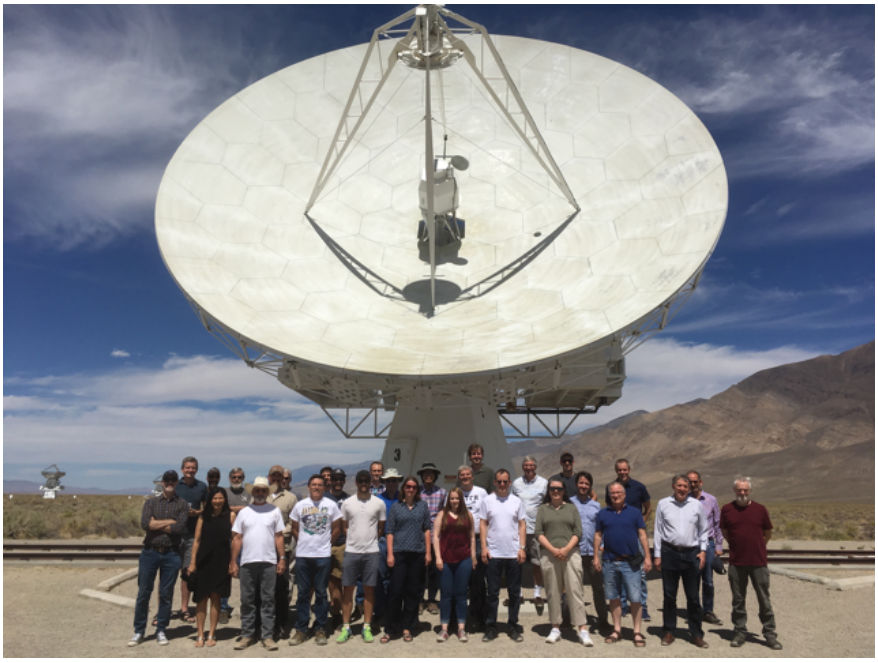

Figure 2.1: The COMAP collaboration standing in front of the COMAP Phase One telescope in August 2018.

The receiver consists of 19 independent detectors, each with a separate signal path from the 19 individual feed horns, through polarizers and low noise amplifiers to downconversion and eventually frequency separation and digitization which happens in two CASPER "Roach2" FPGA-based spectrometers for each signal chain (Cleary, 2015). We usually use the term "feed" to refer to each of these 19 signal chains. For more details on the COMAP instrument see J. Lamb et al. (in prep).

The observed signal in the frequency range 26-34 GHz gets, during the downconversion, split into two different bands, band A coming from 26-30 GHz and band B from 30-34 GHz. One Roach2 module processes the signal from one band, meaning that a total of 38 Roach2 modules are used to collect the full science data from the 19 feeds. In the Roach2 modules the data gets further split into a lower and an upper sideband, each with a bandwidth of 2 GHz.

For the raw science data, the Roach2 modules separate the 2 GHz bandwidth of each sideband into 1024 separate frequency channels, leading to a native frequency resolution of about 2 MHz. The data is also integrated over a period of about 0.02 s, leading to a sample rate of roughly 50 Hz in time.

2.3 Observations

During observations we employ two different scanning strategies, Constant Elevation Scans (CES) and Lissajous scans.

For the CES observations we keep the telescope at a constant elevation while slewing it back and forth in azimuth. During a Lissajous scan the telescope performs independent sinusoidal motion in both elevation and azimuth at the same time as follows

$$\text{Azimuth}(t) = A\sin(at + \phi) \tag{2.1}$$

$$\text{Elevation}(t) = B\sin(bt). \tag{2.2}$$

A Lissajous scan can thus create many different scanning patterns depending on which values used for A, B, a, b and ϕ. We typically use values of $B \approx 0.6$ degrees, while A is scaled with elevation to correspond to roughly the same on-sky distance. The other parameters are randomized in order to not repeat the same scanning pattern each time. Typical values for a and b are in the range 0.3–0.5 Hz, leading to typical periods of 12–25 seconds.

Circular scans, which are sometimes used, are just a special case of the Lissajous scan. In fact, even for the CES observations, we use a sinusoidal motion in the azimutal direction in order to minimize the strain on the telescope, so the CES can also be thought of as a special case of the Lissajous scan (with $a = 0$).

Lissajous scans typically give a more even coverage of the observed field, with better cross-linking, but they have the disadvantage of getting a large signal from the different amount of atmosphere at the different elevations, which needs to be removed. Ground signal is also harder to remove from Lissajous scans, since the scanning pattern is not repeating, in the same way as it is for a CES. A Lissajous scan typically covers a larger area than a CES, so it can recover more of the large scale cosmological signal, at the cost of a higher noise level on small scales.

For the first year of observations, we split the observation time roughly evenly between Lissajous and CES. However, due to the problems we have found in the Lissajous data (see Paper II, Ihle et al., in prep), the observations taken afterwards have been exclusively CES, and unless we can figure out and solve the issues with the Lissajous data, we will continue to rely exclusively on CES observations for the future.

The general observation strategy is to start by pointing the telescope a bit ahead of the on-sky field that you want to observe (i.e. point to where the centre of the field will be in a few minutes), and then performing CES or Lissajous scans centered on this azimuth and elevation position while the field drifts through. Once the field has drifted past the azimuth and elevation you are observing at, you repeat the sequence by moving slightly ahead of the field again etc. Each such period where you stay centered at at a given azimuth and elevation performing CES or Lissajous motion while the field drifts through is called a "scan". When you move the telescope ahead of the field again and repeat the sequence you start a new scan. Figure 2.2 shows the telescope pointing for 27 subsequent scans. This illustrates the general observation strategy, and how the telescope follows the field across the sky. We can also see the randomization of the Lissajous parameters clearly.

A scan typically lasts for 3-10 minutes, and multiple scans of the same field are combined together into an observation. An observation usually contains around an hour

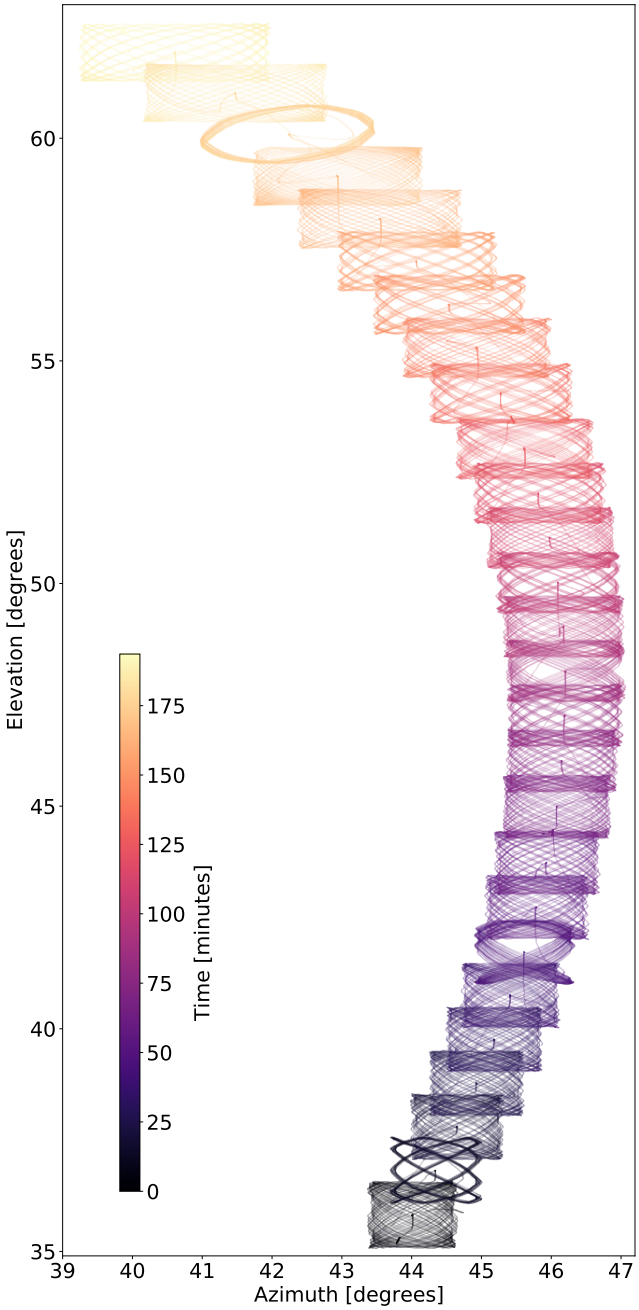

Figure 2.2: Telescope pointing during 27 subsequent Lissajous scans of the same field. Figure courtesy of Jonas Lunde.

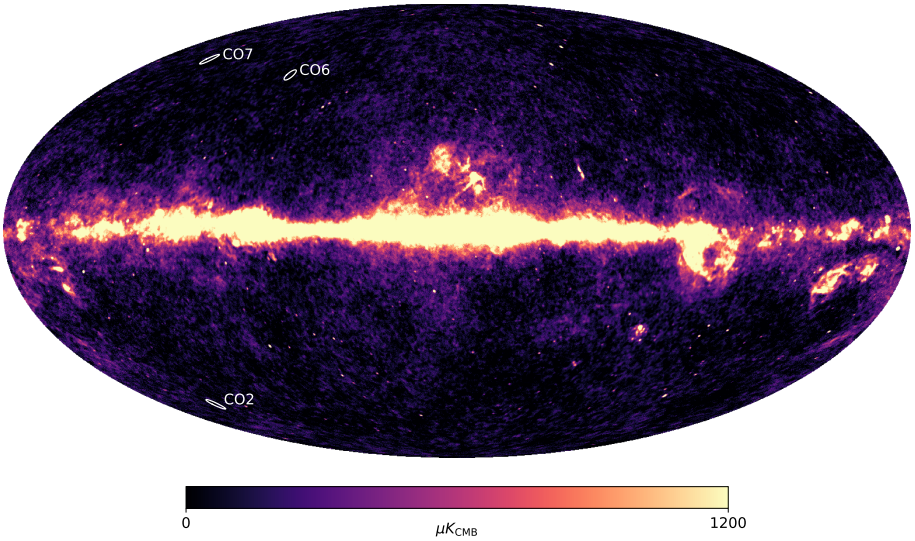

Figure 2.3: The three main CO fields, overplotted on the Planck LFI 30 GHz map. Map downloaded from the PLA, https://pla.esac.esa.int (Planck Collaboration I, 2020). Figure courtesy of Nils O. Stutzer.

Field Name	RA (J2000)	Dec (J2000)
CO2	01:41:44.4	+00:00:00.0
CO6	15:04:00.0	+55:00:00.0
CO7	11:20:00.0	+52:30:00.0

Table 2.1: COMAP main science fields

of data, and is labeled by a, monotonically increasing, integer number called the "obsid" (we refer to the observations as e.g. obsid 7456 etc.). The raw data is collected together into what we call "level 1" files, which contain all the data, from all 19 feeds, for a single obsid, together with pointing information and a large set of housekeeping data.

We observe three main science fields, labeled "co2", "co6" and "co7". These are chosen to be in low foreground regions, and to overlap with other surveys. The position of these fields can be seen in Fig. 2.3 and in Table 2.1. The sizes of the fields are roughly 2 x 2 deg^2.

2.4 Data and Systematics

The basic properties of the raw data are discussed in Paper I (Foss et al., in prep). Here we give a short summary and focus on some of the main systematics present in the data and what we know about them.

2.4.1 Data model

The power picked up by any of the detectors is given by

$$P_{\text{out}} = k_{\text{B}} G \Delta\nu T_{\text{sys}}, \tag{2.3}$$

where G is the instrumental gain, k_{B} is the Boltzmann constant, $\Delta\nu$ is the frequency bandwidth, and T_{sys} is the total system temperature of the instrument.

The system temperature is basically a measure of the overall noise level, and it has many different sources,

$$\begin{aligned} T_{\text{sys}} = {} & T_{\text{receiver}} + T_{\text{atmosphere}} + T_{\text{ground}} \\ & + T_{\text{CMB}} + T_{\text{foregrounds}} + T_{\text{CO}}, \end{aligned} \tag{2.4}$$

where T_{receiver} is the effective noise temperature of the receiver, $T_{\text{atmosphere}}$ is the temperature contribution from the atmosphere, T_{ground} is the ground pickup from far sidelobes, T_{CMB} is the signal from the CMB, $T_{\text{foregrounds}}$ are all continuum foregrounds (mostly from the galaxy), and T_{CO} is the line emission signal from extragalactic CO, which is what we are ultimately trying to measure.

Making some simple assumptions we can make the following model for the system temperature as a function of feed, frequency and time,

$$T^i_{\text{sys},\nu}(t) = \langle T^i_{\text{sys},\nu} \rangle \left[1 + P^i_{\text{cel}} (\Delta s_{\text{cont}} + \Delta s^\nu_{\text{CO}}) + P^i_{\text{tel}} \Delta s_{\text{ground}} + n_{\text{corr}} + n^{\nu i}_{\text{w}} \right]. \tag{2.5}$$

Here $\langle\rangle$ denotes average in time, P^i_{cel} and P^i_{tel} are the pointing matrices in celestial and telescope coordinate systems, respectively; Δs_{cont} denotes the mean subtracted celestial continuum sources, like the CMB or galactic foregrounds; Δs^ν_{CO} is the mean subtracted CO line emission; Δs_{ground} is the mean subtracted ground signal picked up by the far sidelobes; and n_{corr} is the (zero mean) correlated noise component, consisting mostly of atmosphere fluctuations and standing waves. The superscript i is a feed index. Terms with no feed index are assumed to be similar (or at least strongly correlated) between different feeds, while terms with a ν label indicate parts of the model that are assumed to have non-smooth frequency dependence.

Likewise we can make the following simple model for the gain

$$G^i_\nu(t) = \langle G^i_\nu \rangle \left[1 + \delta^i_G(t) \right], \tag{2.6}$$

where $\delta^i_G(t)$ is a correlated noise term coming from gain fluctuations in the low noise amplifiers. These gain fluctuations are assumed to be the same for all frequencies of a single feed, but are independent between different feeds.

We can then combine these into a single model for the power measured by the detectors

$$d^i_\nu(t) \approx \langle d^i_\nu \rangle \left[1 + P^i_{\text{cel}} (\Delta s_{\text{cont}} + \Delta s^\nu_{\text{CO}}) + P^i_{\text{tel}} \Delta s_{\text{ground}} + \delta^i_G + n_{\text{corr}} + n^{\nu i}_{\text{w}} \right], \tag{2.7}$$

where $d^i_\nu(t)$ is the raw time ordered data and we have assumed that the deviations from the mean are small, so that we can neglect second order terms.

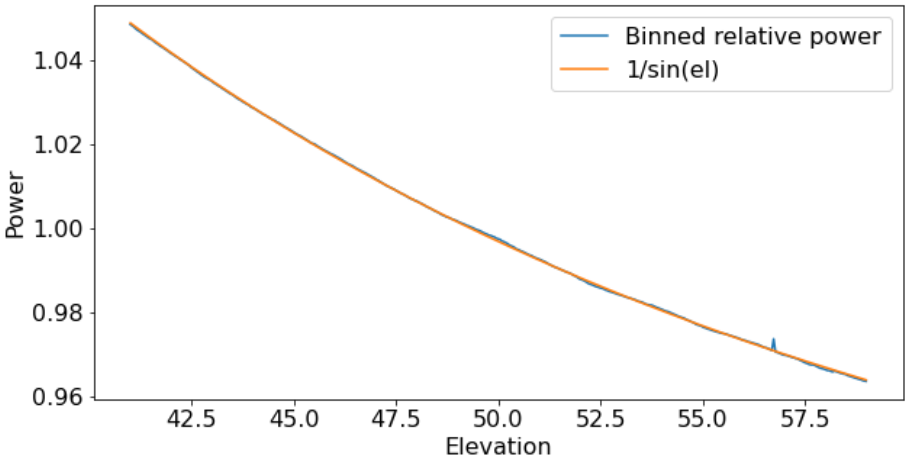

Figure 2.4: Normalized power measured during sky-dip observations compared to a fitted 1/ sin(el) model. Figure courtesy of Jonas Lunde.

2.4.2 Atmosphere

The atmosphere is one of the major contributors to the system temperature as well as to the correlated noise. The brightness temperature of the atmosphere can be roughly modeled as that of an optically thin medium at a certain temperature. The observed brightness temperature from the atmosphere, $T^{\mathrm{b}}_{\mathrm{atm}}$, is then given by

$$T^{\mathrm{b}}_{\mathrm{atm}} = T_{\mathrm{atm}}(1 - e^{-\tau}) \approx T_{\mathrm{atm}}\tau, \tag{2.8}$$

where T_{atm} is the physical temperature of the atmosphere and τ is the optical thickness of the atmosphere.

The optical thickness of the atmosphere depends on the elevation at which we are observing. We can take this elevation dependence into account using a simple model

$$\tau(\mathrm{el}) = \frac{\tau_0}{\sin(\mathrm{el})}, \tag{2.9}$$

where "el" denotes the elevation of the telescope pointing, and τ_0 is the optical thickness at zenith (el = 90 degrees). Figure 2.4 shows data from sky-dip observations (where the telescope quickly sweeps down in elevation and up again) taken with the COMAP telescope and a fitted optical depth model. This is a good illustration that the model works well.

Typical measured values of $T_{\mathrm{atm}}\tau_0$ are around $\sim$ 12 – 15 K. Since we typically observe at elevations between 35 and 65 degrees, the atmosphere typically contributes around 15 – 25 K to the system temperature.

The atmosphere also has a significant contribution to the correlated noise. To a large extent, the atmospheric column observed by the different feeds should be the same, so the correlated noise from the atmosphere should be highly correlated in the different

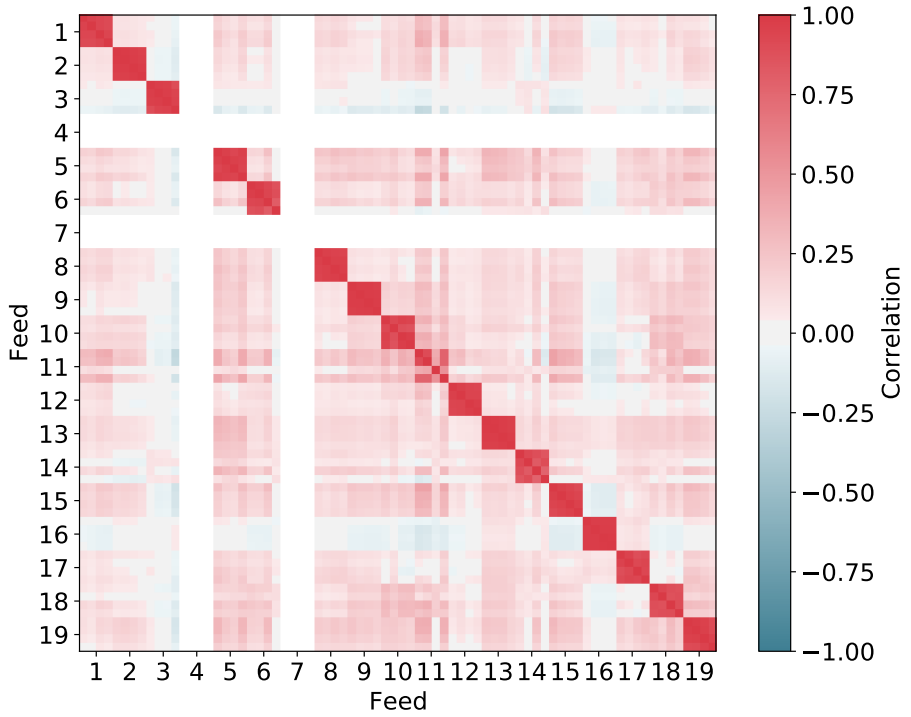

Figure 2.5: Correlation matrix of sideband averaged data from a single CES scan.

feeds, it should also be very smooth in frequency. If this is true, and there are no other major continuum contributions to the correlated noise that are also correlated between the different feeds, then the amount of correlation between the data of different feeds can be used to estimate the the rough magnitude of the atmospheric contribution to the correlated noise.

Figure 2.5 shows the correlation between the raw data averaged over all the frequencies of each sideband. We see that the four sidebands of the same feed are essentially completely correlated, this is because both the gain fluctuations and the atmosphere fluctuations are common to all four sidebands. For the different feeds, however, the gain fluctuations are independent, so the main contribution to the correlation here is from the atmosphere. This suggests that the correlations due to the atmospheric contribution are typically around 10 – 40 %.

2.4.3 Ground Contamination

Figure 2.6 shows a far sidelobe beam model for a single feed of the COMAP instrument. As we can see, we have significant beam sidelobes all the way out to around 70 degrees. This means that, as long as the telescope is pointing at an elevation lower than about 70 degrees, some of our main sidelobes will be hitting the ground. As long as the

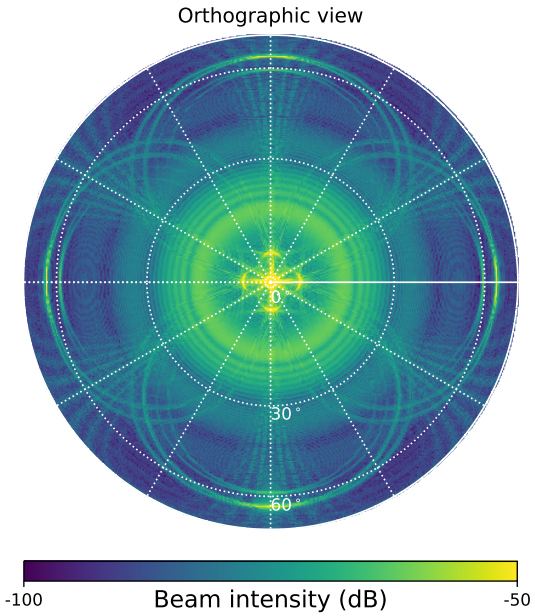

Figure 2.6: Simulated far sidelobe beam model for a single feed of the COMAP instrument. Beam model courtesy of James Lamb.

contribution from the sidelobes is fairly stable in time, this small ground pickup only has the effect of increasing our system temperature by up to a few Kelvin. However, if one of the more sharply defined sidelobes is right at the edge of the horizon, then the contribution from the ground can change quite abruptly, and we get a very strong signal in our time streams.

One thing that makes ground contamination more worrisome than other systematics is that the ground signal is correlated with where the telescope is pointing. This means that it will not necessarily integrate down properly when we add in more data. In general if some systematic is present in the data, this may not be ideal, but as long as the systematic is random and independent between each time we take data, then the contribution of the systematic will generally integrate down roughly as $\sim 1/\sqrt{N_{\text{obs}}}$, where N_{obs} is the number of independent observations.

The ground contribution, however, will be roughly the same each time we are observing at the same point in the sky, which means that it is not so simple. Of course, the ground is roughly constant in AZ/EL coordinates, while the cosmological signal is mapped in RA/DEC, so the signals are not completely degenerate, however, we do observe the same cosmological fields along the same path in AZ/EL every day, so the ground contamination will tend to affect the same modes in RA/DEC again and again.

While the ground itself is fairly close to a blackbody spectrum, which is something we would not be much worried about, the important thing for the ground contamination

is the frequency dependence of the sidelobes themselves. This is because if the sidelobes are very different at different frequencies, the ground contamination seen at different frequencies will also be very different, even though the ground is almost a blackbody. Fortunately, the positions of the sidelobe peaks is something that is not expected to change with frequency, but the sidelobe structure will in general be more sharp at high frequencies, and more diffuse at lower frequencies. The resulting ground contamination is expected to be fairly smooth in frequency, and thus something that can be removed quite effectively by filtering. Note that since the beam model (Fig. 2.6) is fairly expensive to run, we do not have one at another frequency, which means that these are mostly qualitative educated guesses about the beam properties.

While the smooth frequency structure should make it possible to remove the ground contamination quite effectively, we should note that the ground signal is still perhaps five orders of magnitude stronger than the cosmological signal we are trying to measure, so we need to remove it very precisely for it not to affect our measurements. This is what makes ground contamination one of the major challenges for COMAP.

The current version of the COMAP data analysis pipeline uses simple pointing template fitting in time domain in order to remove as much as possible of the ground signal. We use the simple atmosphere model from Sect. 2.4.2 together with using the azimuth position itself as a template

$$d_{\text{after}} = d_{\text{before}} - \frac{g}{\sin(\text{el})} - a\,\text{az} - \left\langle \frac{g}{\sin(\text{el})} + a\,\text{az} \right\rangle, \tag{2.10}$$

here g and a, as well as an overall constant, has been fit to the data. The elevation template is mostly there to remove the effect of the atmosphere, but it will also pick up some ground signal. The azimuth template is basically removing the linear term in the Taylor expansion of the ground signal about the centre of the field. These templates are subtracted from each frequency channel independently. Figure 2.7 shows the average values for these fitted template amplitudes along the paths followed by our three cosmological fields. We see large amplitudes both at high and low elevations. At low elevations (around el $\sim$ 30) the main sidelobe is hitting the edge of the horizon, while at around el $\sim$ 70 the lower of the four far sidelobe peaks (see Fig. 2.6) is just at the edge of the horizon.

The most extreme azimutal amplitudes observed are found in the region az, el $\sim$ $315, 30$ degrees. Here we believe what is happening is that one of the 60-70 degree sidelobes are just at the edge of Black Mountain (see bottom of Fig. 2.7 at az $\sim$ 60). Figure 2.8 shows the feed and frequency averaged raw time stream from a co6 observation of this area, together with the azimuth template. We here see a very strong effect. Even though we can remove most of this effect by fitting and subtracting templates, there is still contamination left, and this will be hard to remove completely at a later stage. This is why we currently do not use any data from below 35 degrees elevation.

In the next iteration of the analysis pipeline, the plan is to model the ground explicitly. This work is led by master student Jonas Lunde, under my supervision. We are using all the raw data to make continuum maps of the ground using the destriping mapmaker (Delabrouille, 1998; Keihänen et al., 2010). These ground maps can then be used to

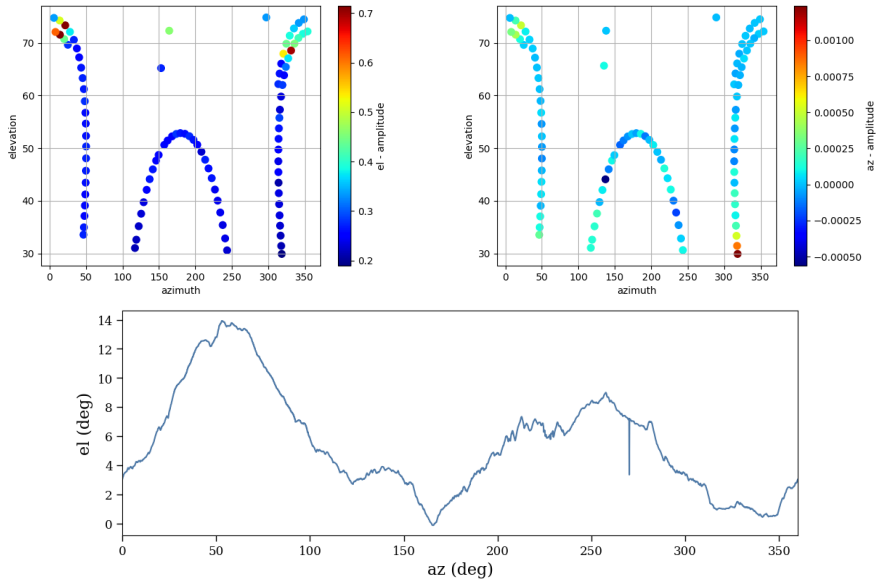

Figure 2.7: Average pointing template amplitudes fitted to the data in the data analysis pipeline (top). Results are binned along the path of the three cosmological fields. Elevation profile of the horizon as seen from the telescope location at OVRO (bottom). Elevation profile courtesy of Duncan Watts and Kieran Cleary.

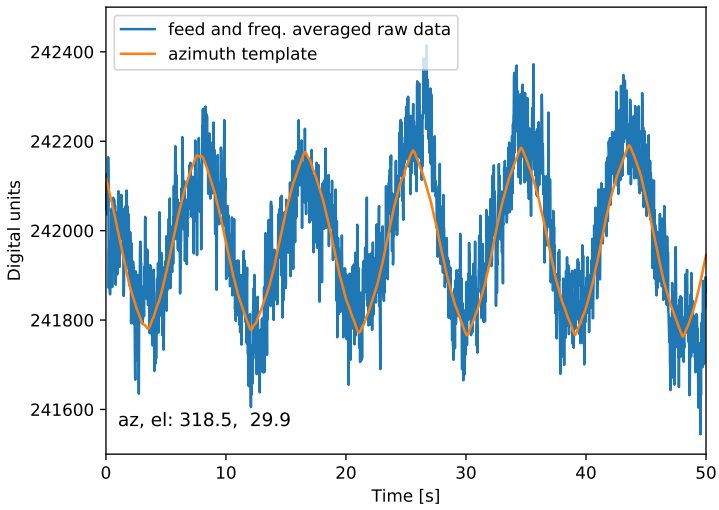

Figure 2.8: Raw time ordered data, averaged over feeds and frequencies, plotted together with the azimuth template.

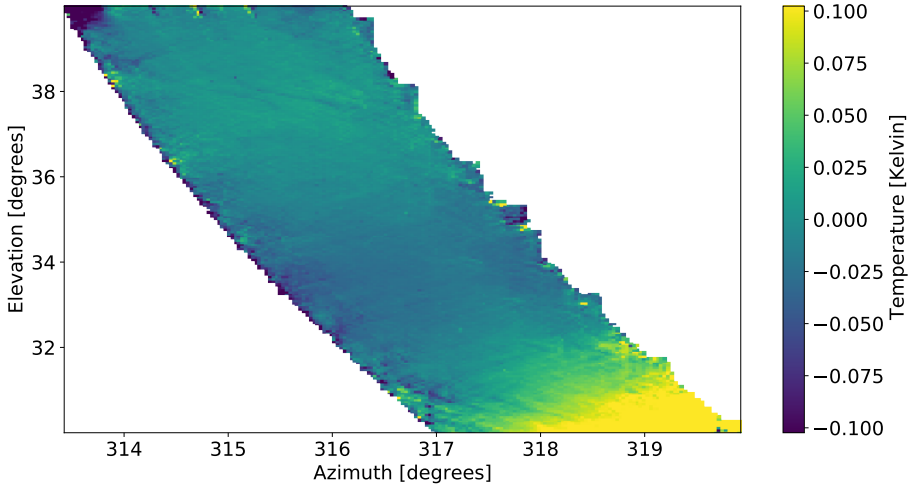

Figure 2.9: Continuum map of ground contamination. Made using a destriper and all the available Lissajous data taken in this region. Figure courtesy of Jonas Lunde.

make more precise templates to remove from the time streams during the regular data analysis. Figure 2.9 shows the previously discussed region with very strong ground contamination. In the lower right corner we see a strong gradient in both azimuth and elevation. By using maps like these, we can produce more precise templates that can remove the ground better than the simple templates we are currently using.

In addition to subtracting the ground contamination from the time streams, we can also design power spectrum methods that are as robust as possible to ground contamination, see Paper II (Ihle et al., in prep) for more details.

2.4.4 Standing Waves

Standing waves are another one of the major systematics in the raw COMAP data. They can form at various stages of the signal chain, from the sky to the digital backend, in the various potential electromagnetic cavities.

A given cavity length D will resonate with frequencies separated by

$$\Delta\nu_{\mathrm{SW}} = \frac{c}{2D}, \tag{2.11}$$

where $\Delta\nu_{\mathrm{SW}}$ is the distance, in frequency between peak resonances and c is the speed of light. A simple model for the standing wave structure in frequency is then given by (D. Chung, 2020)

$$T_{\mathrm{SW}} = T_{\mathrm{sys}} r_{\mathrm{SW}} \sin\left(2\pi \frac{\nu}{\Delta\nu_{\mathrm{SW}}} + \phi\right), \tag{2.12}$$

where T_{SW} is the contribution to the system temperature coming from the standing wave, r_{SW} is the (assumed to be frequency independent) standing wave amplitude and ϕ is some overall phase.

As we will see, a constant standing wave is not an issue for us since it will be removed in the bandpass normalization. It is only if a standing wave changes over time that it will show up in our data. These changes will arise whenever the cavity length changes, due to temperature fluctuations, vibrations etc. In the data the standing waves will then contribute to the correlated noise.

Standing waves can be distinguished from the other sources of correlated noise by their particular structure in frequency. Gain or atmospheric fluctuations are both (in slightly different senses) continuum sources in frequency, while standing waves will contribute with different sign to different frequencies as it goes in and out of resonance, as can bee seen from Eq. 2.12.

If the standing wave cavity is common to all the feeds, like the one between the receiver and the secondary, then the standing wave will change as a single function in time, common to all feeds, that contributes with different amplitudes to different frequencies. As we will see in Sect. 2.6.7, this is a perfect signal to pick up using Principal Component Analysis (PCA). Standing waves that only show up in individual feeds are much harder to measure and remove, but they will also affect a smaller part of the data, so they will integrate down significantly as we combine the different feeds.

Figure 2.10 shows the leading PCA mode from a CES scan of co6. We see that as a function of time, the mode looks like some sort of correlated noise. The amplitudes as a function of frequency show signs of the sinusoidal behavior expected from Eq. 2.12. Largest PCA amplitudes are found in feeds 3 and 5, this is because these feeds have a single stage polarizer, which makes them more susceptible to the standing waves than the other feeds which have two stage polarizers. If we try to estimate the frequency gap between the resonances, we see that we get a bit different results depending on which feeds we use, but the results are in the range $\Delta\nu_{\mathrm{SW}} \sim 0.3$ GHz, which corresponds to a cavity length of about 0.5 m. This fits fairly well with the cavity between the polypropylene weather shield that covers the receiver and the receiver itself (J. W. Lamb, 2020).

This reciever-weather shield standing wave is the one we see most clearly in our time ordered data, and the one we have studied the most. We find it in all the data, but the stability of the standing wave changes significantly, including a clear correlation with the windspeed. In June 2020, we augmented the the weather shield with a polystyrene backing support in order to make the weather shield more stable. This change was successful at suppressing this standing wave, although it is still there at a lower level. The other standing waves we have observed are typically more unstable and sporadic. For an example of another standing wave, with a frequency gap of $\Delta\nu_{\mathrm{SW}} \sim 0.07$GHz, suggestive of the ~ 2 m distance between the receiver and the secondary, see Fig. 2.11. We are still in the early phase of understanding and classifying these standing waves, but we are making progress in this area, especially in understanding which standing waves affect our measured cosmological power spectrum.

2.4.5 Correlated noise

There are three main sources of correlated noise. Atmospheric fluctuations contributes to the correlated noise at timescales longer than one or a few seconds, and these fluctuations are common to all frequencies and feeds. Gain fluctuations are typically the main source

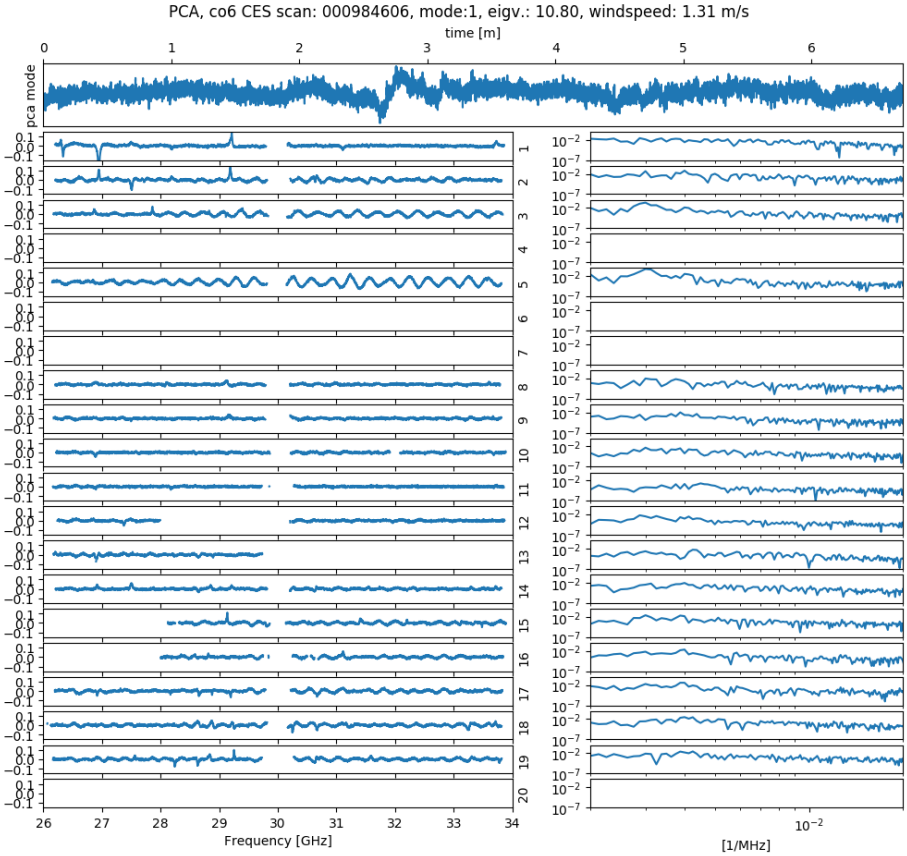

Figure 2.10: Leading PCA mode as a function of time (top panel). The rest of the panels show the corresponding PCA amplitudes as a function of frequency for each of the 19 feeds (left) and power spectral density, in frequency, of these amplitudes (right). Data is from a CES scan of co6.

of correlated noise, but they are also common to all the frequencies within a single feed. Some standing waves contribute significantly to the correlated noise, however, the same frequency structure is typically preserved over time, so these can be removed using a PCA filter.

Since all these sources of correlated noise are also highly correlated between frequencies, they can usually be removed very precisely in the filtering steps of the data analysis pipeline, as can be seen in Fig. 2.12. For this reason, correlated noise is not a systematic that we are very worried about at this stage in COMAP.

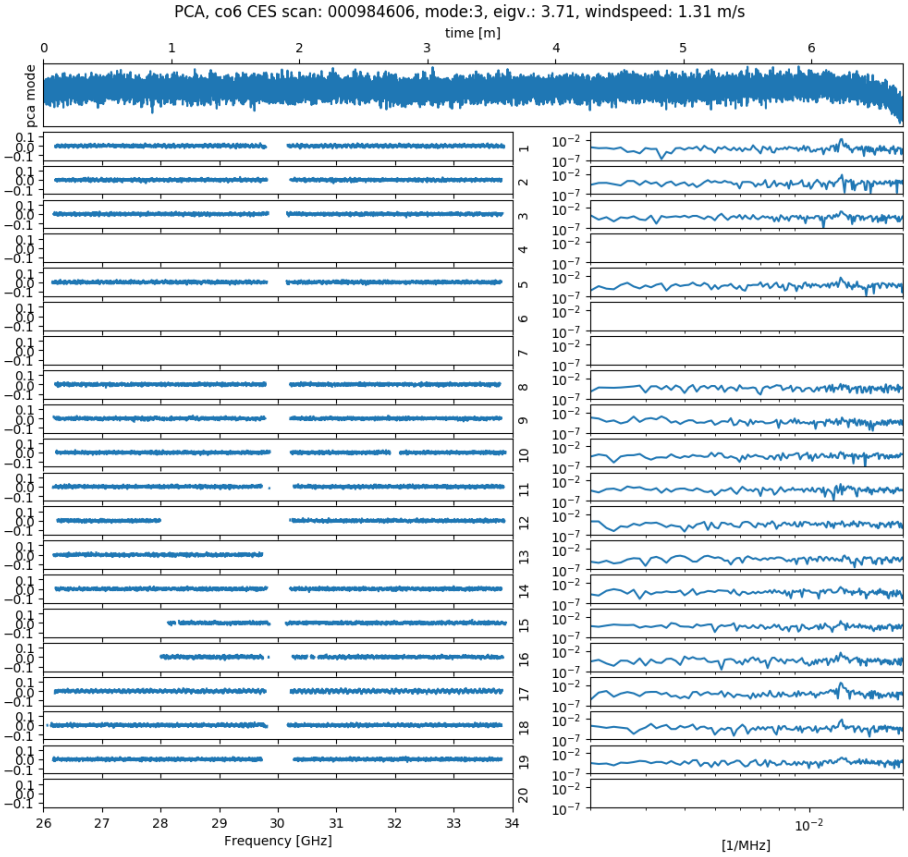

Figure 2.11: Third PCA mode as a function of time (top panel). The rest of the panels show the corresponding PCA amplitudes as a function of frequency for each of the 19 feeds (left) and power spectral density, in frequency, of these amplitudes (right). Data is from a CES scan of co6.

2.4.6 Continuum Foregrounds

Our cosmological fields are selected to be in regions with low galactic foregrounds, for this reason we have not been too worried about these up until now, as they will be fairly effectively removed in the frequency filter. However, the foregrounds are still up to two orders of magnitude stronger than our CO signal, so we do need to suppress them significantly, if we want to measure our signal.

This will be a more and more significant issue the more sensitivity we get, so it is something that we will probably have to focus more on in the future.

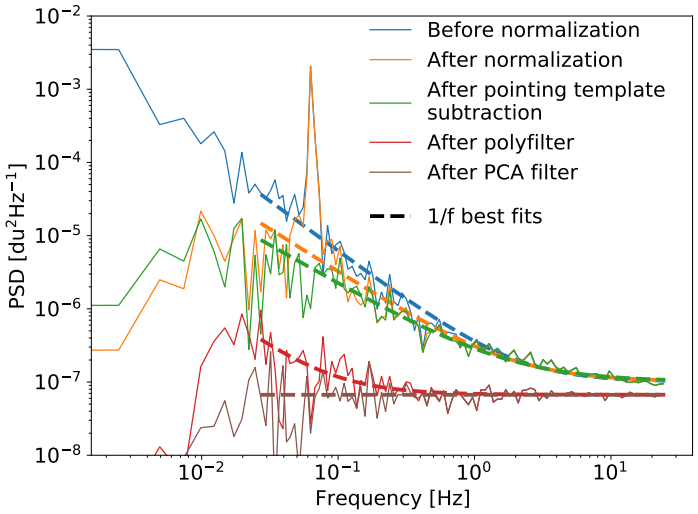

Figure 2.12: Power spectral density of the COMAP time ordered data at various different stages of the analysis pipeline. Figure courtesy of Jonas Lunde.

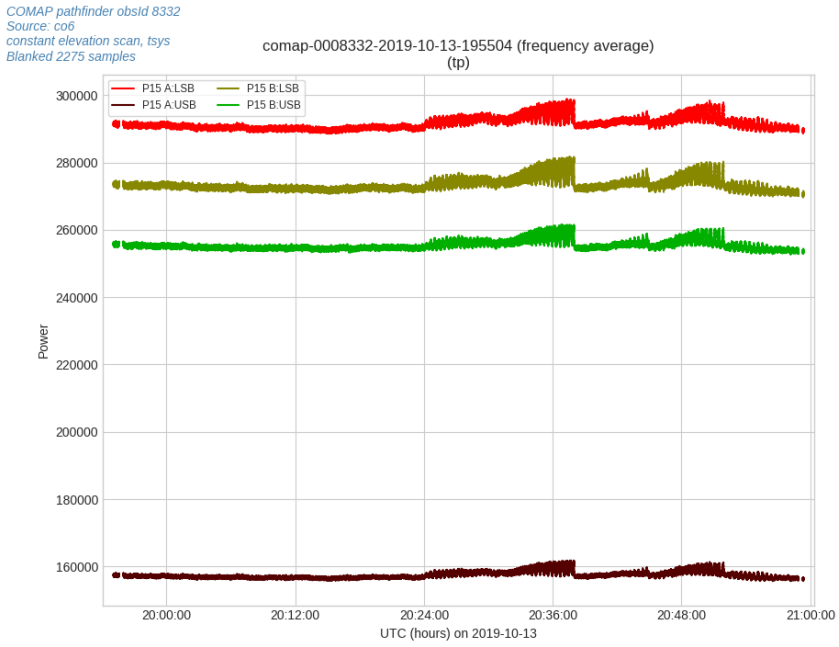

Figure 2.13: Example of time ordered data taken with the sun in the far sidelobe. Here we see data from the four sidebands of feed 15.

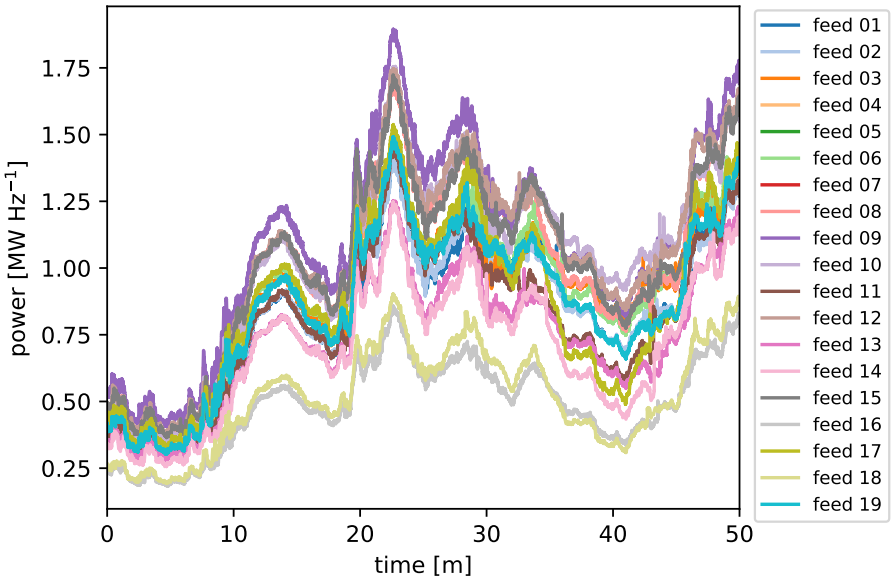

Figure 2.14: Example of data taken during bad weather.

2.4.7 Sun/moon in sidelobes

Another systematic that we did not anticipate the significance of in advance were the times when the sun, or even the moon, is exactly at one of the four peaks in the far sidelobes (see Fig. 2.6 at about 65 degrees). If this happens, we get a very large signal in the time streams, as can be seen in Fig. 2.13 after about 30 minutes. When this effect was first observed, we did not know what was the cause, but after looking at the data in the right way we could see the effect very clearly (see Fig. 2.23). We now track the relative position of the sun carefully and this allows us to remove the affected data during data selection (see Sect.2.7.1).

2.4.8 Weather

If there is significant cloud coverage, or other bad weather, then the data taken is typically not usable. Figure 2.14 shows an extreme example of data taken during bad weather. As we will discuss in Sect. 2.7.1 bad weather actually affects a significant part of the data, and is dealt with during data selection.

2.4.9 Spikes

We occasionally see spikes in the raw data, especially during the summer. We do not know for sure what the cause of the spikes are. Some of them seem to come from insects flying in front of the receiver, but we do not know if all of them are caused by

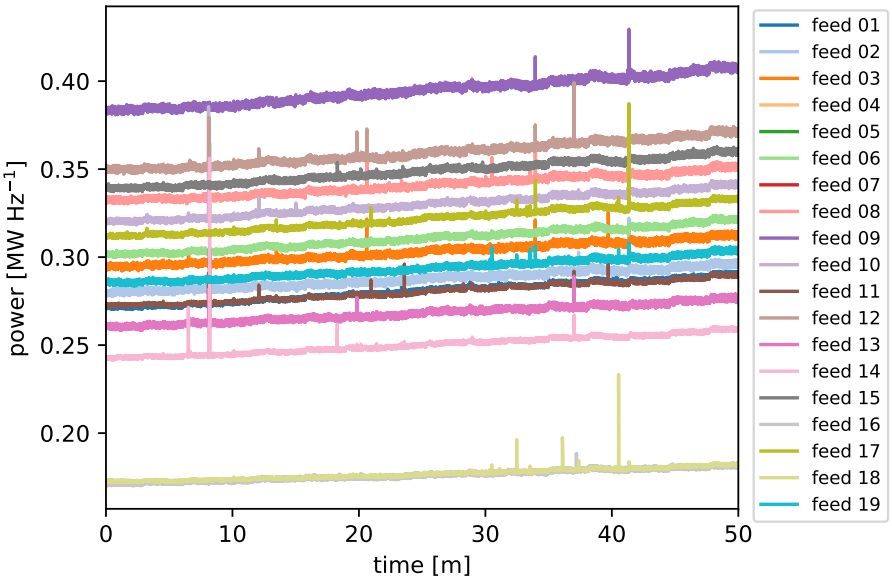

Figure 2.15: Example of an observation with a very large number of spikes.

this. Sometimes the spikes show up sporadically, one or two in an entire obsid, and sometimes the data is flooded by spikes (see Fig. 2.15).

The spikes are continuous in frequency, so we expect most of the signal to be removed in the frequency filters. We have not seen any clear signs that the spikes make a significant impact on the final maps or power spectra, but we are tracking the spikes to look for any such affect.

2.5 Calibration

Calibration is one of the major challenges we need to solve in order to make use of the data. We need to translate the raw power readings of the detectors (given in arbitrary digital units) into brightness temperatures. COMAP uses a main calibration strategy based on a calibration vane at ambient temperature that is periodically moved in front of the receiver. This acts as a hot load reading that is compared to the cold load of the sky measurement.

We can verify the calibration by looking at a source with a known temperature, like Jupiter. Although this will not give a very precise calibration, it is still a useful test to perform, if nothing else than as a sanity check that we are not doing something completely wrong. It will also show if the calibration is consistent over time.

2.5.1 Ambient load calibration

The main method of calibration used for COMAP involves comparing an observation of the cold sky with observations of a calibration vane at ambient temperature. This ambient load is automatically moved in front of the receiver and back during calibration.

The power measured by the telescope at any given time is given by

$$P = GT_{\rm sys}, \tag{2.13}$$

where P is the total power, G is the gain[1] and $T_{\rm sys}$ is the system temperature combining the contribution from the receiver with the external signal coming into the feed horns.

The idea of the calibration method is simple. We observe two different signals at known temperatures, and use these two measurements to determine $T_{\rm sys}$ assuming the gain is constant.

The complication with this is that we want to calibrate to the cosmological signal, and some of this signal gets absorbed by the atmosphere, meaning that only part of the signal makes it down to our receiver. So for a given signal contribution the system temperature changes by

$$\Delta T_{\rm sys} = e^{-\tau}\Delta T_{\rm signal}, \tag{2.14}$$

where $\Delta T_{\rm sys}$ is the contribution to the physical system temperature of the signal, τ is the optical depth of the atmosphere and $\Delta T_{\rm signal}$ is the temperature of the cosmological signal. To account for this effect we define an effective system temperature (Penzias and Burrus, 1973)

$$T'_{\rm sys} \equiv e^{\tau}T_{\rm sys}, \tag{2.15}$$

where $T'_{\rm sys}$ is the effective system temperature and $T_{\rm sys}$, as before, is the physical system temperature measured by the instrument. This new definition ensures that

$$\Delta T'_{\rm sys} = \Delta T_{\rm signal}, \tag{2.16}$$

making the interpretation of our measurements much easier. Another way to think about this is to note that losing some of the cosmological signal in the atmosphere is equivalent to just having a higher noise level in the first place, and the effective temperature, $T'_{\rm sys}$, is just this higher noise level. In the same way we define the effective gain

$$G' \equiv e^{-\tau}G, \tag{2.17}$$

where G' is the effective gain and G is the physical gain. We can now rewrite Eq. 2.13 using our new definitions

$$P = GT_{\rm sys} = G'T'_{\rm sys}. \tag{2.18}$$

When the instrument is looking at the cold sky, the power is given by

$$P_{\rm c} = G\left(T_{\rm rcv} + T^b_{\rm atm} + e^{-\tau}T_{\rm CMB}\right) = GT_{\rm sys}, \tag{2.19}$$

where $P_{\rm c}$ is measured power when looking at the cold sky (the *cold* load), $T_{\rm rcv}$ is the receiver temperature contribution to the physical system temperature, $T^b_{\rm atm}$ =

[1] Note that, for convenience, we have absorbed the conventional constant factors $k_{\rm B}\Delta\nu$ into G for this derivation.

$(1 - e^{-\tau})T_{\rm atm}$ is the brightness temperature of the atmosphere, $T_{\rm atm}$ is the physical temperature of the atmosphere, which we assume to be in thermal equilibrium at a constant temperature, and $T_{\rm CMB}$ is the CMB monopole, which we assume dominates any other cosmological signal. Rewriting this in the effective coordinates we get

$$P_{\rm c} = G'\left(T'_{\rm rcv} + T'^{b}_{\rm atm} + T_{\rm CMB}\right) = G'T'_{\rm sys}, \tag{2.20}$$

where we have defined the effective receiver temperature $T'_{\rm rcv} \equiv e^{\tau}T_{\rm rcv}$ and atmosphere brightness temperature $T'^{b}_{\rm atm} \equiv e^{\tau}T^{b}_{\rm atm}$.

If the instrument is looking at the calibration vane (the *hot* load) the measured power is given by

$$P_{\rm h} = G\left(T_{\rm rcv} + T_{\rm h}\right), \tag{2.21}$$

where $T_{\rm h}$ is the physical temperature of the calibration vane. If we assume that the physical temperature of the calibration vane is the same as the physical temperature of the atmosphere, we can rewrite $T_{\rm h}$ in the following convenient way

$$T_{\rm h} = e^{-\tau}T_{\rm h} + (1 - e^{-\tau})T_{\rm h} = e^{-\tau}T_{\rm h} + T^{b}_{\rm atm}. \tag{2.22}$$

If we now write Eq. 2.21 in the effective units we get

$$P_{\rm h} = G'\left(T'_{\rm rcv} + T_{\rm h} + T'^{b}_{\rm atm}\right) = G'\left(T_{\rm h} + T'_{\rm sys} - T_{\rm CMB}\right). \tag{2.23}$$

Using the measurements of the cold and the hot load, assuming the gain is constant and that we know the temperature of the calibration vane, we can solve for the effective system temperature

$$T'_{\rm sys} = \frac{T_{\rm h} - T_{\rm CMB}}{P_{\rm h}/P_{\rm c} - 1}. \tag{2.24}$$

During this derivation we have neglected the effect of ground spillover. However, a very similar argument can be made to take this into account as well, and if we also assume that the ground has the same physical temperature as the calibration vane and the atmosphere, the result is exactly the same as in Eq. 2.24 with the effective system temperature, $T'_{\rm sys}$, now taking into account both the effect of the atmosphere and the ground spill. Note that whenever the temperature of the system, atmosphere etc. is mentioned elsewhere in this thesis, we are referring to the effective temperatures, not the physical ones.

2.5.2 Jupiter calibration

We do periodic Jupiter observations for calibration and pointing correction. By fitting a Gaussian beam to these observations we can measure the antenna temperature of Jupiter. These measurements can be compared to model expectations depending on the current distance to Jupiter. This work is led by Stuart Harper in the Manchester group.

Figure 2.16 shows antenna temperature measurements of Jupiter from October 2019 to August 2020, for a single feed, compared to the model. We see that apart from an overall constant factor offset, the measurements fit the model very well. The overall offset factor, which we can see is given roughly by 0.7-0.8, is the effect of the aperture

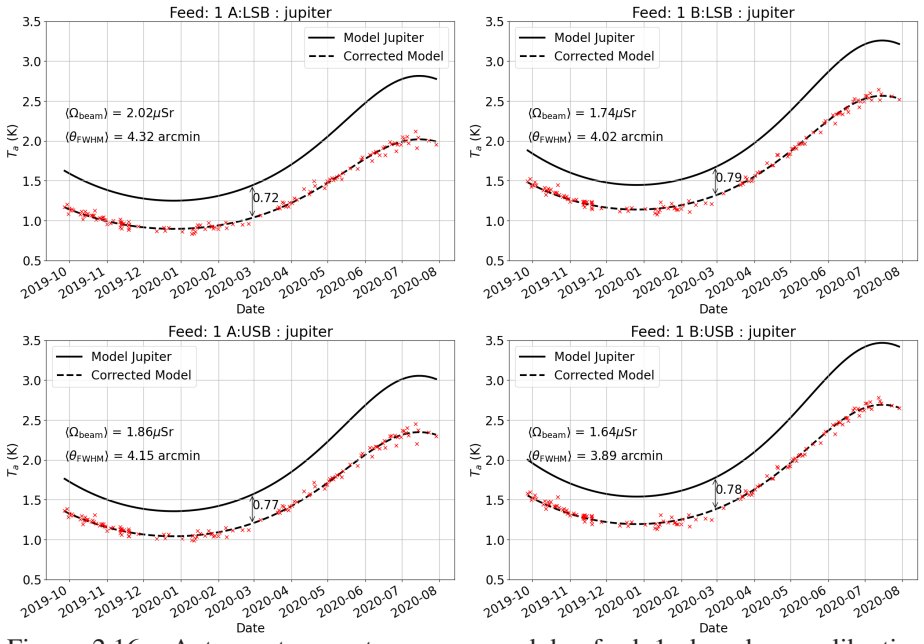

Figure 2.16: Antenna temperatures measured by feed 1, based on calibration observations of Jupiter, compared to model expectations. Figure courtesy of Stuart Harper.

efficiency[2], which is expected. These results show that the calibration is consistent over time, and at a reasonable value.

2.6 Data Analysis Pipeline

The COMAP analysis pipeline takes the raw data, from what we call *level 1* files, does calibration, filtering, performing data selection, making maps and calculating power spectra.

Figure 2.17 shows an overview of the various steps and modules that are part of the COMAP analysis pipeline. The first modules are `scan_detect`, which identifies, classifies and divides our data, and `l2gen`, which performs the low-level calibration, filtering and frequency masking of the data and producing *level 2* files. Level 2 files contain calibrated and filtered time ordered data, as well as a set of diagnostic data. Before the level 2 files are combined into maps, there is a second level of data selection, performed by the module `accept_mod`, which produces an `accept_list` denoting which observations to accept or reject when making the maps. The mapmaker, `tod2comap`, converts the time ordered data from the level 2 files into 3D maps of brightness temperature. The preceding steps are described in Paper I (Foss et al., in

[2]It is slightly different from the aperture efficiency since Jupiter is not a point source.

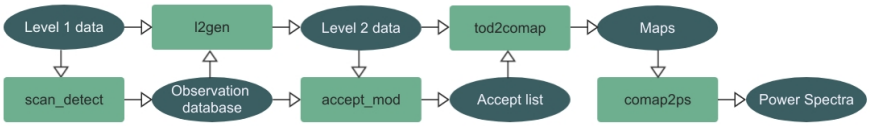

Figure 2.17: Overview of the COMAP CO analysis pipeline. Figure courtesy of Jowita Borowska.

prep). Here we will give a summary overview, while going into more details in some areas.

The last part of the current data analysis pipeline is `comap2ps`. This module takes in the 3D maps in brightness temperatures and calculates various power spectra. These procedures are described in Paper II (Ihle et al., in prep). We will not repeat much of the methods presented there, but will, in later sections, discuss more abstractly the problem of inference from line intensity maps. We will also go into some details on using the voxel intensity distribution (VID) that have not been published.

2.6.1 Scan Detect

The main goal of scan detect is to go through all the raw data files to classify and divide the data into individual scans. This information is then provided in terms of a runlist, which is used when running `l2gen` or `tod2comap`.

A *runlist* is a list of obsids that is divided according to the field or source observed during each obsid. For each obsid, we list all the scans within that obsid. For each scan the modified Julian date (MJD) of the start and end of the scan is provided as well as the scanning strategy (e.g. Lissajous or constant elevation) and some pointing information.

2.6.2 Level 2 File Generator: l2gen

The main goal of l2gen is to take the raw data files, called level 1 files, and turn them into masked and filtered data files, called level 2 files, that are ready for mapmaking. An additional function of l2gen is to keep track of various diagnostics for tracking data quality and for data selection.

The main steps in l2gen are (in chronological order):

- Bandpass normalization
- Removal of pointing templates
- Masking
- Polynomial filter in frequency

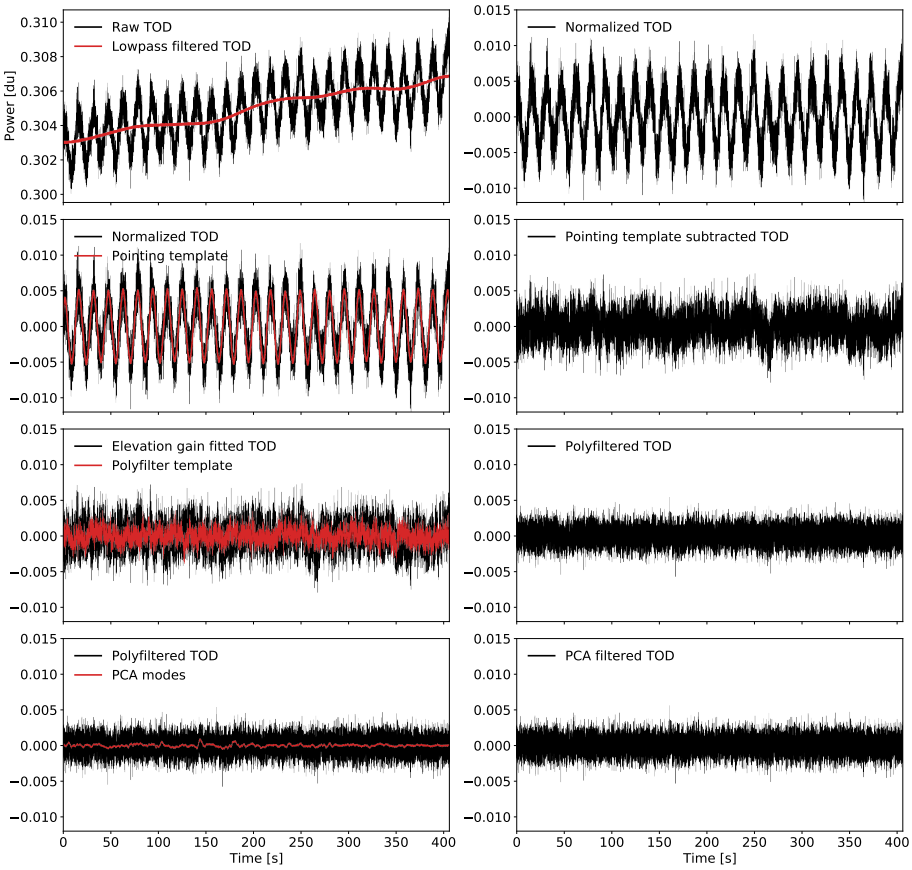

Figure 2.18: Time ordered data plotted along the various steps of the data analysis pipeline. Each row shows the data before (left column) and after (right column) applying the indicated filter. From top to bottom, the filters shown are 1) normalization; 2) pointing template subtraction; 3) polynomial filter in frequency; and 4) PCA filter. Data used is from scan 1445603, feed 5, in a 31.25 MHz band around 27.673 GHz. Figure courtesy of Jonas Lunde.

- PCA-filter
- Calibration
- Decimation

2.6.3 Bandpass normalization

Our model for the total power seen in the detectors is given by

$$P_{\text{out}} = k_{\text{B}} G \Delta\nu T_{\text{sys}}, \tag{2.25}$$

where k_B is Boltzmanns constant, G is the gain and $\Delta\nu$ is the bandwidth of the frequency channel.

The first of the actions that is performed on the data is bandpass normalization. The normalization is performed on each frequency channel separately and is done by dividing the time ordered data (TOD), d, by its *running mean*, $\bar{d}$, and subtracting one. As we see from Eq. 2.25, this normalization will cancel out both the average gain G and the average system temperature $T_{\rm sys}$ leaving us with a time stream with small-scale fluctuations with a variance of $1/\Delta\nu\tau$ as dictated by the radiometer equation, where τ is the duration in time of each sample.

The running mean is calculated by passing the TOD through a lowpass filter. See Fig. 2.18 (top row) for an example of this normalization step. The lowpass filter is implemented in Fourier space by multiplying with the following weight function:

$$W = \left[1 + \left(\frac{f}{f_{\rm knee}}\right)^{\alpha}\right]^{-1}, \tag{2.26}$$

where f is the temporal frequency, $f_{\rm knee} = 0.02$ Hz (corresponding roughly to a 50 s timescale) and $\alpha = 4$. So we have

$$\bar{d} = \mathcal{F}^{-1}\left\{\mathcal{F}\{d\}W\right\}, \tag{2.27}$$

where $\mathcal{F}$ denotes the Fourier transform. The normalized data then becomes

$$d_{\rm norm} = d/\bar{d} - 1. \tag{2.28}$$

2.6.4 Removal of Pointing Templates

As the telescope changes elevation we are looking through different amounts of atmosphere, leading to changes in the power received by the detectors. As we have discussed, this effect can be modeled by a simple expression for the optical depth of the atmosphere, τ

$$\tau(\mathrm{el}) = \tau_0/\sin(\mathrm{el}), \tag{2.29}$$

where τ_0 is the optical depth of the atmosphere at zenith, and el is the elevation.

We also know that sometimes there is significant ground contamination, and we therefore want to subtract an azimuth template, as well as the elevation template. Specifically we assume that

$$d = g/\sin(\mathrm{el}(t)) + a\,\mathrm{az}(t) + c + n, \tag{2.30}$$

where g, a and c are constants and n is some Gaussian noise with constant variance. We then find the best fit values (i.e. least squares) of g, a and c, and use the values of g and a to remove the pointing templates from the time stream:

$$d_{\rm after} = d_{\rm before} - g/\sin(\mathrm{el}(t)) - a\,\mathrm{az}(t) - \langle g/\sin(\mathrm{el}(t)) + a\,\mathrm{az}(t)\rangle, \tag{2.31}$$

where $\langle\ \rangle$ denotes the mean value. See Fig. 2.18, second row.

We fit these templates independently for each frequency in order to try to remove any frequency structure of the ground and atmosphere. For longer scans we divide the data into different segments of roughly 4 minutes each, and perform this template fit and removal separately on each of the segments of data.

2.6.5 Masking

We want to mask out frequency channels that are *bad* in the sense that they are not performing as expected. We do this mostly by using the correlation matrix between different frequency channels.

In order for the data to be as clean as possible before identifying bad channels, we make a copy of the original (level 2) dataset and perform the poly- and PCA filters, described below, on this copied data, before calculating the correlation matrices within each band. We use two main approaches to identifying individual or groups of channels to be masked. The first approach uses the fact that the expected correlation between two independent Gaussian variables (for large n_{samp}) is given by $1/\sqrt{n_{\text{samp}}}$, where n_{samp} is the number of samples the correlation is calculated on.

This means that, after subtracting the expected correlation induced by the polynomial frequency filter, we know the statistics describing *good* data, and can identify bad data as deviations from these statistics. We look at entries within squares of different sizes as well as sets of columns within the correlation matrix and see if the average of the absolute value of the correlations within this region deviates from the values expected, and mask out the corresponding channels if the deviations are outside the acceptable limits.

The second approach is to calculate a set of diagnostics for individual frequency channels (like the average correlation of the channel to all the others in that band or the average of the absolute value of the same). We can then compare the values of these diagnostics for the different channels and look for large outliers. The disadvantage of the first approach is that it is much harder to detect these deviations in a short scan than it is in a long scan, so it is hard to choose the parameters for the acceptable deviations in a way that will mask out enough of a short scan without masking too much of a long scan. The disadvantage of the second approach is that if the overall data quality is bad, then a bad channel will not necessarily be an outlier, and will not be masked out.

In addition to these approaches we look specifically for *edge correlations*, that is, correlations between individual frequencies at the edge of each sideband with the corresponding frequencies at the edge of another sideband. This is to look for and mask out frequencies affected by a known aliasing effect. We also have some overall hard cuts on large individual correlations and on large individual variances (compared to what is expected from the radiometer equation). After this mask has been found we perform all the analysis on the original dataset, but now only using the frequency channels that are not masked.

2.6.6 Polynomial filter in frequency

The goal of the polynomial filter in frequency is to remove the $1/f$-noise specific to each feed, as well as to remove any continuous foregrounds from the data. This filter greatly reduces the correlated noise, and often leaves the TOD close to white noise.

The poly filter is performed separately on each time step, by fitting and removing a polynomial (usually linear) in frequency across each sideband. Specifically, we assume

$$d_\nu = a + b\nu + c\nu^2 .. + n, \tag{2.32}$$

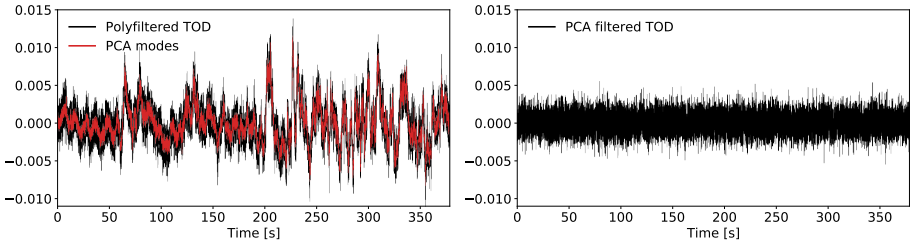

Figure 2.19: Example of time ordered data before and after PCA filter. This is a very extreme case with a very strong standing wave. Figure courtesy of Jonas Lunde.

where d_ν is the data across one sideband at a specific time step, a, b and c etc. are constants and n is some Gaussian noise with constant variance. We then find the best fit values (i.e. least squares) for the constants and subtract the polynomial (see Fig. 2.18, third row)

$$d_\nu^{\text{after}} = d_\nu^{\text{before}} - \left(a + b\nu + c\nu^2 ...\right). \tag{2.33}$$

Our current version of the pipeline uses a linear polynomial for the polyfilter, which appears to be sufficient for our purposes.

One possible issue with the current version of the polyfilter is that, depending on what the origin of the effects we want to remove with this filter, we may want to divide the polynomial in Eq. 2.32 by the T_{sys} corresponding to the frequency channel, since a channel with high T_{sys} is expected to have a relatively lower effect of any signal compared to the noise. This is certainly what we would expect from a continuous foreground.

We are working on improved ways to do this frequency filtering in a way to remove both the gain fluctuations as well as continuum foregrounds in a consistent manner.

2.6.7 PCA-filter

Let us consider all the data corresponding to a single scan as a data matrix, D_{ij}, with the TOD (time stream) corresponding to a single frequency as a row of the matrix. This will be a matrix with dimensions $n_{\text{freq}} \times n_{\text{samp}}$, where $n_{\text{freq}} = 19 \cdot 4 \cdot 1024$, i.e. the total number of frequencies added up over all the sidebands and feeds, and where n_{samp} is the number of samples in time (typically of the order of ~ 20000 for a single scan).

$$D = \begin{bmatrix} D_{11} & \dots & D_{1n_{\text{samp}}} \\ \vdots & \ddots & \vdots \\ D_{n_{\text{freq}}1} & \dots & D_{n_{\text{freq}}n_{\text{samp}}} \end{bmatrix}. \tag{2.34}$$

If we treat the columns (all the different time steps) of the data matrix D as representing a set of n_{samp} random variables, then each frequency of each feed is a new sample of each of these random variables. The covariance matrix for these random

variables is then given by

$$C^{\text{samp}} \propto D^{\text{T}}D = \begin{bmatrix} C_{11} & \dots & C_{1n_{\text{samp}}} \\ \vdots & \ddots & \vdots \\ C_{n_{\text{samp}}1} & \dots & C_{n_{\text{samp}}n_{\text{samp}}} \end{bmatrix} \tag{2.35}$$

The PCA components are the eigenvectors of this matrix with the highest eigenvalues. These eigenvectors are basically just functions of time (one value for each time step). Specifically, the leading PCA-components are the functions of time that explain most of the variance between the different frequencies.

After finding the properly normalized eigenvectors, v_k, corresponding to the four largest eigenvalues of C^{samp} we calculate the PCA amplitude corresponding to each of these eigenvectors for each frequency channel, and subtract it from the data:

$$a_k = d \cdot v_k = \sum_{i=1}^{n_{\text{samp}}} d_i v_k^i$$

$$d_{\text{after}} = d_{\text{before}} - \sum_{k=1}^{4} a_k v_k,$$

where d is the time stream of a single frequency channel, a_k is the PCA amplitude of that same channel corresponding to the eigenvector v_k.

The PCA filter is perfect for removing standing waves that are common to all (or at least many) feeds. This is because, as we discussed in Sect. 2.4.4 the standing waves typically change as a single function of time, but affect different frequencies differently depending on if they are in or out of resonance. The resonance structure of the standing waves will then show up in the PCA amplitudes of the different frequencies, as we see in Figs. 2.10 and 2.11.

The PCA filter also picks up other systematics that do not have the features one would expect from standing waves. These effects are mostly of unknown origin at this point, but the PCA filter can still usually remove them quite effectively.

2.6.8 Calibration

For each frequency we interpolate the measured power and temperature of the hot load (calibration vane), P_{hot} and T_{hot} to the time, t_{scan}, of our current scan, i.e.:

$$P_{\text{hot}}(t_{\text{scan}}) = \frac{P^1_{\text{hot}}(t_2 - t_{\text{scan}}) + P^2_{\text{hot}}(t_{\text{scan}} - t_1)}{t_2 - t_1},$$

$$T_{\text{hot}}(t_{\text{scan}}) = \frac{T^1_{\text{hot}}(t_2 - t_{\text{scan}}) + T^2_{\text{hot}}(t_{\text{scan}} - t_1)}{t_2 - t_1},$$

where 1 and 2 denote the hot load measurement at the start and the end of the current obsid, respectively. We can then calculate T_{sys} for our current frequency

$$T_{\text{sys}} = \frac{T_{\text{hot}}(t_{\text{scan}}) - T_{\text{CMB}}}{P_{\text{hot}}(t_{\text{scan}})/\langle P_{\text{cold}}\rangle - 1}, \tag{2.36}$$

where $\langle P_{\text{cold}}\rangle$ is the mean power of the scan in this frequency channel.

2.6.9 Decimation

All the analysis up until now have been performed on the full frequency grid (1024 channels per sideband). For mapmaking purposes, however, we typically don't need this kind of resolution, so we want to co-add several frequency channels together to a single low resolution channel.

$$d_i^{\text{lowres}} = \frac{1}{\sum_m w_m} \sum_{m=(i-1)n_{\text{dec}}+1}^{in_{\text{dec}}} w_m d_m^{\text{highres}}, \tag{2.37}$$

where d_i^{lowres} is the time stream of frequency channel i in the low resolution frequency grid, d_m^{highres} is the time stream of frequency channel m in the high resolution frequency grid, n_{dec} is the number (usually 16) of high resolution frequencies to be combined in each low resolution frequency channel and $w_m = 1/\sigma_m^2$ is the inverse variance of the time stream in frequency channel m of the high resolution data (w_m is zero for masked frequencies).

2.6.10 Mapmaker, `tod2comap`

The mapmaker takes a set of level 2 files with calibrated and filtered time ordered data, and makes 3D maps of brightness temperature.

The mapmaker we use is very simple. It basically assumes that the input data is dominated by white noise, so that every sample is independent. The value in each pixel of the map is then just the noise weighted sum of all the samples that hit this pixel

$$m_p = \frac{\sum_{d\in p} \sigma_i^{-2} d_i}{\sum_{d\in p} \sigma_i^{-2}}, \tag{2.38}$$

where d_i is sample number i of the time ordered data, σ_i is the white noise level of sample i and where m_p is the value of the map in pixel p. This is typically referred to as a *binned* mapmaker. Figure 2.20 shows single frequency maps based on all the (accepted) data for each of the three main science fields. The maps look white noise dominated and smooth.

A large advantage with using this simple mapmaker is that we do not need to load in all the time-ordered data into memory at once, but can add more data incrementally. The disadvantage is that we lose a lot of the information that is in the TOD. This results in lower sensitivity to the signal on large scales (see Paper I, Foss et al., in prep for more details).

We are working on an improved mapmaker, using a destriper, and this has shown some promising results. This work is lead by master student Nils-Ole Stutzer, under my supervision. Figure 2.21 compares the pipeline transfer function[3] for the binned map and a map created by the destriper. We see that the destriper recovers a lot more of the large scale structure in the angular direction, it is however, much more costly process to run. Note that this destriper still treats each frequency channel independently, so it is still far from an optimal mapmaker.

[3] See Paper I (Foss et al., in prep) for details on how the transfer function is defined and estimated.

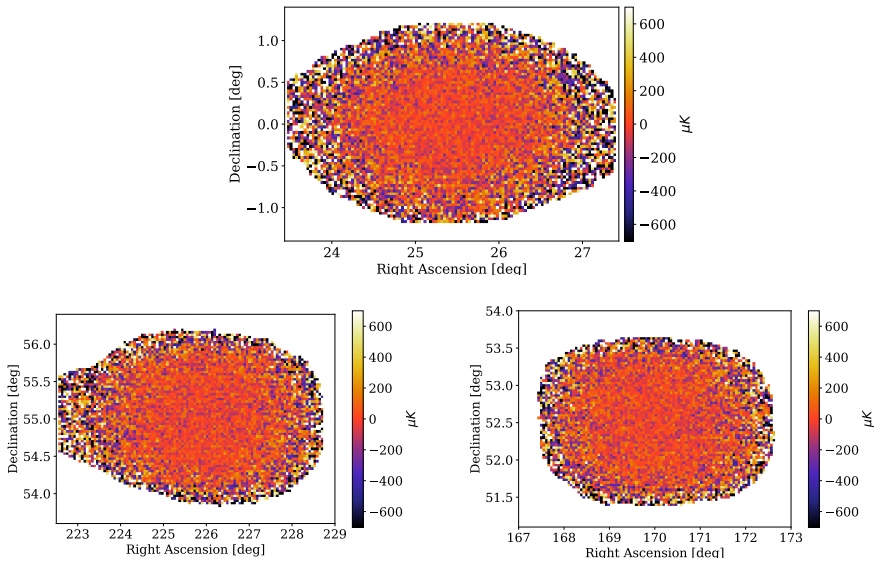

Figure 2.20: Feed averaged maps from of the co2 (top), co6 (bottom left) and co7 (bottom right) fields. These are single frequency maps at 32 GHz with a bandwidth of 31.25 MHz, including all the accepted data. Figure courtesy of Nils O. Stutzer.

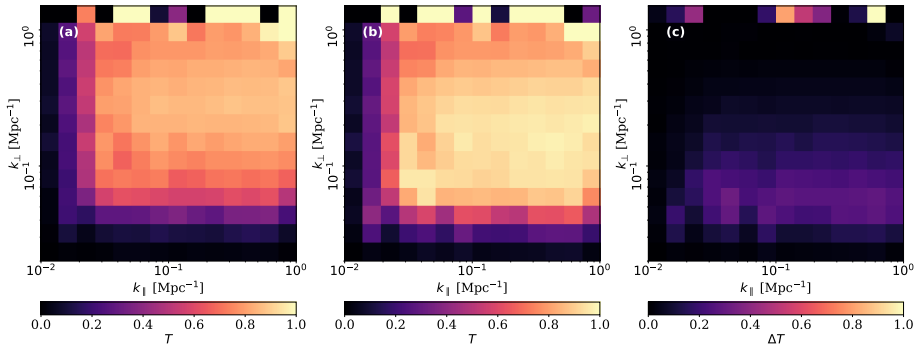

Figure 2.21: Comparison of the transfer function using the current (binned) mapmaker (a), the destriper (b), difference between the two (c). Figure courtesy of Nils O. Stutzer.

2.7 Data selection

There are two main stages of data selection in the COMAP pipeline. First, as discussed in the previous section, we mask out badly performing or outlier frequencies as part of the l2gen process. The second stage happens after we have created level 2 files, but before we run them through the mapmaker. This selection step is performed in the `accept_mod` module, and is based on gathering a large number of statistics for each chunk of data, and then deciding based on these statistics if we should accept or reject that data.

2.7.1 Statistics database

The first and most extensive part of the `accept_mod` module is the generation of the statistics database. This is a database where a number of statistics, or diagnostic variables, are calculated or gathered for each sideband of each feed for each scan. So the entire scan database is an array of dimensions $(n_{\text{scans}}, n_{\text{feeds}}, n_{\text{sidebands}}, n_{\text{stats}})$, where n_{stats} is the number of different diagnostic variables gathered for each chunk of data. Each statistic is represented by a floating point number.

Some of the statistics are calculated from the data itself, like various χ^2-values or the average kurtosis of the data, while others, like air temperature or windspeed are collected from the housekeeping data. In Tab. 2.2 we list all the different statistics currently used for the database, together with a short description of each one.

As the statistics database is a fairly small dataset (in total about 1 GB), it is very useful to look for patterns in this dataset, in order to understand the raw data, and the effect of the data analysis pipeline, better. Perhaps the most basic analysis we can do on a dataset like this is to do a basic correlation analysis. Fig. 2.22 shows the correlation between the different statistics from all the co2 data. Note that we do an automatic removal of extreme outliers, to avoid spurious correlations. Here we see a number of interesting correlations. While the mere correlation between two variables is often not in itself actionable information, the full correlation plot is still very useful to see which variables are related and which distributions to look closer at.

A particularly useful set of statistics are the `ps_chi2`-type statistics. These are measures of the excess power spectrum of the data as compared to what is expected from white noise. These are calculated by making one 3D map for each of the chunks of data, taking the power spectrum of this map, and comparing that to power spectra of simulated white noise maps with the same noise levels, using a normalized χ^2 statistic. The different types of `ps_chi2` statistics corresponds to using chunks of data of different dimensions for this analysis.

The `ps_chi2` statistics measure something that is fairly close to what we are actually interested in measuring, so it serves as a great diagnostic for identifying bad data. If another statistics is correlated with `ps_chi2`, this can help us identify what is the source of the excess power.

An excellent example of this is shown in Fig. 2.23. Here we see the far sidelobe model for the instrumental beam compared to the `ps_chi2` values projected onto a map of the sun position relative to the pointing when that data was taken. We see that if the sun is in one of the 4 main far sidelobe peaks there is a large excess in the average

Table 2.2: List of statistics in the statistics database and a short description of each.

Statistic label	Description	Statistic label	Description
mjd	mean MJD of scan	night	distance in h from 2 AM (UTC - 7)
sidereal	sidereal time in degrees (up to a phase)	az	mean azimuth of scan
el	mean elevation of scan	chi2	χ^2 statistic for all timestreams of whole sb
acceptrate	acceptance rate of sideband	az_chi2	χ^2 of azimuth binned timestreams
max_az_chi2	max χ^2 of az binned single freq TODs	med_az_chi2	median χ^2 of az binned single freq TODs
fbit	feature-bit of scan (indicates scanning mode)	az_amp	average amplitude of fitted az-template
el_amp	average amplitude of fitted el-template	n_spikes	number of spikes
n_jumps	number of jumps	n_anomalies	number of anomalies
n_nan	number of nan-samples	tsys	average T_{sys} value of scan
pca1	average variance of removed PCA mode 1	pca2	average variance of removed PCA mode 2
pca3	average variance of removed PCA mode 3	pca4	average variance of removed PCA mode 4
weather	estimated probability of bad weather	kurtosis	kurtosis of timestreams
skewness	skewness of timestreams	scan_length	length of scan in minutes
saddlebag	saddlebag number (1-4)	sigma_poly0	σ of mean in poly filter
fknee_poly0	f_{knee} of mean in poly filter	alpha_poly0	α of mean in poly filter
sigma_poly1	σ of slope in poly filter	fknee_poly1	f_{knee} of slope in poly filter
alpha_poly1	α of slope in poly filter	power_mean	mean power of sideband mean
sigma_mean	σ of sideband mean	fknee_mean	f_{knee} of sideband mean
alpha_mean	α of sideband mean	airtemp	hk: air temp, C
dewtemp	hk: dewpoint temp, C	humidity	hk: relative humidity, (0-1)
pressure	hk: pressure, millibars	rain	hk: rain today, mm
winddir	hk: az from where wind is blowing, deg	windspeed	hk: windspeed m/s
moon_dist	distance to moon in deg	moon_angle	az of moon relative to pointing
moon_cent_sl	moon close to central sidelobe	moon_outer_sl	moon close to outer (feedleg) sidelobes
sun_dist	distance to sun in deg	sun_angle	az of sun relative to pointing
sun_cent_sl	sun close to central sidelobe	sun_outer_sl	sun close to outer (feedleg) sidelobes
sun_el	elevation of sun	ps_chi2	χ^2 excess in power spectrum (old)
ps_s_sb_chi2	ps_chi2 for single sb single scan	ps_s_feed_chi2	ps_chi2 for single feed single scan
ps_s_chi2	ps_chi2 for all feeds for single scan	ps_o_sb_chi2	ps_chi2 for single sb full obsid
ps_o_feed_chi2	ps_chi2 for single feed full obsid	ps_o_chi2	ps_chi2 for all feeds for full obsid
ps_z_s_sb_chi2	ps_chi2 for avg of z-direction 1D ps	ps_xy_s_sb_chi2	ps_chi2 for avg of xy-direction 2D ps
sw_01	standing wave magnitude $k \approx 0.012\,\mathrm{Mpc}^{-1}$	sw_02	standing wave magnitude $k \approx 0.017\,\mathrm{Mpc}^{-1}$
sw_03	standing wave magnitude $k \approx 0.025\,\mathrm{Mpc}^{-1}$	sw_04	standing wave magnitude $k \approx 0.036\,\mathrm{Mpc}^{-1}$
sw_05	standing wave magnitude $k \approx 0.051\,\mathrm{Mpc}^{-1}$	sw_06	standing wave magnitude $k \approx 0.073\,\mathrm{Mpc}^{-1}$
sw_07	standing wave magnitude $k \approx 0.10\,\mathrm{Mpc}^{-1}$	sw_08	standing wave magnitude $k \approx 0.15\,\mathrm{Mpc}^{-1}$
sw_09	standing wave magnitude $k \approx 0.21\,\mathrm{Mpc}^{-1}$	sw_10	standing wave magnitude $k \approx 0.30\,\mathrm{Mpc}^{-1}$
sw_11	standing wave magnitude $k \approx 0.44\,\mathrm{Mpc}^{-1}$	sw_12	standing wave magnitude $k \approx 0.62\,\mathrm{Mpc}^{-1}$
sw_13	standing wave magnitude $k \approx 0.89\,\mathrm{Mpc}^{-1}$	sw_14	standing wave magnitude $k \approx 1.3\,\mathrm{Mpc}^{-1}$

ps_chi2, indicating clear excess power during these observations. Although we had not anticipated the severity of this effect in advance, the statistics database lets us see this effect very clearly, and it also makes it much easier to reject the relevant part of the data once it has been noticed. The bottom part of the figure shows the sun-sidelobe mask used to reject data that is close to these far sidelobes, as well as the azimutally symmetric main sidelobes.

Another important effect that we have worked a lot on dealing with is weather. As we discussed in Sect. 2.4.8 weather can affect the data significantly. We have therefore implemented a weather classifier using a deep neural network, to detect weather contamination. This work was lead by master student Maren Rasmussen, under my supervision. Her results classified about 38% of the data as affected by bad weather, and this data is thus rejected. For more details and results regarding the weather classifier see Rasmussen, 2020.

2.7.2 Acceptance rates

With the help of the statistics database we set various thresholds for the different statistics, and decide what data to accept or reject based on the values of these statistics.

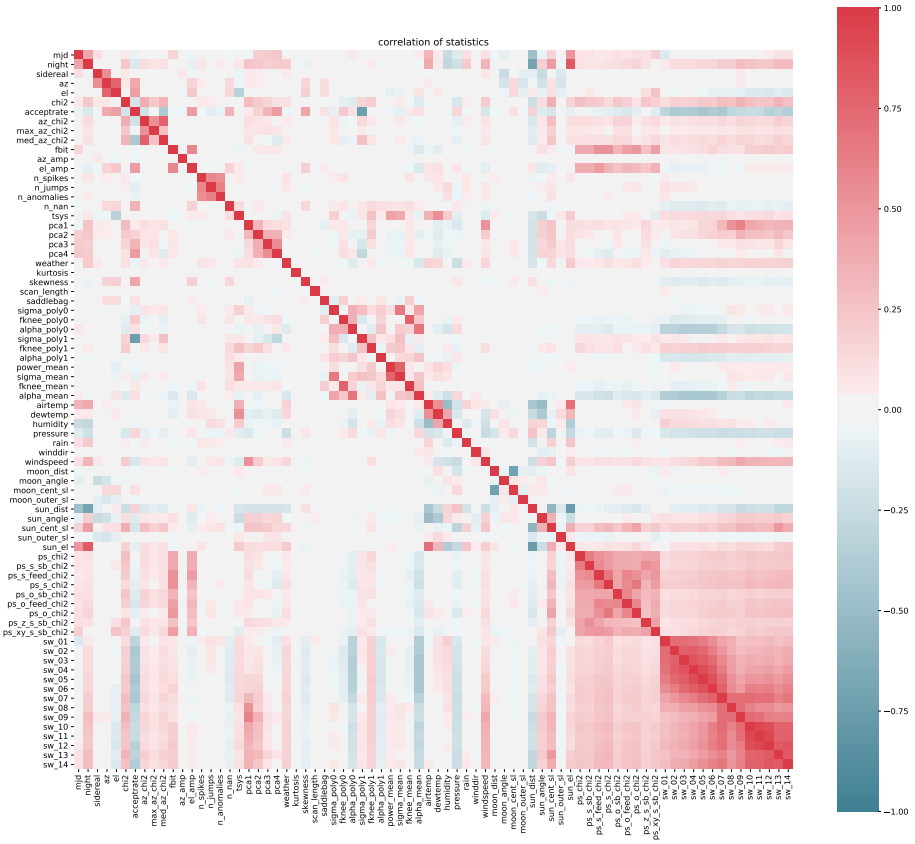

Figure 2.22: Correlation between the different diagnostic variables in the statistic database.

Figure 2.24 shows the fraction of data left over after each cut is applied for the co7 field. For the first year, we lost all the data from feeds 4, 6 and 7, due to very strong systematics, so we lose 3 out of 19 parts of the data right out of the box. After the frequency cuts in `l2gen` we are down to a bit more than 60 % of the data. After observing for close to a year, we understood that data taken at higher than 65 degrees or lower than 35 degrees in elevation was contaminated by ground signal, so we lost a significant fraction of our data from the co6 and co7 fields. After the elevation cut, we see that we are down to a bit more than 40 % of the original data volume. After this, a bunch of smaller cuts due to e.g. weather, large T_{sys} values or the Sun in the sidelobe, we are down to a bit more than 20 %. Now we do the cuts on the `ps_chi2` which gets us down to the roughly 14 % of data. This is the fraction of data that goes into the maps. The corresponding acceptance rates for co2 and co6 is given by about 22 % and 14 %, respectively.

As we see, with the current cuts, only a fraction of the original data volume is retained. We have been deliberately conservative in these cuts, in order to make sure

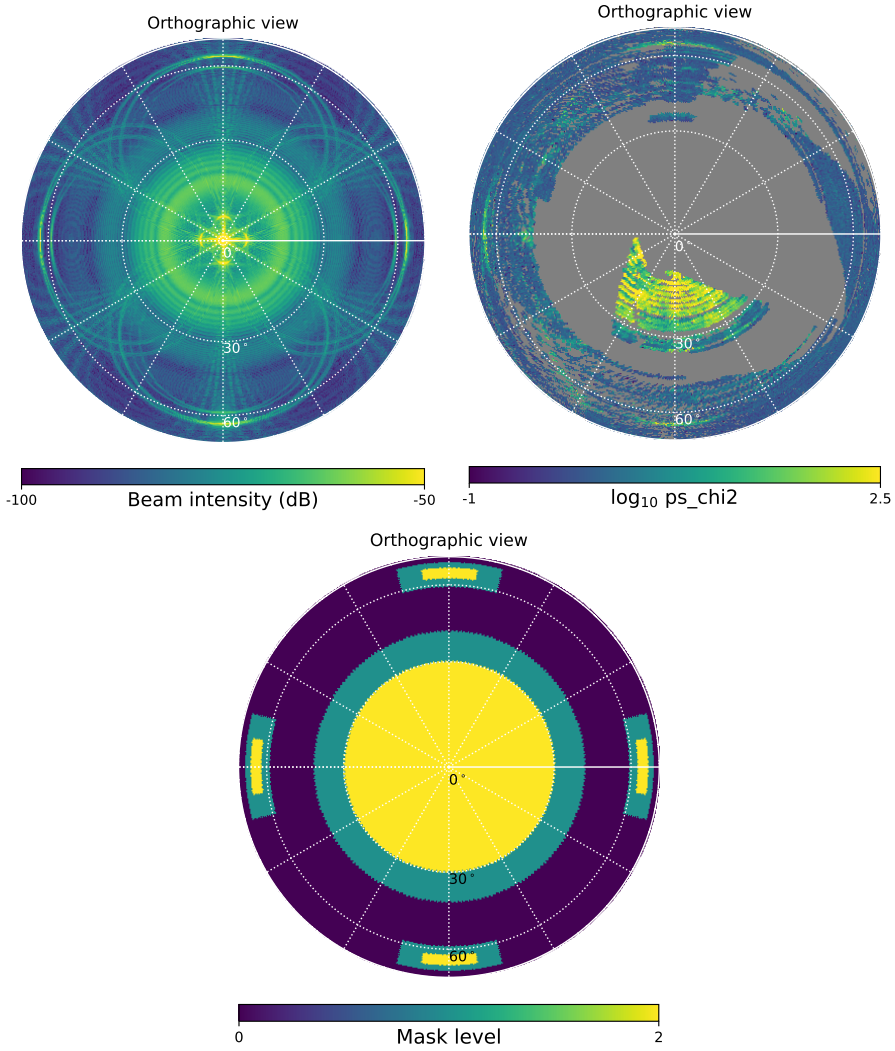

Figure 2.23: Simulated beam model for a single feed (top left), the derived statistic `ps_chi2` binned according to the position of the sun relative to the pointing when the data was taken (top right) and far sidelobe sun mask used for data selection (bottom).

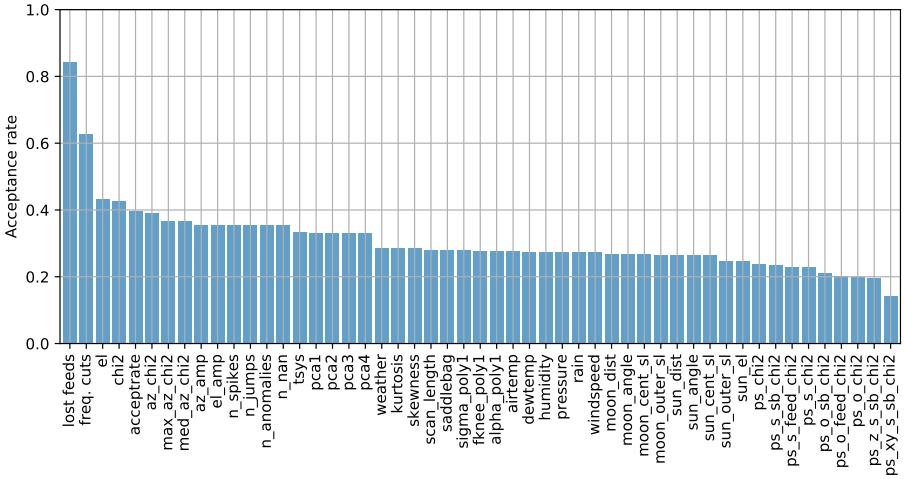

Figure 2.24: Acceptance rate of the data after each individual cut is applied for all the co7 data.

that only the best data makes it through. Once we understand the data better, and have experimented more with different cuts, we will hopefully be able to use more of the data. We now only observe between 35 and 65 degrees in elevation, so we will not lose more data to the elevation cuts in the future, and the lost feeds from the first year have been fixed, so we will be in a significantly better situation for our second year of data. With new improvements in the pipeline, we also hope to clean the data better during `l2gen`, allowing us to use more of the data still. With all this in mind we hope that we can at least increase the fraction of the data used by a factor of 2 or 3 from our current level.

2.7.3 Accept list

All the different data cuts result, in the end, in the accept list, which tells the mapmaker which parts of the data to include or not include when making the maps. The accept list is a boolean array of dimension $(n_{\text{scan}}, n_{\text{feed}}, n_{\text{sideband}})$, which stores the value "True" if the data from a given sideband on a given feed of a given scan is to be included in the map. This allows us to play around with more or less strict data cuts when we make the maps, without having to run the very expensive low level time ordered data filtering.

`accept_mod` can also split the data in any way you specify. This essentially runs the mapmaker independently for the two (or more) different sets of data and produces two (or more) different maps. This is very useful for the cross-spectrum methods we are using (see Paper II (Ihle et al., in prep) for details), and when doing null-tests, jack-knifes etc.

2.8 Inference from line intensity maps

The ultimate goal of observational astronomy and cosmology is to learn about the universe, so the final question we need to ask is how can we infer anything about astrophysics and cosmology from a line intensity map.

This problem is in general very hard. The physics involved in producing the CO signal we are looking for involves many different processes spanning so many different length scales, from large-scale clustering of dark matter, to the physics, chemistry and radiative transfer of cold molecular clouds. On the other hand, this means that there is potentially a lot of information, about all these different processes, in the map, if we can only figure out how to extract it.

2.8.1 Modeling CO line emission

Cosmological simulations of dark matter (DM) only are vastly cheaper to run than simulations including baryonic physics as well. This is because dark matter is only affected by gravity, while baryons have much more complicated interactions. Baryons form structure all the way down to stellar scales, and things like stellar feedback means that what happens at small scales affects what happens at every other scale, especially during reionization. This coupling between scales means that cosmological scale simulations including baryons cannot be run at the resolution required to include all relevant physics.

To include baryonic effects in a cosmological simulations you therefore typically need to employ a model to take into account the physics that happens at smaller scales than you can resolve. Such sub-grid models can then be calibrated using smaller scale simulations and observations, so that it captures as much as possible of the features of the small scale physics that are relevant for the large scale physics.

2.8.1.1 Dark matter halo based models

A simple approach to sub-grid modeling for CO line intensity, is to model the connection between each dark matter halo and the CO line luminosity of the galaxy (or multiple galaxies) within that halo. This way you can run DM only simulations and then replace each DM halo in the simulation with a point source of a given luminosity, which can then be turned into a 3D intensity map, by slicing the DM simulation along a lightcone.

Perhaps the simplest possible such DM halo-$L_{\rm CO}$ model would be to assume a linear relation between the DM halo mass, $M_{\rm halo}$, and the CO luminosity, $L_{\rm CO}$. This would be a one-parameter model with the amplitude of the linear relation being the only free parameter.

A generalization of this model would be a general parameterized relation between $M_{\rm halo}$ and $L_{\rm CO}$

$$L_{\rm CO}(M_{\rm halo}, \theta_i), \tag{2.39}$$

where θ_i is a set of free variables parametrizing the relation between $M_{\rm halo}$ and $L_{\rm CO}$. A more advanced model could also take into account other properties of the DM halos, like the virial velocity, the formation time or the merger history of the halo. It could

also predict not only the CO luminosity, but also for example the line width of the CO emission from each halo.

2.8.1.2 The Li et al., 2016 model

One of the most used models for CO line emission at intermediate redshifts ($z \sim 2-3$) the last years is the model developed in Li et al. (2016). This is a parametric model relating DM halo masses, $M_{\rm halo}$, star formation rates (SFR), infrared (IR) luminosities, $L_{\rm IR}$, and CO luminosities, $L_{\rm CO}$, using five free parameters. The model utilizes the relations between average SFR from DM halo masses derived in Behroozi, Wechsler, and Conroy (2013a) and Behroozi, Wechsler, and Conroy (2013b), and adds an additional log-normal scatter to the resulting SFR, determined by the parameter, $\sigma_{\rm SFR}$.

Given a SFR, the IR luminosity is then found using the linear relation

$$\mathrm{SFR} = \delta_{\rm MF} \times 10^{-10} L_{\rm IR}. \tag{2.40}$$

Further, CO-luminosities are approximated by the following relation

$$\log L_{\rm IR} = \alpha L'_{\rm CO} + \beta, \tag{2.41}$$

before a final step of log-normal scatter is added determined by the parameter, $\sigma_{L_{\rm CO}}$. We then get a total of five parameters $\theta = \{\log \delta_{\rm MF}, \alpha, \beta, \sigma_{\rm SFR}, \sigma_{L_{\rm CO}}\}$, determining the function $L_{\rm CO}(M_{\rm halo})$. See Li et al. (2016) for a detailed discussion of the physical and observational motivation for this model.

2.8.1.3 Simulations

Another approach to modeling CO line intensity is to use smaller hydrodynamical simulations (Hopkins et al., 2014; Schaye et al., 2014; Vogelsberger et al., 2014). These resolve more of the physics related to molecular clouds, star-formation, feedback etc, but they typically cover a much smaller volume than what is observed with COMAP. In order to do forecasts, or use these simulations to interpret the results from a CO intensity map, we would need to somehow extrapolate these results to the full volume, which is hard to do in a statistically rigorous way. For some work along these lines, see Lakhlani (2019) or M. Silva et al. (in prep).

2.8.2 Observables

When doing inference from a line intensity map, the general approach is to have some kind of astrophysical and cosmological model which predicts how we would expect the cosmological signal to look like, and then compare this to the observed map. Of course, we don't expect the model to predict the specific spatial distribution of the CO signal, since that depends on the random initial conditions. Rather, we expect the model to predict the statistical properties of the ensemble that any observed CO intensity map is a specific realization of. This means that instead of directly comparing the map itself to some model prediction, we typically calculate some summary statistics, or observables, from the map, which hopefully captures the statistical information in the map. We can

then try to derive predictions for these different observables from the model, and then compare these predictions to the values derived from the observed map.

In CMB analysis, and (to some degree) analysis of large scale structure, where the signal is extremely close to a Gaussian field, the power spectrum contains all the statistical information in the map. A line intensity map, however, is often significantly non-Gaussian, which means that new statistics need to be developed to extract all the information present in the map.

2.8.2.1 Power spectrum

The power spectrum, $P(\boldsymbol{k})$, sets the variance of the Fourier components of a map

$$P(\boldsymbol{k}) = \frac{V_{\rm vox}}{N_{\rm vox}} \langle |f_{\boldsymbol{k}}|^2 \rangle, \tag{2.42}$$

where k is the wave vector of a given Fourier component, $V_{\rm vox}$ is the volume of each voxel of the map and $N_{\rm vox}$ is the total number of voxels in the map. A voxel is the 3D equivalent of a pixel, it is the small volume element defined by the resolution used for the 3D map.

If the Fourier components have a Gaussian distribution, then the map is called a Gaussian map, in which case the power spectrum fully describes the statistical ensemble that a given map is a single realization from. In the more general case, where the map is not Gaussian, the power spectrum still carries a lot of important information about the map, even if it does not describe all the information.

The cross spectrum, $C(\boldsymbol{k})$, sets the covariance between the Fourier components of two different maps, m_1 and m_2

$$C(\boldsymbol{k}) = \frac{V_{\rm vox}}{N_{\rm vox}} \left\langle \mathrm{Re}\{f^*_{1\boldsymbol{k}} f_{2\boldsymbol{k}}\} \right\rangle \tag{2.43}$$

where f_1 and f_2 are the Fourier components of the two maps and where Re{} denotes the real part of a complex number or function.

For more details on different power spectrum methods and how to robustly estimate the signal power spectrum from an observed line intensity map see Paper II (Ihle et al., in prep). For other recent examples of the use power spectrum analysis on intensity mapping data see e.g. (Keating, Marrone, Bower, and Keenan, 2020; Mertens et al., 2020; Uzgil et al., 2019).

2.8.2.2 Voxel Intensity Distribution (VID)

A useful statistic that is complementary to the power spectrum is the Voxel Intensity distribution (VID) (P. Breysse et al., 2017). The VID, $\mathcal{P}(T)$, is the probability distribution of voxel brightness temperatures. This observable does not use any of the spatial information in the map, but is more sensitive to the bright end of the underlying CO luminosity function and small-scale clustering.

A natural observable related to the VID is the histogram of voxel temperatures, B_i. The expectation value of these are given by the VID

$$\langle B_i \rangle = N_{\rm vox} \int_{T_i}^{T_{i+1}} \mathcal{P}(T) dT, \tag{2.44}$$

where B_i is the number of voxels with a brightness temperature between T_i and T_{i+1} and $N_{\rm vox}$ is the total number of voxels.

2.8.2.3 Pseudo-VID

In the same way as for the auto-spectrum, when a line intensity map has an uneven distribution of noise, then the VID statistic will be dominated by the highest noise parts of the map, so unless you have a large signal to noise, this will affect the usefulness of the statistic significantly. We therefore want to think of a statistic that preserves the nice properties of the VID, but that is less susceptible to uneven noise distribution in the map.

In order to investigate this let us use the following simplified data model

$$d = n + s, \tag{2.45}$$

where d is a vector of the measured temperature in each voxel, n is a vector of the noise in each voxel and s is the signal in each voxel. For simplicity we assume that the noise level, $\sigma_{\rm vox}$, in each voxel is known and that the signal is independent between different voxels.

Let us then define the pseudo-VID as the VID of the following data

$$\tilde{d} = \frac{d}{\sigma_{\rm vox}^{\alpha}}, \tag{2.46}$$

where α is some positive power that we will try to find an optimal value for.

If the signal was just an overall constant number, then $\alpha = 2$ is gives us the maximum likelihood estimate of s, so this is perhaps not a bad initial guess for what value of α to choose. However, this is not a very realistic model of the signal. Another simple model for the signal, which is more realistic as a toy model, is to assume that the signal is an independent Gaussian with a constant standard deviation[4], $\sigma_{\rm signal}$. We can then try to find what is a good value for α to get the best signal to noise estimate of the signal in this model.

We can make a mock observation by generating noise and signal for 10^6 voxels and dividing the data by $\sigma_{\rm vox}^{\alpha}$ and then binning it to obtain the pseudo bin counts, $B_i(\tilde{d})$. The pseudo bin counts follow (under our assumptions) a binomial distribution, and we can thus map out the likelihood of the data, $P(\tilde{d}|\sigma_{\rm signal})$, by using many simulations to determine the expectation, $\langle B_i(\tilde{d})\rangle$, for each value of $\sigma_{\rm signal}$. We can then examine the width of this likelihood as a function of $\sigma_{\rm signal}$ and use this as an estimate of the sensitivity of the pseudo-VID statistic for a given value of α.

Figure 2.25 shows the sensitivity of the pseudo-VID statistics for different values of the exponent, α. These results are obtained in the noise dominated limit (with the signal component of the mock data set to zero), and we show results for different noise level distributions to investigate the effect of the distribution of the noise levels has on the optimal value for α. We find that values in the range $\alpha = 1 - 2$ give consistent good results. Figure 2.26 shows the results from the same analysis using noise level

[4]My work on this toy model is building on some previous work by Patrick Breysse on these issues.

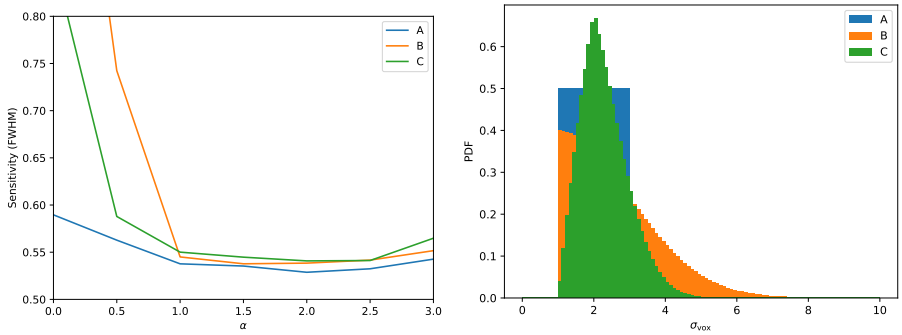

Figure 2.25: Sensitivity of the pseudo-VID statistic for different exponents, α, (left) and the three different noise level distributions used (right).

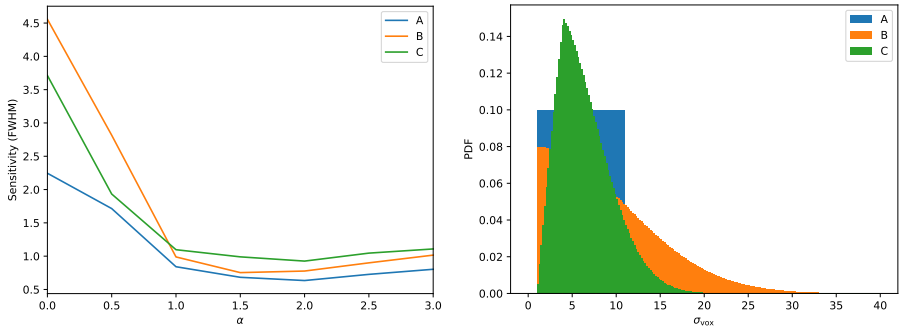

Figure 2.26: Sensitivity of the pseudo-VID statistic for different exponents, α, (left) and the three different noise level distributions used (right).

distributions with significantly more variance, as can be seen in the right panel. The optimal values for α found here are fairly consistent in all cases, suggesting that values between $\alpha = 1 - 2$ are good values for the pseudo-VID in a large range of different noise level distributions.

2.8.3 Likelihood

We can combine all observables into a data vector, d_i. For the case of using the PS and the VID we get

$$d_i = (P_{k_i}, B_i). \tag{2.47}$$

If all the components of d_i were independent, they would (under some simplifying assumptions) have the following variance, which we will denote as the *independent variance*

$$\mathrm{Var_{ind}(P_{k_i})} = \langle P_{k_i} \rangle^2 / N_{\mathrm{modes}}, \tag{2.48}$$

$$\mathrm{Var_{ind}(B_i)} = \langle B_i \rangle. \tag{2.49}$$

2.8.3.1 Gaussian likelihood

Because of spatial clustering, there typically are correlations between the different elements of the data vector. We can take this into account using a full covariance matrix

$$\xi_{ij} = \mathrm{Cov}(d_i, d_j). \tag{2.50}$$

We can then build a Gaussian likelihood of the form (up to a constant)

$$-2\ln P(d|\theta) = \frac{N_s}{N_s+1}\sum_{ij}[d_i - \langle d_i\rangle](\xi^{-1})_{ij}[d_j - \langle d_j\rangle] + \ln|\xi|, \tag{2.51}$$

where $\langle d\rangle(\theta)$ and $\xi(\theta)$ are the mean values and covariance matrix of the observables d_i for specific parameters θ. N_s is the number of simulations used to estimate $\langle d\rangle$, and the factor $N_s/N_s + 1$ takes into account the effect of the uncertainty in the estimate of $\langle d\rangle$.

2.8.3.2 Binomial VID likelihood

In the previous section, we implicitly assumed that all the bin counts, B_i, are Gaussian random variables, with some covariance matrix. However, as we already know, even if the samples are completely independent, the bin counts will follow a binomial distribution. When the expected bin counts are high ($B_i \gg 5$), the binomial distribution can be very well modeled by a Gaussian. When we only expect a few voxels in each bin, however, the Gaussian approximation breaks down.

If we know that the samples are independent, there is no problem, since the likelihood is just given by the joint distribution of n_{bins} independent binomial variables

$$P(B|\theta_l) = \prod_{i=1}^{n_{\mathrm{bins}}}\binom{N_{\mathrm{samples}}}{B_i}p_i^{B_i}(1-p_i)^{N_{\mathrm{samples}}-B_i}, \tag{2.52}$$

where $p_i \equiv p_i(\theta_l) = \langle B_i\rangle(\theta_l)/N_{\mathrm{samples}}$ is the probability of a voxel being placed in bin i. As we have argued, however, when we are not completely noise dominated, the samples will not be independent, and we have no general way to calculate the likelihood. On the other hand, we expect the clustering effect to be dominated by the shot noise when we get to bins with very few samples.

This suggests a possible approximation we can make. We can try to separate the bins according to the expected number of samples. For bins where we expect many samples, $B_i > 5$, we can assume that they are approximately Gaussian, and use simulations to estimate the covariance matrix between these bins. For the bins with few expected samples $B_i \leq 5$ we can assume that they follow independent binomial distributions, because the variance will presumably be dominated by shot noise, and clustering effects are (hopefully) negligible. The full likelihood, in this approximation, is then given by

$$P(B|\theta_l) = P_{\mathrm{Binomial}}(B|\theta_l)P_{\mathrm{Gaussian}}(B|\theta_l), \tag{2.53}$$

where

$$P_{\mathrm{Binomial}}(B|\theta_l) = \prod_{\{i|B_i\leq 5\}}^{n_{\mathrm{bins}}}\binom{N_{\mathrm{samples}}}{B_i}p_i^{B_i}(1-p_i)^{N_{\mathrm{samples}}-B_i}, \tag{2.54}$$

and

$$P_{\text{Gaussian}}(B|\theta_l) = \frac{1}{(2\pi)^{\frac{n_{\text{bins}}}{2}}|\xi_{>5}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\chi^2_{>5}\right), \tag{2.55}$$

and where

$$\chi^2_{>5} \equiv \sum_{\{i,j|B_i,B_j>5\}} [B_i - \langle B_i\rangle(\theta_l)](\xi^{-1})_{ij}[B_j - \langle B_j\rangle(\theta_l)] \tag{2.56}$$

and

$$\xi_{>5} \equiv \xi_{\{i,j|B_i,B_j>5\}}. \tag{2.57}$$

One issue associated with bins with very few expected samples is that if you are estimating p_i from some limited number of simulations, you will sometimes have no samples in the bin from either simulation. In that case it is not clear how to evaluate the likelihood. This problem can occur quite frequently when there is not that much signal to noise in the data, and you are exploring a large region in parameter space. We can deal with this problem by imposing some prior on the expectation value of the binomial distribution, and then update this prior with the values from the simulation. The details for how to do this are discussed in App. A.1.

2.8.4 Exploring the posterior distribution

Using the methods we have sketched in the previous sections we can now try to put them all together. To summarize what we need to do this, here is a list:

- A set of observables, d_i, that can be estimated from a line intensity map.
- A model with parameters θ, which can generate simulated intensity maps, or at least in some way calculate the mean observables, $\langle d_i\rangle(\theta)$, covariance matrix, $\xi_{ij}(\theta)$ (or any other quantity needed in the likelihood) for any given values of θ. Any prior on θ is also assumed to be part of the model.
- A likelihood $P(d_i|\theta)$, that can be evaluated using the model output.
- Some observed (or simulated) line intensity mapping data, summarized in the observables data, d_i^{observed}.

We can then use a sampling procedure, like a Markov-Chain Monte Carlo (MCMC) method, or other methods to explore the posterior distribution $P(\theta|d_i^{\text{observed}})$.

At this abstract level this looks fairly simple, however, the details matter a lot here, and can make things much more complicated. There are several important questions about the model and the parameter space θ. How degenerate are the parameters θ? And how easy are the parameters to interpret? A given set of observed data d_i^{observed} is often good at constraining one combination of parameters, but is completely insensitive to other combinations of parameters, so if you make bad choices of parameters, or simply just use too many parameters, you will not be able to use the data to constrain much. If you make a model that is fairly closely related to the physics, and where each parameter has a simple interpretation, then you typically get a complicated model, with lots of parameters, and also lots of degeneracy between the parameters. A simple

model, however, can avoid any unnecessary degeneracies, but it will in turn often be much harder to connect these parameters back to cosmology and astrophysics. A good model will be one that finds a good compromise between these considerations, for the particular dataset you are considering. For a detailed discussion of these issues and more see D. Chung et al., in prep.

Another set of issues relate to the observables, d_i, and the likelihood. In Ihle et al. (2019) we show that using a combination of the power spectrum and the voxel intensity distribution is better at constraining the CO luminosity function than using either observable individually, for a COMAP-like experiment. While this was a very useful finding, there are many simplifying assumptions made in these (and similar types of) forecasts. If we want to do a similar analysis on actual data from an experiment, we will have to address many issues first. How do you model or estimate the noise in the data? How do you deal with an uneven noise distribution in the map? How do you build a good likelihood function? How do you take into account systematic effects and propagate systematic uncertainties into the likelihood? Some of these issues are discussed in detail for the power spectrum in Paper II (Ihle et al., in prep), and we have made some preliminary investigation into some issues with the VID here, in Sects. 2.8.2.3 and 2.8.3.2. However, we still have quite a way yet to go in order to use the VID on real data.

2.8.5 Future Prospects

CO intensity mapping is in a somewhat special position in that there is a fairly low level of foregrounds, and no significant interloper lines (D. Chung et al., 2017). This makes the analysis and inference from these maps much simpler.

CII intensity mapping, for example, has several interloper CO-lines from lower redshifts. This makes the inference much more complicated (Cheng, Chang, and Bock, 2020), but significantly increases the amount of information in the maps.

As the field matures, more areas of the sky, and more ranges in redshift, will be covered by different experiments, meaning that we will have access to the same galaxy population and large scale structure from multiple tracers. This will allow us to get more robust results, break degeneracies and give unique insights into the astrophysics of galaxies and their chemistry and gas dynamics.

Cross-correlation (D. T. Chung et al., 2019) as well as new tools being developed (Bernal, P. C. Breysse, Gil-Marin, et al., 2019; P. C. Breysse, Anderson, and Berger, 2019; Cheng, Chang, and Bock, 2020; D. T. Chung, 2019; Gong, Chen, and Cooray, 2020; Yang et al., 2020), will be needed to deal with these combined datasets. As we get more different datasets that we want to analyze together the more important it becomes to use a consistent statistical framework for inference, like we use in Paper III (Ihle et al., 2019) and like we have discussed here.

The prospects for doing fundamental cosmology with LIM is very interesting. Getting access to the largely unmeasured large scale modes from the early universe will fill in many gaps in our knowledge of the expansion and evolution of the universe (Bull et al., 2015; Creque-Sarbinowski and Kamionkowski, 2018; Dinda, Sen, and Choudhury, 2018; Bernal, P. C. Breysse, and E. D. Kovetz, 2019; Furlanetto et al., 2019; Liu and

P. C. Breysse, 2020). The ultimate dream, of course, is going to space (M. B. Silva et al., 2019).

Chapter 3

BeyondPlanck

BeyondPlanck (BeyondPlanck Collaboration, 2020) is an ambitious project to develop an end-to-end Bayesian data analysis pipeline for CMB experiments, combining everything from low-level instrumental parameters like gains and bandpasses to foreground components and CMB power spectra into one consistent statistical model. As a demonstration, we apply this framework to the Planck Low Frequency Instrument (LFI) time ordered data. Here I will give a short summary of the general BeyondPlanck approach, and of the two papers that I have been most involved in, on noise modeling (Paper V, Ihle et al., 2020) and calibration (Paper VI, Gjerløw et al., 2020).

3.1 The BeyondPlanck approach

The first part of the BeyondPlanck approach is to define an explicit statistical model of the data, including all the components and effect that we want to take into account. In BeyondPlanck, for each of the three LFI bands, at frequencies of 30, 44, and 70 GHz, we model the raw time-domain data, $\boldsymbol{d}$, as follows,

$$d_{j,t} = g_{j,t}\mathsf{P}_{tp,j}\left[\mathsf{B}^{\mathrm{symm}}_{pp',j}\sum_c \mathsf{M}_{cj}(\beta_{p'}, \Delta^j_{\mathrm{bp}})a^c_{p'} + \mathsf{B}^{\mathrm{asymm}}_{pp',j}\left(s^{\mathrm{orb}}_{j,t} + s^{\mathrm{fsl}}_{j,t}\right)\right] + n^{\mathrm{corr}}_{j,t} + n^{\mathrm{w}}_{j,t}. \quad (3.1)$$

The subscript t is the index in time domain; j is the radiometer index; p is the pixel index; c is the signal component index; $\boldsymbol{g}$ is the gain; P is the pointing matrix; $\mathsf{B}^{\mathrm{symm}}$ and $\mathsf{B}^{\mathrm{asymm}}$ is the symmetric beam matrix and the asymmetric beam matrix, respectively; $\boldsymbol{a}$ are the astrophysical signal amplitudes; $\boldsymbol{\beta}$ are the corresponding spectral parameters; Δ_{bp} are the bandpass corrections; M_{cj} is the bandpass-dependent component mixing matrix; $\boldsymbol{s}^{\mathrm{orb}}$ is the orbital dipole; $\boldsymbol{s}^{\mathrm{fsl}}$ are the far sidelobe corrections; $\boldsymbol{n}^{\mathrm{corr}}$ is the correlated noise; and $\boldsymbol{n}^{\mathrm{w}}$ is the white noise. For more details and discussion of each of these parameters, and how they are modeled, see BeyondPlanck Collaboration (2020) and companion papers.

Given this model, the goal of the Bayesian approach is to explore the full posterior distribution of all the free parameters in the model, given the observed data, $\boldsymbol{d}$,

$$P(\boldsymbol{g}, \boldsymbol{n}_{\mathrm{corr}}, \xi_n, \Delta_{\mathrm{bp}}, \boldsymbol{a}, \beta, C_\ell \mid \boldsymbol{d}). \quad (3.2)$$

This is a really complicated distribution, with a very large number of parameters that depend on each other in complicated ways. The way we deal with this distribution is to use Gibbs sampling. The idea of Gibbs sampling is simple. Instead of sampling directly from the full distribution, we sample iteratively from each of the conditional distributions. In this way we can divide the complex problem of sampling from the whole distribution into a set of smaller manageable steps. We can summarize this

process as follows,

$$\begin{aligned}
\boldsymbol{g} &\leftarrow P(\boldsymbol{g} \quad | \boldsymbol{d}, \quad \xi_n, \Delta_{\rm bp}, \boldsymbol{a}, \beta, C_\ell) && (3.3)\\
\boldsymbol{n}_{\rm corr} &\leftarrow P(\boldsymbol{n}_{\rm corr} | \boldsymbol{d}, \boldsymbol{g}, \quad \xi_n, \Delta_{\rm bp}, \boldsymbol{a}, \beta, C_\ell) && (3.4)\\
\xi_n &\leftarrow P(\xi_n \quad | \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\rm corr}, \quad \Delta_{\rm bp}, \boldsymbol{a}, \beta, C_\ell) && (3.5)\\
\Delta_{\rm bp} &\leftarrow P(\Delta_{\rm bp} \quad | \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\rm corr}, \xi_n, \quad \boldsymbol{a}, \beta, C_\ell) && (3.6)\\
\beta &\leftarrow P(\beta \quad | \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\rm corr}, \xi_n, \Delta_{\rm bp}, \quad C_\ell) && (3.7)\\
\boldsymbol{a} &\leftarrow P(\boldsymbol{a} \quad | \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\rm corr}, \xi_n, \Delta_{\rm bp}, \quad \beta, C_\ell) && (3.8)\\
C_\ell &\leftarrow P(C_\ell \quad | \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\rm corr}, \xi_n, \Delta_{\rm bp}, \boldsymbol{a}, \beta \quad). && (3.9)
\end{aligned}$$

Here, $\leftarrow$ indicates sampling from the distribution on the right-hand side. The challenge then is to implement methods to sample from each of the conditional distributions, and to put everything together into a computational framework.

Perhaps the simplest way to illustrate the appeal of the full end-to-end Bayesian approach of BeyondPlanck is shown in Fig. 3.1. Here we see constraints on the optical depth to reionization, τ, derived in four different cases. The blue curve, labeled "WN", corresponds to fixing all the other parameters in the model, and only taking into account the uncertainty coming from the white noise. The red curve, labeled "TOD + WN", corresponds to fixing the foreground model, but marginalizing over all the instrumental parameters and the correlated noise in addition to the white noise. The green curve, labeled "FG + WN", marginalizes over the foreground parameters and the white noise, but fixes the instrumental parameters and the correlated noise, while the black curve marginalizes over all the parameters, in addition to the white noise. This figure illustrates how important it is to propagate the uncertainties from all the instrumental parameters and the foregrounds simultaneously, and in a statistically rigorous way, which is exactly what the BeyondPlanck pipeline was designed for.

3.2 Noise Modelling and Characterization

Here we will discuss two of the sampling steps, 3.4 and 3.5, from the full BeyondPlanck Gibbs chain in more detail.

The starting point for both conditional distributions is the following parametric data model,

$$\boldsymbol{d}' \equiv \boldsymbol{d} - g\boldsymbol{s}^{\rm tot} = \boldsymbol{n}^{\rm corr} + \boldsymbol{n}^{\rm wn}. \tag{3.10}$$

where we have defined the signal subtracted data, $\boldsymbol{d}'$, and where $\boldsymbol{d}$ is the raw time ordered data (TOD); g is the gain; $\boldsymbol{s}^{\rm tot}$ describes the total sky signal, comprising both CMB and foregrounds, projected into time-domain; $\boldsymbol{n}^{\rm corr}$ represents the correlated noise in time domain; and $\boldsymbol{n}^{\rm wn}$ is white noise.

We assume that both the two noise terms are distributed as a Gaussian, with the following covariance matrices $\mathsf{N}_{\rm corr} \equiv \langle \boldsymbol{n}_{\rm corr}\boldsymbol{n}_{\rm corr}^T\rangle$ and $\mathsf{N}_{\rm wn} \equiv \langle \boldsymbol{n}_{\rm wn}\boldsymbol{n}_{\rm wn}^T\rangle$, respectively. The complete noise power spectral density (PSD) is then given by

$$P(f) = \mathsf{N}_{\rm wn} + \mathsf{N}_{\rm corr} = \sigma_0^2 + \sigma_0^2\left(\frac{f}{f_{\rm knee}}\right)^\alpha, \tag{3.11}$$

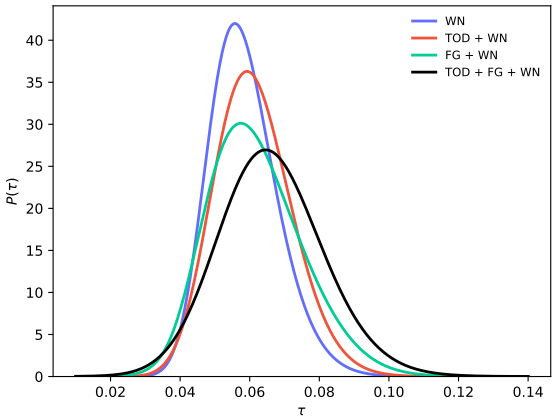

Figure 3.1: Posterior distribution derived for the optical depth to reionization, τ, marginalizing over different combinations of variables. Figure taken from (Paradiso et al., 2020).

where f denotes temporal frequency; σ_0 is the white noise level of the time-ordered data; α is the slope (typically negative) of the correlated noise PSD; and the knee frequency, $f_{\rm knee}$, denotes the (temporal) frequency at which the variance of the correlated noise is equal to the white noise variance. The three PSD parameters are collectively denoted $\xi_n = \{\sigma_0, f_{\rm knee}, \alpha\}$. This noise model is usually referred to as $1/f$-noise (Planck Collaboration II, 2014; Planck Collaboration II, 2020; Tauber et al., 2019).

We can sample the correlated noise, $\boldsymbol{n}^{\rm corr}$, by solving the following equation,

$$\left(\mathsf{N}_{\rm corr}^{-1} + \mathsf{N}_{\rm wn}^{-1}\right)\boldsymbol{n}^{\rm corr} = \mathsf{N}_{\rm wn}^{-1}\boldsymbol{d}' + \mathsf{N}_{\rm wn}^{-1/2}\boldsymbol{\eta}_1 + \mathsf{N}_{\rm corr}^{-1/2}\boldsymbol{\eta}_2, \tag{3.12}$$

where $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are two independent vectors of random variates drawn from a standard Gaussian distribution, $\boldsymbol{\eta}_{1,2} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$. Since we typically use a time-domain mask, this equation cannot be solved by simply going to Fourier space, but needs to be solved by a more general method. For a nice method to solve the problem in this case, see Keihänen et al. (2020). Figure 3.2 shows some examples of how the correlated noise solutions look like. We see that the $1/f$ model seems to be working quite well.

Once we have found the correlated noise, the noise PSD parameters, $f_{\rm knee}$ and α, are sampled from their exact conditional distributions

$$-\ln P(f_{\rm knee}, \alpha \mid \sigma_0, \boldsymbol{n}^{\rm corr}) = \sum_{f=f_{\rm min}}^{f_{\rm max}} \left[\frac{|n_f^{\rm corr}|^2}{\mathsf{N}_{\rm corr}(f)} + \ln \mathsf{N}_{\rm corr}(f)\right] - \ln P(f_{\rm knee}, \alpha), \tag{3.13}$$

where $\mathsf{N}_{\rm corr}(f) = \sigma_0^2\left(\frac{f}{f_{\rm knee}}\right)^{\alpha}$ and $P(f_{\rm knee}, \alpha)$ is an optional prior.

The LFI data are divided into roughly 45 000 pointing periods, denoted PIDs (pointing ID), each with a duration of 30–60 minutes. The official *Planck* LFI Data

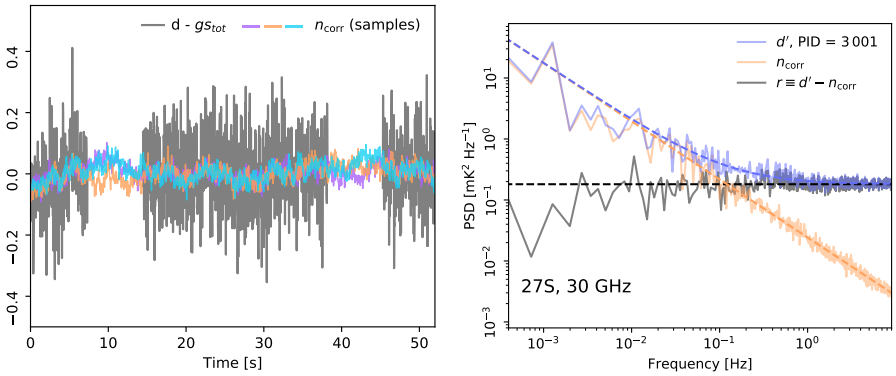

Figure 3.2: Example correlated noise samples in time domain (left) and as PSD (right).

Processing Center (DPC) (DPC, Planck Collaboration II, 2020) assumed that the noise parameters were constant throughout the entire mission. In our analysis, we sample the correlated noise, and the noise parameters independently for each PID, but we use the DPC results as priors for $f_{\rm knee}$ and α. In this way we use some information from the entire mission, but we also allow enough freedom to account for actual changes in noise properties over time.

The results we have found are very interesting. In addition to giving us a whole new insight into the noise properties of the LFI instrument, this detailed work on the noise modeling also highlights some important systematic effects. We demonstrate, for the first time, that the noise properties of most radiometers do indeed change significantly over time, mostly due to changes in the thermal environment of the instrument. Figure 3.3 shows very clearly the intimate relationship between the noise properties of the 70 GHz radiometers and the temperatures measured at the 20K cooling stage.

Although the noise properties of the 70 GHz radiometers change significantly over time, we are still able to fit the noise well using the $1/f$ model. At 30 and 44 GHz, however, some radiometers show signs of a significant power excess at intermediate timescales (0.1–10 Hz), which do not fit well with the $1/f$ model, as is seen in Fig 3.4. This is quite worrisome, since these timescales correspond roughly to 1–60 degrees on the sky, which is in the prime science range, where we don't want an unfitted systematic sloshing around in the model. This is something that will need more work, perhaps with a modified model for the noise PSD.

3.3 Calibration

The calibration step, Eq. 3.3, is a crucial step in any pipeline. We will now describe the approach we use for calibration in BeyondPlanck. We start with the data model,

$$\boldsymbol{d} = g\boldsymbol{s}^{\rm tot} + \boldsymbol{n}^{\rm corr} + \boldsymbol{n}^{\rm wn}. \tag{3.14}$$

The goal is to convert the raw time ordered data, $\boldsymbol{d}$, from voltages to Kelvin, effectively determining g. If we had a signal of known magnitude, then we could estimate g directly.

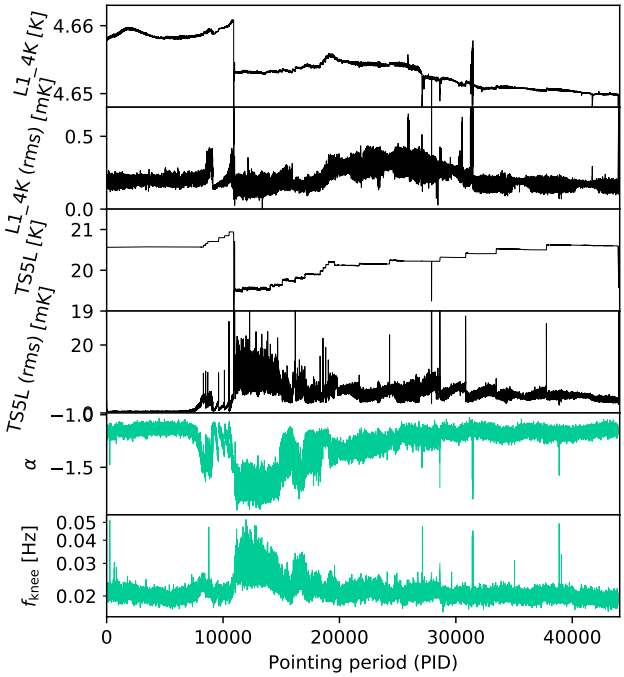

Figure 3.3: Average correlated noise properties of the 70 GHz radiometers (bottom two panels) compared with 4 K (top two panels) and 20 K (middle two panels) temperature sensor read-outs for the full mission.

Thankfully we have a very strong dipole signal which each detector measures, typically with a large signal-to-noise ratio, every minute, as the Planck scanning strategy orbits a great circle on the sky every sixty seconds. The problem, however, is that the overall magnitude of the dipole is not something we know, but one of the things we are trying to measure. On the other hand, there is one signal that we do know very precisely, and that is the component of the overall dipole coming from the movement of the Planck satellite relative to the sun, called the *orbital* dipole. The, much stronger, but unknown, dipole signal from the movement of the sun relative to the CMB rest frame, we denote as the *solar* dipole (see Fig. 3.5).

The idea then is to use the orbital dipole for the overall calibration of the entire system, but to use the solar dipole to measure the relative calibration of the different detectors and the change over time of single detectors. We therefore decompose the gain, g, into the following parts,

$$g_{t,i} = g_0 + \Delta g_i + \delta g_{t,i}, \tag{3.15}$$

where t is a time index and i is an index labeling the different detectors of a given frequency band. g_0 is then the absolute calibration, Δg_i, is the detector specific offset and $\delta g_{t,i}$ tracks the time evolution of the gain for each specific detector. The two last

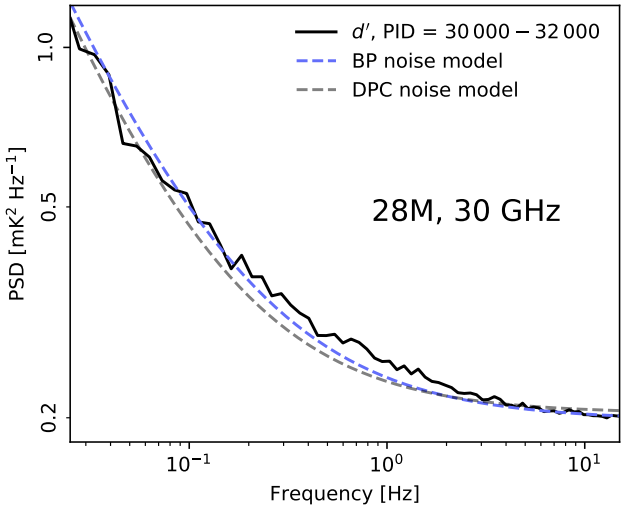

Figure 3.4: PSD of signal-subtracted data from radiometer 28M, averaged over 18 PIDs (black). Compared to the BeyondPlanck (dashed blue) and LFI DPC (dashed gray) noise models.

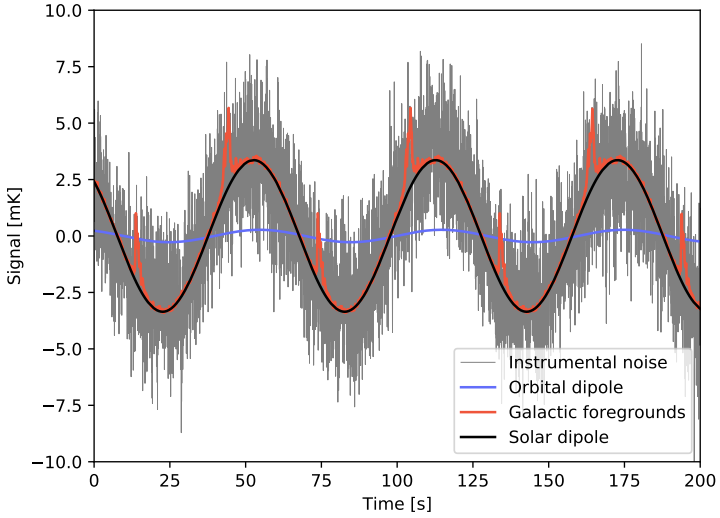

Figure 3.5: Different contributions to the 30 GHz time-ordered data. The orbital and the solar dipole are shown in blue and black, respectively, while the red line shows the modeled small-scale CMB fluctuations and Galactic foregrounds. The residual, given by the instrumental noise, is shown in gray. Figure courtesy of Eirik Gjerløw.

terms are subject to the constraints $\sum_i \Delta g_i = 0$ and $\sum_t \delta g_{t,i} = 0$. We can then use all the data, from all of the detectors, to determine g_0 using the known orbital dipole, and use the full signal (dominated by the solar dipole) to determine Δg_i and $\delta g_{t,i}$ subject to the constraints. See Paper VI (Gjerløw et al., 2020) for the detailed sampling steps, which we will not repeat here.

There are several challenges we need to deal with during calibration. First of all, there is typically not enough signal to noise in a single PID to measure the time dependent gain, $\delta g_{t,i}$, precisely enough. To solve this we apply a smoothing window, essentially using data from a few hundred consecutive PIDs to measure the time dependent gain. Another complication comes from the fact that for some periods the Planck scanning is almost completely orthogonal to the dipole, meaning that it is very hard to use the dipole for calibration. To deal with this we need to periodically employ very wide smoothing windows, making us less sensitive to changes in the gain for these periods. There are also abrupt jumps in the gain, causing the smoothing window approach to break down. If we know when these jumps happen, we can explicitly include them in the model, and fit the gain independently on each side of the jump. Although we know the origin of some of the gain jumps, we do not know the origin of all of them. This leaves us vulnerable to the presence of gain-jumps that we have not included in our model. Any such missing gain-jumps could lead to significant gain errors effecting the entire model (Eq. 3.1).

Figure 3.6 shows the correlated noise in map domain for the Stokes Q 44 GHz channel. We see some very clear stripes along an orbit from the middle of the map and up to the right. While we do not know the origin of these stripes, we suspect they are related to the gain, and quite possibly unmodeled gain-jumps in particular. This is an area we are still working to understand better using the large library of Planck housekeeping data. The stripes we see in the southern hemisphere of the 44 GHz band may be related to the problems we have had using the CMB polarization data from the southern hemisphere (Colombo et al., 2020; Paradiso et al., 2020), which means that this is one of the most important things we are working to figure out (see Paper VI (Gjerløw et al., 2020) for more discussion).

3.4 Future prospects

The future of the CMB is in polarization, particularly in B-mode polarization from primordial gravitational waves. These signals are very weak, which means that we will need to dig deep into both foregrounds and instrumental systematic effects in order to detect it. This is the challenge that BeyondPlanck has set out to tackle, to build a consistent statistical framework to deal with the complex interactions and degeneracies between instrumental effects, foregrounds and the cosmological signal.

The BeyondPlanck results presented here, analysing the Planck LFI data, is a proof of concept, demonstrating that end-to-end analysis from time ordered data to cosmological parameters is actually possible, and that this approach has several advantages over the more traditional, modular, approach. In the future, under the umbrella of *Cosmoglobe*[1], we want to combine as many datasets as possible into the

[1] http://cosmoglobe.uio.no

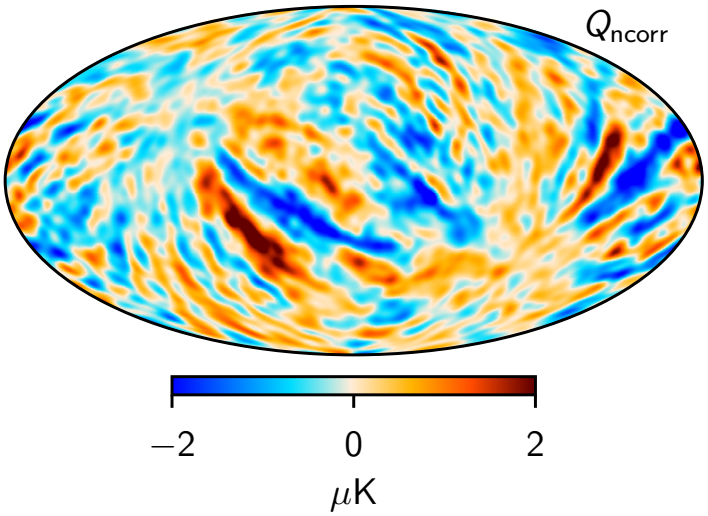

Figure 3.6: Stokes Q-map of the correlated noise, n_{corr}, 44 GHz band, smoothed at an angular resolution of 5 degrees FWHM. Figure courtesy of Eirik Gjerløw.

analysis, to get the most complete and robust sky model, and to be in the best position as possible to analyze data from future experiments like *LiteBIRD*, *PICO* and CMB-S4 (Abazajian et al., 2019; Hanany et al., 2019; Sugai et al., 2020; Suzuki et al., 2018).

Summary and Outlook

In this thesis I have given an introduction to the COMAP experiment and summarized some of mine and others work on the data analysis for COMAP. Papers I and II expand on this and go into more details on the various steps in the data analysis pipeline, and the power spectrum methods used. I also give a short summary of the BeyondPlanck project and the work I have done on noise estimation, systematics and calibration. In addition to discussing the experiments, the data analysis and the results, I have tried to connect the individual experiments with their significance to astrophysics and cosmology more generally.

The COMAP experiment is in a very exciting period right now, where we are working to finalize our analysis of the first year of data and publishing our first results, including an upper limit on the CO clustering power spectrum at around $z \sim 3$. On a longer time horizon we are already working on the next iterations of improvements in the COMAP pipeline, which involves better modeling and mitigation of ground contamination, better characterization of standing waves and their effects and an improvement in the mapmaking techniques, some of which have already been mentioned. The future plans for the COMAP experiments involve an expansion to multiple new telescopes, including instruments observing at lower frequencies, which would give us a whole new window into the epoch of reinonization ($z \sim 6$–8).

A bright future for intensity mapping in general lies in combining the data from multiple different experiments studying different spectral lines of the same galaxies giving us a more complete picture all the way from small scale chemistry and astrophysics to the large scale structure and expansion history of the universe as a whole. Combining together multiple complicated datasets increased the need for robust and statistically rigorous data analysis methods.

Our results in BeyondPlanck is a good demonstration of the unique advantages of the end-to-end approach. By including all the parts of the data analysis into a single model, not only does it help us deal with degeneracies and error propagation consistently, it also highlights problems in the data and where the model does not fit. With BeyondPlanck and *Cosmoglobe* we are laying the groundwork for the CMB community to face its greatest challenge, detecting B-modes from primordial gravitational waves. We believe that the best way to do this is to build an international community focused on global joint end-to-end analysis of cosmological datasets using open source, community built, software. By combining many datasets in this way we can not only get the most accurate sky model possible, but we can also break the degeneracies present in any one dataset, making us less susceptible to the systematic effects in either dataset.

Bibliography

Abazajian, K. et al. (July 2019). "CMB-S4 Science Case, Reference Design, and Project Plan". In: *arXiv e-prints*, arXiv:1907.04473, arXiv:1907.04473. arXiv: 1907.04473 [astro-ph.IM].

Bandura, K. et al. (July 2014). "Canadian Hydrogen Intensity Mapping Experiment (CHIME) pathfinder". In: *Ground-based and Airborne Telescopes V*. Ed. by Stepp, L. M., Gilmozzi, R., and Hall, H. J. Vol. 9145. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, p. 914522. DOI: 10.1117/12.2054950. arXiv: 1406.2288 [astro-ph.IM].

Battye, R. A. et al. (Sept. 2013). "H I intensity mapping: a single dish approach". In: MNRAS vol. 434, no. 2, pp. 1239–1256. DOI: 10.1093/mnras/stt1082. arXiv: 1209.0343 [astro-ph.CO].

Baumann, D. (July 2018). "TASI Lectures on Primordial Cosmology". In: *arXiv e-prints*, arXiv:1807.03098, arXiv:1807.03098. arXiv: 1807.03098 [hep-th].

Behroozi, P. S., Wechsler, R. H., and Conroy, C. (Jan. 2013a). "On the Lack of Evolution in Galaxy Star Formation Efficiency". In: ApJ vol. 762, L31, p. L31. DOI: 10.1088/2041-8205/762/2/L31. arXiv: 1209.3013 [astro-ph.CO].

— (June 2013b). "The Average Star Formation Histories of Galaxies in Dark Matter Halos from z = 0-8". In: ApJ vol. 770, 57, p. 57. DOI: 10.1088/0004-637X/770/1/57. arXiv: 1207.6105 [astro-ph.CO].

Bennett, C. L. et al. (Oct. 2013). "Nine-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Final Maps and Results". In: ApJS vol. 208, no. 2, 20, p. 20. DOI: 10.1088/0067-0049/208/2/20. arXiv: 1212.5225 [astro-ph.CO].

Bernal, J. L., Breysse, P. C., Gil-Marin, H., et al. (Dec. 2019). "User's guide to extracting cosmological information from line-intensity maps". In: Phys. Rev. D vol. 100, no. 12, 123522, p. 123522. DOI: 10.1103/PhysRevD.100.123522. arXiv: 1907.10067 [astro-ph.CO].

Bernal, J. L., Breysse, P. C., and Kovetz, E. D. (Dec. 2019). "Cosmic Expansion History from Line-Intensity Mapping". In: Phys. Rev. Lett. Vol. 123, no. 25, 251301, p. 251301. DOI: 10.1103/PhysRevLett.123.251301. arXiv: 1907.10065 [astro-ph.CO].

BeyondPlanck Collaboration (2020). "**BeyondPlanck I. Global Bayesian analysis of the *Planck* Low Frequency Instrument data**". In: A&A*, in preparation*. arXiv: 2011.05609.

Borriello, A. and Salucci, P. (May 2001). "The dark matter distribution in disc galaxies". In: MNRAS vol. 323, pp. 285–292. eprint: astro-ph/0001082.

Bower, G. C. et al. (Jan. 2016). "Cosmic Structure and Galaxy Evolution through Intensity Mapping of Molecular Gas". In: *American Astronomical Society Meeting Abstracts #227*. Vol. 227. American Astronomical Society Meeting Abstracts, p. 426.04.

Bowman, J. et al. (2012). *The First Billion Years: report of a study program*. Workshop report.

Breysse, P. et al. (May 2017). "Insights from probability distribution functions of intensity maps". In: MNRAS vol. 467, no. 3, pp. 2996–3010. DOI: 10.1093/mnras/stx203. arXiv: 1609.01728 [astro-ph.CO].

Breysse, P. C., Anderson, C. J., and Berger, P. (Dec. 2019). "Canceling Out Intensity Mapping Foregrounds". In: Phys. Rev. Lett. Vol. 123, no. 23, 231105, p. 231105. DOI: 10.1103/PhysRevLett.123.231105. arXiv: 1907.04369 [astro-ph.CO].

Bull, P. et al. (Apr. 2015). "Late-time Cosmology with 21 cm Intensity Mapping Experiments". In: ApJ vol. 803, no. 1, 21, p. 21. DOI: 10.1088/0004-637X/803/1/21. arXiv: 1405.1452 [astro-ph.CO].

Chatrchyan, S. et al. (Sept. 2012). "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physics Letters B* vol. 716, no. 1, pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235 [hep-ex].

Cheng, Y.-T., Chang, T.-C., and Bock, J. J. (Oct. 2020). "Phase-space Spectral Line Deconfusion in Intensity Mapping". In: ApJ vol. 901, no. 2, 142, p. 142. DOI: 10.3847/1538-4357/abb023. arXiv: 2005.05341 [astro-ph.CO].

Chung, D. (2020). "Line-intensity mapping with the CO Mapping Array Pathfinder and beyond". PhD thesis. Stanford University.

Chung, D. et al. (in prep).

Chung, D. et al. (Sept. 2017). "On Estimation of Contamination from Hydrogen Cyanide in Carbon Monoxide Line-intensity Mapping". In: ApJ vol. 846, 60, p. 60. DOI: 10.3847/1538-4357/aa8624. arXiv: 1706.03005.

Chung, D. T. (Aug. 2019). "A Partial Inventory of Observational Anisotropies in Single-dish Line-intensity Mapping". In: ApJ vol. 881, no. 2, 149, p. 149. DOI: 10.3847/1538-4357/ab3040. arXiv: 1905.00209 [astro-ph.CO].

Chung, D. T. et al. (Feb. 2019). "Cross-correlating Carbon Monoxide Line-intensity Maps with Spectroscopic and Photometric Galaxy Surveys". In: ApJ vol. 872, no. 2, 186, p. 186. DOI: 10.3847/1538-4357/ab0027. arXiv: 1809.04550 [astro-ph.GA].

Cleary, K. (2015). *COMAP Memo 5: System Overview*. Internal COMAP memo.

Cleary, K. et al. (Jan. 2016). "The CO Mapping Array Pathfinder (COMAP)". In: *American Astronomical Society Meeting Abstracts #227*. Vol. 227. American Astronomical Society Meeting Abstracts, p. 426.06.

Colombo et al. (2020). "**BeyondPlanck XI. CMB map and power spectrum posteriors**". In: A&A*, in preparation*. arXiv: 201x.xxxxx.

Cooray, A. et al. (Feb. 2016). "Cosmic Dawn Intensity Mapper". In: *arXiv e-prints*, arXiv:1602.05178, arXiv:1602.05178. arXiv: 1602.05178 [astro-ph.CO].

Creque-Sarbinowski, C. and Kamionkowski, M. (Sept. 2018). "Searching for decaying and annihilating dark matter with line intensity mapping". In: Phys. Rev. D vol. 98, no. 6, 063524, p. 063524. DOI: 10.1103/PhysRevD.98.063524. arXiv: 1806.11119 [astro-ph.CO].

Crites, A. T. et al. (2014). "The TIME-Pilot intensity mapping experiment". In: *Millimeter, Submillimeter, and Far-Infrared Detectors and Instrumentation for Astronomy VII*. Ed. by Holland, W. S. and Zmuidzinas, J. Vol. 9153. International Society for Optics and Photonics. SPIE, pp. 613–621. DOI: 10.1117/12.2057207.

DeBoer, D. R. et al. (Apr. 2017). "Hydrogen Epoch of Reionization Array (HERA)". In: PASP vol. 129, no. 974, p. 045001. DOI: 10.1088/1538-3873/129/974/045001. arXiv: 1606.07473 [astro-ph.IM].

Delabrouille, J. (Feb. 1998). "Analysis of the accuracy of a destriping method for future cosmic microwave background mapping with the PLANCK SURVEYOR satellite". In: A&AS vol. 127, pp. 555–567. DOI: 10.1051/aas:1998119.

Dinda, B. R., Sen, A. A., and Choudhury, T. R. (Apr. 2018). "Dark energy constraints from the 21 cm intensity mapping surveys with SKA1". In: *arXiv e-prints*, arXiv:1804.11137, arXiv:1804.11137. arXiv: 1804.11137 [astro-ph.CO].

Dodelson, S. (2003). *Modern Cosmology*. ISBN: 9780122191411. Academic Press.

Doré, O. et al. (Dec. 2014). "Cosmology with the SPHEREX All-Sky Spectral Survey". In: *arXiv e-prints*, arXiv:1412.4872, arXiv:1412.4872. arXiv: 1412.4872 [astro-ph.CO].

Foss, M. K. et al. (in prep). *First Season COMAP Results: CO Data Processing*.

Furlanetto, S. et al. (May 2019). "Fundamental Cosmology in the Dark Ages with 21-cm Line Fluctuations". In: BAAS vol. 51, no. 3, 144, p. 144.

Gjerløw et al. (2020). "**BeyondPlanck VII. Gain and absolute calibration**". In: A&A, *in preparation*. arXiv: 2011.08082.

Gong, Y., Chen, X., and Cooray, A. (May 2020). "Cosmological Constraints from Line Intensity Mapping with Interlopers". In: ApJ vol. 894, no. 2, 152, p. 152. DOI: 10.3847/1538-4357/ab87a0. arXiv: 2001.10792 [astro-ph.CO].

Guth, A. H. (Jan. 1981). "Inflationary universe: A possible solution to the horizon and flatness problems". In: Phys. Rev. D vol. 23, no. 2, pp. 347–356. DOI: 10.1103/PhysRevD.23.347.

Hanany, S. et al. (Feb. 2019). "PICO: Probe of Inflation and Cosmic Origins". In: *arXiv e-prints*, arXiv:1902.10541, arXiv:1902.10541. arXiv: 1902.10541 [astro-ph.IM].

Hill, G. J. et al. (Oct. 2008). "The Hobby-Eberly Telescope Dark Energy Experiment (HETDEX): Description and Early Pilot Survey Results". In: *Panoramic Views of Galaxy Formation and Evolution*. Ed. by Kodama, T., Yamada, T., and Aoki, K. Vol. 399. Astronomical Society of the Pacific Conference Series, p. 115. arXiv: 0806.0183 [astro-ph].

Hopkins, P. F. et al. (Nov. 2014). "Galaxies on FIRE (Feedback In Realistic Environments): stellar feedback explains cosmologically inefficient star formation". In: MNRAS vol. 445, no. 1, pp. 581–603. DOI: 10.1093/mnras/stu1738. arXiv: 1311.2073 [astro-ph.CO].

Huterer, D. and Shafer, D. L. (Jan. 2018). "Dark energy two decades after: observables, probes, consistency tests". In: *Reports on Progress in Physics* vol. 81, no. 1, 016901, p. 016901. DOI: 10.1088/1361-6633/aa997e. arXiv: 1709.01091 [astro-ph.CO].

Ihle et al. (2020). "**BeyondPlanck VI. Noise characterization and modelling**". In: A&A, *in preparation*. arXiv: 2011.06650.

Ihle, H. T. et al. (in prep). *First Season COMAP Results: Power spectrum methodology and preliminary data quality assessment*.

Ihle, H. T. et al. (Jan. 2019). "Joint Power Spectrum and Voxel Intensity Distribution Forecast on the CO Luminosity Function with COMAP". In: ApJ vol. 871, no. 1, 75, p. 75. DOI: 10.3847/1538-4357/aaf4bc. arXiv: 1808.07487 [astro-ph.CO].

Jungman, G., Kamionkowski, M., and Griest, K. (Mar. 1996). "Supersymmetric dark matter". In: Phys. Rep. Vol. 267, pp. 195–373. DOI: 10.1016/0370-1573(95)00058-5. arXiv: hep-ph/9506380 [hep-ph].

Keating, G. K., Bower, G. C., et al. (Dec. 2015). "First Results from COPSS: The CO Power Spectrum Survey". In: ApJ vol. 814, no. 2, 140, p. 140. DOI: 10.1088/0004-637X/814/2/140. arXiv: 1510.06744 [astro-ph.GA].

Keating, G. K., Marrone, D. P., Bower, G. C., and Keenan, R. P. (Oct. 2020). "An Intensity Mapping Detection of Aggregate CO Line Emission at 3 mm". In: ApJ vol. 901, no. 2, 141, p. 141. DOI: 10.3847/1538-4357/abb08e. arXiv: 2008.08087 [astro-ph.GA].

Keating, G. K., Marrone, D. P., Bower, G. C., Leitch, E., et al. (Oct. 2016). "COPSS II: The Molecular Gas Content of Ten Million Cubic Megaparsecs at Redshift z ∼ 3". In: ApJ vol. 830, no. 1, 34, p. 34. DOI: 10.3847/0004-637X/830/1/34. arXiv: 1605.03971 [astro-ph.GA].

Keihänen et al. (2020). "**BeyondPlanck II. Optimal mapmaking by Gibbs sampling**". In: A&A*, in preparation*. arXiv: 2011.06024.

Keihänen, E. et al. (Feb. 2010). "Making cosmic microwave background temperature and polarization maps with MADAM". In: A&A vol. 510, A57, A57. DOI: 10.1051/0004-6361/200912813. arXiv: 0907.0367 [astro-ph.CO].

Klypin, A. A. et al. (Apr. 1987). "Limits on Microwave Background Anisotropies - the Relikt Experiment". In: *Soviet Astronomy Letters* vol. 13, p. 104.

Koopmans, L. et al. (Apr. 2015). "The Cosmic Dawn and Epoch of Reionisation with SKA". In: *Advancing Astrophysics with the Square Kilometre Array (AASKA14)*, p. 1. arXiv: 1505.07568 [astro-ph.CO].

Kovetz, E. D. et al. (Sept. 2017). "Line-Intensity Mapping: 2017 Status Report". In: *arXiv e-prints*, arXiv:1709.09066, arXiv:1709.09066. arXiv: 1709.09066 [astro-ph.CO].

Kovetz, E. et al. (May 2019). "Astrophysics and Cosmology with Line-Intensity Mapping". In: BAAS vol. 51, no. 3, 101, p. 101. arXiv: 1903.04496 [astro-ph.CO].

Lagache, G., Cousin, M., and Chatzikos, M. (Jan. 2018). "The [CII] 158 μm line emission in high-redshift galaxies". In: A&A vol. 609, A130, A130. DOI: 10.1051/0004-6361/201732019. arXiv: 1711.00798 [astro-ph.GA].

Lakhlani, G. (Jan. 2019). "The Structure and Dynamics of the Interstellar Medium in the FIRE Simulations". PhD thesis. University of Toronto (Canada).

Lamb, J. W. (2020). *Standing waves due to optics and cables in the COMAP system*. Internal COMAP memo.

Lamb, J. et al. (in prep).

Li, T. Y. et al. (Feb. 2016). "Connecting CO Intensity Mapping to Molecular Gas and Star Formation in the Epoch of Galaxy Assembly". In: ApJ vol. 817, 169, p. 169. DOI: 10.3847/0004-637X/817/2/169. arXiv: 1503.08833.

Liu, R. H. and Breysse, P. C. (Feb. 2020). "Coupling parsec and gigaparsec scales: primordial non-Gaussianity with multi-tracer intensity mapping". In: *arXiv e-prints*, arXiv:2002.10483, arXiv:2002.10483. arXiv: 2002.10483 [astro-ph.CO].

Madau, P., Meiksin, A., and Rees, M. J. (Feb. 1997). "21 Centimeter Tomography of the Intergalactic Medium at High Redshift". In: ApJ vol. 475, pp. 429–444. DOI: 10.1086/303549. eprint: astro-ph/9608010.

Masui, K. W. et al. (Jan. 2013). "Measurement of 21 cm Brightness Fluctuations at z ~0.8 in Cross-correlation". In: ApJ vol. 763, no. 1, L20, p. L20. DOI: 10.1088/2041-8205/763/1/L20. arXiv: 1208.0331 [astro-ph.CO].

Mather, J. C. et al. (Jan. 1994). "Measurement of the Cosmic Microwave Background Spectrum by the COBE FIRAS Instrument". In: ApJ vol. 420, p. 439. DOI: 10.1086/173574.

Mertens, F. G. et al. (Apr. 2020). "Improved upper limits on the 21 cm signal power spectrum of neutral hydrogen at z ≈ 9.1 from LOFAR". In: MNRAS vol. 493, no. 2, pp. 1662–1685. DOI: 10.1093/mnras/staa327. arXiv: 2002.07196 [astro-ph.CO].

Milgrom, M. (June 2010). "New Physics at Low Accelerations (MOND): an Alternative to Dark Matter". In: *American Institute of Physics Conference Series*. Ed. by Alimi, J.-M. and Fuözfa, A. Vol. 1241. American Institute of Physics Conference Series, pp. 139–153. arXiv: 0912.2678 [astro-ph.CO].

Moustakas, L. A. and Metcalf, R. B. (Mar. 2003). "Detecting dark matter substructure spectroscopically in strong gravitational lenses". In: MNRAS vol. 339, pp. 607–615. eprint: astro-ph/0206176.

Newburgh, L. B. et al. (Aug. 2016). "HIRAX: a probe of dark energy and radio transients". In: *Ground-based and Airborne Telescopes VI*. Ed. by Hall, H. J., Gilmozzi, R., and Marshall, H. K. Vol. 9906. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, p. 99065X. DOI: 10.1117/12.2234286. arXiv: 1607.02059 [astro-ph.IM].

Paradiso et al. (2020). "**BeyondPlanck XII. Cosmological parameter posteriors**". In: A&A, *in preparation*. arXiv: 201x.xxxxx.

Parsons, A. R. et al. (June 2014). "New Limits on 21 cm Epoch of Reionization from PAPER-32 Consistent with an X-Ray Heated Intergalactic Medium at z = 7.7". In: ApJ vol. 788, no. 2, 106, p. 106. DOI: 10.1088/0004-637X/788/2/106. arXiv: 1304.4991 [astro-ph.CO].

Pen, U.-L. et al. (Oct. 2009). "The GMRT EoR experiment: limits on polarized sky brightness at 150 MHz". In: MNRAS vol. 399, no. 1, pp. 181–194. DOI: 10.1111/j.1365-2966.2009.14980.x. arXiv: 0807.1056 [astro-ph].

Penzias, A. A. and Burrus, C. A. (Jan. 1973). "Millimeter-Wavelength Radio-Astronomy Techniques". In: ARA&A vol. 11, p. 51. DOI: 10.1146/annurev.aa.11.090173.000411.

Planck Collaboration II (2014). "*Planck* 2013 results. II. Low Frequency Instrument data processing". In: A&A vol. 571, A2. DOI: 10.1051/0004-6361/201321550. arXiv: 1303.5063.

Planck Collaboration I (2020). "*Planck* 2018 results. I. Overview, and the cosmological legacy of *Planck*". In: A&A vol. 641, A1. DOI: 10.1051/0004-6361/201833880. arXiv: 1807.06205.

Planck Collaboration II (2020). "*Planck* 2018 results. II. Low Frequency Instrument data processing". In: A&A vol. 641, A2. DOI: 10.1051/0004-6361/201833293. arXiv: 1807.06206.

Planck Collaboration VI (2020). "*Planck* 2018 results. VI. Cosmological parameters". In: A&A vol. 641, A6. DOI: 10.1051/0004-6361/201833910. arXiv: 1807.06209.

Planck Collaboration et al. (Sept. 2020). "Planck 2018 results. I. Overview and the cosmological legacy of Planck". In: A&A vol. 641, A1, A1. DOI: 10.1051/0004-6361/201833880. arXiv: 1807.06205 [astro-ph.CO].

Rasmussen, M. (2020). "Data selection for intensity mapping experiments using machine learning". MA thesis. University of Oslo.

Riess, A. G. et al. (Sept. 1998). "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant". In: AJ vol. 116, no. 3, pp. 1009–1038. DOI: 10.1086/300499. arXiv: astro-ph/9805201 [astro-ph].

Santos, M. G. et al. (Sept. 2017). "MeerKLASS: MeerKAT Large Area Synoptic Survey". In: *arXiv e-prints*, arXiv:1709.06099, arXiv:1709.06099. arXiv: 1709.06099 [astro-ph.CO].

Saro, A. et al. (July 2013). "Toward Unbiased Galaxy Cluster Masses from Line-of-sight Velocity Dispersions". In: ApJ vol. 772, 47, p. 47. arXiv: 1203.5708.

Schaye, J. et al. (Nov. 2014). "The EAGLE project: simulating the evolution and assembly of galaxies and their environments". In: *Monthly Notices of the Royal Astronomical Society* vol. 446, no. 1, pp. 521–554. DOI: 10.1093/mnras/stu2058. eprint: https://academic.oup.com/mnras/article-pdf/446/1/521/4139718/stu2058.pdf.

Silva, M. et al. (in prep).

Silva, M. B. et al. (Aug. 2019). "Mapping Large-Scale-Structure Evolution over Cosmic Times". In: *arXiv e-prints*, arXiv:1908.07533, arXiv:1908.07533. arXiv: 1908.07533 [astro-ph.CO].

Square Kilometre Array Cosmology Science Working Group et al. (Mar. 2020). "Cosmology with Phase 1 of the Square Kilometre Array Red Book 2018: Technical specifications and performance forecasts". In: PASA vol. 37, e007, e007. DOI: 10.1017/pasa.2019.51. arXiv: 1811.02743 [astro-ph.CO].

Stacey, G. J. et al. (July 2018). "CCAT-Prime: science with an ultra-widefield submillimeter observatory on Cerro Chajnantor". In: *Ground-based and Airborne Telescopes VII*. Ed. by Marshall, H. K. and Spyromilio, J. Vol. 10700. Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, p. 107001M. DOI: 10.1117/12.2314031. arXiv: 1807.04354 [astro-ph.GA].

Sugai, H. et al. (Jan. 2020). "Updated Design of the CMB Polarization Experiment Satellite LiteBIRD". In: *Journal of Low Temperature Physics* vol. 199, no. 3-4, pp. 1107–1117. DOI: 10.1007/s10909-019-02329-w. arXiv: 2001.01724 [astro-ph.IM].

Susskind, L. (Nov. 1979). "Dynamics of spontaneous symmetry breaking in the Weinberg-Salam theory". In: Phys. Rev. D vol. 20, no. 10, pp. 2619–2625. DOI: 10.1103/PhysRevD.20.2619.

Suzuki, A. et al. (Dec. 2018). "The LiteBIRD Satellite Mission: Sub-Kelvin Instrument". In: *Journal of Low Temperature Physics* vol. 193, no. 5-6, pp. 1048–1056. DOI: 10.1007/s10909-018-1947-7. arXiv: 1801.06987 [astro-ph.IM].

Tauber, J. A. et al. (2019). "Characterization of the in-flight properties of the *Planck* telescope". In: A&A vol. 622, A55. DOI: 10.1051/0004-6361/201833150.

The OST mission concept study team (Sept. 2018). "The Origins Space Telescope (OST) Mission Concept Study Interim Report". In: *arXiv e-prints*, arXiv:1809.09702, arXiv:1809.09702. arXiv: 1809.09702 [astro-ph.IM].

Tingay, S. J. et al. (Jan. 2013). "The Murchison Widefield Array: The Square Kilometre Array Precursor at Low Radio Frequencies". In: PASA vol. 30, e007, e007. DOI: 10.1017/pasa.2012.007. arXiv: 1206.6945 [astro-ph.IM].

Tristram, M. et al. (2021). "Planck constraints on the tensor-to-scalar ratio". In: *A&A* vol. 647, A128. DOI: 10.1051/0004-6361/202039585.

Uzgil, B. D. et al. (Dec. 2019). "The ALMA Spectroscopic Survey in the HUDF: Constraining Cumulative CO Emission at $1 \lesssim z \lesssim 4$ with Power Spectrum Analysis of ASPECS LP Data from 84 to 115 GHz". In: ApJ vol. 887, no. 1, 37, p. 37. DOI: 10.3847/1538-4357/ab517f. arXiv: 1911.00028 [astro-ph.GA].

van Haarlem, M. P. et al. (Aug. 2013). "LOFAR: The LOw-Frequency ARray". In: A&A vol. 556, A2, A2. DOI: 10.1051/0004-6361/201220873. arXiv: 1305.3550 [astro-ph.IM].

van Uitert, E. et al. (Sept. 2012). "Constraints on the shapes of galaxy dark matter haloes from weak gravitational lensing". In: A&A vol. 545, A71, A71. arXiv: 1206.4304.

Vogelsberger, M. et al. (Oct. 2014). "Introducing the Illustris Project: simulating the coevolution of dark and visible matter in the Universe". In: MNRAS vol. 444, no. 2, pp. 1518–1547. DOI: 10.1093/mnras/stu1536. arXiv: 1405.2921 [astro-ph.CO].

Yang, S. et al. (Sept. 2020). "Multitracer cosmological line intensity mapping mock lightcone simulation". In: *arXiv e-prints*, arXiv:2009.11933, arXiv:2009.11933. arXiv: 2009.11933 [astro-ph.GA].

Zwicky, F. (Jan. 1933). "Die Rotverschiebung von extragalaktischen Nebeln". In: *Helvetica Physica Acta* vol. 6, pp. 110–127.

Aad, G. et al. (Sept. 2012). "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physics Letters B* vol. 716, no. 1, pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214 [hep-ex].

Appendix A

Appendix

A.1 Binomial likelihood and mean estimation

Say you have some model for a statistic with a binomial distribution. The binomial mean, $\mu(\theta) = Np$, is then a function of the model parameters θ. Say you don't know the function $\mu(\theta)$, but you can simulate samples from the distribution at fixed θ. After one sample x_1, the posterior distribution of $\mu(\theta)$ is given by

$$P_1(\mu) \equiv P(\mu|x_1) = \frac{P(x_1|\mu)P_0(\mu)}{\int_{\mu'} P(x_1|\mu')P_0(\mu')d\mu'}, \tag{A.1}$$

where

$$P_0(\mu) = \int_\theta P(\mu|\theta)P(\theta)d\theta, \tag{A.2}$$

where $P(\theta)$ is the prior on the model parameters. In general, after n, samples we get

$$P_n(\mu) = P(\mu|x_n, \cdots, x_1) = \frac{P(x_n|\mu)P_{n-1}(\mu)}{\int_{\mu'} P(x_n|\mu')P_{n-1}(\mu')d\mu'}. \tag{A.3}$$

The estimated likelihood of some data d given by the model at parameters θ is then given by

$$P(\theta|d) = \frac{P(\theta)}{P(d)} \int_{\mu(\theta)} P(d|\mu(\theta))P_n(\mu(\theta))d\mu. \tag{A.4}$$

In practice $P_0(\mu)$ could perhaps be set to some weak prior, although the details will matter. Choosing an exponential prior we get

$$P_0(\mu) = \frac{\exp\left(-\frac{\mu}{\mu_0}\right)}{\int_0^\infty d\mu' \exp\left(-\frac{\mu'}{\mu_0}\right)} = \frac{\exp\left(-\frac{\mu}{\mu_0}\right)}{\mu_0}, \tag{A.5}$$

where μ_0 is the scale parameter of the prior, if needed, this parameter could be tuned to approximate Eq. A.2, or simply chosen at a value we think is reasonable. We can now

perform the integral in equation A.1

$$I_1 = \frac{1}{\mu_0}\int P(x_1|\mu')\exp\left(-\frac{\mu'}{\mu_0}\right)d\mu', \tag{A.6}$$

$$= \frac{1}{\mu_0}\int \binom{x_1}{N}\left(\frac{\mu'}{N}\right)^{x_1}\left(1-\frac{\mu'}{N}\right)^{N-x_1}\exp\left(-\frac{\mu'}{\mu_0}\right)d\mu', \tag{A.7}$$

$$\approx \int \frac{\binom{x_1}{N}}{N}\left(\frac{\mu'}{N}\right)^{x_1} e^{-\mu'}\exp\left(-\frac{\mu'}{\mu_0}\right)d\mu', \tag{A.8}$$

$$= \frac{N}{\mu_0}\binom{x_1}{N}\left(\frac{\mu_0}{N(\mu_0+1)}\right)^{x_1+1}\int u^{x_1}e^{u}\,du, \tag{A.9}$$

$$= \frac{N}{\mu_0}\binom{x_1}{N}\left(\frac{\mu_0}{N(\mu_0+1)}\right)^{x_1+1}\Gamma(x_1+1), \tag{A.10}$$

where we have assumed $N \gg \mu, d$, and used the substitution $u = \mu'\left(\frac{\mu_0}{\mu_0+1}\right)$.

Inserting this we can now find the analytic expression for Eq. A.1,

$$P_1(\mu) \approx \frac{\frac{1}{\mu_0}\binom{x_1}{N}\left(\frac{\mu'}{N}\right)^{x_1}e^{-\mu}\exp\left(-\frac{\mu'}{\mu_0}\right)}{\frac{N}{\mu_0}\binom{x_1}{N}\left(\frac{\mu_0}{N(\mu_0+1)}\right)^{x_1+1}\Gamma(x_1+1)} \tag{A.11}$$

$$= \frac{\mu^{x_1}e^{-\mu}\exp\left(-\frac{\mu'}{\mu_0}\right)}{\left(\frac{\mu_0}{(\mu_0+1)}\right)^{x_1+1}\Gamma(x_1+1)}. \tag{A.12}$$

In general, we get

$$P_n(\mu) = \frac{\mu^{x_1+x_2+\cdots+x_n}e^{-n\mu}\exp\left(-\frac{\mu'}{\mu_0}\right)}{\left(\frac{\mu_0}{(n\mu_0+1)}\right)^{x_1+x_2+\cdots+x_n+1}\Gamma(x_1+x_2+\cdots+x_n+1)}. \tag{A.13}$$

The posterior distribution then becomes

$$P(\theta|d) = \frac{P(\theta)}{P(d)}\int P(d|\mu)P_n(\mu)d\mu, \tag{A.14}$$

$$= \frac{P(\theta)}{P(d)}\binom{d}{N}\frac{1}{N^d}\frac{1}{\left(\frac{\mu_0}{(n\mu_0+1)}\right)^{x+1}\Gamma(x+1)}\int \mu^{x+d}e^{-(n+1)\mu}\exp\left(-\frac{\mu'}{\mu_0}\right), \tag{A.15}$$

$$= \frac{P(\theta)}{P(d)}\binom{d}{N}\left(\frac{\mu_0}{N(n+1)\mu_0+1}\right)^{d}\left(\frac{n\mu_0+1}{(n+1)\mu_0+1}\right)^{x}\frac{\Gamma(x+d+1)}{\Gamma(x+1)}, \tag{A.16}$$

where we have defined $x \equiv x_1 + x_2 + \cdots + x_n$.

Papers

Paper I

First Season COMAP Results: CO Data Processing (draft)

M. K. Foss, H. T. Ihle and the COMAP collaboration

ApJ, in preparation.

I

Draft version February 4, 2021
Typeset using LaTeX **twocolumn** style in AASTeX62

First Season COMAP Results: CO Data Processing

Marie K. Foss,[1] Håvard T. Ihle,[1] Jowita Borowska,[1] Morgan Catha,[2] Kieran A. Cleary,[3] Hans Kristian Eriksen,[1] Stuart Harper,[4] James W. Lamb,[2] Jonas G. S. Lunde,[1] Nils-Ole Stutzer,[1] Duncan J. Watts,[1] Ingunn K. Wehus,[1] David P.Woody,[2] and the COMAP collaboration

[1] *Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, N-0315 Oslo, Norway*
[2] *Owens Valley Radio Observatory, California Institute of Technology, Big Pine, CA 93513, USA*
[3] *California Institute of Technology, Pasadena, CA 91125, USA*
[4] *Jodrell Bank Centre for Astrophysics, School of Physics and Astronomy, The University of Manchester, Oxford Road, Manchester, M13 9PL, U.K.*

ABSTRACT

We describe the data processing pipeline used to analyze the first season of COMAP observations, converting raw detector readouts to calibrated sky maps. This pipeline implements five main steps: gain calibration, filtering, noise characterization, data selection, and mapmaking. Absolute gain calibration relies on a combination of instrumental and astrophysical sources, while relative gain calibration exploits real-time total-power variations. High efficiency filtering is achieved through spectroscopic common-mode rejection within and across receivers, resulting in nearly uncorrelated white noise within single-frequency channels. Consequently, near-optimal but biased maps are produced by binning the filtered time stream into a pixelized map structure; a corresponding signal bias transfer function is estimated through simulations. Data selection is performed automatically through a series of goodness-of-fit statistics, including χ^2 and multi-scale correlation tests. Applying this pipeline to the first-season COMAP data, we find that night-time observations using a constant elevation scanning strategy are sufficiently stable for science extraction, whereas both day-time and Lissajous scanning observations fail critical null-tests. In total, only 10 % of the full data volume is currently retained for scientific analysis, which is substantially lower than the original goal of 35 %. However, explicit sidelobe and ground modelling are likely to increase this significantly in future analyses. All second-season COMAP observations will employ the constant elevation scanning strategy, and this alone will immediately double the effective mapping speed. Power spectrum results derived from the first-season COMAP maps are presented in a companion paper by Ihle et al. (2021).

Keywords: stars: formation, galaxies: star formation,radio lines: galaxies, methods: data analysis

1. INTRODUCTION

Understanding the evolution of galaxies and the intergalactic medium (IGM) over the largest spatial and temporal scales is one of the principle goals of cosmology. Galaxy surveys address this challenge by resolving and detecting individual galaxies, a technique that necessarily favors brighter galaxies and smaller cosmic volumes. Spectral line intensity mapping (LIM) is a complementary technique (see Kovetz et al. 2017 or Kovetz et al. 2019 for a review) that holds the potential to characterize the global properties of galaxies and their evolution by surveying the aggregate emission from all galaxies over large volumes.

This technique uses redshifted line emission (e.g., Lyα, CO, or C II) as a tracer for the underlying density field. Large volumes along a given line-of-sight may be surveyed simultaneously with a single spectrometer at relatively low spatial resolution, and by scanning this spectrometer across the sky a full 3D density map may be derived. Despite multiple different modeling efforts (Li et al. 2016; Padmanabhan 2018; Chung et al. 2021) and significant progress on the observational front (Pavesi et al. 2018; Keating et al. 2016, 2020), the overall level of the CO signal, especially in the clustering regime, is still unknown.

Corresponding author: Marie K. Foss
m.k.foss@astro.uio.no

The CO Mapping Array Pathfinder (COMAP; Cleary et al. 2021) is an intensity mapping experiment that targets carbon monoxide (CO) in the frequency range between 26 and 34 GHz, using a 19-element receiver mounted on a 10.4 m single-dish telescope, with each element coupled to a 8096-channel ROACH2-based spectrometer. In this frequency range, the receiver is sensitive to CO(1–0) at $z = 2.4$–3.4, with a fainter contribution from CO(2–1) at $z = 6$–8. Phase I of the experiment aims to detect the CO(1–0) signal and use it to constrain the properties of galaxies at the Epoch of Galaxy Assembly. A future phase will add a second receiver at Ku-band in order to detect CO(1–0) from around $z = 6$–8, cross-correlating with the CO(2–1) signal from the 26–34 GHz receiver and constraining the properties of galaxies towards the end of the Epoch of Reionization.

The receiver's detector chain is based on cryogenically cooled HEMT low-noise amplifiers (LNA) which contribute to a mean system temperature of 35–45 K across the full frequency range. The predicted signal from high-redshift CO emission is expected to be no more than a few microkelvins per COMAP spatial/spectral resolution element (or "voxel"). Thus, the raw instrumental noise must be reduced by many orders of magnitude before a statistically significant detection may be achieved. In practice, this is done by repeatedly observing the same part of the sky using multiple detectors, and thereby gradually increasing the sensitivity per voxel. For this to succeed, however, it is necessary to suppress systematic contributions from atmospheric temperature variations, sidelobe contamination, ground pickup, standing waves, Galactic foregrounds etc. by a corresponding amount.

This paper describes the first-season COMAP data analysis pipeline, which aims to produce clean maps from raw time-ordered COMAP observations. This includes calibration, data selection, filtering, and mapmaking. The rest of this paper is organized as follows: First, in order to establish useful notation and conventions, we give a brief introduction to the COMAP instrument in Sect. 2, while referring the interested reader to for full details. Next, we provide a high-level overview of the analysis pipeline in Sect. 3.1, before specifying each step in Sects. 3.3–3.8. Consistency tests are considered in Sects. 4–4.3, and we summarize and conclude in Sect. 5.

2. INSTRUMENT AND DATA MODEL

Before describing the COMAP analysis pipeline, we provide a brief overview of the instrument itself, and define an explicit data model. A more detailed description of the instrument can be found in a separate paper (Lamb et al. 2021).

Table 1: Frequency range of each COMAP sideband (SB).

Band	SB	Freq. (GHz)
A	LSB	26–28
	USB	28–30
B	LSB	30–32
	USB	32–34

2.1. *Instrument overview*

The COMAP Phase I instrument observes in the K_a band, at 26–34 GHz and is located at the Owens Valley Radio Observatory (OVRO) in California, USA. It is mounted on a 10.4 m telescope that was originally built for the Millimeter Array at OVRO, then used as a part of the Combined Array for Research in Millimeter-wave Astronomy (CARMA) experiment, and has now been repurposed for COMAP. The telescope's primary and secondary reflectors have diameters of 10.4 m and 1.1 m, respectively, and the beam FWHM is 4.5 arcmin at 30 GHz.

The receiver comprises 19 independent detector chains, called "pixels" or "feeds". Each feed consists of the following elements, ordered according to their position in the signal path:

1. A feedhorn that collects incoming photons.

2. A polarizer that separates the radiation into two orthogonal circular polarization states. For the observations described in this paper, 15 feeds have a two-stage polarizer, two feeds have a single-stage polarizer, and two feeds have no polarizer.

3. A low-noise amplifier (LNA) module based on high electron-mobility transistors (HEMT) integrated in the form of a monolithic microwave integrated circuit (MMIC). Each LNA module contains two MMIC LNAs and is cooled to 15–18 K.

4. Two down-converter modules (DCMs). DCM1 shifts the 26–34 GHz frequency band down to 2–10 GHz. DCM2 splits this 8 GHz bandwidth into two bands: A (2–6 GHz) and B (6–10 GHz), each of which is quadrature downconverted to produce an "in-phase" (I) and "quadrature" (Q) signal.

5. Two CASPER "Roach2" FPGA-based spectrometers, each of which takes the I and Q signals

Field Name	RA (J2000)	Dec (J2000)	Notes
CO2	01:41:44.4	+00:00:00.0	CO science field - lies within the HETDEX fall field
CO6	15:04:00.0	+55:00:00.0	CO science field - lies outside the main HETDEX survey
CO7	11:20:00.0	+52:30:00.0	CO science field - lies within the HETDEX spring field
FG4	00:42:44.4	+41:16:08.6	M31, the Andromeda Galaxy
FG6	18:44:56.4	−01:59:52.0	W43, star forming region
FG7	03:40:49.0	+31:54:38.0	AME-emitting region of Perseus molecular cloud
Galactic Plane survey			Survey covering Galactic longitudes $22.5^\circ < l < 50^\circ$
TauA	05:34:31.9	+22:00:52.2	Pointing calibrator - supernova remnant (Crab Nebula)
CasA	23:23:24.0	+58:48:54.0	Pointing calibrator - supernova remnant
CygA	19:59:28.4	+40:44:02.1	Pointing calibrator - radio galaxy
Jupiter			Pointing calibrator

Table 2: COMAP fields and calibrators.

as input and performs digital sideband separation. The output of the two spectrometers is four 2 GHz-wide sidebands (SB), each of which has 1024 frequency channels, resulting in a native frequency resolution of approximately 2 MHz. The sidebands are labelled "lower" (LSB) or "upper" (USB), as shown in Table 1. These data are then written to computer disk with a sample frequency of 50 Hz, resulting in a total data production rate of 43 GB/hour or 377 TB/year. Before being processed by the pipeline, each full 8 GHz-wide spectrum is averaged in software into 1024 ∼2 MHz-wide frequency channels.

Different feeds have different polarizer solutions for development and testing purposes. As discussed later, one of the main systematic effects observed in the COMAP system are standing waves, i.e., electromagnetic waves reflected between the boundaries of various components in the receiver chain. The magnitude and impact of this effect depends sensitively on the polarization properties of the signal, and different solutions were therefore explored in the early phases of the experiment. An important finding from these tests was that feeds without a polarizer are significantly more susceptible to the standing wave between the focal plane and the secondary than the rest of the feeds, and the two pixels in question are therefore excluded from the analysis. Similarly, a 20th pixel was added as a blind feed, to help distinguish internally and externally generated signals, and is also excluded from the science analysis.

The receiver cryostat containing the cooled LNAs is situated between the primary and secondary reflector. To protect the instrument from the environment, a metal shroud with a polypropylene foam window covers the cryostat. This window was augmented by a polystyrene backing support in June 2020 in order to minimize the variation of standing waves between the window and the receiver caused by movement of the window.

To support frequent and accurate gain estimation, COMAP employs an ambient temperature load that is directly attached to the environmental shroud housing. This "calibration vane" is automatically moved in front of the feedhorn array at the beginning and end of each observation (each lasting for about one hour; see Sect. 2.2), fully filling the field of view of each pixel. The temperature of the calibration vane is monitored with sensors, allowing the system temperature to be calculated and applied to calibrate the gain.

2.2. *Observation Strategy and Field Selection*

COMAP observes three areas of the sky, selected to maximize the observing efficiency, avoid bright 30 GHz point sources, and overlap with coverage of a galaxy survey targeting Lyα emission from galaxies in the same redshift range, the Hobby-Eberly Telescope Dark Energy eXperiment (HETDEX; Hill et al. 2008). Table 2 lists the details of these fields.[1] Although COMAP's observing strategy has been designed to permit the direct detection of CO fluctuations from galaxies at $z = 2.4 - 3.4$, cross-correlation with a galaxy survey such as HETDEX can increase the detection significance by at least a factor of two (Chung et al. 2019; Silva et al. 2021) as well as provide validation for the origin of detected signal in galaxies at the target redshift.

In addition to the main science fields, we observe several other fields for ancillary science applications, in particular to constrain the spectrum of anomalous microwave emission (AME), or spinning dust emission,

[1] Since COMAP began observing, the boundaries of the HETDEX spring field coverage changed, with the result that one COMAP field no longer overlaps with the main HETDEX survey although we hope to also fill in this field with additional HETDEX observations.

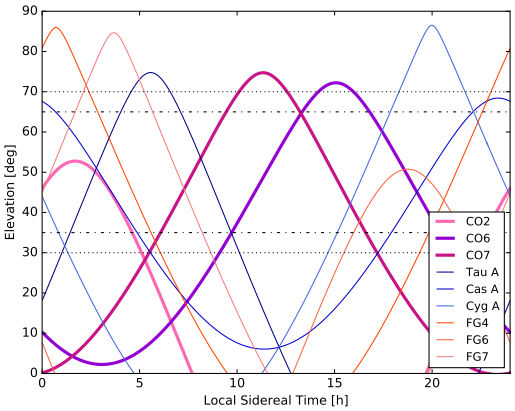

Figure 1: Elevation of CO (pink/purple), Galactic (orange) and calibration (blue) fields as a function of Local Sidereal Time.

around 30 GHz. These include M31, the star forming region W43, and a region within the Perseus molecular cloud. We are also conducting a survey of the Galactic plane covering longitudes $22.5° < l < 50°$, the results of which can be found in Harper et al. (2021).

To facilitate calibration with astrophysical sources, we also observe a handful of radio sources, including Jupiter, the supernova remnants Taurus A (TauA) and Cassiopeia A (CasA), and the radio galaxy Cygnus A (CygA), all of which are somewhat extended compared to the beam except for Jupiter. Analysis of these observations is discussed in Sect. 3.4.

Table 2 lists all science fields and calibrators. In Fig. 1 we plot the elevation of each field as a function of Local Sidereal Time, indicating when the fields are available for observation.

Telescope scans of the science fields follow a harmonic motion described by

$$\mathrm{az} = A\sin(at+\phi); \quad \mathrm{el} = B\sin(bt), \tag{1}$$

where A, B are amplitude parameters that define the size of the field, the ratio a/b determines the shape of the curve, and ϕ is a phase parameter. Three different scan types were used: "circular" ($A = B$, $a = b$, $\phi = \pi/2$), "constant elevation scans (CES)" ($b = 0$) and "Lissajous" (varying parameters). After some early observations with circular scans, we adopted CES and Lissajous as the preferred patterns, alternating between each on a daily basis. At the start of a scan, the telescope is positioned at the leading edge of the field. The telescope then executes the scan while the field drifts through the pattern. This typically takes 5–10 minutes, after which the telescope is repointed to the leading edge of the field again in preparation for the next scan. An example of the scanning path for 3 hours of continuous observations with a Lissajous scan is shown in Figure 4. Testing the relative performance of the CES and Lissajous scanning strategies in terms of final data quality is an important goal of the first-season COMAP survey.

2.3. *Data model*

As described by Lamb et al. (2021), the COMAP detector readout for a single frequency channel may be modelled as

$$P_{\mathrm{out}} = k_{\mathrm{B}} G \Delta\nu T_{\mathrm{sys}}, \tag{2}$$

where G is the gain, $\Delta\nu$ is the bandwidth, and T_{sys} is the system temperature of the instrument. The system temperature may be further modeled as

$$\begin{aligned} T_{\mathrm{sys}} = {} & T_{\mathrm{reciever}} + T_{\mathrm{atmosphere}} + T_{\mathrm{ground}} \\ & + T_{\mathrm{CMB}} + T_{\mathrm{foregrounds}} + T_{\mathrm{CO}}, \end{aligned} \tag{3}$$

where T_{reciever} is the effective noise temperature of the receiver, $T_{\mathrm{atmosphere}}$ is the noise contribution from the atmosphere, T_{ground} is ground pickup from far sidelobes, T_{CMB} is the contribution from the CMB, $T_{\mathrm{foregrounds}}$ are continuum foregrounds (typically from the galaxy), and T_{CO} is the line emission signal from extragalactic CO, which is the main scientific target of the COMAP instrument.

To understand the challenge involved in measuring the cosmological CO signal, it is instructive to consider the order of magnitude and stability of each term in Eq. (3). By far the largest single contribution is that of the receiver temperature, which is usually about 10–30 K for COMAP. For the COMAP receiver, with HEMT LNA technology, this is very stable.

The second-largest contribution is from the atmosphere, which typically adds 15–25 K. This term varies significantly on all time scales longer than a few seconds, and depends on external conditions including elevation, humidity, cloud coverage, ambient temperature and wind speed. It is also strongly correlated between detectors and frequencies, since all feeds observe through essentially the same atmospheric column at any given time; fortunately, the phase structure of the atmospheric fluctuations are uncorrelated on long time scales.

Next, ground pickup typically accounts for 1.5–2 K, and this term can be particularly problematic because it depends sensitively on the instrument pointing: If a sidelobe happens to straddle a strong signal gradient, such as the horizon or the Sun, several mK variations may be measured on very short timescales and with a time-dependency that appears nearly sky synchronous.

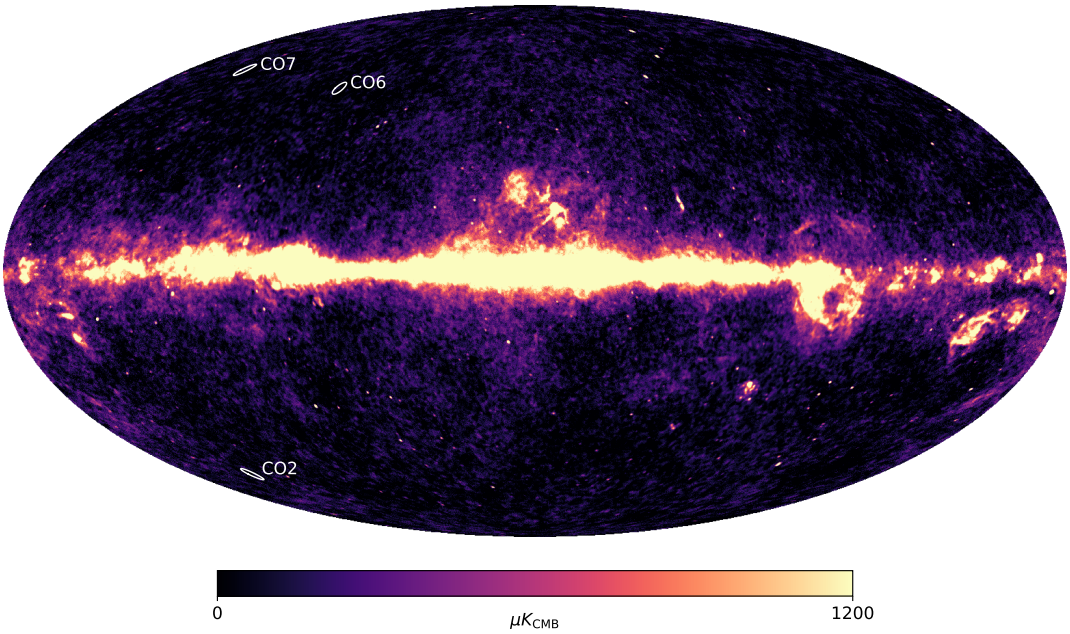

Figure 2: The three main CO fields observed by the telescope overplotted as circles of radius 2° on top of the *Planck* LFI 30 GHz full-mission map (downloaded from the *Planck Legacy Archive* Planck Collaboration et al. 2020).

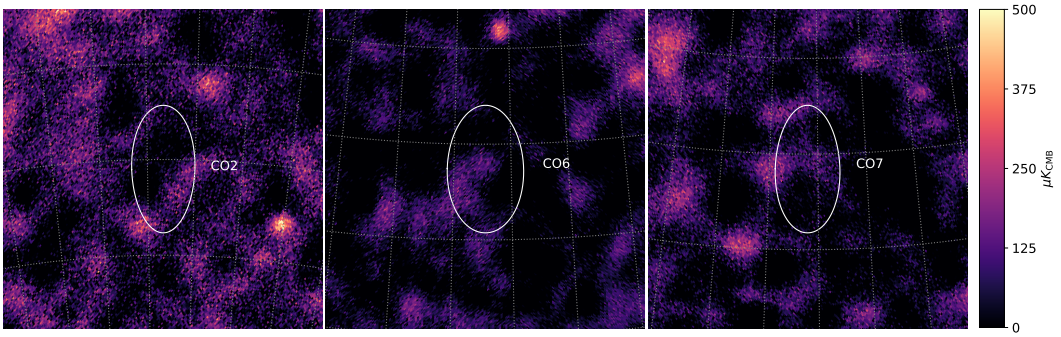

Figure 3: The three main CO fields observed. The circles, illustrating the rough coverage of each field, have radii of 2° and are centered at the field centers (in Galactic coordinates) $(\mathrm{lon}, \mathrm{lat}) = (149.0°, -60.3°), (91.35°, 53.22°)$ and $(150.64°, 59.53°)$, for the CO2, CO6 and CO7 fields respectively. Each of the three plots represents a $10° \times 10°$ sky area around each field with the *Planck* LFI 30 GHz full-mission map (downloaded from the PLA, Planck Collaboration et al. 2020) in the background showing the foreground levels at each field.

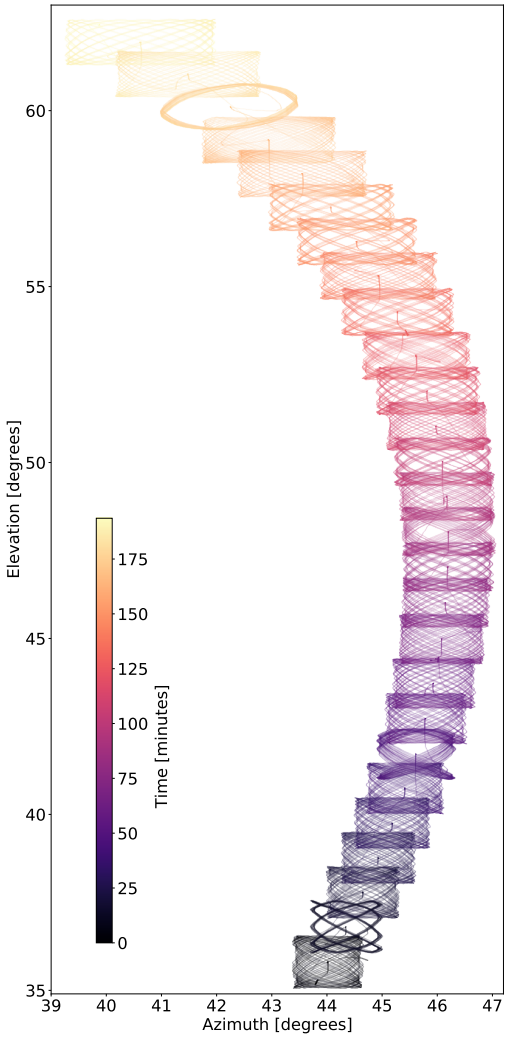

Figure 4: Movement of the telescope in azimuth and elevation for three consecutive observations (27 individual scans, 194 minutes) of the same field, with the Lissajous scanning strategy.

The fourth term represents the CMB temperature of 2.7 K, which is both isotropic and stationary, while the fifth term represents astrophysical foregrounds, expected to contribute at most 1 mK, for instance synchrotron, free-free, and dust emission from the Galaxy. Although these are sky synchronous, and in principle could confuse potential CO measurements, they also have very smooth frequency spectra, and are therefore relatively easy to distinguish from the cosmological CO signal, which varies rapidly with frequency. An important potential exception is line emission from other molecules redshifted to our band from galaxies at other epochs. The hydrogen cyanide (HCN) line is expected to be one of the brightest such lines. Emission from HCN in galaxies towards our CO fields at redshift $z = 1.6$–2.4 will appear in our frequency range. However, this contribution is expected to be an order of magnitude lower than that from CO (Chung et al. 2017)

Finally, the cosmological CO line emission signal is expected to account for $\mathcal{O}(1\,\mu\mathrm{K})$. Whether it is possible to detect such a weak signal depends directly on the stability and sensitivity of the instrument. In this respect, the fundamental quantity of interest is the overall noise level of the experiment, which is dominated by random thermal Johnson noise caused by thermal motions of electrons within the electronics.

The magnitude of these random thermal fluctuations is proportional to T_{sys}, with a standard deviation that is given by the so-called radiometer equation,

$$\sigma_{\mathrm{N}} = \frac{T_{\mathrm{sys}}}{\sqrt{\Delta\nu\,\tau}}, \tag{4}$$

where τ is the integration time. Thus, since both the system temperature and the bandwidth are essentially fixed experimental parameters, the only way of reducing the total uncertainty is by increasing the integration time. As a concrete and relevant example, we note that an integration time of 45 hours is required to achieve a standard deviation of $20\,\mu$K with a system temperature of 45 K and a bandwidth of 31.25 MHz.

In addition to the thermal and uncorrelated noise described by the radiometer equation, there are three main sources of correlated noise, namely gain fluctuations in the low-noise amplifiers, atmospheric temperature fluctuations, and time-dependent standing waves. All of these are expected to have a roughly $1/f$-type spectrum, although with different particular properties[2]. The fact that these sources of correlated noise are also strongly correlated between frequencies is very useful in order to filter out this noise in the analysis.

Equation (2) describes the detector output at any given time. To connect this to the actual measurements recorded by the detector, we adopt the following data

[2] There are several different sources of standing waves, some of the main ones give rise to $1/f$-like spectra, but others do not.

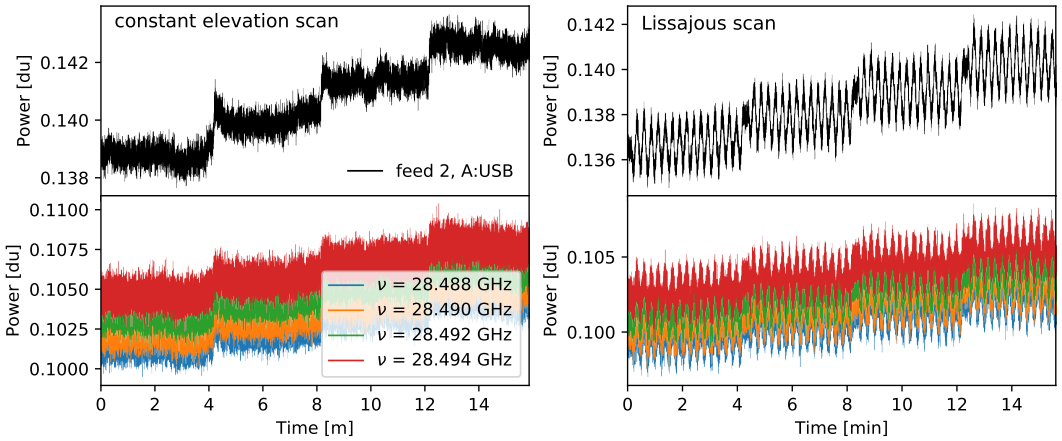

Figure 5: Raw data from the COMAP instrument (in arbitrary digital units of power Here we see data averaged over a single 2 GHz-wide sideband (top) and examples of data from four individual frequency channels in that sideband (bottom). These data were taken using two different scan patterns: CES (left) and Lissajous (right).

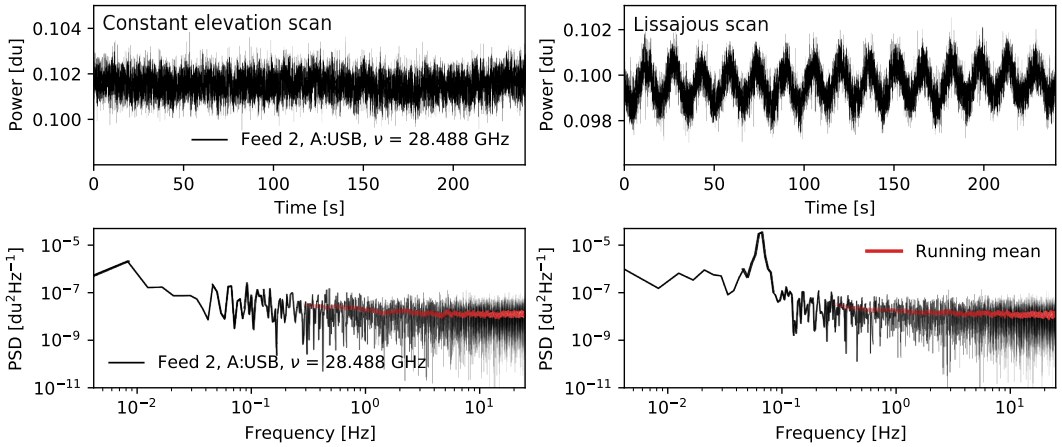

Figure 6: Raw data from an individual frequency channel of the COMAP instrument. Power is shown as a function of time (top), and the corresponding power spectral density (PSD) is also shown (bottom). We show data from a CES scan (left) and a Lissajous scan (right).

model,

$$d^i_\nu(t) = \langle d^i_\nu \rangle (1 + \delta^i_G) \left[1 + P^i_{\rm cel}(s_{\rm cont} + s^\nu_{\rm CO}) + P^i_{\rm tel} s_{\rm ground} + n_{\rm corr} + n^{\nu i}_{\rm w}\right]. \quad (5)$$

Here $d^i_\nu(t)$ denotes the raw data recorded at time t for frequency channel ν in feed i; $\langle d^i_\nu \rangle$ represents the corresponding time average; δ^i_G denotes feed dependent gain fluctuations; $P^i_{\rm cel}$ and $P^i_{\rm tel}$ are pointing matrices in celestial and telescope coordinate systems, respectively; $s_{\rm cont}$ denotes the celestial continuum sources, mainly from the CMB and Galactic foregrounds; $s^\nu_{\rm CO}$ is the CO line emission; $s_{\rm ground}$ is the ground signal picked up by the far sidelobes; and $n_{\rm corr}$ is the correlated noise component, mostly consisting of atmosphere fluctuations and standing waves. Factors with no feed or frequency index are assumed to be similar (or at least strongly correlated) at different frequencies and feeds, while factors with a ν label indicate parts of the model that are assumed to

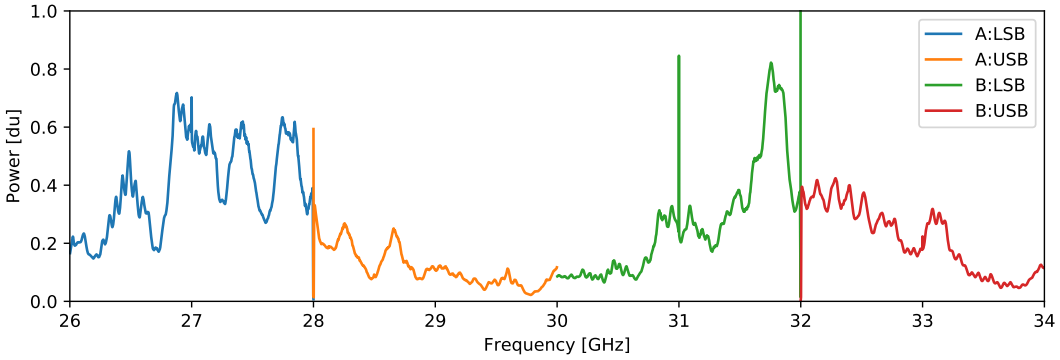

Figure 7: Time-averaged raw data from each frequency channel on single feed of the COMAP instrument. The colors represent the four 2 GHz-wide sidebands. Note that a few of the frequency channels at at the edges and middle of sidebands tend to be unstable and are masked out in the analysis.

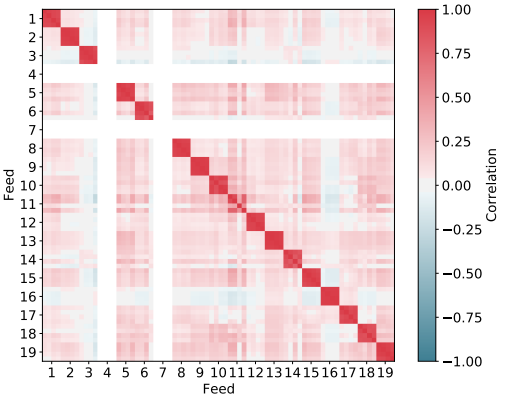

Figure 8: Correlation between the sideband-averaged data from the 20 feeds of the COMAP instrument for a single constant elevation scan. For this observation, as for much of the observing campaign, the LNAs for feeds 4 and 7 were turned off because those feeds, as a test, did not have a polarizer and so had large standing waves due to reflections between the receiver and the secondary reflector.

have non-smooth frequency dependence. The main purpose of the COMAP analysis pipeline is to characterize $s^{\nu}_{\rm CO}$ given $d^{i}_{\nu}(t)$.

2.4. *Data overview*

Before presenting the analysis pipeline, we provide a preview of the raw time-ordered data (TOD) generated by the COMAP instrument, with the goal of building intuition that will be useful for understanding the purpose of each component of the analysis pipeline described in this paper. Figures 5 and 6 show examples of such raw time-ordered data (TOD) from the instrument using the CES (left column) and Lissajous (right column) scanning strategies. Perhaps the most obvious features in these plots are step-wise changes in power as the telescope changes elevation during repointings between scans; see Sect. 2.2. The Lissajous scans additionally show oscillations in power as the telescope changes elevation during the scan, since the telescope looks through a thicker slab of atmosphere at lower elevations, and this increases the atmospheric contribution to the system temperature.

The top panels in Fig. 6 show an individual frequency channel for a single scan (i.e., stationary observation period), while the bottom panel shows the corresponding power spectral density (PSD). For the CES case, the PSD is relatively featureless, with an overall shape that looks consistent with a typical $1/f$ noise spectrum. For the Lissajous case, an additional strong peak is seen around 0.007 Hz, which matches the scanning period of 14 sec, and this corresponds to the periodic atmospheric variations seen in the panels above.

Figure 7 shows the time averaged data for all frequency channels of a single feed. The spectral shape is mostly determined by the average gain as a function of frequency, due to the combined effect of the various components of the receiver chain. This average gain is a purely instrumental effect, not associated with the true sky signal, and therefore simply corresponds to a normalization factor that should be calibrated out before higher-level analysis. However, some of the spectral shape is also determined by the fact that the system temperature also changes with frequency, and in some cases

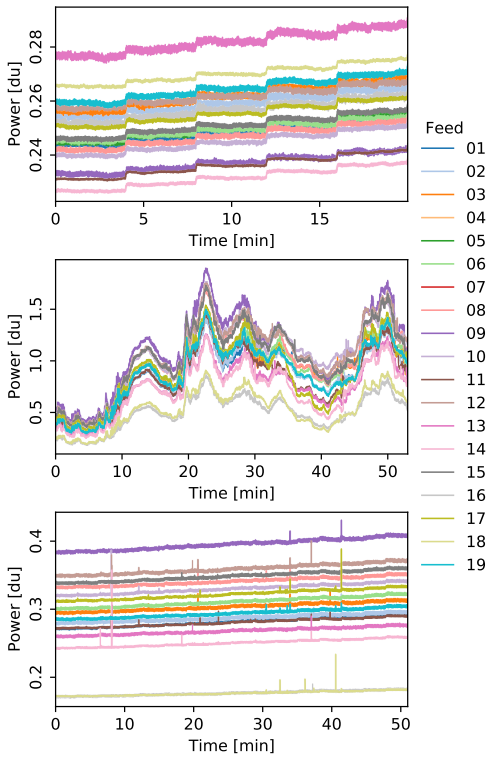

Figure 9: Feed averaged COMAP TOD recorded under various observing conditions. The top panel shows data observed under normal conditions, and is dominated by instrumental noise. The middle panel shows data observed under poor weather conditions with a thick cloud coverage, resulting in large coherent power fluctuations observed by all feeds. This third panel shows data with strong spikes, which may for instance happen during periods with high insect activity.

results in large spikes within specific frequency ranges. Separating these two effects is a main goal of the calibration procedures described below.

In Fig. 8 we plot the correlation,

$$C_{ij} = \frac{\left\langle \hat{d}^i \hat{d}^j \right\rangle}{\sqrt{\left\langle \hat{d}^i \hat{d}^i \right\rangle \left\langle \hat{d}^j \hat{d}^j \right\rangle}}, \tag{6}$$

between the power, $\hat{d}^i$ recorded by any two feeds, i and j, after averaging over all frequencies within each sideband for each radiometer. Here we first note that the data from different sidebands of the same feed are strongly

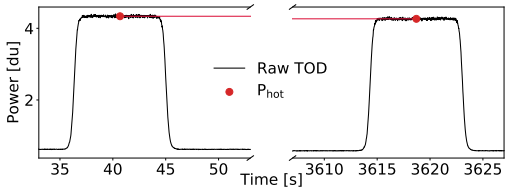

Figure 10: The calibration vane is inserted in front of the receiver at the beginning and end of one observation of a CO science field. The time between calibration vane insertions is typically about an hour, a period set by the preferred data file size for the CO field observations.

correlated. This is because both main sources of correlated noise in the COMAP data, namely gain fluctuations and atmospheric fluctuations, are common for sidebands within a given pixel. In contrast, sidebands for different feeds share the atmospheric fluctuations (and also some standing waves), but have mostly independent gain fluctuations, and this results in lower overall correlations, but still typically in the 10-40% range. Accounting for and mitigating such correlations will clearly be essential in order to extract robust science from these observations.

The quality of the COMAP data depends strongly on the observing conditions, as illustrated in Fig. 9. The top panel shows an observation made under normal conditions, while the middle panel shows an observation made during poor weather, with thick cloud coverage. The bottom panel shows a data segment with strong "spikes", a feature of some data taken in summer, possibly associated with insects flying in front of the focal plane. Automatic identification and removal of problematic data is clearly an important and necessary component of the pipeline.

Finally, Fig. 10 shows the calibration vane observations that are made at the beginning and end of each observation period. Since the ambient temperature is about one order of magnitude higher than $T_{\rm sys}$, the measured power is also correspondingly about one order of magnitude higher, and this bright and known signal allows for a precise estimate of $T_{\rm sys}$. Note that, unless specified otherwise, these data segments are always removed prior to data analysis, as they would otherwise compromise any filtering that may be applied to the data.

3. COMAP ANALYSIS PIPELINE

3.1. *Pipeline Overview*

We are now ready to present the COMAP analysis pipeline, which is designed to process the raw data dis-

cussed in Sect. 3.1 into calibrated and cleaned CO maps. The main steps of this pipeline are schematically illustrated in Fig. 11.

The processing starts with raw "Level-1" files, which contain raw data as recorded by the instrument, together with pointing information and house-keeping data. These files typically each contain about one hour of observations, including calibration vane observations at the beginning and end, as well as a corresponding estimate of the system temperature, $T_{\rm sys}$, based on the same calibration measurements. We denote each one hour of data as one observation, and assign it an individual observation ID (abbreviated obsID). Each observation consists of several scans, where one scan is the period between two re-pointings of the telescope, during which the telescope performs the same motions around a fixed point in azimuth and elevation while the target field drifts through. The instrumental properties are consequently assumed to be stationary within each scan. The module denoted `scan_detect` in Fig. 11 indicates a dedicated code that partitions each observation into individual scans, based on pointing information, and records information on each scan in a database.

The main processing takes part in the `l2gen` module, which generates calibrated and cleaned TOD and stores them in so-called "Level-2" files. This is achieved through the application of a series of filters (see Sect. 3.3) and a time-varying gain normalization (see Sect. 3.4). This stage also evaluates basic goodness-of-fit statistics and defines a frequency channel mask that excludes missing or broken data for the current scan, before reducing the spectral resolution of the data to a spectral resolution suitable for mapmaking. In most analyses, we typically reduce the resolution from 2 MHz to 32 MHz, resulting in the computational speed-up of subsequent steps and a memory saving for storing final maps by a factor of 16.

Next, the `accept_mod` module reads in the statistics (including goodness-of-fit) and basic frequency mask produced by `l2gen` and produces a list of accepted observations as defined by user-specified thresholds for each statistic. Examples of relevant statistics used for this purpose are χ^2 per observation, $f_{\rm knee}$, and Solar elongation. The output from this process is called an *accept list*, which is used to define data splits.

Converting time-ordered data into pixel-ordered data is done by a mapmaker called `tod2comap` (see Sect. 3.8). As shown in the following, the adopted filters result in very nearly uncorrelated white noise, and the current implementation of `tod2comap` accordingly adopts simple binning into voxels. Finally, from these maps we can extract an estimate of the CO power spectrum, or the voxel intensity distribution, or perform general astrophysics and cosmology inference, all of which happens in `comap2ps`.

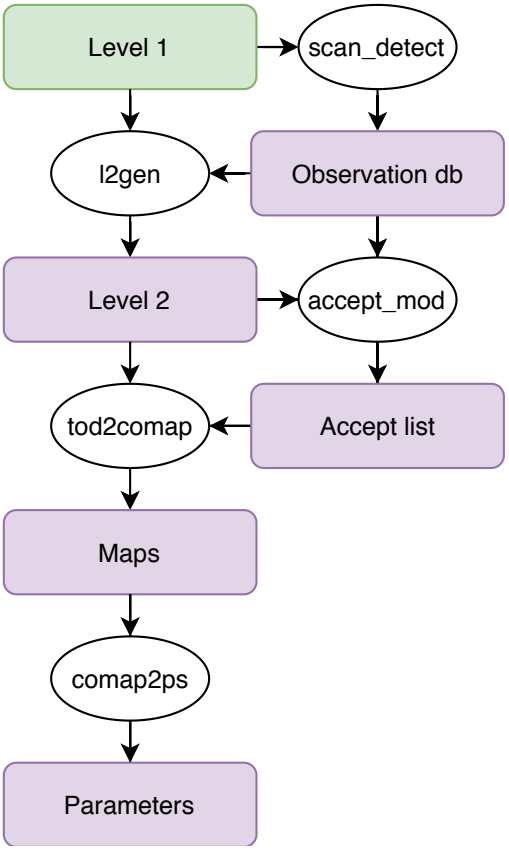

Figure 11: Flow diagram of the analysis pipeline. The purple boxes are products of this pipeline, whereas the green is the input data (raw data from the telescope and some housekeeping data). The white ellipses represents different modules of the code.

3.2. *Data Segmentation*

As described above, we define a *scan* to be the observing period between re-pointings of the telescope, during which the telescope executes a scan pattern around a fixed azimuth and elevation for 3–10 minutes while the observed source or field drifts through. In the subsequent analysis each scan is analyzed independently, and is converted into an individual Level-2 file.

The purpose of the `scan_detect` code is to identify all scans within all observation periods, and sort these ac-

cording to astrophysical target. This selection is based on telescope house-keeping information, and in particular on an observation flag that records the state of the telescope at any given time. However, recognizing that human errors do happen, automatic simple sanity tests are also performed, for instance to validate that the specified field name agrees with the recorded pointing.

The basic result from this selection process is an observation database, which is a list of obsIDs sorted according to source. For each obsID, we list all scans within that obsID, including basic information such as the modified Julian date (MJD) of the start and end of the scan, as well as the scanning mode (e.g. Lissajous or CES) and mean pointing information. The calibration vane observations are defined to be the first and last scans within each obsID, but flagged as calibration measurements.

3.3. *Filtering*

As described in Sect. 3.1, the COMAP TOD exhibit a wide range of non-CO-related contributions, both of instrumental and external origin. These must be suppressed by orders of magnitude prior to map-making in order to extract the astrophysically valuable signal. With this goal in mind, we introduce four specific filters, each targeting one class of artefacts.

Figure 12 shows the evolution of the data as it passes through each of the filters.

3.3.1. *Normalization*

The first filtering operation we introduce is data normalization. This is done simply by dividing the raw TOD, $P_{\rm out}$, by its own running mean, and then subtracting the overall average,

$$d(\nu,t) = \frac{P_{\rm out}(\nu,t)}{\langle P_{\rm out}(\nu,t)\rangle} - 1. \tag{7}$$

Here t is a time sample index and ν denotes frequency channel. This operation is performed separately on each frequency channel.

The main purpose of this step is to equalize (i.e., "flatten") the instrumental bandpass, as illustrated in Fig. 7, and effectively establish data with appropriate relative calibration. The main practical advantage of doing so is that the amplitude of common-mode contaminants, such as gain-induced correlated noise or atmospheric fluctuations, become comparable across all frequencies within a single sideband, and therefore much easier to filter out. The same also holds true for broadband astrophysical contributions, such as the CMB or foregrounds, which also must be removed prior to signal extraction. See top panel of Fig. 12 to see the effect of the normalization step. We can see that long timescale fluctuations are removed, and that the data now fluctuates around zero.

Note also that with the definition in Eq. (7), the noise level of $d(\nu,t)$ is given by the sample rate and bandwidth alone in the ideal case, and should equal $1/\sqrt{\tau\Delta\nu}$. Calibration into physical units is performed simply by multiplying $d(\nu,t)$ with $T_{\rm sys}$. We find that $d(\nu,t)$ is a particularly convenient function for goodness-of-fit tests and it will serve as our main object of interest in the following.

3.3.2. *Removal of Pointing Templates*

The second filter we apply is designed to suppress signals that are correlated with local pointing (azimuth and elevation), as opposed to sky-correlated signals. The two main effects of this type are elevation-correlated atmospheric contributions and azimuth-correlated sidelobe contributions. The first of these effects may be modelled by a simple expression for the optical depth of the atmosphere of the form

$$\tau(\mathrm{el}) = \frac{\tau_0}{\sin(\mathrm{el})}, \tag{8}$$

where τ_0 is the optical depth of the atmosphere at zenith, and el is the elevation, while the second effect may be approximated through a low-order polynomial in azimuth. We therefore filter the data by fitting and subtracting the following simple model to each normalized frequency channel separately,

$$d = \frac{g}{\sin(\mathrm{el}(t))} + a\,\mathrm{az}(t) + c + n. \tag{9}$$

Here g, a and c are fitting constants, and n denotes Gaussian noise with an assumed constant variance. We find the best-fit values for the free parameters by minimizing a χ^2 statistic, and use g and a to clean the TOD with respect to the pointing templates,

$$d_{\rm after} = d_{\rm before} - \frac{g}{\sin(\mathrm{el})} - a\,\mathrm{az} - \left\langle \frac{g}{\sin(\mathrm{el})} + a\,\mathrm{az} \right\rangle. \tag{10}$$

In this expression, $\langle\,\rangle$ denotes the mean value in time for a specific frequency channel, and this term ensures that the TOD has vanishing mean also after subtraction of pointing templates. For long-duration scans we divide the TOD into disjoint segments of roughly 4 minutes each, and perform the template fit and removal separately on each data segment, in order to improve the tracking of temporal variations.

The effect of the pointing template removal can be seen in the second row of Fig. 12.

3.3.3. *Common-mode polynomial filter*

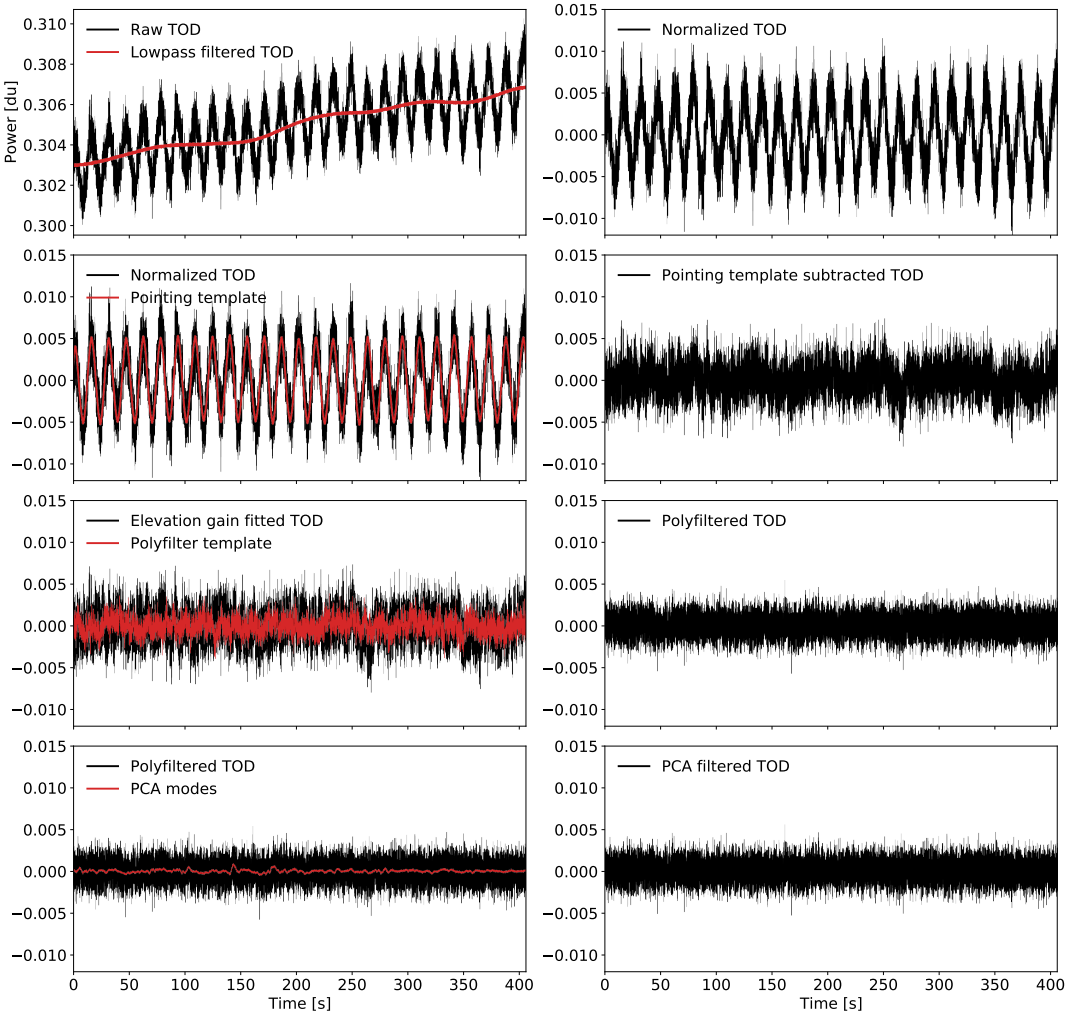

Figure 12: Effect of each filter in time-domain. Each row shows the data before (left column) and after (right column) applying the indicated filter. From top to bottom, the filters shown are 1) normalization; 2) elevation gain subtraction; 3) poly-filtering; and 4) PCA filtering. Data used is from scan 14456.03, feed 5, in a 31.25 MHz band around 27.673 GHz.

Our third filter is designed to remove all common-mode signals observed simultaneously by all frequency channels within a given sideband. Specifically, for each timestep we fit and subtract a low-order (and typically linear) polynomial to the normalized and pointing-subtracted TOD in frequency space for each sideband.

Specifically, we assume

$$d_\nu = c_0 + c_1\nu + c_2\nu^2 + \dots\,, \tag{11}$$

where d_ν are the data across one sideband at a specific timestep, c_0, c_1 and c_2, etc., are constants that are fitted independently for each sideband. We then remove the fitted polynomial from the data. In the third row of Fig. 12, we can see an example of how this filter removes the majority of the correlated noise from the data.

The main target of this filter is $1/f$ noise from gain variations in the receiver electronics and atmospheric

temperature fluctuations, and is strongly correlated between frequency channels within each sideband. Indeed, the fact that this noise is so tightly correlated between channels is one of the key instrumental features of the COMAP instrument that makes CO measurements feasible in the first place, effectively reducing the final noise level by more than one order of magnitude.

As a bonus, this polynomial filter also removes any slowly varying astrophysical signal, and in particular broadband signals such as CMB, synchrotron, free-free or anomalous microwave emission. In contrast, the cosmological CO signal is expected to vary on the scale of adjacent frequency channels, and is therefore only mildly affected by this filter. However, some CO signal is indeed lost on the largest longitudinal scales due to this filter, and this effect will later be quantified in terms of an effective transfer function. Fortunately, the modes that are suppressed correspond mostly to physical scales of around 50–100 Mpc/h and larger, and are therefore mostly modes that we do not expect to be able to measure with COMAP.

3.3.4. *Principal Component Analysis (PCA) filter*

While the previous filter removes common-mode signals within each sideband, our fourth and final filter targets common-mode signals seen simultaneously by the entire focal plane. The two most prominent examples of such contaminants are atmospheric variations and standing waves, both of which have strongly correlated time variations across all feeds and frequencies. To suppress these signals, we perform a so-called Principal Component Analysis (PCA) on the whole data set, and subtract the leading modes. Intuitively speaking, this amounts to identifying some functions of time that explain the largest amount of the variance between the different frequencies across all the different feeds. These functions are often called the leading PCA components.

To formulate this idea in a mathematical language, let us organize all data in given scan into a data matrix D, where each row contains the TOD corresponding to a single frequency channel on a single feed. Thus D is a matrix with dimensions $n_{\rm freq} \times n_{\rm samp}$, where $n_{\rm freq} = n_{\rm feeds} \cdot n_{\rm sidebands} \cdot n_{\rm freq\ per\ sideband} = 19 \cdot 4 \cdot 1024$ is the total number of frequencies added up from all sidebands and feeds, and $n_{\rm samp}$ is the number of samples in time, such that

$$D = \begin{bmatrix} D_{11} & \dots & D_{1n_{\rm samp}} \\ \vdots & \ddots & \vdots \\ D_{n_{\rm freq}1} & \dots & D_{n_{\rm freq}n_{\rm samp}} \end{bmatrix}. \tag{12}$$

The empirical data covariance matrix, C, may then be written as

$$C = D^T D, \tag{13}$$

and the eigenvectors, $\mathbf{v}_k$, of this matrix that correspond to the highest eigenvalues are precisely the PCA components we are looking for. In practice, we identify the few leading PCA components through a standard iterative method.

For each frequency (in each feed) we compute the relative PCA amplitudes by projecting the observed data vector, $\mathbf{d}$, onto the PCA eigenvector,

$$a_k = \mathbf{d} \cdot \mathbf{v}_k = \sum_{i=1}^{n_{\rm samp}} d_i v_k^i, \tag{14}$$

where $\mathbf{d}$ is now the normalized, pointing-corrected, and polynomial-filtered data described above. The leading PCA components are then subtracted from the data,

$$\mathbf{d}_{\rm after} = \mathbf{d}_{\rm before} - \sum_{i=1}^{n_{\rm comp}} a_k \mathbf{v}_k, \tag{15}$$

where $n_{\rm comp}$ is the number of leading components removed (typically four).

Figure 13 shows the three leading PCA components for a typical scan. For each component, its variation with time is shown for the duration of the scan, as well as its contribution to the overall variance for each feed. A more extreme example is shown in Fig. 14.

3.3.5. *Masking*

Sometimes individual frequency channels or groups of nearby frequency channels show artefacts, even after applying all the filters described above. This could manifest in a significant excess noise that is correlated in time, or in correlations between different frequency channels. We wish to mask these frequency channels so that their contribution does not contaminate the final results.

To determine which frequencies should be masked we first perform the poly- and PCA filters on a copy of the original dataset. We then use two main approaches to identifying individual or groups of frequency channels to be masked. The first approach uses the fact that the expected correlation between two independent Gaussian variables (for large $n_{\rm samp}$) is given by $1/\sqrt{n_{\rm samp}}$, where $n_{\rm samp}$ is the number of samples used to calculate the correlation. Thus, after accounting for the expected correlation induced by the polyfilter, we know the statistics describing *good* data, and can identify bad data as deviations from these statistics. Specifically, we consider groups of elements within the frequency-frequency correlation matrix (either squares of different sizes or sets of

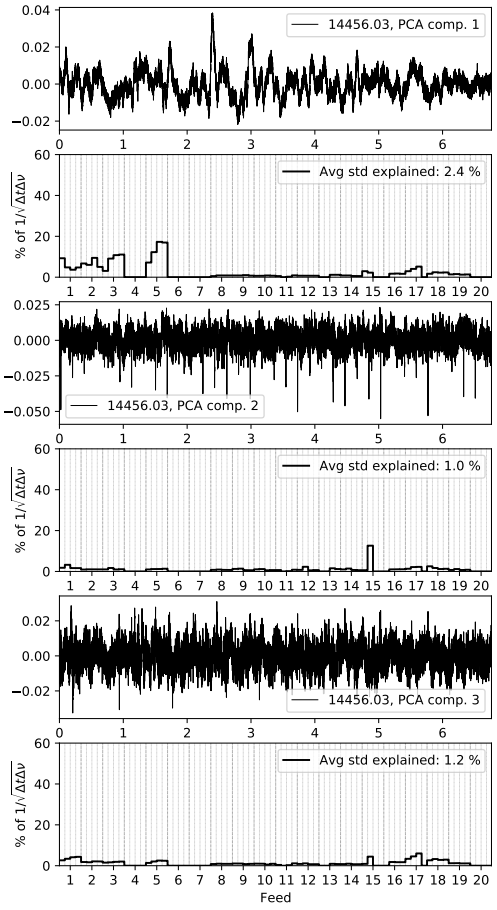

Figure 13: The three leading PCA components of a typical scan, and which feeds are affected.

columns), and compare the average absolute correlation within this group with the scatter expected from white noise alone. Any channel with an absolute correlation larger that 5σ is removed from further analysis.

Our second approach is to calculate a set of diagnostics for individual frequency channels, for instance the average correlation of the channel in question to all the others in the same sideband, or the average absolute value of the same. We then compare the values of these diagnostics for the different channels and remove significant outliers.

In addition to these approaches we look specifically for *edge correlations*, that is, correlations between individual frequencies at the edge of each sideband with the corresponding frequencies at the edge of another sideband. This is motivated by a known aliasing effect (Lamb et al. 2021), and we want to remove frequencies where this aliasing is large.

After the full mask has been determined, we apply this to the original dataset, and repeat the filtering described above, but now only using the unmasked data. This prevents bad data from contaminating good data through the various non-local filters.

3.4. *Calibration*

With cleaned and co-added TOD in hand, the final basic step before mapmaking is calibration; that is, assigning a noise temperature scale to the detector readout. From Eq. (3), the overall noise level is proportional to $T_{\rm sys}$.

3.4.1. *Calibration using a reference load*

Ideally we would put a load of a known temperature in front of the telescope and above the atmosphere and compare the measured output power with the output power measured with no load. A good approximation is to use an ambient temperature load that covers the receiver feedhorn. Assuming that the telescope, ground spillover, and the atmosphere have the same temperature as the ambient load, the output power will be the same as if the load was above the atmosphere (Penzias & Burrus 1973). Taking into account the vertical temperature profile and the distribution of the absorbing components in the atmosphere, the corrections are only a few percent for the relevant wavelengths. To measure the system temperature we compare the readout when we have a vane $P_{\rm amb}$ and when we look at the cold sky $P_{\rm cold}$. From Eq. (2) we can estimate $T_{\rm sys}$ as

$$T_{\rm sys} = \frac{T_{\rm amb} - T_{\rm CMB}}{P_{\rm amb}/P_{\rm cold} - 1}, \tag{16}$$

where $T_{\rm amb}$ is the ambient temperature and $T_{\rm CMB}$ is the cold sky temperature. We then multiply the data $d(\nu, t)$ with the $T_{\rm sys}$ measurement to go from (normalized) detector units to temperature

$$d(\nu,t)\Big|_{\rm K} = d(\nu,t)\langle T_{\rm sys}(\nu,t)\rangle. \tag{17}$$

As both the atmosphere and the receiver gain varies over time, the measurements of $T_{\rm sys}$ vary over time. To get the most accurate estimation, we make use of the vane at the beginning and end of each observation. Figure 16 shows a typical example of how the estimated $T_{\rm sys}$ looks for a single obsID, as function of frequency. The temperature usually behaves as a relatively smooth function, with large spikes at specific frequencies. Many of

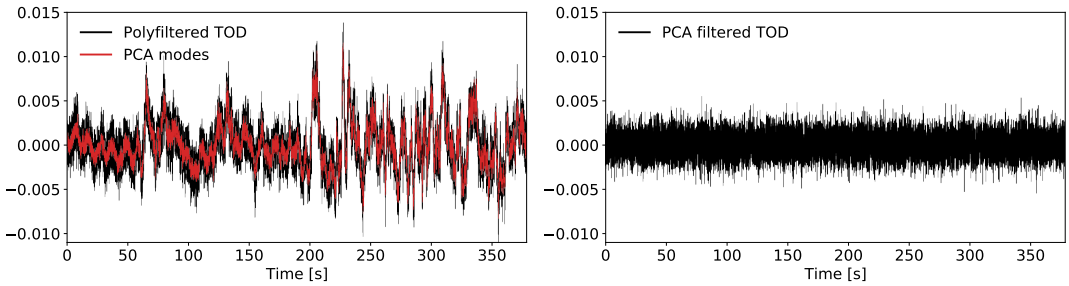

Figure 14: Effect of PCA filter on a "bad" scan, with especially heavy weather or standing wave contributions.

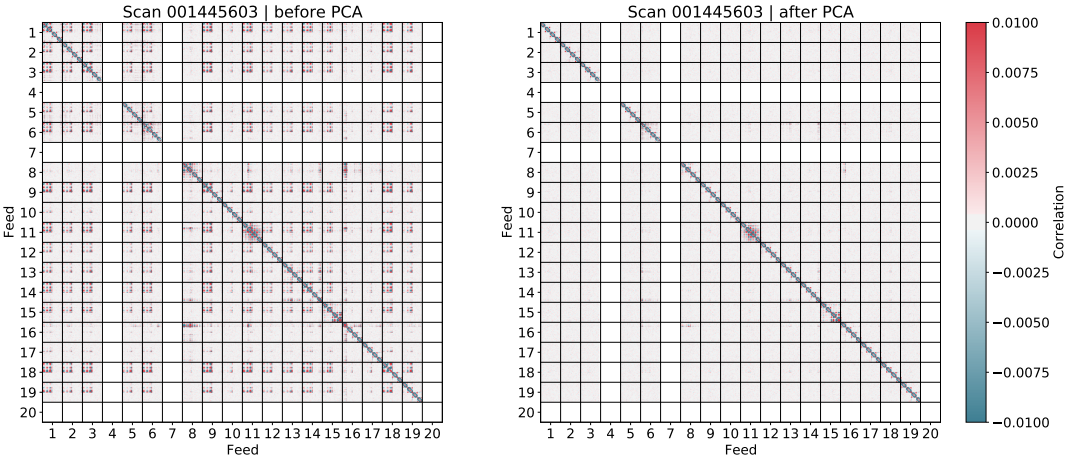

Figure 15: Comparison between channel–channel correlation matrices before (left panel) and after (right panel) filtering

these spikes are at frequency channels known to be bad, such at those close to the edges of the sidebands, but the cause of the others are largely unknown.

3.5. *Downsampling*

Until now, all steps have been performed at full frequency resolution, i.e., 1024 channels per sideband or 2 MHz channel bandwidth. For mapmaking purposes, however, we typically do not require such high resolution, since the signal-to-noise ratio of each channel is very low. To save both memory and computing time, we therefore co-add several neighboring frequency channels into a single low resolution channel as follows,

$$d_i^{\text{lowres}} = \frac{1}{\sum_m w_m} \sum_{m=(i-1)n_{\text{dec}}+1}^{in_{\text{dec}}} w_m d_m^{\text{highres}}, \qquad (18)$$

where d_i^{lowres} is the clean TOD of frequency channel i in the low resolution frequency grid, d_m^{highres} is the cleaned

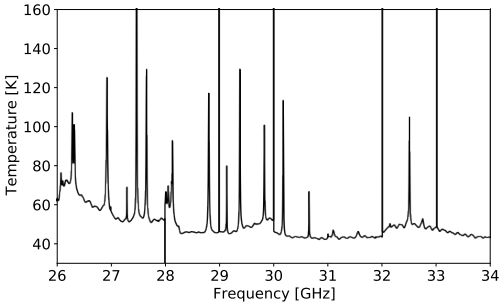

Figure 16: T_{sys} measurement from feed 1 of obsID 15117 across the 4096 frequency channels.

TOD of frequency channel m in the high resolution frequency grid, n_{dec} is the number (usually 16, corresponding to a final bandwidth of 31.25 MHz) of high resolu-

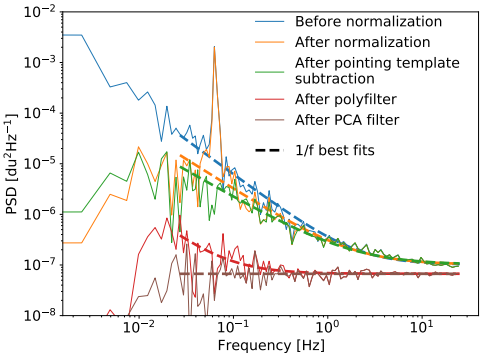

Figure 17: Power spectral distribution of a single scan from a 31.25 MHz band around 26.329 GHz at different stages in the pipeline, with $1/f$ noise curves fitted. The power spectral distribution is binned with logarithmic bin-sizes towards higher frequencies for clarity. Lower frequencies have been excluded from the fit, as these scales are greatly suppressed at the normalization stage.

tion frequencies to be combined in each low resolution frequency channel, and $w_m = 1/\sigma_m^2$ is the inverse variance of frequency channel m of the high resolution data (w_m is zero for masked frequencies).

3.6. *Noise characterization*

One of the most important functions of the COMAP data pipeline is the removal of correlated noise. As the CO brightness temperature is many orders of magnitudes below the telescope system temperature, any significant deviations from a white noise spectrum can be attributed to some systematic. The actual CO signal is also not expected to be correlated across feeds or frequencies in a significant fashion, giving us powerful leverage to separate signal from correlated noise in our data. In this section we describe our noise characterization model, and demonstrate how effectively the pipeline removes correlated noise from the data.

We can often approximate the correlated noise using a spectral density on the form

$$N(f) = \sigma_0^2 \left(1 + \left(\frac{f}{f_{\rm knee}}\right)^{\alpha}\right), \tag{19}$$

where σ_0 is the white noise level.

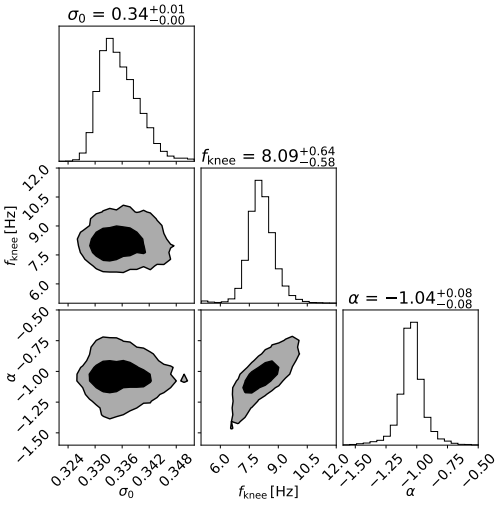

Figure 18: Distribution of noise parameters σ_0, $f_{\rm knee}$ and α for the c_0 coefficient of the polyfilter. All available scans of feed 1, sideband A:LSB was used.

The second term in Eq. (19) is known as $1/f$ noise,[3] which is characterized by a knee-frequency $f_{\rm knee}$, representing the transition frequency between the flat white noise and the sloped $1/f$ noise, and the exponent α, giving the slope of the spectral density in the $1/f$ dominated regime. The white noise level is estimated by calculating the variance between neighboring samples in the TOD, as

$$\sigma_0 = \sqrt{\frac{\mathrm{Var}(d_i - d_{i-1})}{2}}. \tag{20}$$

Figure 17 shows the $1/f$ behavior of the TOD throughout different steps in the pipeline, and clearly demonstrates the effect of each filter. The normalization filter heavily suppresses the low-frequency end of the specturm. The pointing template knocks out the strong ∼0.7 Hz correlation caused by the Lissajous scanning strategy. The polyfilter significantly reduces the noise power across the entire power spectrum, even lowering the white noise limit. This is possible because even though the white noise is uncorrelated in time, parts of it is still correlated in frequency. Finally, the PCA filter further reduces the noise left over by the polyfilter. By the end of the pipeline, the TOD is al-

[3] Keep in mind that f refers to the temporal frequency of the time ordered signal, not the observed photon frequencies, to which we consistently refer to as ν.

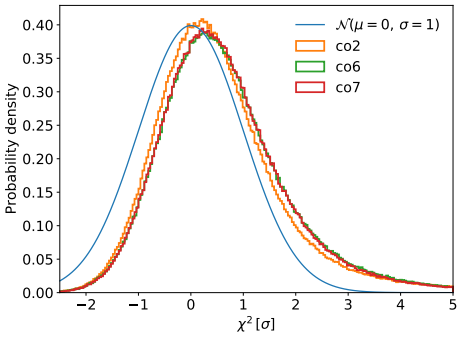

Figure 19: χ^2 distributions of filtered data for the three main fields, with a standard normal distribution for comparison.

most completely dominated by white noise. It should be noted that while the polyfilter typically suppresses much more noise power than the PCA filter in an average scan, this is not always the case. In scans with significant noise contamination (like bad weather), the PCA filter may suppress even more noise power than the polyfilter. Such a scan is shown in Fig. 14.

3.6.1. *Polyfilter noise properties*

As discussed in Sect. 3.3.3, the polyfilter involves fitting and subtracting a low order polynomial in frequency space from each sideband at each individual timestep. The polyfilter is the first filter targeting correlated noise except on the very larges timescales, and the resulting coefficients are therefore highly informative regarding the noise properties of the data. In the current analysis setup, we only use a first order polynomial filter, such that each timestep of each sideband are associated with two coefficients, c_0 and c_1. These coefficients, treated as functions of time, turn out to have $1/f$-like power spectra. Figure 18 shows the distribution of noise parameters of $1/f$ fits performed on c_0 for all available scans of the A:LSB sideband of feed 1. As dicussed in Sect. 2.3 the correlated noise common to each sideband is mostly dominated by gain fluctiations of the individual low noise amplifiers at each feed. We therefore expect, and find, that each feed has its own characteristic noise properties. Since we can use the polyfilter to remove this correlated noise, the individual noise properties of the different feeds are less important when measuring the CO line emission than if we were measuring continuum sources, in which case these properties would become crucial.

3.6.2. *Goodness-of-fit, χ^2-test*

The main goal of our pipeline is to remove both correlated noise and continuum foregrounds, while leaving as much as possible of the CO line intensity signal intact. In the ideal case, and assuming that the cosmological CO signal is so weak that it can not be measured in a single scan, our cleaned TOD should therefore be described by white noise alone. We therefore need statistics to measure potential deviations from white noise. We use a standard χ^2 statistic per scan for this purpose, defined as follows,

$$\chi^2 = \frac{\sum_{i=0}^{N} \left(\frac{d_i}{\sigma_0}\right)^2 - N}{\sqrt{2N}}. \tag{21}$$

Here d_i are the N samples of the scan, and σ_0 is the white noise level defined in Eq. (20). For a perfect white noise TOD, we expect $\chi^2 \sim \mathcal{N}(\mu = 0, \sigma = 1)$.

Figure 19 shows the χ^2 distribution for the set of scans available at the time of writing of this paper, comprising about 5 000 hours of observations, divided by observational field. Here we have combined all the datapoints for each sideband, such that the $N = n_{\rm samp} \cdot n_{\rm freq}$, where $n_{\rm samp}$ is the number of samples in time (typically $n_{\rm samp} = 10$–$20\,000$) and $n_{\rm freq} = 64$ is the number of frequencies per sideband. As seen in Fig. 19, the data are indeed very close to white noise, with only a small shift and a positive tail. We also note that the CO2 field outperforms the two other fields by a small margin. Given that $N \sim \mathcal{O}(10^5)$, a mean bias of only $1\,\sigma$ per scan suggests that the typical residual variance from correlated noise is very low.

3.7. *Data Selection and splits*

While typical cleaned COMAP data appears to be well described by white noise per scan, this does not necessarily imply that they are free from systematics to a level required for scientific analysis. On the contrary, many effects may only be discovered when co-adding both over time and frequency. For this reason, we apply several data cuts based on criteria that have been established through joint analysis of full-season observations, using null-tests as an important tool.

The main data selection step happens after we have already generated Level-2 files. In `accept_mod` we first scan the full data set to make a database of various statistics and housekeeping data for each sideband of each feed and each scan. This database allow us to find patterns and correlations at levels far below the noise level of an individual scan. For example we can exclude all data with an elevation lower than 35 degrees, or any sideband in which the raw data exhibits many spikes in a

particular scan. This procedure produces an accept list that simply lists all acceptable scans. We note that these cuts are made after the main filtering steps, which by far are the most computationally expensive part of the pipeline, and this allows efficient exploration of different data cuts with minimal computational costs.

In addition to data cuts, we use the statistics database also to define data splits, which later are used for null-tests and cross-spectrum evaluation. Examples include a half-mission split, where the full-season data are simply divided chronologically into the first and self half of the observation period; a sidereal time split; or various instrumental statistics splits. The mapmaker uses the accept list and split list (if such are requested) provided by `accept_mod`, and outputs two maps for each splits and one co-added all-data map, each of which only includes accepted data within its own split definition.

3.8. *Mapmaking*

The last step in the pipeline is mapmaking, which is implemented in a code called `tod2comap`. This reads in cleaned TOD and pointing information, applies a high-pass filter, and produces temperature sky maps for each frequency channel. The highpass filter removes the noise caused by the lower frequencies. This is done by Fourier transforming the TOD, and removing the part with frequency below a set value, typically 0.02 Hz, before transforming back to TOD.

Ideally, the TOD can be written as a sum of the signal $\mathbf{s}$ and the noise $\mathbf{n}$,

$$\mathbf{d} = \mathbf{P}\mathbf{s} + \mathbf{n}, \tag{22}$$

where $\mathbf{P}$ is the pointing matrix, which connects each time sample to a pixel on the sky. Our goal is to estimate $\mathbf{s}$ given $\mathbf{d}$. Assuming that the noise is Gaussian distributed with a time-domain covariance matrix $\mathbf{N}$, the log-likelihood function corresponding to Eq. (22) may be written as

$$\log \mathcal{L} \propto (\mathbf{d} - \mathbf{P}\mathbf{s})^{\mathrm{T}} \mathbf{N}^{-1} (\mathbf{d} - \mathbf{P}\mathbf{s}) . \tag{23}$$

Setting the derivative of this log-likelihood to 0, we obtain the standard mapmaker equation,

$$\hat{\mathbf{s}} = \left(\mathbf{P}^{\mathrm{T}}\mathbf{N}^{-1}\mathbf{P}\right)^{-1} \mathbf{P}^{\mathrm{T}}\mathbf{N}^{-1}\mathbf{d}. \tag{24}$$

As discussed above, the COMAP noise after filtering is very close to white, and this implies that $\mathbf{N}$ may be approximated as diagonal.[4] In that case, Eq. (24) may be solved explicitly and independently for each pixel p as follows,

$$\hat{\mathbf{s}}_p = \frac{\sum_{t\in p} \sigma_t^{-2} d_t}{\sum_{t\in p} \sigma_t^{-2}}. \tag{25}$$

Here σ_t is the noise standard deviation of sample t, and samples with lower noise are thus weighted more strongly than the samples with higher noise. The corresponding map-domain noise standard deviation is given by

$$\sigma_p = \left(\sum_{t\in p} \frac{1}{\sigma_t^2}\right)^{-1/2}. \tag{26}$$

4. HIGHER-LEVEL ANALYSIS

Scientific exploitation and interpretation of the full first-season COMAP data set are presented in a series of companion papers, including Chung et al. (2021); Cleary et al. (2021); Harper et al. (2021); Ihle et al. (2021); Silva et al. (2021). In this section, however, we analyze a small sub-set of the full data volume for demonstration purposes and to assess the performance of the pipeline itself.

4.1. *3D power spectrum and white noise simulations*

We restrict this preliminary analysis to the spherically averaged power spectrum of the CO signal, defined as follows,

$$\mathcal{P}(k) = \frac{\langle |f_{\mathbf{k}}|^2 \rangle}{n_x n_y n_z} V_{\mathrm{vox}}, \tag{27}$$

This is extracted from the temperature sky maps by, first, computing the 3D Fourier transform of the maps; binning the squared Fourier coefficients according to the wave number, k; and averaging over all contributions to a given k-bin. Finally, they are multiplied by the co-moving voxel volume, V_{vox}, and divided by the total number of voxels, $n_x n_y n_z$. Each voxel is inverse variance weighted by σ_p^{-2} as given by Eq. (26) before computing the Fourier transform.

The same noise RMS map is also used to simulate the noise level in this spectrum. Specifically, we generate $n_{\mathrm{sim}} \approx 10$ random white noise map-domain simulations according to Eq. (26), compute the 3D power spectrum from each, and define the uncertainty in each k-bin to the be standard deviation among these simulations. For applications in which multiple independent sky maps are available, either from different regions of the sky or

[4] This is not strictly correct for long time-scales. As such, the current mapmaker is statistically slightly sub-optimal, and the resulting transfer function is lower than strictly necessary. Future implementations of the COMAP mapmaker will therefore instead rely on well-established destriping or maximum-likelihood algorithms, which are often able to recover slightly more large-scale information than a binning mapmaker.

from different data splits, a mean power spectrum can be calculated as

$$\bar{\mathcal{P}}(k) = \frac{\sum_i^N \sigma_{i,k}^{-2} P_i(k)}{\sum_i^N \sigma_{i,k}^{-2}}, \tag{28}$$

where $\mathcal{P}_i(k)$ and $\sigma_{i,k}$ are the spherically averaged power spectrum of a single map and its uncertainty, respectively. The associated error bars at a given scale, k, become

$$\bar{\sigma}_k = \left(\sum_{i=1}^{N} \frac{1}{\sigma_{i,k}^2} \right)^{-1/2}. \tag{29}$$

The same approach can also be used for cross spectra.

4.2. *Transfer functions*

As the raw data passes through our filtering and map-making procedures, some of the signal is typically lost at each stage, and the maps described in Sect. 3.8 are therefore biased. In order to estimate and correct for this bias at each scale, $\vec{k}$, we need to estimate the so-called pipeline transfer function, which is simply defined as the power ratio between the recovered and original signal. We can estimate this transfer function by adding a signal-only simulation to a pure noise TOD, and then comparing the combined signal-plus-noise simulation output to the true signal-only input. We adopt the raw COMAP TOD as a model for the noise, which in power units are denoted $P_{\rm N}$. The signal-only contribution is produced by scanning a pre-computed 3D N-body simulation of brightness $T_{\rm sim}(p)$ with the telescope pointing, and we denote this $P_{\rm S}$. We then add these together in power units,

$$P_{\rm S+N} = P_{\rm N} + P_{\rm S} = k_B G \Delta\nu T_{\rm sys} \left(1 + \frac{T_{\rm sim}}{T_{\rm sys}}\right). \tag{30}$$

We then separately generate 3D voxel maps from $P_{\rm S}$, $P_{\rm N}$ and $P_{\rm S+N}$, and from these we compute corresponding 3D power spectrum, $\mathcal{P}_{\rm S}$, $\mathcal{P}_{\rm N}$, and $\mathcal{P}_{\rm S+N}$, following the above procedure. Based on these three spectra, we can finally estimate a scale-dependent transfer function $T(\vec{k})$ as

$$T(\vec{k}) = \frac{P_{\rm S+N}(\vec{k}) - P_{\rm N}(\vec{k})}{P_{\rm S}(\vec{k})}. \tag{31}$$

Noting that the pipeline filters have very different impact in the spatial and frequency directions, it is in some cases useful to decompose $\vec{k}$ into parallel and orthogonal modes, $\vec{k} = (k_\parallel, k_\perp)$, and we will consider both 1D and 2D transfer functions in the following. First, to understand the impact of the various filters in terms of signal-loss, we estimate 1D transfer functions for a range of different pipeline configurations. Specifically, we analyze six obsIDs (three CES and three Lissajous obsIDs), where we consider different combinations of PCA and polyfilter, enabling or disabling each filter in turn. For the polyfilter, we additionally consider two cases, namely a constant fitting term or a linear fitting function. The results from these calculations are summarized in Fig. 20. The black solid line shows the default pipeline configuration.

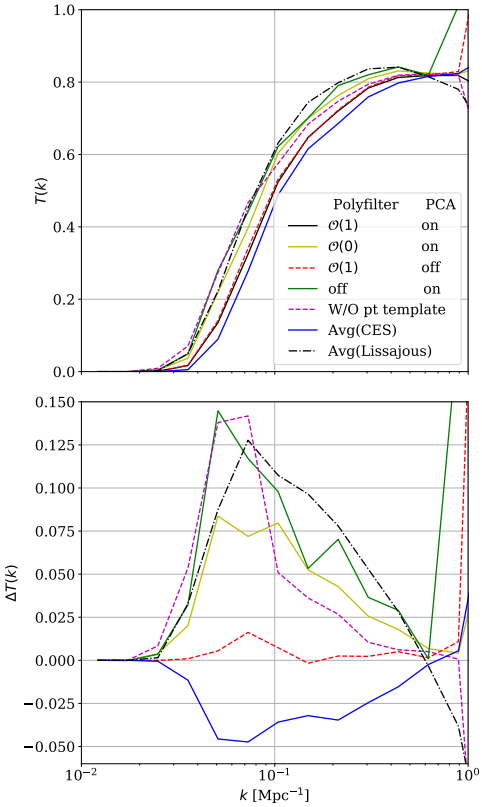

Figure 20: (*Upper panel*:) 1D transfer functions, $T(k)$, for different filter options and scanning modes as a function of scale, k. The default combination used in the COMAP pipeline is shown as a solid black line. (*Lower panel*:) Difference between the various filter and scanning options and the default configuration.

One can see that the default settings, i.e., a polyfilter of $\mathcal{O}(1)$ and PCA filtering turned on, yields almost the same transfer function as the case where the PCA filter is turned off. The PCA filter is not expected to remove much of the actual input signal, as it only removes the

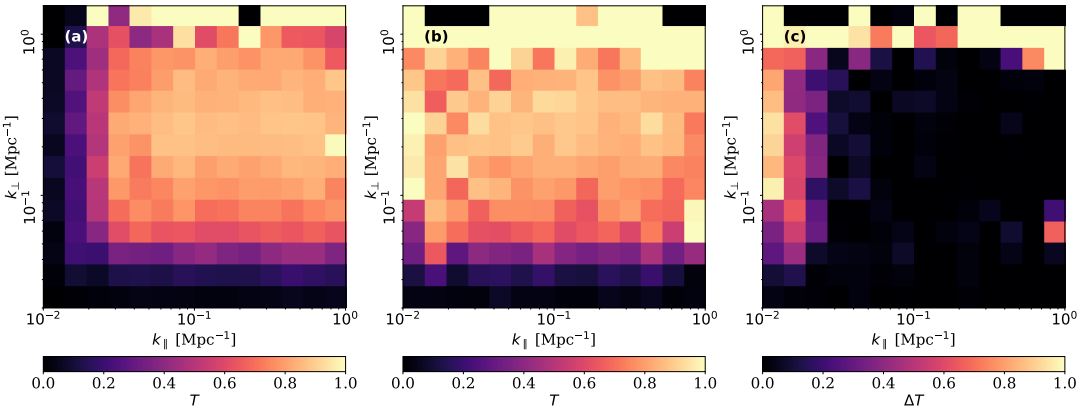

Figure 21: Comparison of 2D transfer function estimates with (*left panel*) and without (*middle panel*) the polyfilter. (*Right panel*:) Difference between the two previous cases.

components of the TOD that are the most correlated over all frequencies and feeds, thus potentially removing only the structures of the input signal that are common over the entire survey volume observed at any given time.

When it comes to the case with polyfilter of $\mathcal{O}(0)$ or turned completely off there are, however, large differences seen from the results using the default settings. Using a polyfilter of $\mathcal{O}(0)$ or $\mathcal{O}(1)$, a considerable fraction of the input signal is removed by the pipeline on scales above $k \sim 0.04\,\mathrm{Mpc}^{-1}$. We see that the $\mathcal{O}(0)$ polyfilter yields a similar result as without the polyfilter near the peak regions of the transfer functions, however a non-negligible portion of signal between $k \sim 0.04\,\mathrm{Mpc}^{-1}$ up to the peak region is taken out when turning on the $\mathcal{O}(0)$ polyfiltering. The low transfer efficiency on low k for any of the shown filter combinations is likely due to the limited area covered in each scan, and the highpass filter imposed in the mapmaker, as well as the polynomial filter in frequency.

If we turn off the subtraction of the pointing templates, we can also see in Fig. 20 that more signal is let through the pipeline on scales $k \leq 0.3\,\mathrm{Mpc}^{-1}$. The effect of the pointing template subtraction is however especially noticable on scales $k < 0.1\,\mathrm{Mpc}^{-1}$, which is expected as the structures in the power spectra induced removed by the pointing templates are of a larger scale in the pixel domain.

Note also that when computing these transfer functions for different filter combinations, we used the combined maps of three obsIDs of type Lissajous scan and three with constant elevation scans. However, we found that there were significant differences between the transfer functions from a Lissajous and CES scan type and have therefore also included the average of the three transfer function of each type in Fig. 20. As one can see, the Lissajous scan type results in a transfer function that is larger on most scales, which probably is a result of the Lissajous scan being more efficient in covering the sky in the same time as the constant elevation scans. The Lissajous scans, as opposed to the ones with CES, also seem to result in a transfer function that drops a bit down from its peak at high k. The reason for this difference is not yet fully understood at this point.

When looking at the 2D version of the transfer function, as shown in Fig. 21, the effects of the polyfilter on the transfer function become more evident because we can then distinguish between what is happening in the angular directions ($k_\perp$) and the spectral dimension ($k_\parallel$). As the polyfilter is designed to remove the $1/f$-noise as well as continuum foreground emission along the frequency dimension on each sideband, we expect the changes in the transfer function to be most visible in the large line-of-sight scales $k_\parallel$. This is indeed what is seen in the difference $\Delta T(k)$ between the transfer functions without and with an $\mathcal{O}(1)$ polyfilter in Fig. 21 for low $k_\parallel$, where we note a $50-90\,\%$ relative loss in power when using a $\mathcal{O}(1)$ polyfilter. Meanwhile on all other scales the difference ΔT is left mostly unchanged.

4.3. *Cross-spectra and null-tests*

As already discussed in Sect. 3.7, we split the original TOD into two or more disjoint parts with respect to some feature before computing final power spectra. Example splits include, for instance, even versus odd numbered obsIDs; daytime versus nighttime obsIDs; first versus second half of the mission; or a split in sidereal time. To minimize the impact of potential residual sys-

tematic errors, any final power spectrum estimate only includes cross-correlations between independent maps of these types, computed as follows,

$$\mathcal{C}(k) = \frac{\left\langle \mathrm{Re}\left(f_{1,\mathbf{k}} f^*_{2,\mathbf{k}}\right)\right\rangle}{n_x n_y n_z} V_{\mathrm{vox}}. \tag{32}$$

Here $f_{1,\mathbf{k}}$ and $f_{2,\mathbf{k}}$ denote the Fourier modes of two independent maps within some split group, and, as usual, we bin this quantity according to k. We compute individual maps for each feed in each split group, and we may therefore individually cross-correlate each combination of feeds from the first and second split. This results in a total of 16×16 individual cross-spectra, after excluding feeds that were turned off for at least some duration of the experiment. From these, we calculate the associated mean and uncertainty according to Eqs. (28) and (29).

We perform null-tests using the same data organization, but in this case we subtract pairwise mean cross-spectra. Specifically, for each data split we first compute all possible cross-spectra of the form

$$\mathcal{C}_{ij} = \left\langle \mathrm{Re}\left(f_{1,i} f^*_{2,j}\right)\right\rangle, \tag{33}$$

where $f_{1,i}$ and $f_{2,j}$ are the Fourier coefficients derived for feeds $i, j = 1, \ldots, n_{\mathrm{feed}}$ for each of the two data splits, and we have dropped normalization constants for notational convenience. We then divide these spectra into two disjoint sets ($i < j$ and $i > n_{\mathrm{feed}} - j$), and compute the mean cross-spectrum over each set separately, $\bar{\mathcal{C}}_1$ and $\bar{\mathcal{C}}_2$. Finally, we compute the half-difference between these spectra, $(\bar{\mathcal{C}}_1(k) - \bar{\mathcal{C}}_2(k))/2$, and the associated uncertainty, $\sigma_{\mathcal{C}} = \sqrt{\bar{\sigma}_1^2 + \bar{\sigma}_2^2}/2$. To account for filtering effects, we divide both the mean and uncertainty estimates by the product of two independent transfer functions. The first accounts for the pipeline filtering, as described in Sect. 4.2, and this increases both the mean estimates and uncertainties on large scales. The second transfer function affects the smallest scales, and accounts for beam smoothing and the intrinsic line width of the CO emission. The impact of aperture efficiency has not been included yet. One example of such a calculation is shown in Fig. 22, derived from the day vs night data split.

4.4. *Result preview*

For completeness, we end this paper by briefly showing a few examples of the final results derived using the COMAP pipeline.

Figure 23 shows a single frequency map of the CO7 patch based on the nightime data from the first year of observations. This data set results in a sensitivity of a few tens of μK per 2×2 arcmin^2 pixel for a single

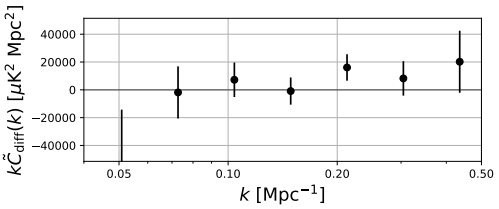

Figure 22: Example of null cross-spectrum, used for goodness-of-fit evaluation and pipeline validation. This particular example is derived analyzing data from the CO2 field, and using the day versus night data split.

31.25 MHz channel. At least at a visual level, the map appears largely dominated by white noise.

Figure 24 shows the noise weighted average of all feed-feed cross spectra for the time-of-day split of the nightime-only constant elevation scan of the CO2 field. We see that COMAP is mostly sensitive to scales with $k \in 0.05 - 0.5$ Mpc^{-1}. The error bars on smaller scales increase because of the instrumental beam and intrinsic line width of CO emission, while the error bars on large scales increase because of our limited field size and the filtering transfer function. This cross spectrum is statistically consistent with zero, which is expected given the limited amount of data used for this calculation. Most importantly, however, is the fact that we do not see evidence of any significant systematic errors, and this suggests that the noise level is indeed integrating down as expected from the measured white noise level and transfer function. (Note that we have not yet taken into account the aperture efficiency when estimating the transfer function, and we therefore expect the uncertainties to increase by an additional $\mathcal{O}(1)$ factor.)

5. SUMMARY AND CONCLUSIONS

We have presented the data analysis pipeline used to process the first-season COMAP observations with respect to high-redshift CO emission, from raw time-ordered data to final calibrated maps. This pipeline implements four main steps (calibration, filtering, data selection and mapmaking), each of which are designed to optimally exploit the unique instrumental capabilities of the COMAP instrument. For instance, calibration is performed using a combination of frequent comparison with a hardware calibrator and real-time total power measurements. The filtering procedure explicitly exploits the multi-feed and -frequency design of the COMAP instrument to reject common-mode contaminants, resulting in data that are strongly dominated by uncorrelated white noise after filtering. Finally, both the data selection and mapmaking processes directly use

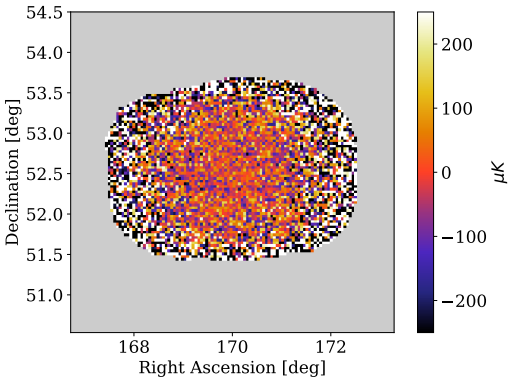

Figure 23: Co-added COMAP CO7 map using night-time data only. Regions that are either not observed by the telescope, or have a noise level $\sigma_{\rm rms} > 1000\,\mu$K, are masked out.

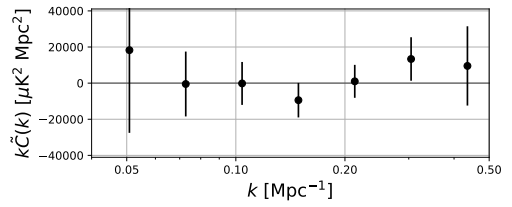

Figure 24: Weighted mean of cross spectra calculated for all feed-feed combinations from time-of-day map-split for the constant elevation scans of the CO2 field night-time data.

this fact to produce near-optimal goodness-of-fit statistics and pixelized sky maps with very limited computational costs.

Based on preliminary analyses, the performance of this pipeline appears good, not only in terms of data quality per scan, but also in terms of full-season co-added maps. No obvious systematic artefacts are seen either in the final maps or power spectra. Indeed, demonstrating a system with precisely this property—that the noise level integrates down with time as expected for pure noise—was the defining goal of the first phase of the COMAP pathfinder experiment. This has been achieved through the current work. The next operational goal of the COMAP experiment is to increase the overall mapping speed by adding more telescopes and detectors, and then observe for a longer time. Comparing the demonstrated COMAP performance with current state-of-the-art theoretical models, a first positive robust detection seems fully realistic with one or two more years of observing time (Ihle et al. 2021).

A second important goal for the first COMAP phase was to optimize the experimental setup, and thereby both reduce systematic effects and increase the overall mapping speed. In that respect, we conclude that two-stage polarizers are highly preferrable over single-stage (or no) polarizers, as they significantly reduce the impact of standing waves. Secondly, we find that a constant elevation-based scanning strategy is preferrable over a Lissajous-based scanning strategy, and results in significantly better data quality. Simply installing dual-stage polarizers in all feeds, and always employing CES scanning, will by itself increase the effective mapping speed by more than a factor of 2.5.

We acknowledge support from the Research Council of Norway through grant 251328. Thanks to graduated master students Erik Levén and Maren Rasmussen for their contribution to this pipeline.

REFERENCES

Bernal, J. L., Breysse, P. C., Gil-Marín, H., et al. 2019, PhRvD, 100, 123522.

Breysse, P. C., Anderson, C. J., & Berger, P. 2019, PhRvL, 123, 231105.

Breysse, P. C., Kovetz, E. D., Behroozi, P. S., et al. 2017, MNRAS, 467, 2996.

Chung, D. T. 2019, ApJ, 881, 149.

Chung, D. T. et al. 2021, (in preparation)

Chung, D. T., Li, T. Y., Viero, M. P., et al. 2017, ApJ, 846, 60.

Chung, D. T., Viero, M. P., Church, S. E., et al. 2019, ApJ, 872, 186.

Cleary, K. et al. 2021, (in preparation)

Harper, S. et al. 2021, (in preparation)

Hill, G. J., Gebhardt, K., Komatsu, E., et al. 2008, Panoramic Views of Galaxy Formation and Evolution, 399, 115

Ihle, H. T., Chung, D., Stein, G., et al. 2019, ApJ, 871, 75.

Ihle, H. T. et al. 2021, (in prep)

Keating, G. K., Marrone, D. P., Bower, G. C., et al. 2016, ApJ, 830, 34.

Keating, G. K., Marrone, D. P., Bower, G. C., et al. 2020, ApJ, 901, 141.
Kovetz, E., Breysse, P. C., Lidz, A., et al. 2019, BAAS, 51, 101
Kovetz, E. D., Viero, M. P., Lidz, A., et al. 2017, arXiv:1709.09066
Lamb, J. W. et al. 2021, (in preparation)
Li, T. Y., Wechsler, R. H., Devaraj, K., et al. 2016, ApJ, 817, 169
Padmanabhan, H. 2018, MNRAS, 475, 1477.
Papovich, C., Shipley, H. V., Mehrtens, N., et al. 2016, ApJS, 224, 28
Pavesi, R., Sharon, C. E., Riechers, D. A., et al. 2018, ApJ, 864, 49.
Penzias, A. A. & Burrus, C. A. 1973, ARA&A, 11, 51.
Planck Collaboration, Aghanim, N., Akrami, Y., et al. 2020, A&A, 641, A1.
Silva, M. et al. 2021, (in preparation)
Weiland, J. L., Odegard, N., Hill, R. S., et al. 2011, ApJS, 192, 19

Paper II

First Season COMAP Results: Power spectrum methodology and preliminary data quality assessment (draft)

H. T. Ihle and the COMAP collaboration

ApJ, in preparation.

II

Draft version February 4, 2021
Typeset using LaTeX **twocolumn** style in AASTeX63

First Season COMAP Results: Power spectrum methodology and preliminary data quality assessment

Håvard T. Ihle[1] and the COMAP collaboration

[1] *Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, N-0315 Oslo, Norway*

ABSTRACT

We present the power spectrum methodology used for the first-season COMAP analysis, and assess the quality of the current data set. The main results are derived through a Feed-feed Pseudo-Cross-Spectrum (FPXS) method, which is a robust estimator with respect to both noise modeling errors and experimental systematics. We use effective transfer functions to take into account the effects of instrumental beam smoothing and various filter operations applied during the low-level data processing. We find that power spectrum results derived from constant elevation scan observations are consistent with white noise expectations, and this demonstrates that the COMAP experiment and analysis pipeline are able to suppress both correlated noise and systematics below the white noise level. However, results derived from Lissajous scan observations, for which the telescope motion scan continuously in elevation during observations, suggest residual systematics on large angular scales that are not sufficiently suppressed to support astrophysical constraints at this time. We speculate that these residuals may be caused by residual atmospheric and/or sidelobe contamination, and these effects will be considered further in future publications.

1. INTRODUCTION

This paper, one of a set associated with the first-season COMAP analysis, presents the methodology used to constrain the CO power spectrum with COMAP data. An overview of the COMAP experiment is presented by Cleary et al. (in prep), while the COMAP instrument is described by Lamb et al. (in prep).

The low-level COMAP data processing pipeline is summarized by Foss et al. (in prep). This pipeline converts raw uncalibrated observations into three-dimensional maps, using redshifted CO line emission from distant galaxies as a tracer of the cosmic density field. Since the first-season COMAP instrument observes at frequencies between 26 and 34 GHz, and the rotational CO 1-0 transition has a rest frequency of 115 GHz, the current measurements trace galaxy formation at redshifts between $z = 2.4$ and 3.4, which often is referred to as "the epoch of galaxy assembly".

One common and powerful quantity used to characterize the statistical properties of such three-dimensional cosmic maps is the so-called power spectrum (or the two-point correlation function), which measures the strength of fluctuations as a function of physical distance (e.g., Bernal et al. 2019; Ihle et al. 2019; Chung 2019; Gong et al. 2020; Yang et al. 2020). For an isotropic and Gaussian random field, this function quantifies all statistically relevant information in the original data set, but with a far smaller number of data points, and it thus represents a dramatic compression of the full data set. Even for non-Gaussian fields, such as the galactic density field, the power spectrum encapsulates a large fraction of the important information, and it is therefore an efficient tool even for such fields.

However, while compressing hundreds of terabytes of raw data into a handful of power spectrum coefficients certainly makes the interpretation of the data easier in terms of theoretical comparisons, it also makes the final estimates highly sensitive to small systematic effects and instrumental noise. To guide our intuition, we note that current theories predict a intrinsic CO standard deviation per resolution element of no more than a few microkelvins, which is to be compared with a typical system temperature of 45 K for the COMAP instrument; or atmospheric fluctuations of a few kelvins; or sidelobe contributions of a few millikelvins. All such effects must therefore be suppressed by many orders of magnitude in order to establish robust astrophysical constraints.

As described by Lamb et al. (in prep); Foss et al. (in prep), the COMAP instrument has many unique features that makes this process more robust. A few important examples include highly efficient spectroscopic

Corresponding author: Håvard T. Ihle
h.t.ihle@astro.uio.no

rejection of common-mode signals, a strongly interconnected and configurable scanning strategy, and frequent usage of hardware calibrators. Still, systematics rejection at the microkelvin level is highly challenging, and the current paper describes several algorithmic methods that can be applied to improve the robustness of the final data.

The rest of the paper is organized as follows. In Sect. 2 we review the adopted COMAP power spectrum estimators. Results as derived from the actual data are presented in Sect. 3, and we conclude in Sect. 4.

2. METHODS

We begin our discussion with an overview of the fundamental algorithms used for COMAP power spectrum estimation. For other recent examples of the use power spectrum analysis on intensity mapping data see e.g. (Uzgil et al. 2019; Mertens et al. 2020; Keating et al. 2020).

2.1. *Auto-spectrum analysis*

Let m_{ijk} denote a three-dimensional map, and let us call each resolution element in this map a voxel. We define the power spectrum $P(\vec{k})$ of this map to be the variance of its Fourier components, $f_{\vec{k}}$,

$$P(\vec{k}) = \frac{V_{\rm vox}}{N_{\rm vox}} \langle |f_{\vec{k}}|^2 \rangle, \tag{1}$$

where k is the wave vector of a given Fourier mode, $V_{\rm vox}$ is the volume of each voxel, and $N_{\rm vox}$ is the total number of voxels.

If we assume that the map is statistically isotropic, then the power spectrum will only be a function of the magnitude of the wave vector, $P(\vec{k}) = P(k)$. In observational cosmology we often want to distinguish between the angular directions (denoted by the x and y coordinates) from the line-of-sight direction (denoted by the z coordinate). This is because the map typically has different properties in the different directions, for example due to instrumental beam effects or redshift space distortions. It is therefore often useful to define the power spectrum in terms of parallel (line-of-sight) modes, $k_\parallel \equiv |k_z|$, and the perpendicular (angular) modes, $k_\perp \equiv \sqrt{k_x^2 + k_y^2}$. We can estimate the power spectrum in a given set of $\vec{k}$-bins, $\{\vec{k}_i\}$, from a given map as

$$P(\vec{k}_i) \approx \frac{V_{\rm vox}}{N_{\rm vox} N_{\rm modes}} \sum_{j=1}^{N_{\rm modes}} |f_{\mathbf{k_j}}|^2 \equiv P_{\vec{k}_i}. \tag{2}$$

Assuming that foreground and systematic contributions have already been removed to negligible levels through pre-processing, the power spectrum of a cleaned line intensity map is typically modeled as a sum of a signal and noise component,

$$P(\vec{k}) = P_{\rm signal}(\vec{k}) + P_{\rm noise}(\vec{k}). \tag{3}$$

If we are able to estimate the noise power spectrum through independent means, for example using a noise model or simulations, we can extract the signal power spectrum simply by subtracting the estimated noise,

$$P_{\rm signal}(\vec{k}_i) \approx P_{\vec{k}_i} - P^{\rm est}_{\rm noise}(\vec{k}_i), \tag{4}$$

where $P^{\rm est}_{\rm noise}(\vec{k}_i)$ is the estimated noise power spectrum in bin number i.

If the map consists of uniformly distributed white noise, then the noise power spectrum is independent of $\vec{k}$ and given by

$$P_{\rm noise} = V_{\rm vox} \sigma_T^2, \tag{5}$$

where σ_T is the white noise standard deviation in each voxel (in units of kelvin). In our case, this magnitude of the white noise level is determined by the radiometer equation,

$$\sigma_T = \frac{T_{\rm sys}}{\sqrt{\delta_\nu \tau}}, \tag{6}$$

where $T_{\rm sys}$ is the system temperature of the detector, δ_ν is the frequency resolution of each voxel, and τ is the total time each pixel is observed.

In addition to this instrumental noise contribution, there is an intrinsic uncertainty when estimating the signal power spectrum from a map, often alled cosmic variance, that arises from the limited number of Fourier modes in the map. This contribution is given by

$$\sigma_P^2 = \langle (P_{\vec{k}_i} - P(\vec{k}_i))^2 \rangle \approx \frac{P(\vec{k}_i)^2}{N_{\rm modes}}, \tag{7}$$

where $N_{\rm modes}$ is the number of Fourier modes in bin number i, and the last approximation is exact when the Fourier modes are assumed to be independent Gaussian variables.

If the power spectrum is noise dominated, we can reduce this intrinsic uncertainty in two ways. First, we can observe for a longer time on the same area of the sky, thus decreasing the noise power spectrum contribution to the uncertainty. Alternatively, we can cover a larger sky area, and thus increase the number of measured Fourier modes. As long as we are noise dominated, a simple analysis suggests that observing a small area for a long time is more efficient for making a first detection than spreading the observations over a larger area. In a realistic situation, however, there are several other factors that must be taken into account, including the

choice of angular resolution and scanning strategy constraints, and these will typically limit how small fields it is possible to observe.

Another source of uncertainty in estimating the signal is the accuracy of the estimated noise power spectrum model. If this model is biased or uncertain, then the associated residuals will propagate directly into the estimate of $P_{\rm signal}(\vec{k}_i)$.

2.2. *Pseudo-spectrum analysis*

There are several challenges with an auto-spectrum analysis as described above, as discussed both in this and the following sections. First of all, if the noise in the map is not uniform, which it generally is not, the noise power spectrum will be dominated by the parts of the map with the highest noise level. In order to address this, it is necessary to devise a method that puts more weight on the parts of the map with low noise, and less weight the parts of the map with high noise.

The standard method of accounting for this is through inverse noise variance weights. That is, we weigh the map, m, by the noise level map, σ_m, before we compute the power spectrum,

$$\tilde{P}_{\vec{k}_i} = \frac{V_{\rm vox}}{N_{\rm vox} N_{\rm modes}} \sum_{j=1}^{N_{\rm modes}} |\tilde{f}_{\mathbf{k_j}}|^2, \tag{8}$$

where $\tilde{P}$ denotes the *pseudo*-spectrum, and $\tilde{f}$ are the Fourier components of the noise weighted map,

$$\tilde{m} \equiv wm, \tag{9}$$

and

$$w \equiv \mathcal{N} \frac{1}{\sigma_m^2}. \tag{10}$$

$\mathcal{N}$ is a single overall normalization constant (which we will get back to), and σ_m is, as usual, the noise level map.

On a general note, the term 'pseudo-spectrum' typically refers to a power spectrum estimator that is computed from a biased estimator of the true sky map, and is as such itself biased; see Hivon et al. (2002). This may be contrasted to more conventional power spectrum estimators that aim to estimate the power spectrum of the true sky signal. The statistical information content of the pseudo-spectrum and unbiased power spectrum is identical, and the main difference between the two classes of estimators concerns their ease of interpretation; while the unbiased power spectrum may be directly compared with theoretical models and other literature results, the pseudo-spectrum is experiment dependent, and typically requires simulations for proper statistical interpretation.

In our setting, we use the pseudo-spectrum to take into account both masked voxels (by setting $\sigma_m \to \infty$ for voxels that are excluded from further analysis) and varying noise levels across the map. Both these operations lead to *mode mixing*, i.e., different signal Fourier modes map are mixed together, and the estimated signal pseudo-spectrum is therefore a slightly distorted version of the true signal power spectrum. However, since we know exactly how the signal map has been distorted, we can, at least in principle, calculate the exact mode mixing matrix that is needed to reconstruct the mode mixing and obtain an unbiased signal spectrum from the pseudo-spectrum (Hivon et al. 2002). How feasible this is for a specific case depends on the details of the map dimensions and computational resources For more details on mode mixing, see App. C.

Although mode mixing does complicate the physical interpretation of the pseudo-spectrum, there are several ways of dealing with this without having to calculate and invert the full mode mixing matrix. First of all, if the analysis involves comparisons with signal simulations, then one may simply apply the same weight matrix to each simulation, making the observed and simulated power spectra statistically compatible. Second, if the level of mode mixing is modest, then the pseudo-spectrum may be an adequate estimator for the signal power spectrum for a given application. This typically holds particularly well for noise-dominated applications, for which a single power estimate covering a large range in k is desired; in that case, the mode mixing often has minimal effect on the estimates, and the pseudo-spectrum often is a perfectly valid estimate in its own right. The accuracy of this approximation must be assessed for each use case.

In cases for which the pseudo spectrum is intended to be used as a direct estimator, it is necessary to set the normalization factor $\mathcal{N}$ in Eq. 10 properly. To establish the *formally correct* value for this normalization is not entirely well defined. However, we can make a simple and fairly reasonable choice as follows

$$\mathcal{N} = \frac{1}{\sqrt{\left\langle \frac{1}{\sigma_m^4} \right\rangle}}, \tag{11}$$

where $\langle\rangle$ denotes average over the whole map. To make the results easier to interpret, we therefore apply this normalization to all results shown in this paper. For analyses that employ the full mode-mixing matrix, or in which the pseudo-spectrum is compared directly to simulations, this normalization is completely irrelevant.

To roughly estimate the expected level of mode mixing we calculate the ratio of the pseudo-spectrum and the

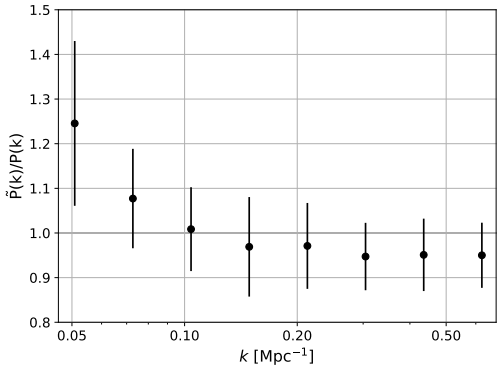

Figure 1: Ratio of the signal pseudo-spectrum to the signal auto-spectrum based on ten signal realizations.

auto-spectrum for ten signal realization maps. Figure 1 shows the mean and standard deviation of the mode mixing in each of the main power spectrum bins. Overall, we see that at the scales where we have most of our sensitivity, the effect of mode mixing is fairly modest, typically in the 5-30% range. Thus, even at face value, the pseudo-spectrum does provide a reasonable order-of-magnitude estimate of the true power spectrum, even if it may not be appropriate for precision analysis.

We leave it for future work to estimate the mode mixing matrix and undo the mode-mixing bias in the pseudo-spectra. For the rest of this paper we will interpret the pseudo-spectra at face value.

2.3. *Cross-spectrum analysis*

A general challenge when using either the auto- or pseudo-spectrum is that highly accurate estimates of the noise contribution are required to estimate the signal power spectrum. In many cases this can be very challenging, and any systematic error will directly bias the final signal estimate.

One way to avoid this complication is to use the so-called cross spectrum, $C(\vec{k})$. While the power spectrum quantifies the variance of the Fourier components of a single map, the cross spectrum quantifies the covariance between the Fourier modes of two different maps,

$$C(\vec{k}) = \frac{V_{\rm vox}}{N_{\rm vox}}\left\langle \mathrm{Re}\{f^*_{1\vec{k}} f_{2\vec{k}}\}\right\rangle \approx \frac{V_{\rm vox}}{N_{\rm vox} N_{\rm modes}} \sum_{j=1}^{N_{\rm modes}} \mathrm{Re}\{f^*_{1\vec{k}_j} f_{2\vec{k}_j}\} \equiv C_{\vec{k}_i}. \quad (12)$$

Here Re{} denotes the real part of a complex number, and f_1 and f_2 are the Fourier components of two maps m_1 and m_2.

Clearly, if m_1 and m_2 are identical, then the cross-spectrum is equivalent to the auto-spectrum. The advantage of the cross-spectrum, however, is that, if the maps m_1 and m_2 are independent, then the noise contributions are also independent, and they do not contribute to the mean of the cross-spectrum, but only to its variance. Therefore, it is not necessary to estimate and subtract the noise power spectrum to obtain an unbiased signal estimate, but rather

$$\langle C_{\vec{k}_i}\rangle = P_{\rm signal}(\vec{k}_i). \quad (13)$$

Of course, a proper noise estimate is still necessary for uncertainty estimation, but the requirements on this are typically far less stringent than for the estimator mean.

Although the cross spectrum strongly alleviates the need for estimating the noise power spectrum, we do pay a price in the form of somewhat lower intrinsic sensitivity. For instance, when splitting the data into two independent parts, and cross-correlate these, we do lose a factor of at least $\sqrt{2}$ from the fact that we do not exploit the auto-correlations within each data set separately. Fortunately, this problem can be remedied by splitting the data into more independent maps, and averaging the cross-spectra of all possible combinations. A lower limit on the cross-spectrum sensitivity is given by

$$\sigma_C^{N_{\rm split}} \geq \sqrt{\frac{1}{1-1/N_{\rm split}}}\sigma_P, \quad (14)$$

where $N_{\rm split}$ is the number of different map splits, and σ_P is the optimal sensitivity of the auto-spectrum derived from the full data set.

The cross-spectrum has some other very important advantages with respect to the auto-spectrum as well. As discussed in the introduction, one of the major challenges for an experiment like COMAP, in which we have to integrate down the noise several orders of magnitude in order to measure a small signal, are systematic errors. However, since the cross-spectrum may only be biased by common-mode contributions among the two maps, one is well advised to ensure that any known systematic effect contribute independently to the two maps. In that case, the systematic effects will not bias the signal estimate. Combining this insight with splitting the data into multiple parts allows us to design a power spectrum statistic that is far more robust to systematics than the auto-spectrum.

We define a pseudo-cross-spectrum in an analogous manner as for the pseudo-auto-spectrum. The only subtlety is that we make sure to apply the same weight map, w, for both maps. Explicitly, we adopt the follow-

ing weight map,

$$w_{1,2} \propto \frac{1}{\sigma_{m_1}\sigma_{m_2}}, \tag{15}$$

for both m_1 and m_2 when calculating the pseudo-cross-spectrum, $\tilde{C}_{\vec{k}_i}$.

2.4. *The Feed-feed Pseudo-Cross-Spectrum*

The idea of the Feed-feed Pseudo-CROSS-Spectrum (FPXS) method is to combine all the insights from the preceding sections to construct a single statistic for the CO signal that has a high intrinsic sensitivity, uses proper noise weighting, and that is robust against instrumental and other systematics. In that respect, we first note that the COMAP focal plane consists of 19 feeds, each with its own amplifiers and detectors. Furthermore, many systematics are particular to each feed, due to different bandpasses, amplifiers, cables, beams etc. We may therefore split the data according to feeds (i.e., make one map per feed), and then compute cross-spectra of all different feed combinations, while never correlating two maps from the same feed.

Second, we also note that one of the most troublesome systematics for COMAP is ground pickup. This is mainly because the ground contamination correlates with the pointing, and it therefore does not average down the same way as any systematic that is random in time-domain (and hence independent in different observations). We can make ourselves as robust as possible to any residual ground signal in our map by also splitting the data by the elevation of the observations, so that we never take the cross-spectrum of two different datasets taken at the same elevation.

With these considerations in mind, we define the following procedure for calculating the FPXS:

1. We split the data into disjoint sets sorted according to elevation. For simplicity we assume for now that we split the data into two sets, A and B, where A contains all the observations taken at elevations below the median elevation, and B contains all the observations from the higher elevations. We can easily generalize this to a case where we split the data into more than two sets.

2. For each set, A and B, we generate maps for each of the 19 feeds. We denote the different maps according to dataset and feed, such that A_{13} indicates the map that combines all data from dataset A for feed number 13.

3. We then calculate the pseudo-cross-spectrum, $\tilde{C}^{ij}_{\vec{k}_i}$ for all different map combinations of A_i and B_j where $i \neq j$.

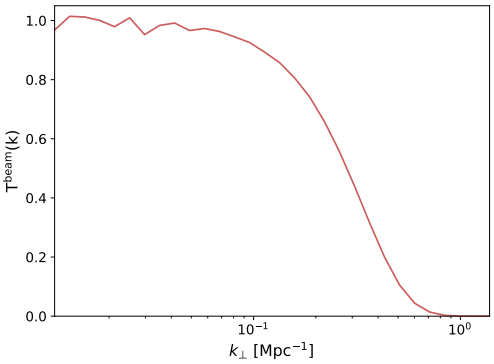

Figure 2: Instrumental beam transfer function.

4. Next, we compute the average pseudo-cross-spectrum, $\tilde{C}_{\vec{k}_i}$, by noise weighting all different cross spectra,

$$\tilde{C}^{\mathrm{FPXS}}_{\vec{k}_i} = \frac{1}{\sum_{i\neq j}\frac{1}{\sigma^2_{\tilde{C}^{ij}_{\vec{k}_i}}}} \sum_{i\neq j} \frac{\tilde{C}^{ij}_{\vec{k}_i}}{\sigma^2_{\tilde{C}^{ij}_{\vec{k}_i}}}. \tag{16}$$

Here $\sigma_{\tilde{C}^{ij}_{\vec{k}_i}}$ is the uncertainty (standard deviation) in $\vec{k}$ bin number i of the pseudo-cross-spectrum of the maps A_i and B_j, and the sum is over all combinations of i and j except the cases where $i = j$. Under the naive assumption that all cross-spectra are independent, the uncertainty of the combined cross-spectrum is given by

$$\sigma_{\tilde{C}^{\mathrm{FPXS}}_{\vec{k}_i}} = \sqrt{\frac{1}{\sum_{i\neq j}\frac{1}{\sigma^2_{\tilde{C}^{ij}_{\vec{k}_i}}}}}. \tag{17}$$

Similar procedures can of course be done by splitting the data in other ways, to make ourselves less susceptible to other systematics, but we have found that using the feeds and elevation splits yields good results for the current dataset.

2.5. *White noise simulations*

Until now we have not discussed how to estimate the noise power spectrum and the corresponding noise uncertainty of the power spectrum. In general, estimating the noise power spectrum precisely is very difficult, since one needs to take into account not only the intrinsic white noise level of the data, but also the effect of the different filtering procedures in the low-level data analysis, as well as any correlated noise contribution.

Since we use a cross-spectrum method, however, the noise spectrum is only used to estimate the uncertainties of the power spectrum, not its mean level, and the requirements on the absolute noise spectrum are therefore somewhat alleviated. Explicitly, if we make an error of a few percent in our noise estimate, we will not bias the estimated signal spectrum, only only misestimate the error bars by a few percent. While clearly not ideal, this is usually not critical, considering all the other simplifying assumptions introduced in the analysis. On the other hand, if we had adopted an auto-spectrum method, an error of a few percent on the noise power spectrum could easily have rendered our signal estimate unusable, even in the case of very high intrinsic sensitivity.

For this reason, we therefore adopt a simple approach to noise power spectrum estimation: We assume that the noise in the maps is uncorrelated white noise, and generate noise simulations, m_i, by drawing random samples in each voxel from a Gaussian distribution with zero mean and a standard deviation given by the value of the noise level map, σ_m. We then estimate error bars by generating a large number of noise simulation maps, calculating the power spectrum from each, and, and finally taking the standard deviation in each $\vec{k}$ bin of interest. This gives us error bars of the noise contribution to each power spectrum bin, but neglects the intrinsic uncertainty in the signal power spectrum itself. However, as we are still completely noise dominated, this intrinsic uncertainty of the signal spectrum should be negligible.

2.6. *Transfer functions*

Until now we have assumed that the sky maps produced by the low-level analysis pipeline are unbiased. For multiple reasons, this is not the case. First of all, the instrument does not have infinite resolution, and the instrumental beam will therefore smooth out the signal on small angular scales. The same effect happens due to the finite spectral resolution of the instrument in the frequency dimension. Secondly, the various filters and mapmaking procedures in the analysis pipeline generally remove some of the signal, mostly on larger angular and spectral scales. In the following, we take these effects into account through so-called transfer functions. These are functions in the $k_\parallel$-$k_\perp$ plane that quantify the fraction of the signal power that is retained in each $\vec{k}$-bin, and allow us to establish unbiased estimates of the power spectrum from biased sky maps.

In general a transfer function, $T(\vec{k})$, is defined through the following relation,

$$\langle P_{\vec{k}} \rangle = T(\vec{k}) P_{\rm signal}(\vec{k}) + P_{\rm noise}(\vec{k}), \tag{18}$$

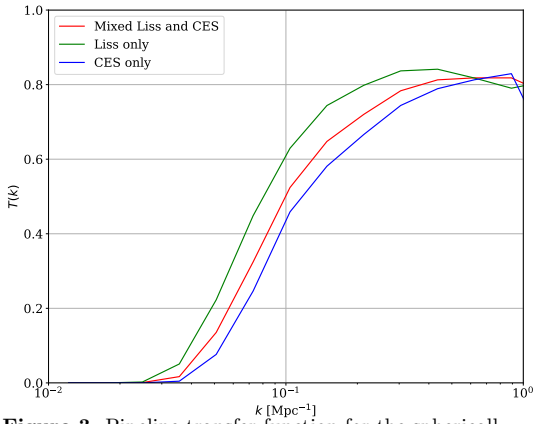

Figure 3: Pipeline transfer function for the spherically averaged power spectrum from Lissajous scans (green), constant elevation scans (blue) and combined data (red). These transfer functions are based on a single signal realization and roughly three hours of Lissajous data, and 3 hours of constant elevation data.

where $P_{\vec{k}}$ is the power spectrum calculated from the final map and $P_{\rm signal}(\vec{k})$ is the actual physical signal power spectrum. We decompose the full transfer function into different parts, and derive each separately. We then multiply the transfer functions together to get the full transfer function.

2.6.1. *Instrumental beam transfer function*

Due to the finite resolution of the instrument, we cannot measure the cosmological signal on the smallest angular scales. In order to take this effect into account we introduce a beam transfer function. For now, we assume the beam to be both achromatic (i.e., constant in frequency) and azimuthally symmetric, and we assume that it may be approximmated well by a 2D Gaussian with a full-width-half-maximum (FWHM) of $\delta_\theta^{\rm FWHM} = 4.5$ arcminutes (Lamb et al. in prep).

The corresponding beam transfer function is estimated using signal-only simulations. That is, Wwe generate a large number of 3D signal realizations and apply a Gaussian beam smoothing to the angular dimensions of the map. We then calculate the power spectrum of each of the signal realization maps with and without beam smoothing. The estimated transfer function is given as the average ratio of these,

$$T^{\rm beam}(\vec{k}) \approx \left\langle \frac{P_{\vec{k}}^{\rm signal,beam}}{P_{\vec{k}}^{\rm signal}} \right\rangle, \tag{19}$$

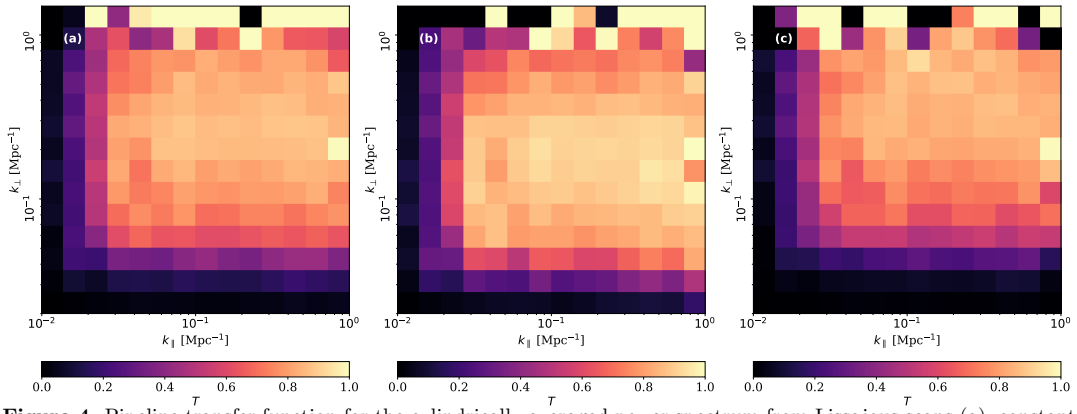

Figure 4: Pipeline transfer function for the cylindrically averaged power spectrum from Lissajous scans (a), constant elevation scans (b) and combined data (c). These transfer functions are based on a single signal realization and roughly three hours of Lissajous data, and 3 hours of constant elevation data.

where $P_{\vec{k}}^{\rm signal,beam}$ is the power spectrum calculated from a beam smoothed signal realization map and $P_{\vec{k}}^{\rm signal}$ is the power spectrum calculated from non-smoothed one.

By using a single Gaussian beam, we implicitly assume that all sky signal power enters the instrument through the main beam, i.e., that the *aperture efficiency* is equal to one. In reality a significant fraction of the power is located in near sidelobes, and a smaller fraction is located in the far sidelobes. We leave it for future work to take these effects into account, which will then allow us to estimate the beam transfer function more accurately.

In this paper, we do not consider the effect of the finite spectral resolution on the recovered signal power, but leave this for future work. As a result, the reported power spectra will be biased at small scales. However, since the reference simulations are biased in precisely the same manner, this does not affec the interpretation as measured in terms of statistical significance.

Figure 2 shows the beam transfer function derived using 100 signal simulations. We see that the 4.5' FWHM beam corresponds roughly to a physical scale of 8 Mpc (comoving), so any angular scales smaller than this will be significantly suppressed.

2.6.2. *Pipeline transfer function*

Each step of the analysis pipeline, including low-level filtering, calibration and mapmaking, affects how much of the true sky signal is present in the final maps and power spectra. We estimate the transfer function of these operations by processing the sum of the raw data and a known signal-only time-ordered simulation through the analysis pipeline, following the exact same procedure as for the raw data alone. The pipeline transfer function may then be estimated as

$$T^{\rm pipeline}(\vec{k}) \approx \left\langle \frac{P_{\vec{k}}^{\rm full} - P_{\vec{k}}^{\rm noise}}{P_{\vec{k}}^{\rm signal}} \right\rangle, \tag{20}$$

where $P_{\vec{k}}^{\rm full}$ is the power spectrum calculated from the maps derived from the raw data with added signal, $P_{\vec{k}}^{\rm noise}$ is the spectrum derived from the same data but without the added signal and $P_{\vec{k}}^{\rm signal}$ is the power spectrum derived from the raw signal simulation that was added to the raw data.

2.6.3. *Unbiased signal estimate*

Correcting the FPXS with the above transfer functions, we can establisy an unbiased estimate of the signal pseudo-spectrum,

$$\tilde{P}_{\rm signal}(\vec{k}) \approx \frac{\tilde{C}_{\vec{k}}^{\rm FPXS}}{\tilde{T}_{\vec{k}}^{\rm full}}, \tag{21}$$

where $\tilde{P}_{\rm signal}(\vec{k})$ is the signal pseudo-spectrum and $\tilde{T}_{\vec{k}}^{\rm full} = \tilde{T}_{\vec{k}}^{\rm beam}\tilde{T}_{\vec{k}}^{\rm pipeline}$ is the full estimated transfer function for the pseudo-spectrum. The uncertainty of this signal estimate is given by

$$\sigma_{\tilde{P}_{\rm signal}(\vec{k})} = \frac{\sigma_{\tilde{C}_{\vec{k}}^{\rm FPXS}}}{\tilde{T}_{\vec{k}}^{\rm full}}. \tag{22}$$

To illustrate the process, we estimate the pipeline transfer function for both Lissajous and CES scans, using three obsIDs (of about one hour observation time each) in each case, as well as the combined case. However, as it is quite expensive to produce these transfer functions, only a single signal simulation outcome

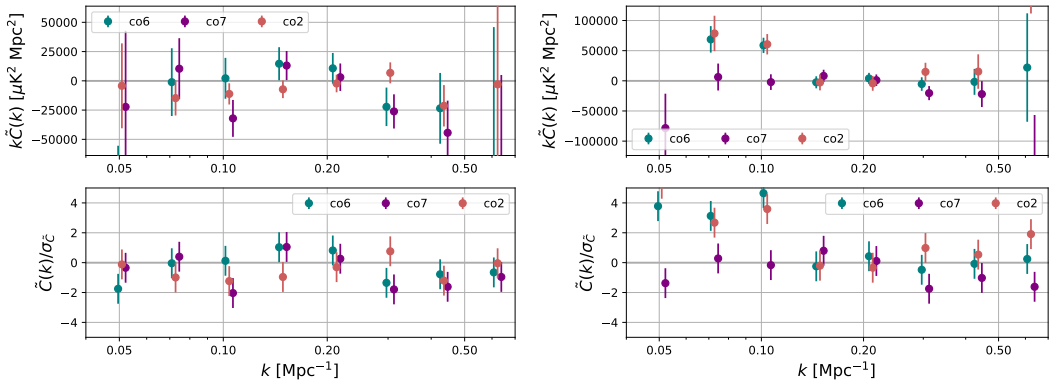

Figure 5: Spherically averaged mean pseudo cross spectra for the CES observations (left) and Lissajous observations (right) of co2 (red), co6 (blue) and co7 (purple) fields. These spectra were generated from all the accepted data using the FPXS cross spectrum statistic. In addition the full transfer function has been applied, to de bias the signal estimate.

was used. It thus remains for future work to compute transfer functions using several more simulations and subsequently take an ensemble average to get a better estimate of the true filter transfer function.

In Fig. 3 the resulting 1D binned transfer functions are shown for both the mixed Lissajous/CES case, as well as the individual Lissajous- and CES-only cases. As can be seen in the figure, the transfer function resulting from Lissajous scans has a somewhat higher maximum compared to the transfer function of the CES-only scans. Also, we note that the Lissajous transfer function peaks at lower k than the CES equivalent. At the same time, the transfer function resulting from a map of mixed Lissajous and CES scans is close to the mean between the two single-type transfer functions.

This pattern is also seen in the corresponding 2D binned transfer function in Fig. 4, visualizing the effect of filtering separately on transverse and perpendicular scales. As can be seen in the middle panel of Fig. 4, the Lissajous-only transfer function peaks at intermediate to high k's, with efficiencies of $\sim 0.9-0.95$ around the peak region, while the the CES equivalent seen in the right panel peaks at higher k with efficiencies of $\sim 0.8-0.85$. The mixed transfer function in this case also lies in between that of the Lissajous- and CES transfer function with efficiencies at $\sim 0.85-0.88$ near its peak region at intermediate to high k. The reason for why the Lissajous transfer function is a little higher at its peak than that of the CES case is most probably due to the fact that a Lissajous scans typically cover a larger area of the sky in a single scan, and it therefore retains more larger angular scale information. As the various filters remove large-scale structure, the low $k_\perp$ and $k_\parallel$ region shows a very low transfer efficiency, which greatly reduces our sensitivity at these scales.

3. POWER SPECTRUM RESULTS

As described in Foss et al. (in prep), after the COMAP time-ordered data have been filtered and calibrated, and bad observations have been removed, the cleaned data set is compressed into a set of 3D maps. We make separate maps for the Lissajous scans and the CES, since these tend to have different systematics and statistical properties.

3.1. *FPXS results*

We estimate the cross-spectrum separately for the Lissajous and CES data. In each case we split the data in two parts according to the elevation of the observations, and use the FPXS method on these two sets of feed maps in order to minimize systematics. We also calculate a χ^2 statistic for each of the 19×18 different feed-feed cross spectra, $\tilde{C}^{ij}_{k_i}$, and reject any spectrum with more than a 5 sigma excess, before we calculate the FPXS mean spectrum. For the CES data, about 3–10 % of the spectra are removed by this cut, while for the Lissajous data about 10–20 % is cut.

The resulting spherically averaged pseudo cross spectra are shown in Fig. 5. We see that the results for the CES data appear largely flat, with fluctuations that are consistent with our white noise estimate. This demonstrates that we are in fact averaging down the noise as expected for uncorrelated noise for the current data set, and the various potential systematics are suppressed to a

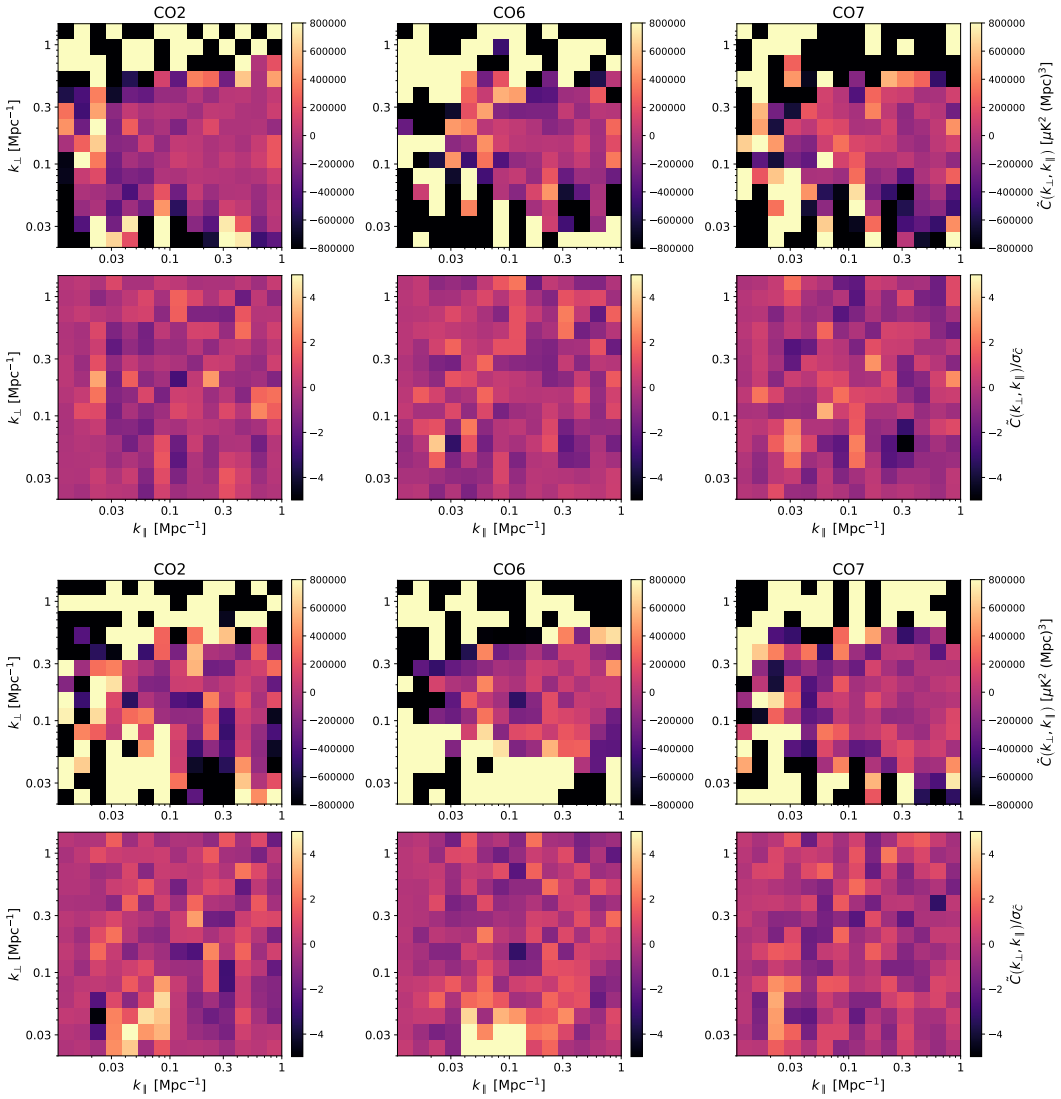

Figure 6: Cylindrically averaged mean pseudo cross spectra for the CES observations (top 2 rows) and Lissajous observations (bottom two rows). Second and fourth row shows the spectra divided by the corresponding white noise uncertainty.

level well below the noise. At this point in the COMAP survey, this is the ideal outcome, given that most theoretical models predict a signal level that is lower than the current white noise sensitivity. At the same time, these data are in fact already ruling out some of the most optimistic models (see Chung et al. (in prep) for more details).

In contrast, the Lissajous data do not appear equally well behaved. For these, we see clear signs of excess power on large scales in both the CO2 and CO6 fields, although not in the CO7 observations. These excesses suggest that large scale systematics are still present in for the Lissajous scans, and may for instance be caused by either atmospheric variations or complex sidelobe contamination.

These residuals are even more prominent when considering the 2D $k_\parallel \times k_\perp$ power spectrum, as shown in Fig. 6. Here we see that while the CES results look mostly consistent with the white noise expectation, the Lissajous results clearly exhibit regions with a systematic power excess. This is seen most clearly in the fourth row of the figure, which shows the power spectrum divided by the expected white noise fluctuations, and thus correspond to power measured in units of standard devations. In particular, we see a bright region on the largest angular scales, and on scales between $k_\parallel \sim 0.03$–0.1 Mpc^{-1} in the frequency direction. This is seen very clearly for both the CO2 and CO6 fields, although not for the CO7 case, corresponding to the behaviour seen in the 1D case.

3.2. *Null tests*

Given that the current data appear to be largely consistent with white noise, the importance of null-tests is less critical than if a potential detection had been made. Still, null-tests may be useful to identify and highlight specific systematic errors, and they may potentially provide hints regarding the nature of the Lissajous excesses.

In the following, we split the data in two disjoint sets according to various criteria specified below. For each disjoint set, we compute independent maps, and compute the power spectrum of the difference map. Any common signal or systematic should then ideally cancel, leaving only noise; any surviving signal would be indicative of a systematic associated with the split in question. In the following, we consider the four different data splits, as defined by 1) season (winter versus summer), 2) first versus last half of the mission, 3) day versus night, and 4) even versus odd observation number. In the first three cases we adjust the split threshold such that there are roughly equal amounts of data in each of the two parts of the split.

Figure 7 shows the results from these calculations, and we see that most of the null spectra are consistent with white noise expectations. This is true even for the Lissajous data which showed a large excess in the regular spectra, suggesting that the excess is present in all the data and thus is removed in the difference maps. The main exception to this is a slight positive bias in the CES null tests for the CO2 field. However, the statistical significance of this excess is marginal, with an overall significance level of 3–4σ even with the simple white noise approximation used in this paper. For the purposes of the current analysis, we conclude that the data are largely consistent with the white noise approximation, both as measured through co-added and null maps.

4. DISCUSSION

In this paper, we have introduced the FPXS as a robust method for estimating the CO signal power spectrum from 3D intensity maps produced by the COMAP data analysis pipeline. We have discussed how to account for signal loss both due to both filtering and beam smoothing, and we have estimated their magnitudes for the first-year COMAP observations with simulations. Computing the FPXS from the actual COMAP data, we find that the current data set is consistent with white noise for constant elevation scan data, and the uncertainties average down with time as expected for ideal data. Equivalently, these results suggest that all systematics are significantly lower than the white noise level in our main sensitivity range.

In contrast, the FPXS results from the Lissajous scan data show clear signs of excess systematics at large scale, as is most clearly seen in the 2D power spectrum. Further modelling and analysis work is required before these data can be used for astrophysical analysis.

Null tests largely seem consistent with zero. However, we do note that some of the most worrisome systematics may not show up in the current null test suite. One prominent example is ground contamination, which tends to be similar in different observations, for example when a given field is observed at similar telescope coordinates. Thus, the ground contamination could be, at least partially, canceled out along with the signal when forming the difference map, at least for some null-maps. This could potentially explain why the large scale systematic effects in the Lissajous data seem to disappear in our current null tests. We also note that null tests are in general significantly less sensitive than the corresponding co-added spectra, since these are formed through straight difference maps, as opposed to noise weighted

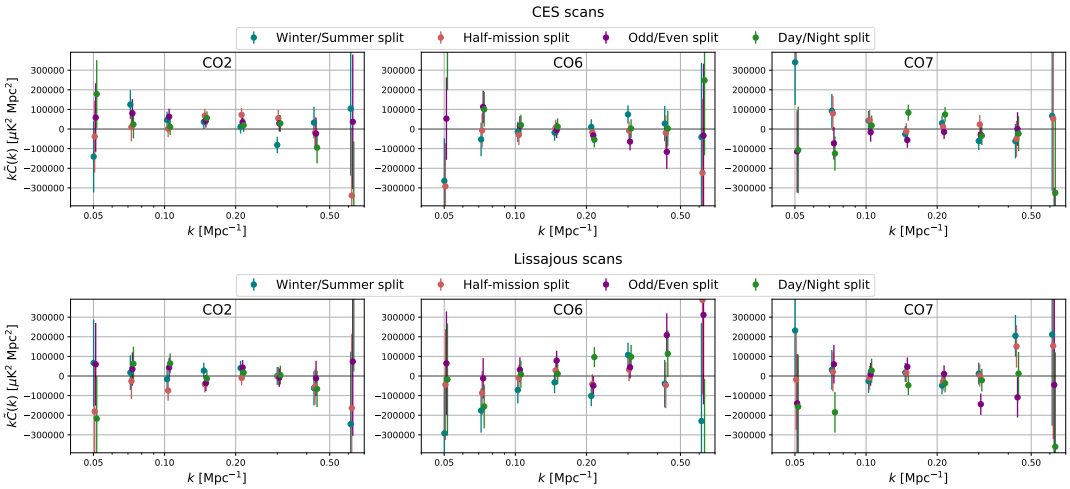

Figure 7: FPXS spectra of difference maps between data split according to Winter/Summer (blue), Half-mission (red), Odd/Even obsid (purple) and Day/Night (green). The full transfer functions have been applied.

averages; this is necessary to ensure that the signal vanishes in the difference map.

We also note that there are several $\mathcal{O}(1)$ effects that have not yet been accounted for that are relevant for deriving an actual upper limit on the CO intensity power spectrum. First of all, we recall that our current estimate is a pseudo-spectrum, and hence subject to mode mixing. Even though we have estimated that this mode mixing should modest at the scales we are considering, it will still probably affect our results at the 5–30 % level. Second, we also have not accounted for the aperture efficiency, but have rather assumed that all the power is in the main beam; this effect could also give a significant suppression of the signal at small angular scales. Third, we have also neglected the effect of a finite spectral resolution, which will also suppress power on small scales in the line-of-sight dimension. For these reasons, we do not yet combine our power spectrum bins into an upper limit on the CO power spectrum. Rather, the main result of the current paper is a demonstration of the fact that the COMAP instrument produces data that are statistically consistent with white noise, and systematics are suppressed to below the white noise for the CES scan type. All of the above-mentioned effects are multiplicative in nature, and does not change the statistical significance of the measurements, but only their physical normalization.

APPENDIX

A. FOURIER CONVENTIONS

In this Appendix, we present the conventions for the discrete Fourier transformations used in this paper. All the conventions are consistent with the default conventions in numpy's FFT library. The forward transformation is given by

$$f_l = \sum_{m=0}^{n-1} x_m \exp\left(-2\pi i \frac{ml}{n}\right), \quad l = 0, \cdots, n-1$$

where x_m are the discrete values of the function in real space, and f_l are the Fourier coefficients. The inverse transformation is then given by

$$x_m = \frac{1}{n} \sum_{l=0}^{n-1} f_l \exp\left(2\pi i \frac{ml}{n}\right),$$

We define the physical wave number

$$k \equiv \frac{2\pi j}{\Delta x n}, \quad j \in \left\{-\frac{n}{2}, \cdots, -1, 0, 1, \cdots, \frac{n}{2}\right\}$$
$$= 2\pi \cdot \text{np.fft.fftfreq}(n, \Delta x).$$

B. DEFINITION OF COSMOLOGICAL MAP GRID

Since Fourier transforms require a rectangular grid, we assume that the 3D temperature maps can be approximated by a rectangular grid in co-moving cosmological parameters. We assume that all the voxels have the same shape and size as the middle voxel at redshift $z_{\rm mid} \approx 2.9$.

The comoving length corresponding to an angular separation $\delta\theta$, for a given redshift z, is given by

$$\delta l_\perp = r(z)\delta\theta = \delta\theta \int_0^z \frac{cdz'}{H(z')}, \tag{B1}$$

where $r(z)$ is the comoving distance travelled by light emitted from redshift z to us.

The comoving radial distance corresponding to a small redshift interval $\delta z = z_1 - z_2 = \nu_0/\nu_1^{\rm obs} - \nu_0/\nu_2^{\rm obs} \approx (1+z)^2\delta\nu^{\rm obs}/\nu_0$, where $z_1 > z_2$, is given by

$$\delta l_\parallel = \int_{z_2}^{z_1} \frac{cdz}{H(z)} \approx \frac{c\delta z}{H(z)} \approx \frac{c}{H(z)} \frac{(1+z)^2\delta\nu^{\rm obs}}{\nu_0}, \tag{B2}$$

where $\nu_0 \approx 115.27$ is the emission frequency of the CO 1→0 line we are studying and $\delta\nu^{\rm obs} = 31.25$ MHz is the resolution of our frequency bins.

C. MODE MIXING AND THE MASTER ALGORITHM

In order to understand the mode-mixing effect, let us consider in more detail the Fourier transform of a weighted map[1],

$$\tilde{f}_{k_1k_2} = \sum_{m_1=0}^{n-1}\sum_{m_2=0}^{n-1} x_{m_1m_2} W_{m_1m_2} \exp\left(-2\pi i \frac{m_1k_1 + m_2k_2}{n}\right). \tag{C3}$$

Here $x_{m_1m_2}$ is the map, $W_{m_1m_2}$ is the weight map and $\tilde{f}_{k_1k_2}$ is the Fourier transform of the weighted map. We can insert the expression for the inverse Fourier transform of x and W,

$$\tilde{f}_{k_1k_2} = \frac{1}{n^4}\sum_{k_1'=0}^{n-1}\sum_{k_2'=0}^{n-1} f_{k_1'k_2'} \sum_{k_1''=0}^{n-1}\sum_{k_2''=0}^{n-1} f^W_{k_1''k_2''} \sum_{m_1=0}^{n-1}\sum_{m_2=0}^{n-1}$$
$$\times \exp\left(-2\pi i \frac{m_1(k_1' + k_1'' - k_1) + m_2(k_2' + k_2'' - k_2)}{n}\right), \tag{C4}$$

$$\tag{C5}$$

where f and f^W are the Fourier transforms of x and W respectively. Working through the algebra, we get

$$\tilde{f}_{k_1k_2} = \frac{1}{n^2}\sum_{k_1'=0}^{n-1}\sum_{k_2'=0}^{n-1} f_{k_1'k_2'} \sum_{k_1''=0}^{n-1}\sum_{k_2''=0}^{n-1} f^W_{k_1''k_2''}$$
$$\times \delta_{k_1''(k_1-k_1')\%n, k_2''(k_2-k_2')\%n}$$
$$\tilde{f}_{k_1k_2} = \frac{1}{n^2}\sum_{k_1'=0}^{n-1}\sum_{k_2'=0}^{n-1} f_{k_1'k_2'} f^W_{(k_1-k_1')\%n(k_2-k_2')\%n}$$
$$\tilde{f}_{k_1k_2} = \sum_{k_1'=0}^{n-1}\sum_{k_2'=0}^{n-1} f_{k_1'k_2'} \underbrace{\frac{1}{n^2} f^W_{(k_1-k_1')\%n(k_2-k_2')\%n}}_{\equiv K_{k_1,k_2,k_1',k_2'}}, \tag{C6}$$

[1] We work in 2D here to save some indices; the generalization to 3D is straightforward.

where % denotes the modulo operation and where we have defined the mode mixing amplitude $K_{\vec{k},\vec{k}'}$.

Adopting vector notation, we may now write the pseudo-spectrum as follows,

$$\tilde{P}(\vec{k}) = \frac{V_{\rm vox}}{N_{\rm vox}} \langle \tilde{f}_{\vec{k}} \tilde{f}^*_{\vec{k}} \rangle \tag{C7}$$

$$= \frac{V_{\rm vox}}{N_{\rm vox}} \frac{1}{n^{2D}} \sum_{\vec{k}'} \sum_{\vec{k}''} \langle f_{\vec{k}'} f^*_{\vec{k}''} \rangle K_{\vec{k},\vec{k}'} K^*_{\vec{k},\vec{k}''} \tag{C8}$$

$$= \frac{1}{n^{2D}} \sum_{\vec{k}'} \sum_{\vec{k}''} P(\vec{k}') \delta_{\vec{k}',\vec{k}''} K_{\vec{k},\vec{k}'} K^*_{\vec{k},\vec{k}''} \tag{C9}$$

$$= \sum_{\vec{k}'} P(\vec{k}') \underbrace{\frac{1}{n^{2D}} |K_{\vec{k},\vec{k}'}|^2}_{M_{\vec{k},\vec{k}'}}, \tag{C10}$$

where D is the number of dimensions of the map, and where we have defined the mode mixing matrix, $M_{\vec{k},\vec{k}'}$. We see that the auto-spectrum and the pseudo-spectrum is related by a linear transformation, so all the information in one is also there in the other.

Within the CMB field, accounting for mode mixing by explicitly calculating and inverting $M_{\vec{k},\vec{k}'}$ is often referred to as the MASTER algorithm (Hivon et al. 2002). Doing this requires that we calculate the mode mixing between each Fourier mode and all the other Fourier modes, so for a 3D maps this scales poorly with the map dimension. On the other hand, the algorithm parallelizes trivially, and the matrix must only be computed once for a given weight map, after which the same operation may be applied efficiently to any number of simulations. Whether this is feasible depends on the details of the individual use case. Some methods exists in the literature to approximate this procedure in a faster way, see e.g. Louis et al. (2020).

ACKNOWLEDGMENTS

REFERENCES

Bernal, J. L., Breysse, P. C., Gil-Marín, H., & Kovetz, E. D. 2019, PhRvD, 100, 123522, doi: 10.1103/PhysRevD.100.123522

Chung, D., et al. in prep

Chung, D. T. 2019, ApJ, 881, 149, doi: 10.3847/1538-4357/ab3040

Cleary, K., et al. in prep

Foss, M. K., et al. in prep

Gong, Y., Chen, X., & Cooray, A. 2020, ApJ, 894, 152, doi: 10.3847/1538-4357/ab87a0

Hivon, E., Górski, K. M., Netterfield, C. B., et al. 2002, ApJ, 567, 2, doi: 10.1086/338126

Ihle, H. T., Chung, D., Stein, G., et al. 2019, ApJ, 871, 75, doi: 10.3847/1538-4357/aaf4bc

Keating, G. K., Marrone, D. P., Bower, G. C., & Keenan, R. P. 2020, ApJ, 901, 141, doi: 10.3847/1538-4357/abb08e

Lamb, J., et al. in prep

Louis, T., Naess, S., Garrido, X., & Challinor, A. 2020, PhRvD, 102, 123538, doi: 10.1103/PhysRevD.102.123538

Mertens, F. G., Mevius, M., Koopmans, L. V. E., et al. 2020, MNRAS, 493, 1662, doi: 10.1093/mnras/staa327

Uzgil, B. D., Carilli, C., Lidz, A., et al. 2019, ApJ, 887, 37, doi: 10.3847/1538-4357/ab517f

Yang, S., Somerville, R. S., Pullen, A. R., et al. 2020, arXiv e-prints, arXiv:2009.11933. https://arxiv.org/abs/2009.11933

Paper III

Joint Power Spectrum and Voxel Intensity Distribution Forecast on the CO Luminosity Function with COMAP

H. T. Ihle, D. Chung, G. Stein and the COMAP collaboration

Published in ApJ, Jan 2019, volume 871, no. 1 p. 75. DOI: 10.3847/1538-4357/aaf4bc.

III

The Astrophysical Journal, 871:75 (11pp), 2019 January 20 https://doi.org/10.3847/1538-4357/aaf4bc
© 2019. The American Astronomical Society. All rights reserved.

Joint Power Spectrum and Voxel Intensity Distribution Forecast on the CO Luminosity Function with COMAP

H. T. Ihle[1], D. Chung[2], G. Stein[3,4], M. Alvarez[5], J. R. Bond[4], P. C. Breysse[4], K. A. Cleary[6], H. K. Eriksen[1], M. K. Foss[1], J. O. Gundersen[7], S. Harper[8], N. Murray[4], H. Padmanabhan[9], M. P. Viero[2], and I. K. Wehus[1]
(COMAP collaboration)
[1] Institute of Theoretical Astrophysics, University of Oslo, P.O. Box 1029 Blindern, NO-0315 Oslo, Norway; h.t.ihle@astro.uio.no
[2] Kavli Institute for Particle Astrophysics and Cosmology and Physics Department, Stanford University, Stanford, CA 94305, USA
[3] Department of Astronomy and Astrophysics, University of Toronto, 50 St. George St., Toronto, ON, M5S 3H4, Canada
[4] Canadian Institute for Theoretical Astrophysics, University of Toronto, 60 St. George St., Toronto, ON M5S 3H8, Canada
[5] Berkeley Center for Cosmological Physics, University of California, Berkeley, CA 94720, USA
[6] California Institute of Technology, Pasadena, CA 91125, USA
[7] Department of Physics, University of Miami, 1320 Campo Sano Avenue, Coral Gables, FL 33146, USA
[8] Jodrell Bank Centre for Astrophysics, School of Physics and Astronomy, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK
[9] Institute for Particle Physics and Astrophysics, ETH Zurich, Wolfgang-Pauli-Strasse 27, CH 8093 Zurich, Switzerland
Received 2018 August 22; revised 2018 November 8; accepted 2018 November 27; published 2019 January 23

Abstract

We develop a framework for joint constraints on the CO luminosity function based on power spectra (PS) and voxel intensity distributions (VID) and apply this to simulations of CO Mapping Array Pathfinder (COMAP), a CO intensity mapping experiment. This Bayesian framework is based on a Markov chain Monte Carlo (MCMC) sampler coupled to a Gaussian likelihood with a joint PS + VID covariance matrix computed from a large number of fiducial simulations and re-calibrated with a small number of simulations per MCMC step. The simulations are based on dark matter halos from fast peak patch simulations combined with the $L_{\rm CO}(M_{\rm halo})$ model of Li et al. We find that the relative power to constrain the CO luminosity function depends on the luminosity range of interest. In particular, the VID is more sensitive at large luminosities, while the PS and the VID are both competitive at small and intermediate luminosities. The joint analysis is superior to using either observable separately. When averaging over CO luminosities ranging between $L_{\rm CO} = 10^4$–$10^7\,L_\odot$, and over 10 cosmological realizations of COMAP Phase 2, the uncertainties (in dex) are larger by 58% and 30% for the PS and VID, respectively, when compared to the joint analysis (PS + VID). This method is generally applicable to any other random field, with a complicated likelihood, as long a fast simulation procedure is available.

Key words: diffuse radiation – galaxies: high-redshift – large-scale structure of universe

1. Introduction

Intensity mapping (Madau et al. 1997; Battye et al. 2004; Peterson et al. 2006; Loeb & Wyithe 2008) appears promising for mapping large 3D volumes cheaply in a relatively short period of time, using specific bright emission lines as matter tracers. This is an interesting avenue for advancing precision cosmology, with a multitude of ongoing efforts (Kovetz et al. 2017), following on the successes of the CMB field in the last few decades. One such line intensity mapping experiment currently under construction is called the CO Mapping Array Pathfinder (COMAP; Cleary et al. 2016; Li et al. 2016), which aims to observe frequencies between 26 and 34 GHz, corresponding to redshifted CO line emission from the epoch of galaxy assembly (redshifts between $z = 2.4$ and 3.4) for the CO $J = 1 \rightarrow 0$ line at 115 GHz rest frequency, and CO emission from the epoch of reionization ($z = 5.8$–6.7) for the CO $J = 2 \rightarrow 1$ line at 230 GHz rest frequency.

One important scientific target for studying and understanding the epoch of galaxy assembly, the main goal of the first COMAP phase, is the so-called CO luminosity function, which measures the number density of CO emitters as a function of luminosity. Several methods for extracting this function from real data have already been suggested in the literature, the most prominent being the power spectrum (PS) approach, for instance as implemented by Li et al. (2016). A second complementary method is the one-point function, or voxel intensity distribution (VID), $\mathcal{P}(T)$, as suggested by Breysse et al. (2016, 2017).

In this paper, we consider the prospect of combining the VID and PS approaches when constraining the CO luminosity function, and we study this approach within the context of the COMAP experiment. To do so, we first define a joint likelihood that includes both the VID and the PS and construct a joint covariance matrix for both observables. This covariance matrix is constructed from a large set of dark matter (DM) light-cone halo catalogs from so-called "peak patch" cosmological simulations (Bond & Myers 1996, Stein et al. 2019), coupled to an empirical $L_{\rm CO}(M_{\rm halo})$ model (Li et al. 2016) that infers CO luminosities, $L_{\rm CO}$, from DM halo masses, $M_{\rm halo}$. We then investigate the posterior distribution of the resulting model parameters for each of the first two anticipated phases of the COMAP experiment (see Table 1). Finally, we compare the constraints on the CO luminosity function derived from joint PS and VID measurements to those obtained from the PS or VID separately.

2. Idealized Simulations of the COMAP Experiment

We start our discussion by reviewing some central properties of the COMAP experiment, focusing in particular on those required for generating representative yet computationally

Table 1
Experiment Setup for the Two COMAP Phases

Parameter	COMAP1	COMAP2
System temperature, $T_{\rm sys}$ [K]	40	40
Number of feeds	19	95
Beam FWHM (arcmin)	4	4
Frequency band [GHz]	26–34	26–34
Channel width, $\delta\nu$ (MHz)	15.6	15.6
Observing time [hr]	6000	9000
Noise per voxel [μK]	11.0	8.0
Field size [deg^2]	2.25	2.25
Number of fields	1	4

affordable simulations. For convenience, these properties are summarized in Table 1.

In Phase 1, COMAP will employ a single telescope equipped with 19 single-polarization detectors, each with 512 frequency channels with width $\delta\nu \approx 15.6$ MHz[10] covering frequencies between 26 and 34 GHz. The system temperature is expected to be around $T_{\rm sys} \approx 40$ K and the angular resolution corresponds to a Gaussian beam with $4'$ full width at half maximum (FWHM). We anticipate two years of observation time targeting a single field of $1.^\circ5 \times 1.^\circ5$ close to the north celestial pole, and we assume a conservative observing efficiency of 35% for a total of 6000 hr of total integration time on the field.

In Phase 2, the experiment will be expanded to five telescopes with the same setup as in Phase 1 and will observe for three additional years. In this phase, we assume that the observation time will be split between four fields of the same size as in Phase 1. The two COMAP phases will be referred to as COMAP1 and COMAP2 in the following.

2.1. Noise

The simulations used in this work consist of two components only, namely the target CO signal and random white noise with properties corresponding to the parameters described above. Explicitly, the noise per voxel is given by

$$\sigma_T = \frac{T_{\rm sys}}{\sqrt{\tau\ \delta\nu}} = \frac{T_{\rm sys}\sqrt{N_{\rm pixels}}}{\sqrt{\tau_{\rm tot} e_{\rm obs} N_{\rm feeds}\delta\nu}}, \quad (1)$$

where $T_{\rm sys}$ is the system temperature, τ is the observation time per pixel, $\tau_{\rm tot}$ is the total observation time, $e_{\rm obs}$ is the observation efficiency, $N_{\rm feeds}$ is the number of feeds, $N_{\rm pixels}$ is the number of pixels, and $\delta\nu$ is the frequency resolution. This gives us $\sigma_T \approx 11\,\mu$K and $8\,\mu$K for the COMAP1 and COMAP2 phases, respectively. For simplicity we assume that the noise is evenly distributed over all voxels.

A voxel is the 3D equivalent of a pixel. Two of the dimensions correspond to a regular pixel on the sky, while the third dimension corresponds to a small range of redshifts from where line emission would redshift into a given frequency bin of our instrument.

Both instrumental systematics and astrophysical foreground contamination are neglected in the following. However, since our estimator is inherently simulation based, these effects can be added at a later stage when a sufficiently realistic instrument model is available. For discussion of foreground contamination in similar line intensity surveys see, e.g., da Cunha et al. (2013), Breysse et al. (2015, 2017), and Chung et al. (2017).

[10] Higher spectral resolutions are available, but these are most likely useful only for systematics mitigation rather than science due to limited signal-to-noise per voxel.

2.2. DM Simulations

The signal component is based on the peak patch DM halo approach described by Bond & Myers (1996) and Stein et al. (2019), coupled to the $L_{\rm CO}(M_{\rm halo})$ model presented by Li et al. (2016). Additionally, we adopt the same cosmological parameters as the Li et al. (2016) analysis for the DM simulations: $\Omega_m = 0.286$, $\Omega_\Lambda = 0.714$, $\Omega_b = 0.047$, $h = 0.7$, $\sigma_8 = 0.82$, and $n_s = 0.96$.

The DM simulations in this paper were created using the peak patch method of Bond & Myers (1996) and Stein et al. (2019). To cover the full redshift range of the COMAP experiment we simulated a volume of $(1140\,{\rm Mpc})^3$ using a particle-mesh resolution of $N_{\rm cells} = 4096^3$. Projecting this onto the sky results in a $9.^\circ6 \times 9.^\circ6$ field of view covering the redshift range $2.4 < z < 3.4$, with a minimum DM halo mass of $2.5 \times 10^{10}\,M_\odot$.

The resulting halo catalog contains roughly 54 million halos, each with a position, velocity, and mass. The peak patch method can simulate continuous light cones on-the-fly, so stitching snapshots together was not required to create the light cone. Although peak patch simulations result in quite accurate halo masses, the DM halo catalogs were additionally mass corrected by abundance matching along the light cone to Tinker et al. (2008) to ensure statistically the same mass function as the simulations used in the Li et al. (2016) analysis. For a detailed study of the clustering properties of peak patch simulations and other approximate methods, see Lippich et al. (2019), Blot et al. (2018), and Colavincenzo et al. (2019).

A single run required 900 s of computation time on 2048 Intel Xeon EE540 2.53 GHz CPU cores of the Scinet-GPC cluster, with a memory footprint of $\simeq$2.4 TB. This efficiency of the peak patch method allowed for 161 independent realizations of the full 1140 Mpc, $N_{\rm cells} = 4096^3$ volume, taking a total computation time of only $\sim$82,000 CPU hours, over three orders of magnitude faster when compared to an N-body method of equivalent size.

2.3. Converting to CO Brightness Temperature

There are many approaches in the literature for estimating the expected CO signal based on DM halos (e.g., Righi et al. 2008; Obreschkow et al. 2009; Visbal & Loeb 2010; Carilli 2011; Gong et al. 2011; Lidz et al. 2011; Fu et al. 2012; Carilli & Walter 2013; Pullen et al. 2013; Breysse et al. 2014; Greve et al. 2014; Mashian et al. 2015; Li et al. 2016; Padmanabhan 2018), with resulting estimates of the CO luminosities spanning roughly an order of magnitude.

Here we adopt the model described by Li et al. (2016) to convert from simulated light cones populated with DM halos to observed CO brightness temperature. This model is defined by a set of parametric relations between DM halo masses, star formation rates (SFR), infrared (IR) luminosities, $L_{\rm IR}$, and CO luminosities, $L_{\rm CO}$.

The model uses the results from Behroozi et al. (2013a, 2013b) to obtain average SFR from DM halo masses and adds an additional log-normal scatter on top of the average, determined by $\sigma_{\rm SFR}$. IR luminosities are then obtained through

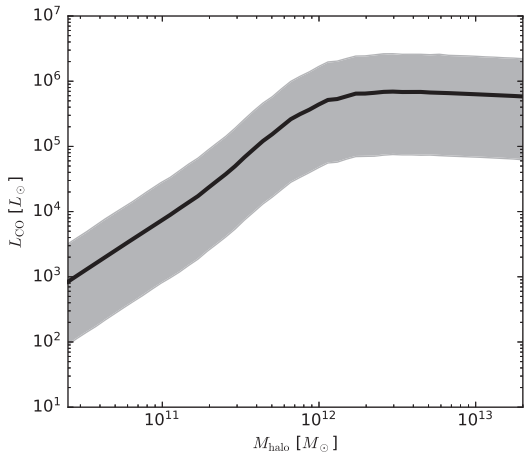

Figure 1. Plot of CO luminosity, $L_{\rm CO}$, as a function of dark matter halo mass, $M_{\rm halo}$, in the Li et al. (2016) model. Here, ($\sigma_{\rm SFR}$, $\log\delta_{\rm MF}$, α, β, $\sigma_{L_{\rm CO}}$) = (0.3, 0.0, 1.17, 0.21, 0.3) (our fiducial model), and we have evaluated the function at redshift 2.9. The solid line corresponds to the mean relation with no scatter added, while the shaded region corresponds to the 95% confidence intervals after adding log-normal scatter at the two appropriate steps.

the relation

$$\mathrm{SFR} = \delta_{\rm MF} \times 10^{-10}\, L_{\rm IR}. \tag{2}$$

Further, to obtain CO luminosities, the relation

$$\log L_{\rm IR} = \alpha \log L'_{\rm CO} + \beta \tag{3}$$

is used before a second round of log-normal scatter is added, determined by the parameter $\sigma_{L_{\rm CO}}$.

This gives us a $L_{\rm CO}(M_{\rm halo})$ model with five free parameters, $\theta = \{\sigma_{\rm SFR}, \log\delta_{\rm MF}, \alpha, \beta, \sigma_{L_{\rm CO}}\}$. The relation between $L_{\rm CO}$ and $M_{\rm halo}$ for our fiducial model parameters is shown in Figure 1. For more discussion of the physical and observational motivation for this model, see the original paper, Li et al. (2016).

This model is applied to each DM halo separately and we create high-resolution maps from the resulting CO luminosities by converting the total luminosity in a given voxel into brightness temperature. These maps were created using the publicly available `limlam_mocker` code.[11] Next, we convolve these maps with the (Gaussian) instrumental beam profile, degrade to the low-resolution voxel size used in the analysis, and, finally, we add Gaussian uncorrelated noise with standard deviation σ_T, as specified above.

3. Algorithms

The ultimate goal of this work is to constrain cosmological and astrophysical parameters from CO line intensity observations. The computational engine for this work is a standard Metropolis Markov chain Monte Carlo (MCMC) sampler (see, e.g., Gilks et al. 1995), coupled to a posterior distribution with a corresponding likelihood and prior. For this task to be computationally tractable, though, the full CO line intensity data set must first be compressed to a smaller set of observables that may be modeled in terms of the desired astrophysical parameters, fully analogous to how CMB sky maps are compressed to a CMB power spectrum from which cosmological parameters are derived (e.g., Bond et al. 2000). As described above, we adopt the power spectrum and the VID as representative observables, each of which may be approximated in terms of multivariate Gaussian random variables. However, in order to perform a joint analysis of these two observables, we need to construct their joint covariance matrix, and that is the primary goal of this section. Before doing that, however, we review for completeness each observable individually, and our posterior sampler of choice, referring to relevant literature for full details.

3.1. The Power Spectrum

The estimated power spectrum, $P(k_i)$, is calculated simply by taking the 3D Fourier transform of the temperature cube, binning the absolute squared values of the Fourier coefficients according to the magnitude of corresponding wave number k, and averaging over all the contributions within each bin. For a Gaussian map, the Fourier components within each bin follow a perfect normal distribution with mean zero and variance given by the value of the power spectrum. For a non-Gaussian field, the distribution of the Fourier components is more complicated, and thus the power spectrum does not contain all the statistical information in the map. We expect the CO signal to form a highly non-Gaussian map, therefore, in this paper we simply consider the power spectrum as a useful observable that carries some, but far from all, of the statistical information in the map.

As an observable, the power spectrum needs to be accompanied by a covariance matrix $\xi^P_{ij} \equiv \mathrm{Cov}(P(k_i), P(k_j))$ in the analysis, since there are correlations between the power spectrum at different k values.

3.2. The Voxel Intensity Distribution

We consider the VID as another observable, complementary to the PS and more closely related to the luminosity function.

Unlike in many other works on $\mathcal{P}(D)$ analysis (e.g., Lee et al. 2009; Glenn et al. 2010; Vernstrom et al. 2014; Breysse et al. 2017; Leicht et al. 2019), we do not try to estimate the VID analytically, rather we estimate it based on simulations. This allows us to fully take into account the effects of the beam, clustering, and covariance between temperature bins in a very straightforward manner.

We consider two contributions to the VID, namely the CO signal itself and the instrumental noise. Together they result in the the full probability distribution of voxel temperatures, $\mathcal{P}(T)$, where T is the observed brightness temperature from a voxel. Since we assume the noise to be uniformly distributed over all voxels in the observed field and assume that the CO signal itself is statistically homogeneous and isotropic, the total probability distribution, $\mathcal{P}(T)$, is the same across all voxels.

The basic observable related to the VID are the temperature bin counts (i.e., the histogram of voxel temperatures), B_i. The expectation value of these are given by the VID itself,

$$\langle B_i \rangle = N_{\rm vox} \int_{T_i}^{T_{i+1}} \mathcal{P}(T) dT, \tag{4}$$

where $N_{\rm vox}$ is the number of voxels observed and B_i is the number of voxels with a temperature between T_i and T_{i+1}.

[11] https://github.com/georgestein/limlam_mocker

If the temperatures of all the voxels that we sample were completely independent, then each of the voxel bins would be approximately independent and would follow a binomial distribution with variance $\mathrm{Var}_{\mathrm{ind}}(B_i) = \langle B_i\rangle(1 - \langle B_i\rangle/N_{\mathrm{vox}})$. However, even in this ideal case they would not be perfectly independent. We only have a finite number of voxels, and, therefore, if one bin contains a number of voxels above average, then the other bins must have a number lower than average.

In general, the samples will not be independent for many reasons, including correlated sky or noise structures and processing effects, and we therefore need the full covariance matrix between bins, $\xi_{ij}^B \equiv \mathrm{Cov}(B_i, B_j)$. This covariance matrix will depend on the DM density field, the CO bias, and the luminosity function, and we will estimate it using simulations.

3.3. The Joint PS+VID Covariance Matrix

The main missing component in the above method is definition of a joint power spectrum and VID covariance matrix. By having access to the computationally cheap yet realistic Monte Carlo simulations described above, we can approximate this matrix with simulations. In addition to giving us covariance matrices to do our analysis, this also allows us to check under what conditions the full covariance matrix is necessary and when we can get away with assuming that individual samples are independent.

In this paper, we start with 161 independent simulated light-cone cubes of DM halos, each covering about $9.^{\circ}6 \times 9.^{\circ}6$ on the sky and a frequency range between 26 and 34 GHz, corresponding to redshifts between 2.4 and 3.4. The frequency dimension is divided equally into 512 frequency bins, each spanning $\delta\nu \approx 15.6$ MHz, corresponding to a redshift resolution of $\delta z \approx 0.002$. Since the COMAP field only spans $1.^{\circ}5 \times 1.^{\circ}5$ on the sky, we sub-divide each of the $9.^{\circ}6 \times 9.^{\circ}6$ light-cone cubes, after beam convolution, into 36 square fields, each covering $1.^{\circ}5 \times 1.^{\circ}5$, resulting in a total of 5796 semi-independent sky realizations. The final pixelization of these maps is a 22×22 grid of square pixels, resulting in a pixel size of $\delta\theta \approx 4.'1$. To these maps, we add uniformly distributed white noise at the appropriate levels for the COMAP1 and COMAP2 experiment setups described above.

When choosing the pixel size to use for the analysis, we follow Vernstrom et al. (2014). They show that, for $\mathcal{P}(D)$ analysis, choosing a pixel size to be equal to the FWHM of the beam is a good tradeoff between picking a small pixel size to include the maximal information, and choosing a larger pixel size to reduce the pixel to pixel correlations induced by the beam.

We combine our two observables into a joint one-dimensional vector of the form

$$d_i = (P_{k_i}, B_i), \tag{5}$$

where P_{k_i} is the binned power spectrum and B_i are the temperature bin counts. Let us first consider the ideal case in which all elements in this vector are independent and the Fourier components are approximately Gaussian. In that case we can compute the expected variance, which we will simply call the *independent variance*, analytically,

$$\mathrm{Var}_{\mathrm{ind}}(P_{k_i}) = \langle P_{k_i}\rangle^2/N_{\mathrm{modes}}(k_i), \tag{6}$$

$$\mathrm{Var}_{\mathrm{ind}}(B_i) = \langle B_i\rangle(1 - \langle B_i\rangle/N_{\mathrm{vox}}) \approx \langle B_i\rangle, \tag{7}$$

where $N_{\mathrm{modes}}(k_i)$ is the number of modes in the ith k bin and where we have introduced the notation $\mathrm{Var}_{\mathrm{ind}}(d_i)$ for this conditionally independent variance.

With this notation in hand, we define a "pseudo-correlation matrix" as

$$c_{ij} \equiv \frac{\xi_{ij}}{\sqrt{\mathrm{Var}_{\mathrm{ind}}(d_i)\mathrm{Var}_{\mathrm{ind}}(d_j)}}, \tag{8}$$

where, as in Section 3.4, ξ_{ij} is the full covariance matrix. Note that c_{ij} is the exact correlation matrix in the limit that $\mathrm{Var}_{\mathrm{ind}}(d_i)$ is the true full variance. An important advantage of the pseudo-correlation matrix, however, is the fact that $\mathrm{Var}_{\mathrm{ind}}(d_i)$ may be estimated directly from the average data itself, and this is required for our MCMC procedure to be sufficiently fast.

The full covariance matrix ξ is estimated for the model described by Li et al. (2016), adopting the fiducial parameters θ_0, using the set of 5796 simulations described above. However, for the MCMC sampler described in Section 3.4, we actually need the full covariance matrix, corresponding to different model parameters θ, at each step in the Markov chain. Generating the full covariance matrix with the above procedure at each MC step is clearly not computationally feasible and we therefore need to approximate this somehow.

With regard to this last point, we introduce the following proposal: we assume that the full covariance matrix scales, under a change of model parameters from θ_0 to θ, the same way as the independent variance, $\mathrm{Var}_{\mathrm{ind}}(d_i)$,

$$\hat{\xi}_{ij}(\theta) \approx \xi_{ij}(\theta_0)\frac{\sqrt{\mathrm{Var}_{\mathrm{ind}}^{\theta}(d_i)\mathrm{Var}_{\mathrm{ind}}^{\theta}(d_j)}}{\sqrt{\mathrm{Var}_{\mathrm{ind}}^{\theta_0}(d_i)\mathrm{Var}_{\mathrm{ind}}^{\theta_0}(d_j)}}, \tag{9}$$

where $\mathrm{Var}_{\mathrm{ind}}^{\theta_0}(d_i)$ is the independent variance for the fiducial model and $\mathrm{Var}_{\mathrm{ind}}^{\theta}(d_i)$ is the independent variance for arbitrary parameters θ. Since this latter function only depends on the average quantities $\langle d_i\rangle$, it is computationally straightforward to compute $\hat{\xi}_{ij}(\theta)$ at any position in an MCMC sampler. Note also that $\hat{\xi}_{ij}(\theta)$ is, by construction, positive definite, as required for a proper covariance matrix.

For a noise-dominated experiment, where all samples are approximately independent, the independent variance, $\mathrm{Var}_{\mathrm{ind}}(d_i)$, is the correct variance and Equation (9) is the correct scaling of the covariance matrix. However, we use this scaling as a first approximation even in cases where there is some covariance in the data.

Intuitively, Equation (9) is equivalent to postulating that the pseudo-correlation matrix, c_{ij}, is approximately constant (i.e., independent of the specific parameters in question). For real-world applications, we recommend testing this assumption explicitly by computing the covariance matrix by brute force simulation for a few extreme parameter combinations drawn from the Markov chains produced during the analysis.

The above prescription applies straightforwardly to single-field observations as, for instance, planned for COMAP1. In contrast, COMAP2 will, under our assumptions, span $N = 4$ independent but statistically identical fields. Since the mean vector of observables evaluated across those four fields equals the average of the four corresponding independent observable vectors, the full covariance matrix is simply given by the

single-field covariance matrix divided by the number of fields:

$$\xi_{ij}^{N\ \mathrm{field}} = \frac{\xi_{ij}^{1\ \mathrm{field}}}{N}. \quad (10)$$

Note that c_{ij} then, assuming the fields are of the same size, only depends on the noise level per field, so for a given noise level per field, c_{ij} is independent of the number of fields.

Finally, we note that the total number of degrees of freedom in our joint PS+VID statistic is in this paper equal to 45, corresponding to 20 power spectrum bins and 25 VID bins. For this number of degrees of freedom, a set of 5796 (semi-independent) simulations provides a very good estimate of all numerically dominant components of the covariance matrix, including both the diagonal and the leading off-diagonal modes, and ξ_{ij} is well conditioned.

3.4. Posterior Mapping by MCMC

As previously mentioned, we use an MCMC algorithm to sample from the posterior distribution of the $L_{\mathrm{CO}}(M_{\mathrm{halo}})$ model parameters, $\theta = \{\sigma_{\mathrm{SFR}}, \log \delta_{\mathrm{MF}}, \alpha, \beta, \sigma_{L_{\mathrm{CO}}}\}$. This posterior distribution is, as usual, given by Bayes' theorem,

$$P(\theta|d) \propto P(d_i|\theta)P(\theta), \quad (11)$$

where d represents our compressed data set, $P(d|\theta)$ is the likelihood defined below, and $P(\theta)$ is some set of priors. We use the *emcee* package (Foreman-Mackey et al. 2013) and its implementation of an affine-invariant ensemble MCMC algorithm, with 142 walkers.

We use a burn-in period of 1000 steps, and use the next 1000 steps for the posterior estimation.

We assume a Gaussian likelihood for our observables d_i of the form (up to an additive constant)

$$-2\ln P(d|\theta) = \sum_{ij}[d_i - \langle d_i\rangle](\xi^{-1})_{ij}[d_j - \langle d_j\rangle] + \ln|\xi|, \quad (12)$$

where the means $\langle d_i\rangle$ depend on the model parameters θ, and the covariance matrix ξ_{ij} is approximated by the expression given in Equation (9). (Note that we do not need to assume that the low-level data are Gaussian, but only that the compressed observables may be well approximated by a multivariate Gaussian distribution. Due to the central limit theorem, this is in practice very often an excellent approximation.)

For both the power spectrum and the low and intermediate temperature VID bins, for which there is a large number of voxel counts per bin, this Gaussian approximation holds to a high degree. However, for the highest VID temperature bins, where there are only a few voxel counts per bin, the discrete nature of the bin count may become relevant and the full binomial distribution should, in principle, be taken into account. However, this effect can also be easily remedied by increasing the bin width, albeit at the cost of a slight loss of information, as is suggested in Vernstrom et al. (2014), and we therefore neglect it in the following, since our primary focus is the dominant Gaussian component of the likelihood. A more thorough analysis may take this issue into account either analytically or by simulations.

An advantage of using a Gaussian likelihood for the VID is that it gives us a straightforward way to take into account the correlations between temperature bins apparent in the covariance matrix, ξ_{ij} (e.g., in Figure 2). For another approach to building up a $\mathcal{P}(D)$ likelihood, see Glenn et al. (2010).

To estimate $\langle d_i\rangle$, we compute 10 maps of the survey volume at each step in the MCMC chain using the current model parameters θ with different DM halo realizations (randomly drawn from 252 independent catalogs corresponding to the survey volume). The specific number of realizations, 10 in our case, represents a compromise between minimizing the sample variance in the estimate of $\langle d_i\rangle$ and maintaining a reasonable computational cost per MC step. Finally, we bin all of the halos in the 10 realizations according to their luminosity and use this histogram to estimate the luminosity function at the current values of θ. This way the MCMC procedure gives us the luminosity function at different points in parameter space, sampled according to the posterior of the model parameters, which we can use to derive constraints on the luminosity function itself.

We adopt the same physically motivated priors as discussed by Li et al. (2016). Specifically, these read

$$P(\sigma_{\mathrm{SFR}}) = \mathcal{N}(0.3, 0.1) \quad (13)$$

$$P(\log \delta_{\mathrm{MF}}) = \mathcal{N}(0.0, 0.3) \quad (14)$$

$$P(\alpha) = \mathcal{N}(1.17, 0.37) \quad (15)$$

$$P(\beta) = \mathcal{N}(0.21, 3.74) \quad (16)$$

$$P(\sigma_{L_{\mathrm{CO}}}) = \mathcal{N}(0.3, 0.1), \quad (17)$$

where $\mathcal{N}(\mu, \sigma)$ corresponds to a Gaussian distribution with mean μ and standard deviation σ. Additionally, we require the two logarithmic scatter parameters, σ_{SFR} and $\sigma_{L_{\mathrm{CO}}}$, to be positive. We choose the mean of all these distributions as the fiducial model, θ_0.

To quantify the importance of joint PS+VID analysis, we perform the above analysis both with each observable separately and with the joint analysis. The main result in this paper may then be formulated in terms of the relative improvement on the CO luminosity function uncertainty derived from the joint analysis to those found in the independent analyses.

When calculating our observables (PS and VID), we assume that our survey volume can be treated as a rectangular grid of voxels with constant co-moving volume. We also neglect the evolution of our observables over redshifts between $z = 2.4$ and 3.4. That is, we assume that samples from different redshifts are drawn from the same distribution, whether they are power spectrum modes or voxel temperatures. We also assume that the instrument beam is achromatic and is equal to the value at the central frequency. This is of course just an approximation that we make in order for the analysis to be simple. If we were doing experiments with higher signal to noise, we might divide our data into two different redshift regions and do an independent analysis of each region. This could allow us to study the redshift evolution of the observables. For COMAP (1 and 2), however, we are probably best off combining all the data, like we do here, in order to increase the overall signal to noise.

Finally, since COMAP will not measure absolute zero levels, we subtract the mean from all maps. For the power spectrum, this has a negligible impact, as it simply removes one out of

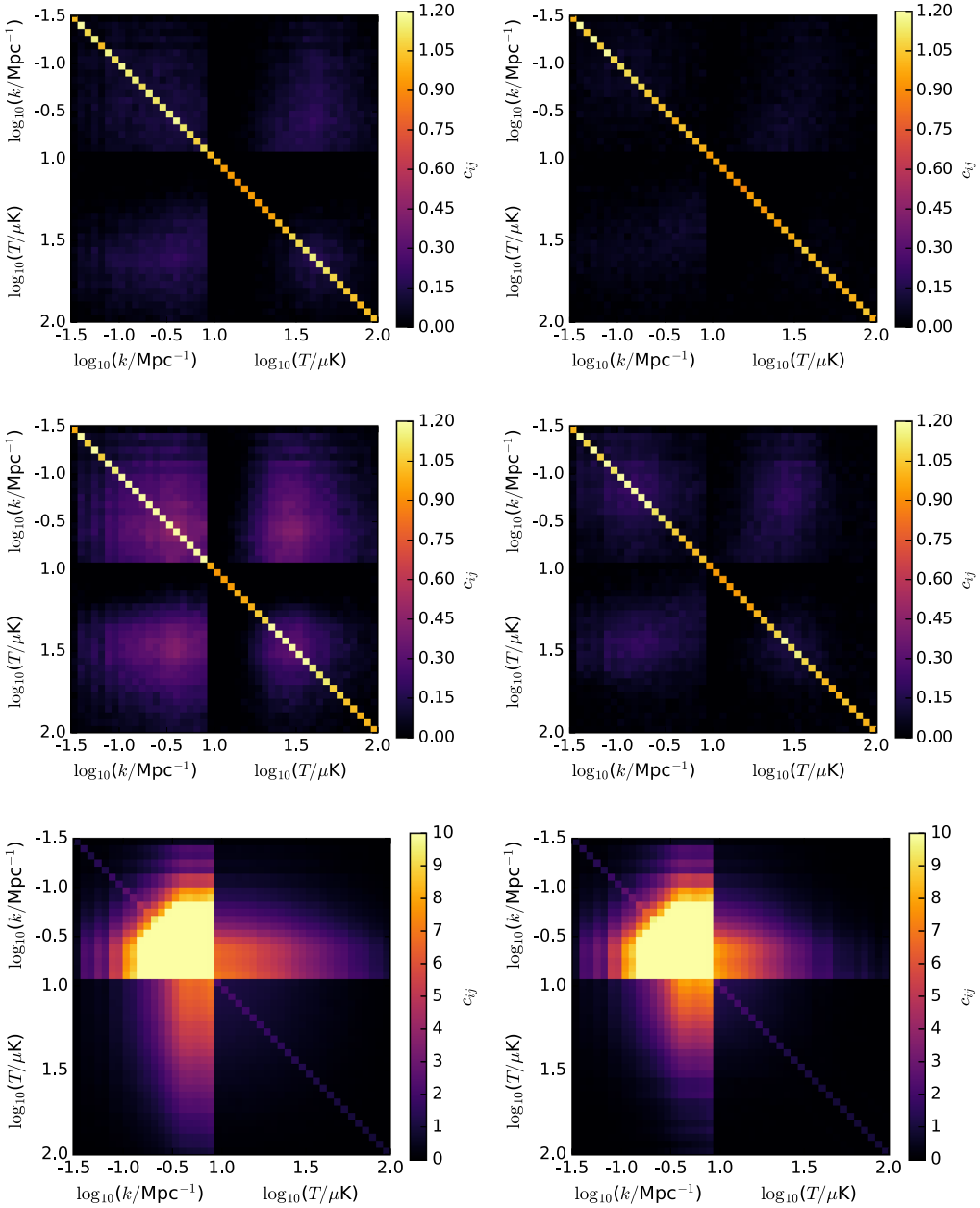

Figure 2. Estimated pseudo-correlation matrix of observables d_i, $c_{ij} = \mathrm{Cov}(d_i, d_j)/(\sqrt{\mathrm{Var}_{\mathrm{ind}}(d_i)\mathrm{Var}_{\mathrm{ind}}(d_j)})$, based on simulated maps with and without noise. The first block in each matrix corresponds to the power spectrum and the second block to the VID. Top: signal plus white noise corresponding to the COMAP1 experiment ($\sigma_{\mathrm{voxel}} \approx 11\ \mu$K). Middle: signal plus white noise corresponding to the COMAP2 experiment ($\sigma_{\mathrm{voxel}} \approx 8\ \mu$K). Bottom: signal alone. Note that here we have changed the color scale. Left: covariance matrices without beam smoothing. Right: covariance matrices with $\theta_{\mathrm{FWHM}} = 4'$ beam smoothing.

N_{vox} modes. However, it has a significantly higher impact for the VID. Specifically, it makes it much harder to distinguish a potential background of weak sources from noise. Indeed, as shown by Breysse et al. (2017), removing the monopole makes it much harder to detect a possible low luminosity cutoff in the CO luminosity function using the VID.

4. Results

We are now ready to present the main numerical results from our analysis, starting with an inspection of the joint PS+VID covariance matrix itself.

4.1. Visual Inspection of the PS+VID Covariance Matrix

Figure 2 shows the pseudo-correlation matrices, c_{ij}, for our two experimental setups, as well as for pure signal alone, for reference. In order to illustrate the effect of the beam, we show covariance matrices from maps both without and with beam smoothing in the left and right columns, respectively.

The first thing to notice is that instrumental noise significantly reduces the numerical values of the normalized covariance matrices, bringing it closer to the independent white noise case for which $c_{ij} = \delta_{ij}$. This agrees with intuition, since the noise itself is white and uncorrelated.

Beam smoothing also leads to weaker correlations. This is mainly due to the beam diluting the signal at small scales, where the correlation is otherwise strongest.

Next, we notice that the cross-correlations between the power spectrum and VID are of the same order of magnitude as the correlations internal to each observable itself. Thus, it is essential to account for all these correlations in any joint PS and VID analysis, as is done in the present paper.

Finally, we note that when designing an experiment like COMAP, one of the important trade-offs involves observation time per field. To obtain a fast signal detection it is in general advantageous to observe deep on the smallest possible field. However, this only holds true while the signal-to-noise per voxel is significantly less than unity. When the noise starts to become comparable to the signal, the signal-induced voxel–voxel correlations starts to become important, and the effective uncertainties no longer scale as $\mathcal{O}(1/\sqrt{\tau})$, where τ is the observation time per pixel. Generally, in such a tradeoff, any significant correlations between different power spectrum modes or voxel temperatures will tend to favor larger survey area or multiple fields, both effectively leading to more independent samples, and thereby higher overall integration efficiency.

4.2. Luminosity Function Constraints

We are now ready to present both individual and joint PS +VID constraints on the CO luminosity function, which are summarized in Figure 3 for COMAP1 (*left column*) and COMAP2 (*right column*). The top row shows the constraints obtained from the power spectrum alone; the middle row shows the constrains obtained from the VID alone; and the third row shows the constraints from the joint analysis. In each panel, the shaded colored region shows the 95% credibility region from the MCMC samples and the solid line with the same color shows the posterior median. The purple solid line shows the average luminosity function obtained from the mean of all available halo catalogs, and thus represents the ensemble average of our input model. Note that the colored regions correspond to one single realization and the uncertainties therefore contain contributions from instrumental noise, cosmic variance, and sample variance. The agreement between the estimated confidence regions and the ensemble mean is quite satisfactory in all cases, with uncertainties that appear neither too large nor too small.

Considering first the individual PS and VID estimates, shown in the top two rows, we see that the two observables are indeed complementary. In particular, the VID primarily constrains the high luminosity end of the luminosity function, while the power spectrum imposes relatively stronger constraints on the low luminosity end. This makes sense intuitively, since the VID is essentially optimized to look for strong outliers above the noise, whereas the power spectrum represents a weighted mean across the full field for each physical scale. It is interesting to note, however, that the VID provides, on average, stronger constraints on the luminosity function than the power spectrum does.

Due to this complementarity, the joint estimator provides the strongest constraints of all. To make this point more explicit, the fourth row compares the uncertainties of the independent power spectrum and VID analyses to the joint constraints. Of course, there is a significant amount of cosmic variance in each of these functions and the precise numerical value of the uncertainty ratio therefore varies significantly with luminosity; but the mean trend is clear: The individual analyses typically result in 20%–70% larger uncertainties than the joint analysis when averaged over luminosities between $L_{\rm CO} = 10^4$–$10^7\,L_\odot$. Over 10 cosmological realizations, the PS and VID resulted in, on average, 58% and 30% larger uncertainties (in dex) individually, than the joint analysis. This is the main novel result presented in this paper.

4.3. Posterior Distribution of Model Parameters

Lastly we present the constraints of the model parameters themselves. When doing the MCMC posterior mapping we explore the parameter space of the Li et al. (2016) $L_{\rm CO}(M_{\rm Halo})$ model. Figure 4 shows the posterior distribution for these parameters derived from one realization of the COMAP2 experiment (the same realization as the COMAP2 results in Figure 3).

Results for PS, VID, and joint PS+VID analysis are shown in blue, red, and black, respectively. Prior distributions are shown in green. The two curves of each color correspond 68% and 95% credibility regions.

We see that the two parameters that are mainly constrained are α and β, the two parameters from the average $L_{\rm CO}$–$L_{\rm IR}$ relation. These two parameters are fairly degenerate, and the direction in which they are degenerate is given roughly by the line $\alpha = -0.1\beta + 1.19$ (Li et al. 2016). In Figure 5, we show the luminosity function for different points on this line. For the figure, the values of $\sigma_{\rm SFR}$, $\sigma_{L_{\rm CO}}$, and $\log\delta_{\rm MF}$ are fixed at 0.3, 0.3, and 0.0, respectively. Although the overall signal strength, at least in terms of detectability, is fairly constant along this line, the shape of the luminosity function changes significantly. Lower values of alpha imply a more steep power-law relation between $L_{\rm CO}$ and $L_{\rm IR}$ leading to more sources with very high or very low luminosities. We see this as a flattening of the luminosity function. In such a case, a larger fraction of the overall signal will be given by high-luminosity sources.

The other parameter that is also slightly constrained is the log-normal scatter parameter from the $L_{\rm CO}$–$L_{\rm IR}$ relation, $\sigma_{L_{\rm CO}}$. This parameter is only slightly more constrained compared to the prior, with the highest values of $\sigma_{L_{\rm CO}}$ being disfavored. The posterior of the other scatter parameter, $\sigma_{\rm SFR}$, is basically given by the corresponding prior (i.e., this parameter is not very well constrained by the experiment), although, as expected from the

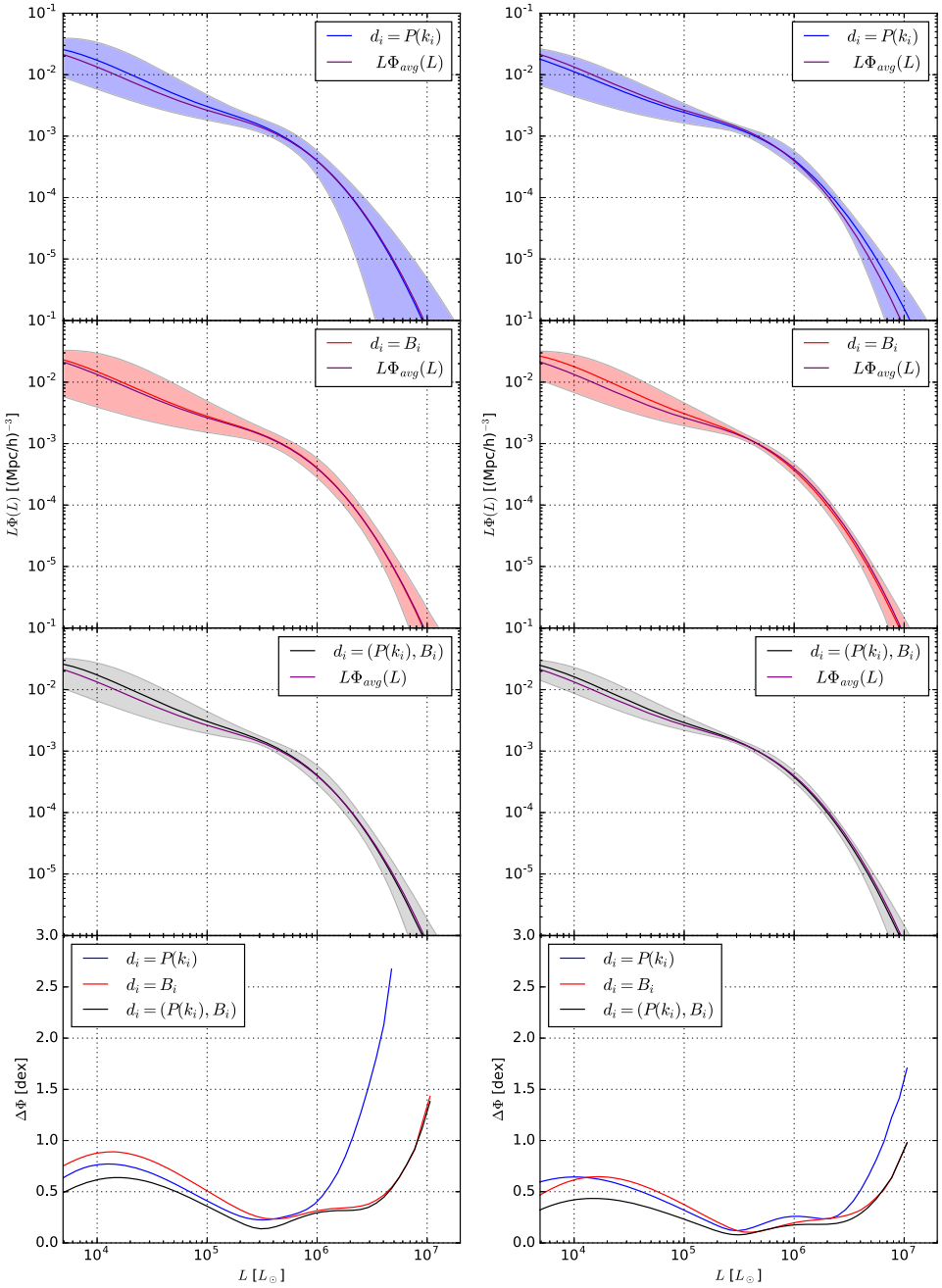

Figure 3. Constraints on the luminosity function from simulated experiments COMAP1 (left) and COMAP2 (right). The shaded area corresponds to 95% credibility intervals, solid lines correspond to the median, while the purple curve corresponds to the average luminosity function derived from all the available halo catalogs (i.e., the ensemble mean). Top row: constraints derived using only the power spectrum $P(k_i)$ as the observable. Middle row: constraints derived using only the temperature bin counts B_i as the observable. Bottom row: constraints derived by a joint analysis using both the power spectrum $P(k_i)$ and the temperature bin counts B_i as observables. Bottom: comparison of the uncertainty of the luminosity function constraints in dex, i.e., $\Delta\Phi \equiv \log_{10}\Phi_{97.5\%} - \log_{10}\Phi_{2.5\%}$.

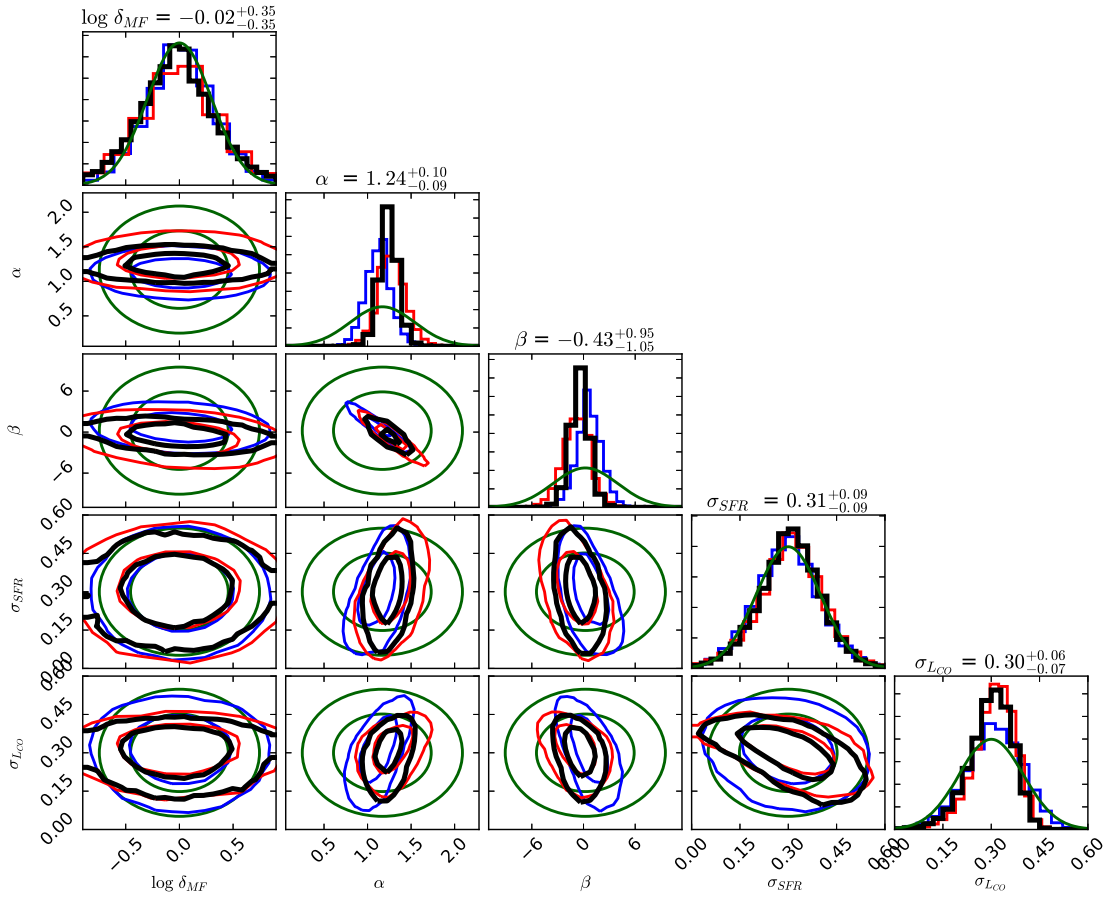

Figure 4. Posterior distributions for the Li et al. (2016) model parameters for a single realization of the COMAP2 experiment (the same realization as the COMAP2 results in Figure 3). Results for PS, VID, and joint PS+VID analysis are shown in blue, red, and (slightly bolder) black, respectively. Prior distributions are shown in green. The two curves of each color correspond to 68% and 95% credibility regions. The numbers on top of each column correspond to the 68% credibility interval for each parameter from the PS+VID analysis. We see that while the posterior of the two scatter parameters, $\sigma_{\rm SFR}$ and $\sigma_{L_{\rm CO}}$, is mostly set by the prior, the posterior on $\log \delta_{\rm MF}$, from the SFR–$L_{\rm IR}$ relation, is actually slightly wider than the prior, suggesting a significant intrinsic scatter in estimates of this parameter. These results are consistent with the corresponding results in Figure 7 in Li et al. (2016). The two parameters that are actually strongly constrained by the simulated experiment are α and β, the two parameters from the $L_{\rm CO}$–$L_{\rm IR}$ relation, and this figure shows that, at least for this realization, the constraints on these two get significantly improved in the combined analysis (PS+VID) as compared to analysis using the individual observables. This figure was created using the publicly available code (https://github.com/dfm/corner.py) `corner` (Foreman-Mackey 2016).

fact that the scatter parameters have basically the same effect, we see signs of the degeneracy between them in the posterior.

Interestingly, the normalization parameter in the SFR–$L_{\rm IR}$ relation, $\log \delta_{\rm MF}$, actually has a posterior that is wider than the prior. This may be because the best fit of this parameter from each of the different patches have an intrinsic scatter larger than the scatter in the prior. We note that we see the same effect in Li et al. (2016; their Figure 7).

From the mean relations in the Li et al. (2016) model, we have $\log L_{\rm CO} \sim -\beta - \log \delta_{\rm MF}$. Intuitively, we would then expect $\log \delta_{\rm MF}$ to be completely degenerate with β. However, since the SFR–$L_{\rm IR}$ is much better constrained by observations than the $L_{\rm CO}$–$L_{\rm IR}$ relation, the prior on $\log \delta_{\rm MF}$ is much tighter than the one on β. The degeneracy thus prevents us from constraining $\log \delta_{\rm MF}$ *until* β is constrained to a comparable level.

5. Discussion

We have developed a joint power spectrum and VID analysis for the CO luminosity function in the context of the COMAP CO intensity mapping experiment. We have implemented an efficient approach to estimating the joint covariance matrix for these two observables and have shown that accounting for both one- and two-point correlations leads to 20%–70% smaller uncertainties on the CO luminosity function for both COMAP1 and COMAP2.

The critical computational engine in our approach is the construction of fast yet semi-realistic simulations of the signal

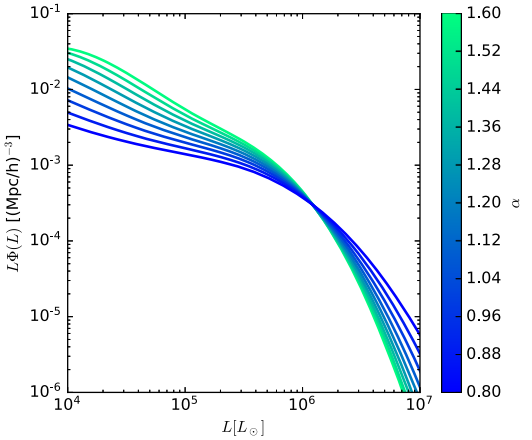

Figure 5. Plot of the CO luminosity function in the Li et al. (2016) model for different values of α and β. The colors of the lines indicate the values of α, the values of $\sigma_{\rm SFR}$, $\sigma_{L_{\rm CO}}$ and $\log \delta_{\rm MF}$ are fixed at 0.3, 0.3, and 0.0, respectively, while the value of β is determined from the relation $\alpha = -0.1\beta + 1.19$. This line corresponds roughly to the direction along which α and β are degenerate. Although the overall detectability of the signal remains roughly constant along this line, we see that the shape of the luminosity function changes significantly. We see that lower values of α correspond to less steep high-luminosity tails in the luminosity function, meaning that a larger proportion of the overall signal comes from high-luminosity sources.

in question. These simulations are based on the computationally cheap peak patch DM halo simulations produced by Bond & Myers (1996) and Stein et al. (2019), coupled to the semi-analytic CO luminosity model of Li et al. (2016). Of course, the results we derive are correspondingly limited by how well the model reproduces the true cosmological signal. If the true signal is significantly more complex than the model predicts, the constraints in Figure 3 will not be reliable.

The strength of the constraints on the CO luminosity function will depend on the overall level of the CO signal, which is highly uncertain. However, given the same rough level of signal, we expect the constraints on the luminosity function at the high luminosities to be less model dependent than the constraints on the $L_{\rm CO}$–$M_{\rm halo}$ relation or the luminosity function at lower luminosities. This is because the high-luminosity sources leave a fairly unique imprint on the maps that does not depend on the specific model used.

Additionally, we expect that the relative merits of using the PS or the VID as observables will change depending on the properties of the signal. In particular, anything that increases the shot noise of the signal, like a a strong galactic duty cycle, a large intrinsic scatter in luminosities or just a more top-heavy luminosity function, will make the resulting map more non-Gaussian, tending to favor observables like the VID more as compared to the PS. We can see this effect directly in Figure 4. The VID is better, compared to the PS, at ruling out low values of α and high values of $\sigma_{L_{\rm CO}}$, both of which correspond to cases where we would have a more top-heavy luminosity function and thus more shot noise.

We also expect the map to be more non-Gaussian on small scales than on large, so a wide survey with low resolution will tend to favor the PS, relative to the VID, more than a narrower high-resolution survey.

While the issues of model dependence are less relevant for low signal-to-noise measurements, where we are just trying to establish the rough level of the signal, they will become more important as the measurements improve.

Another potential issue with the simulations used in this paper is the minimum DM halo mass of $2.5 \times 10^{10}\,M_\odot$. While the model used here predicts that only a small fraction of the CO signal would come from halos lighter than this (see Li et al. 2016 and Chung et al. 2017), other models could disagree. If fact, searching for a low luminosity cutoff in the CO luminosity function is an interesting target for CO intensity mapping, and simulated halo catalogs with a smaller minimum DM halo mass would be useful both for forecasts and inference in such a scenario.

In general, it will be important to continuously improve the simulation pipeline as the experiment proceeds in order to account for more and more cosmological, astrophysical, and instrumental effects. However, the most important point in our approach is the fact that all such effects may be seamlessly accounted for, as long as the simulation procedure is sufficiently fast in order to be integrated into the MCMC procedure.

It should also be noted that our approach may be generalized in many different directions. For instance, the CO luminosity function does not play any unique role in our analysis, but is rather simply one specific worked example of a particularly interesting astrophysical function to be constrained. Many other functions may be constrained in a fully analogous manner, including, for instance, non-parametric $L_{\rm CO}(M_{\rm halo})$ models, or any of the parameters that are involved in converting the DM halo distributions to CO luminosities. The method is also not specific to CO intensity mapping, but should be equally well suited for other lines, or a combination of lines (Chung et al. 2018). Indeed, it should work for any type of random fields for which the covariance matrix must be estimated by simulations. Finally, we also note that there is nothing special about the power spectrum or VID as observables, but any other efficient data compression can be equally well included in the analysis, as long as the required compression step is sufficiently computationally efficient.

Support for the COMAP instrument and operation comes through the NSF cooperative agreement AST-1517598. Parts of this work were performed at the Jet Propulsion Laboratory (JPL) and California Institute of Technology, operating under a contract with the National Aeronautics and Space Administration. H.T.I., H.K.E., M.K.F., and I.K.W. acknowledge support from the Research Council of Norway through grant 251328. Research in Canada is supported by NSERC and CIFAR. These calculations were performed on the GPC supercomputer at the SciNet HPC Consortium. SciNet is funded by the Canada Foundation for Innovation under the auspices of Compute Canada; the Government of Ontario; Ontario Research Fund—Research Excellence; and the University of Toronto. Work at Stanford University is supported by NSF AST-1517598 and by a seed grant from the Kavli Institute for Particle Astrophysics and Cosmology. J.O.G. acknowledges support from the Keck Institute for Space Studies, NSF AST-1517108, and the University of Miami. H.P.'s research is supported by the Tomalla Foundation.

We thank Sarah E. Church, Tim Pearson, and other members of the COMAP collaboration for useful discussion and comments on a draft of this paper.

ORCID iDs

D. Chung https://orcid.org/0000-0003-2618-6504

References

Battye, R. A., Davies, R. D., & Weller, J. 2004, MNRAS, 355, 1339
Behroozi, P. S., Wechsler, R. H., & Conroy, C. 2013a, ApJL, 762, L31
Behroozi, P. S., Wechsler, R. H., & Conroy, C. 2013b, ApJ, 770, 57
Blot, L., Crocce, M., Sefusatti, E., et al. 2018, arXiv:1806.09497
Bond, J. R., Jaffe, A. H., & Knox, L. 2000, ApJ, 533, 19
Bond, J. R., & Myers, S. T. 1996, ApJS, 103, 1
Breysse, P. C., Kovetz, E. D., Behroozi, P. S., Dai, L., & Kamionkowski, M. 2017, MNRAS, 467, 2996
Breysse, P. C., Kovetz, E. D., & Kamionkowski, M. 2014, MNRAS, 443, 3506
Breysse, P. C., Kovetz, E. D., & Kamionkowski, M. 2015, MNRAS, 452, 3408
Breysse, P. C., Kovetz, E. D., & Kamionkowski, M. 2016, MNRAS, 457, L127
Carilli, C. L. 2011, ApJL, 730, L30
Carilli, C. L., & Walter, F. 2013, ARA&A, 51, 105
Chung, D. T., Li, T. Y., Viero, M. P., Church, S. E., & Wechsler, R. H. 2017, ApJ, 846, 60
Chung, D. T., Viero, M. P., Church, S. E., et al. 2018, ApJ, submitted (arXiv:1809.04550)
Cleary, K., Bigot-Sazy, M.-A., Chung, D., et al. 2016, AAS Meeting, 227, 426.06
Colavincenzo, M., Sefusatti, E., Monaco, P., et al. 2019, MNRAS, 482, 4883
da Cunha, E., Groves, B., Walter, F., et al. 2013, ApJ, 766, 13
Foreman-Mackey, D. 2016, JOSS, 1, 24
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, PASP, 125, 306
Fu, J., Kauffmann, G., Li, C., & Guo, Q. 2012, MNRAS, 424, 2701
Gilks, W., Richardson, S., & Spiegelhalter, D. 1995, Markov Chain Monte Carlo in Practice, Chapman Hall/CRC Interdisciplinary Statistics (London: Taylor and Francis)
Glenn, J., Conley, A., Béthermin, M., et al. 2010, MNRAS, 409, 109
Gong, Y., Cooray, A., Silva, M. B., Santos, M. G., & Lubin, P. 2011, ApJL, 728, L46
Greve, T. R., Leonidaki, I., Xilouris, E. M., et al. 2014, ApJ, 794, 142
Kovetz, E. D., Viero, M. P., Lidz, A., et al. 2017, arXiv:1709.09066
Lee, S. K., Ando, S., & Kamionkowski, M. 2009, JCAP, 7, 007
Leicht, O., Uhlemann, C., Villaescusa-Navarro, F., et al. 2019, MNRAS, 484, 269
Li, T. Y., Wechsler, R. H., Devaraj, K., & Church, S. E. 2016, ApJ, 817, 169
Lidz, A., Furlanetto, S. R., Oh, S. P., et al. 2011, ApJ, 741, 70
Lippich, M., Sánchez, A. G., Colavincenzo, M., et al. 2019, MNRAS, 482, 1786
Loeb, A., & Wyithe, J. S. B. 2008, PhRvL, 100, 161301
Madau, P., Meiksin, A., & Rees, M. J. 1997, ApJ, 475, 429
Mashian, N., Sternberg, A., & Loeb, A. 2015, JCAP, 11, 028
Obreschkow, D., Klöckner, H.-R., Heywood, I., Levrier, F., & Rawlings, S. 2009, ApJ, 703, 1890
Padmanabhan, H. 2018, MNRAS, 475, 1477
Peterson, J. B., Bandura, K., & Pen, U. L. 2006, arXiv:astro-ph/0606104
Pullen, A. R., Chang, T.-C., Doré, O., & Lidz, A. 2013, ApJ, 768, 15
Righi, M., Hernández-Monteagudo, C., & Sunyaev, R. A. 2008, A&A, 489, 489
Stein, G., Alvarez, M. A., & Bond, J. R. 2019, MNRAS, 483, 2236
Tinker, J., Kravtsov, A. V., Klypin, A., et al. 2008, ApJ, 688, 709
Vernstrom, T., Scott, D., Wall, J. V., et al. 2014, MNRAS, 440, 2791
Visbal, E., & Loeb, A. 2010, JCAP, 11, 016

Paper IV

BeyondPlanck I. Global Bayesian analysis of the Planck Low Frequency Instrument data

The BeyondPlanck collaboration

A&A, to be submitted, arXiv: 2011.05609,

IV

Astronomy & Astrophysics manuscript no. ms
November 12, 2020
©ESO 2020

arXiv:2011.05609v1 [astro-ph.CO] 11 Nov 2020

BeyondPlanck I. Global Bayesian analysis of the *Planck* Low Frequency Instrument data

BeyondPlanck Collaboration⋆: K. J. Andersen[11], R. Aurlien[11], R. Banerji[11], M. Bersanelli[4,9,10], S. Bertocco[8], M. Brilenkov[11], M. Carbone[14], L. P. L. Colombo[4], H. K. Eriksen[11], J. R. Eskilt[11], M. K. Foss[11], C. Franceschet[4,10], U. Fuskeland[11], S. Galeotta[8], M. Galloway[11], S. Gerakakis[14], E. Gjerløw[11], B. Hensley[2], D. Herman[11], M. Iacobellis[14], M. Ieronymaki[14], H. T. Ihle[11], J. B. Jewell[12], A. Karakci[11], E. Keihänen[3,7], R. Keskitalo[1], G. Maggio[8], D. Maino[4,9,10], M. Maris[8], A. Mennella[4,9,10], S. Paradiso[4,9], B. Partridge[6], M. Reinecke[13], M. San[11], A.-S. Suur-Uski[3,7], T. L. Svalheim[11], D. Tavagnacco[8,5], H. Thommesen[11], D. J. Watts[11], I. K. Wehus[11], and A. Zacchei[8]

[1] Computational Cosmology Center, Lawrence Berkeley National Laboratory, Berkeley, California, U.S.A.
[2] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, U.S.A.
[3] Department of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland
[4] Dipartimento di Fisica, Università degli Studi di Milano, Via Celoria, 16, Milano, Italy
[5] Dipartimento di Fisica, Università degli Studi di Trieste, via A. Valerio 2, Trieste, Italy
[6] Haverford College Astronomy Department, 370 Lancaster Avenue, Haverford, Pennsylvania, U.S.A.
[7] Helsinki Institute of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland
[8] INAF - Osservatorio Astronomico di Trieste, Via G.B. Tiepolo 11, Trieste, Italy
[9] INAF-IASF Milano, Via E. Bassini 15, Milano, Italy
[10] INFN, Sezione di Milano, Via Celoria 16, Milano, Italy
[11] Institute of Theoretical Astrophysics, University of Oslo, Blindern, Oslo, Norway
[12] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California, U.S.A.
[13] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85741 Garching, Germany
[14] Planetek Hellas, Leoforos Kifisias 44, Marousi 151 25, Greece

November 12, 2020

ABSTRACT

We describe the BeyondPlanck project in terms of motivation, methodology and main products, and provide a guide to a set of companion papers that describe each result in fuller detail. Building directly on experience from ESA's *Planck* mission, we implement a complete end-to-end Bayesian analysis framework for the *Planck* Low Frequency Instrument (LFI) observations. The primary product is a full joint posterior distribution $P(\omega \mid \boldsymbol{d})$, where ω represents the set of all free instrumental (gain, correlated noise, bandpass etc.), astrophysical (synchrotron, free-free, thermal dust emission etc.), and cosmological (CMB map, power spectrum etc.) parameters. Some notable advantages of this approach compared to a traditional pipeline procedure are seamless end-to-end propagation of uncertainties; accurate modeling of both astrophysical and instrumental effects in the most natural basis for each uncertain quantity; optimized computational costs with little or no need for intermediate human interaction between various analysis steps; and a complete overview of the entire analysis process within one single framework. As a practical demonstration of this framework, we focus in particular on low-ℓ CMB polarization reconstruction, paying special attention to the LFI 44 GHz channel. We find evidence of several significant residual systematic effects that are still not accounted for in the current processing, but must be addressed in future work. These include a break-down of the $1/f$ correlated noise model at 30 and 44 GHz, and scan-aligned stripes in the Southern Galactic hemisphere at 44 GHz. On the Northern hemisphere, however, we find that all results are consistent with the ΛCDM model, and we constrain the reionization optical depth to $\tau = 0.067 \pm 0.016$, with a low-resolution χ^2 probability-to-exceed of 16 %. The marginal CMB dipole amplitude is $3359.5 \pm 1.9\,\mu$K. This analysis framework can play a central role in the analysis of several current and future high-sensitivity CMB experiments, including *LiteBIRD*. All software is made publicly available under an OpenSource license, and both codes and products may be obtained through `http://beyondplanck.science`.

Key words. ISM: general – Cosmology: observations, polarization, cosmic microwave background, diffuse radiation – Galaxy: general

Contents

1 Introduction 2
1.1 CMB cosmology . . . 2
1.2 *Planck* . . . 3
1.3 Large-scale CMB polarization, the reionization optical depth, and systematic errors . . . 4
1.4 Lessons learned from *Planck* . . . 5
1.5 The next frontier: Primordial gravitational waves 6
1.6 The BeyondPlanck program . . . 6

2 BeyondPlanck analysis strategy and organization 7
2.1 End-to-end Bayesian CMB analysis . . . 7
2.2 `Commander` . . . 7
2.3 Paper organization . . . 7

3 Parameterizing the microwave sky 8
3.1 Conventions: Stokes parameters, pixelization, spherical harmonics, and units . . . 8

⋆ Corresponding author: H. K. Eriksen; `h.k.k.eriksen@astro.uio.no`

3.2 Cosmic microwave background anisotropies . . . 9
3.3 Galactic foreground emission . . . 10
3.3.1 Synchrotron emission . . . 10
3.3.2 Free-free emission . . . 10
3.3.3 Thermal dust emission . . . 11
3.3.4 Spinning dust (or anomalous microwave) emission . . . 11
3.3.5 Carbon monoxide emission . . . 12
3.4 Extra-galactic foreground emission . . . 12
3.4.1 Extra-galactic compact sources . . . 12
3.4.2 Sunyaev-Zeldovich effect . . . 13
3.4.3 Cosmic infrared background . . . 13
3.5 Zodiacal light emission . . . 13
3.6 Default sky model . . . 14

4 Instrument characterization 14
4.1 Ideal instrument model . . . 14
4.2 Spectral response, bandpass averaging, and unit conversion . . . 15
4.2.1 Bandpass uncertainties and corrections . 15
4.3 Beam and pixel window convolution . . . 16
4.4 Gain and analog-to-digital conversion . . . 18
4.5 Instrumental noise . . . 18

5 Data 19
5.1 LFI instrument overview . . . 19
5.1.1 Instrument configuration . . . 19
5.1.2 Stabilization . . . 20
5.1.3 LFI signal model . . . 20
5.1.4 Naming convention . . . 21
5.2 Implementation details . . . 22
5.2.1 Unprocessed Level-1 data . . . 22
5.2.2 Pre-processed Level-2 data . . . 22
5.2.3 1 Hz spike correction . . . 22
5.2.4 Analog-to-digital conversion correction . 22
5.3 Pixel-domain data . . . 22
5.3.1 *Planck* HFI data . . . 23
5.3.2 *Wilkinson Microwave Anisotropy Probe* . 23
5.3.3 Low-frequency surveys . . . 23

6 Bayesian analysis and MCMC sampling 24
6.1 Metropolis sampling . . . 24
6.2 Metropolis-Hastings sampling . . . 25
6.3 Gibbs sampling . . . 25

7 Global model specification 26
7.1 Global parametric model . . . 26
7.2 Deterministic quantities . . . 27
7.2.1 Frequency maps and leakage corrections 27
7.2.2 Spurious leakage maps . . . 28
7.2.3 Orbital dipole . . . 29
7.2.4 Far sidelobe corrections . . . 29

8 The BeyondPlanck Gibbs sampler 29
8.1 Global posterior distribution . . . 29
8.2 Overview of Gibbs chain . . . 29
8.3 Specification of conditional sampling steps . . . 30
8.3.1 Gain and calibration sampling . . . 30
8.3.2 Correlated noise sampling . . . 31
8.3.3 Noise PSD sampling . . . 33
8.3.4 Bandpass sampling . . . 33
8.3.5 Diffuse component spectral parameter sampling . . . 34
8.3.6 Diffuse component amplitude sampling . 35
8.3.7 Compact source sampling . . . 36
8.3.8 C_ℓ and cosmological parameter sampling 36
8.4 Computational requirements and optimization . . 38
8.4.1 Low-level optimization . . . 39
8.4.2 High-level parallelization and optimization 40

9 Results 40
9.1 Instrumental parameters . . . 42
9.2 Frequency maps . . . 47
9.3 Residual maps and masking . . . 55
9.4 Astrophysical component posteriors . . . 56
9.5 CMB posteriors . . . 57
9.5.1 The CMB solar dipole . . . 58
9.5.2 CMB maps . . . 62
9.5.3 Low-ℓ CMB likelihood . . . 62
9.5.4 High-ℓ CMB likelihood . . . 65
9.5.5 Cosmological parameters . . . 69

10 Reproducibility and Open Science 70
10.1 Reproducibility . . . 70
10.2 Software . . . 70

11 Conclusions, summary and outlook 71

A Review of frequently used textbook sampling algorithms 76
A.1 Univariate and low-dimensional Gaussian sampling . . . 76
A.2 High-dimensional Gaussian sampling . . . 76
A.3 Inversion sampling . . . 77

1. Introduction

1.1. CMB cosmology

According to the current cosmological concordance model, the observable universe came into existence some 13.8 billion years ago in a process often referred to as the Big Bang. While the physical laws underpinning this singular event remain unknown, it is a testament to the success of modern cosmology that physicists today are able to measure and model the evolution and energy content of the universe to exquisite precision, starting only a fraction of a second after the Big Bang.

Among the most important cosmological observables is the cosmic microwave background (CMB), first detected by Penzias & Wilson (1965). This leftover heat from the Big Bang fills the entire universe, and may today be observed as a nearly isotropic blackbody signal with a temperature of 2.7255 K (Fixsen 2009). Initially, CMB photons were trapped locally within a dense electron–proton plasma by Thomson scattering. However, once the mean plasma temperature fell below 3000 K as the Universe expanded, electrons and protons combined into neutral hydrogen atoms, and the photons were free to move throughout the entire observable universe, with almost no further scattering. This event took place some 380 000 years after the Big Bang, at a time often referred to as the *recombination epoch*. To any observer, the resulting photons appear to come from a so-called *last-scattering surface*, a sphere corresponding to a light travel distance of just under the entire history of the universe.

While the CMB field is very nearly isotropic, it does exhibit small spatial variations at the $O(30\ \mu\mathrm{K})$ level (e.g., Hu & Dodelson 2002, and references therein). These fluctuations are produced primarily by variations in the local gravitational potential, temperature, density, and velocity at the last-scattering surface.

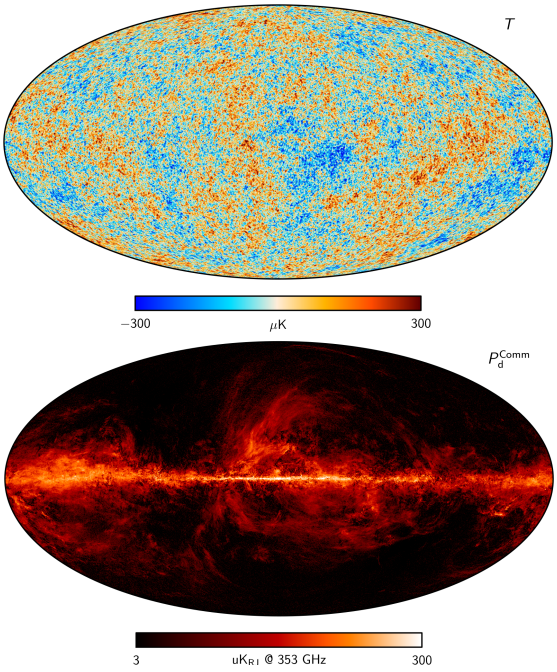

Fig. 1. *Planck* 2018 `Commander` CMB temperature (*top panel*) and thermal dust polarization amplitude (*bottom panel*) maps. Reproductions from Planck Collaboration IV (2020).

Smaller amplitude fluctuations arise from various secondary interactions taking place after the photons leave the last-scattering surface, for instance through gravitational lensing or Thomson scattering in the hot, ionized medium in clusters of galaxies. It is precisely by measuring and modelling all these small variations that cosmologists are able to decipher the history of the universe in ever greater detail. The currently best-fit cosmological model derived from this work is often referred to as *the ΛCDM model*, which posits that the universe is isotropic and homogeneous on large scales; that it started in a hot Big Bang; that it underwent a brief period of exponential expansion called inflation that seeded the universe with Gaussian random density fluctuations drawn from a scale-invariant power spectrum; and that the energy contents of the universe comprise 4.9 % baryonic matter, 26.5 % cold dark matter, and 68.5 % dark energy (Planck Collaboration VI 2020). Flat spatial curvature is also frequently assumed.

The rich cosmological information embedded in the CMB is not, however, easy to extract. Even the most dominant physical effects produce only $O(10^{-5})$ temperature fluctuations in the CMB. A primary goal for next-generation CMB experiments is the detection of primordial gravitational waves through the subtle polarization they imprint on the CMB (e.g., Kamionkowski & Kovetz 2016, and references therein). These so-called *B*-modes are likely to have an amplitude no larger than 30 nK, or a relative amplitude smaller than $O(10^{-8})$.

The fact that current CMB observations reach the μK level in the face of instrument noise and systematics is a testament to the effort of many scientists and engineers in this field, and to the time and money they have spent. Here, we list only a few of the results of five decades of observational milestones. NASA's *COBE* mission produced the first highly accurate measurement of the thermal spectrum of the CMB (Mather et al. 1994) and the first detection of large scale fluctuations in the CMB (Smoot et al. 1992). The first high-fidelity map of smaller scale CMB fluctuations was made by the BOOMERanG team (de Bernardis et al. 2000), and the first detection of polarized fluctuations by DASI (Kovac et al. 2002). These are among the more than 50 past and present CMB projects, all of which have contributed to technological innovations or scientific breakthroughs.

Two space missions, however, are primarily responsible for today's cosmological concordance model. They are NASA's *Wilkinson Microwave Anisotropy Probe* (*WMAP*; Bennett et al. 2013) and ESA's *Planck* (Planck Collaboration I 2020) satellite missions. *WMAP* was the first experiment to take full advantage of the exquisite thermal stability at Earth's second Lagrange point (L2), and observed the CMB sky for nine years (2001–2010) in five frequency bands (23–94 GHz) with precision unprecedented at the time.

1.2. Planck

The state-of-the-art in all-sky CMB observations as of 2020 is defined by *Planck* (Planck Collaboration I 2020). *Planck* observed the CMB sky for four years (2009–2013) in nine frequency bands (30–857 GHz), with three times higher angular resolution and ten times higher sensitivity than *WMAP*. Its original design goal was to measure the primary CMB temperature fluctuations with a precision limited only by fundamental physical processes, including cosmic variance, not by instrumental sensitivity (Planck Collaboration 2005).

Planck comprised two separate instruments within a common focal plane. One was the *Low Frequency Instrument* (LFI; Planck Collaboration II 2020), which employed coherent High Electron Mobility Transistor (HEMT) radiometers with center frequencies near 30, 44 and 70 GHz, each with a fractional bandwidth of roughly 20 %. The other was the *High Frequency Instrument* (HFI; Planck Collaboration III 2020), which employed spider-web and polarization sensitive bolometers with center frequencies of 100, 143, 217, 353, 545 and 857 GHz, each with a fractional bandwidth of 25 %.

Two different detector technologies were required to span *Planck*'s frequency range. The use of two very different detector technologies also provided crucial cross-checks against some subtle instrumental errors. *Planck*'s wide frequency range fully covered most of the spectrum of a 2.7255 K blackbody, but more crucially allowed for the removal of contaminating foreground signals (e.g., Leach et al. 2008). These arise from synchrotron emission from relativistic electrons moving in the magnetic field of the Galaxy, thermal emission from warm Galactic dust and bremsstrahlung emission from ionized gas, as well as microwave emission from extra-galactic sources. This list is not exhaustive; but each mechanism for foreground emission has a unique spatial distribution on the sky and a unique, non-blackbody spectrum which allows it to be distinguished from the CMB. The preferred method for separating cosmological fluctuations in the CMB from astrophysical foreground signals is to map the sky at multiple frequencies, and then perform a joint fit to this set of maps while taking into account the particular spatial and spectral behaviour of each foreground. These considerations drove the design of *Planck* (Planck Collaboration 2005). The capability to detect polarized signals was added at the seven lowest frequency bands, from 30 to 353 GHz. Figure 1 shows the CMB temperature fluctuation and the polarized thermal dust emission maps as

derived from *Planck* observations, which rank among the most iconic results from the experiment.

The *Planck* satellite was launched to L2 on May 14th 2009 and deactivated on October 23rd 2013; it thus completed in total almost 4.5 years of observations (Planck Collaboration I 2020). Unlike the case for *WMAP*, both *Planck* instruments were cryogenically cooled. The last 18 months of operation included only LFI measurements, as HFI exhausted its cooling capacity in January 2012.

The first *Planck* data release (denoted either "PR1" or 2013 here; Planck Collaboration I 2014) took place in March 2013, and was based on the first 15.5 months of data, covering the full sky twice. By and large, these measurements confirmed the cosmological model presented by *WMAP* and other previous experiments, but with significantly higher signal-to-noise ratio. This higher sensitivity also supported several truly groundbreaking results, two of which were a 25σ detection of gravitational lensing of CMB anisotropies (Planck Collaboration XVII 2014), and a revolutionary new image of polarized thermal dust emission in the Milky Way (Planck Collaboration XI 2014).

The 2013 release, however, did not include any CMB polarization results. In addition, the initial angular power spectrum of CMB anisotropies exhibited a ∼2 % shift in amplitude compared to the earlier *WMAP* power spectrum (Planck Collaboration XV 2014). Both of these issues had a common origin, namely incompletely controlled systematic errors arising from instrumental effects. As noted earlier, CMB observations are not easy: even small errors in assumptions made about foregrounds or instrumental behaviour can have dramatic effects on the recovered CMB signal. Examples of instrumental effects include: uncertainties in the beam shape and far sidelobes; mis-estimation of the frequency response of detectors, which can introduce temperature to polarization leakage; unaccounted-for non-linearity in the analog-to-digital converters (ADCs) used in each detector chain; and uncertainties in the polarization properties of detectors.

The *Planck* team grappled with all of these, as well as uncertainties in foreground contamination, in the years between 2013 and the release of the final *Planck* results in 2020 (Planck Collaboration I 2020). Very substantial investments of time and money were made to develop increasingly accurate models of the two *Planck* instruments; these allowed for more precise and robust science results. We emphasize that the official LFI and HFI pipelines evolved step-by-step in the post-launch period as instrument-specific effects emerged due to increased calibration accuracy. BeyondPlanck builds on all this accumulated experience in implementing a global approach to the data analysis problem.

A major milestone in this iterative process was the second *Planck* data release ("PR2" or 2015; Planck Collaboration I 2016), which for the first time included the full set of *Planck* observations (50 months of LFI data and 27 months of HFI data). At this point, the polarization properties of both the LFI and HFI instruments were sufficiently well understood to allow for a direct measurement of CMB polarization on intermediate and small angular scales (Planck Collaboration XI 2016). For HFI, however, accurate large-scale polarization was still out of reach due to systematic errors, and only LFI provided such constraints. The original power spectrum discrepancy relative to *WMAP* was tracked down to inaccuracies in the calibration procedure and reference dipole values used for the *Planck* 2013 analysis, and these were subsequently corrected in the 2015 release. With this second data release, *Planck* finally fulfilled its promise of measuring the primary CMB temperature fluctuations to the limits set by astrophysical and cosmological effects (Planck Collaboration I 2016).

1.3. Large-scale CMB polarization, the reionization optical depth, and systematic errors

Planck analysis continued beyond 2015, with a particular emphasis on reducing large-scale polarization systematics (Planck Collaboration I 2020). Both the importance and difficulty of this specific issue may be summarized in terms of the reionization optical depth, τ (e.g., Planck Collaboration Int. XLVII 2016). This parameter is directly related to the epoch during which the first stars were born, often called the *epoch of reionization* (e.g., Loeb & Barkana 2001, and references therein). According to detailed measurements of the abundance of neutral hydrogen in the universe from quasar spectra (the so-called "Lyman alpha forest"; Gunn & Peterson 1965), this event cannot have happened later than about 1 billion years after the Big Bang, corresponding to an optical depth of $\tau \gtrsim 0.048$. However, an independent measurement of τ may also be derived through CMB observations, by noting that the first stars or galaxies ionized their surrounding medium, and thereby released large numbers of free electrons off which CMB photons could scatter. Detailed models predict a CMB polarization signal at the level of $O(0.5\ \mu\mathrm{K})$ on angular scales larger than 10° (e.g., Alvarez et al. 2006, and references therein).

While the scientific potential in establishing robust large-scale polarization measurements is very high, potentially pinpointing a critical epoch in the history of the universe, the technical challenges are massive. The expected curl-free E-mode polarization signal is only about 1 % of the corresponding CMB temperature fluctuations, and the signal is only visible on large angular scales. Among all parameters in the cosmological concordance model, the reionization optical depth is the most susceptible to systematic errors, and for this reason it is often adopted as a monitor for residual errors.

To illustrate the difficulties associated with measuring τ, it is interesting to consider its value as reported in the literature as a function of time. The first CMB constraint was reported in the first-year *WMAP* release, which claimed $\tau = 0.17 \pm 0.04$ corresponding to a reionization epoch of $t_{\mathrm{r}} = 180^{+220}_{-80}$ Myr (Kogut et al. 2003). Such an early reionization epoch imposed strong limits on galaxy formation processes, and was not immediately compatible with standard theories. However, this preliminary measurement was based on the cross-correlation between temperature and polarization fluctuations for which uncertainties and degeneracies are large. Furthermore, it also did not account for bias introduced by foreground emission.

After adding more data, and, critically, allowing more time for understanding the data and controlling systematic errors, the 3-year *WMAP* data release resulted in a significantly revised estimate of $\tau = 0.089 \pm 0.03$, nearly doubling the time allowed for structure formation (Page et al. 2007). This estimate was derived directly from polarization-only measurements, and included proper foreground corrections. Based on further improvements and additional data, the reported 5-year *WMAP* posterior mean value was $\tau = 0.085 \pm 0.016$ (Komatsu et al. 2009), while in the 7-year release it was $\tau = 0.088 \pm 0.015$ (Larson et al. 2011), before finally settling on $\tau = 0.089 \pm 0.014$ in the 9-year release (Hinshaw et al. 2013). This represented the state-of-the-art before *Planck* in terms of large-scale CMB polarization measurements.

As already mentioned, no CMB polarization measurements were included in the first *Planck* 2013 release (Planck Collabo-

ration I 2014). However, from temperature measurements alone, the best-fit optical depth was constrained to $\tau = 0.097 \pm 0.038$, in seemingly excellent agreement with the final *WMAP* polarization results (Planck Collaboration XVI 2014). Then, in the *Planck* 2015 release, the LFI data allowed the first independent constraint on large-scale CMB polarization since *WMAP* (Planck Collaboration XI 2016). At the same time, the HFI polarization observations provided new and powerful constraints on Galactic polarized thermal dust (Planck Collaboration X 2016), to which the *WMAP* experiment was only marginally sensitive. The combination of LFI CMB and HFI thermal dust polarization measurements alone resulted in $\tau = 0.064^{+0.022}_{-0.023}$, or 1.1σ lower than the 9-year *WMAP* value. Furthermore, when combining the *WMAP* large-scale polarization CMB data with the same HFI polarization foreground data, the best-fit value was $\tau = 0.067 \pm 0.013$, in full agreement with LFI.

The HFI large-scale CMB polarization data were not considered sufficiently mature for scientific analysis until 2016, when new calibration, mapmaking, and simulation procedures had been implemented in a code called SROLL (Planck Collaboration Int. XLVIII 2016). Taking advantage of these new developments, and leveraging the higher statistical power of the HFI data, the reported estimate of the reionization optical depth was adjusted further down by HFI to $\tau = 0.055 \pm 0.009$. In parallel, the LFI procedure was improved by merging calibration and component separation into one framework. Combined, these new analysis procedures formed the basis for the third and final official *Planck* release (Planck Collaboration I 2020), for which a final value of $\tau = 0.053 \pm 0.009$ was reported. The good agreement with the lower limit imposed by quasar measurements, $\tau > 0.048$, implies both that reionization by the first generation of stars occurred relatively late, and that we can pin down the epoch of reionization with precision.

While a stable and internally consistent ΛCDM model, including τ, had emerged by the official end of the *Planck* consortium in 2018, one could still see clear signatures of residual systematics present in various subsets of the data. For HFI, several internal cross-correlations did not agree with each other to statistical precision (Planck Collaboration III 2020). For LFI the 44 GHz channel failed internal null tests (Planck Collaboration II 2020), and there were clear discrepancies between the raw frequency maps as seen by LFI and *WMAP* (Planck Collaboration IV 2020), indicating that there were still issues to be resolved within either LFI or *WMAP*, or both.

The last effort of the *Planck* collaboration to resolve these questions was organized within the so-called NPIPE pipeline (Planck Collaboration Int. LVII 2020). This name is short for "NERSC pipeline", a name deriving from the computer facilities at which it is executed, namely the National Energy Research Scientific Computing Center (NERSC). One unique feature of this pipeline is its ability to analyze both LFI and HFI jointly within the same framework. Combining some of the most powerful features from each of the instrument analysis pipelines, this approach has led to further reduction of systematic errors in both data sets, as reported in Planck Collaboration Int. LVII (2020). The resulting best-fit estimate of the reionization optical depth from NPIPE reads $\tau = 0.058 \pm 0.006$ (Tristram et al. 2020).

An independent initiative to improve the *Planck* processing was SROLL2(Delouis et al. 2019), which was a direct continuation of the HFI SROLL effort (Planck Collaboration III 2020). A defining feature of this approach is improved ADC corrections, which in particular leads to more robust large-scale polarization estimates. From the SROLL2 polarization analysis alone, the current best-fit estimate of the reionization optical depth is $\tau = 0.0566^{+0.0053}_{-0.0062}$ (Pagano et al. 2020).

A second independent initiative is called BeyondPlanck, and this is the primary focus of the current paper and suite of companion papers. The scope of this project is significantly different than the previous efforts, in that BeyondPlanck primarily aims at building a complete integrated end-to-end analysis pipeline for current and future CMB experiments. The current work focuses in particular on the *Planck* LFI data set, although significant effort is spent ensuring that the tools are generalizable to other experiments. Indeed, two examples of this are already presented within the current project, with preliminary applications to *WMAP* (Bennett et al. 2013; Watts et al. 2020) and *LiteBIRD* (Sugai et al. 2020; Aurlien et al. 2020).

Because instrumental systematics and residual foreground contamination have such a dramatic impact on the large-scale CMB polarization estimates, we will in this paper use the reionization optical depth as a direct demonstration of the BeyondPlanck framework, and our ultimate scientific goal is to estimate the posterior distribution $P(\tau \mid \boldsymbol{d})$ from *Planck* LFI and *WMAP* observations, $\boldsymbol{d}$. The posterior summarizes our knowledge about τ in the form of a probability distribution, and we will estimate $P(\tau \mid \boldsymbol{d})$ within a strict Bayesian framework, with as few approximations and little data selection as possible. We avoid the use of cross-spectrum techniques, which generally reduce the sensitivity of the final products to instrumental systematics. In this project, we aim to do the opposite, and *highlight* the impact of residual systematics, such that, if needed, they can be addressed at a lower level of the analysis. As such, internal consistency, goodness-of-fit and χ^2 tests will play critical roles.

1.4. Lessons learned from Planck

Historically speaking, to understand the background and motivation for the BeyondPlanck program, it is useful to revisit the "Lessons learned from *Planck*,"[1] as compiled by the *Planck* consortium in 2016. In Section 9.6 ("Understanding the data") one can read the following:

> *In a project like Planck, "understanding the data" is certainly the most significant driver of the quality of the final products and science it can produce. This activity must be at the core of the data processing. It covers a lot of ground – photometry, optical response, time response, calibration, systematic effects, etc. – all interlinked issues that can be diagnosed at many different levels in the data processing pipelines, from raw data streams to finished maps and scientific products.*
>
> *(. . .) In the early phases of Planck, much of the strategy was based on separating the various elements of the problem into independent parts. This was adequate for a first treatment of the data. However, as the quality of the data improved, it became harder to find and analyse subtler non-ideal effects, and to do so required a more integrated approach, where a variety of effects were treated simultaneously.*
>
> *(. . .) An example is the influence of foregrounds on calibration: initially model foreground templates were used to isolate the CMB dipole signal (the calibrator), but in later stages the template had to be iterated within the calibration pipeline to include and self-consistently reduce the effects of polarization, sidelobes, dipoles, etc.*

[1] https://www.cosmos.esa.int/web/planck/lessons-learned

(. . .) As understanding of the data progresses, analysis – and the teams doing it – need to become more and more integrated, pulling in parts of the pipeline which initially could be separated out.

As described in these paragraphs, the analysis approach adopted by *Planck* became gradually more and more integrated as the effective sensitivity of the data set improved through more refined analysis, and new systematic effects were uncovered. Indeed, only toward the end of the *Planck* mission period did it become evident that the single most limiting factor for the overall analysis was neither instrumental systematics nor astrophysical foregrounds as such, but rather the *interplay* between the two. Intuitively speaking, the problem may be summarized as follows: *One cannot robustly characterize the astrophysical sky without knowing the properties of the instrument, and one cannot characterize the instrument without knowing the properties of the astrophysical sky.* The calibration and component separation procedures are intimately tied together. By the time this issue was fully understood, there were neither sufficient resources nor time to redesign a complete *Planck* analysis pipeline from bottom-up. An important organizational goal of the BeyondPlanck program has therefore been to provide a financial structure that allows the team to consolidate this experience into practical computer code, and make this publicly available to the general community.

1.5. The next frontier: Primordial gravitational waves

While a statistically coherent analysis of existing data is undoubtedly both interesting and useful in its own right, it is important to emphasize that none of the developments detailed in this work are likely to impact the overall cosmological concordance model to any significant degree. Indeed, looking at the big picture, the cosmological model has been remarkably stable even before *WMAP* and *Planck* provided their high-precision measurements; see, e.g., Wang et al. (2003) for a summary of pre-*WMAP* measurements and constraints. The main achievement of *WMAP* and *Planck* has been to refine this model to the level at which cosmology now is a high-precision science within which competing theoretical models can be tested and rejected at high statistical significance.

Planck has for all practical purposes completed the study of primary CMB temperature fluctuations. Currently, however, another frontier is driving the CMB field, namely the search for primordial gravitational waves created during inflation. These are predicted to exist by most inflationary theories, although their predicted amplitudes can vary by many orders of magnitude, depending on the precise details of the assumed inflationary model (e.g., Kamionkowski & Kovetz 2016). Typically, this amplitude is quantified in terms of the tensor-to-scalar ratio, r, which measures the ratio in fluctuation power attributable to gravitational waves and ordinary density perturbations, respectively.

If such gravitational waves do exist, one generically expects a specific imprint in the CMB polarization field in the form of a large-scale "divergence-free" or B-mode polarization signal. The observational challenges associated with gravitational wave detection are essentially the same as those for measuring τ. However, the state-of-the-art upper limit on the tensor-to-scalar ratio is $r < 0.044$ at 95 % confidence (Tristram et al. 2020), which immediately implies that the B-mode signal must be more than one order of magnitude smaller than the E-mode signal, and thus no more than a few tens of nK in amplitude.

With such a small target amplitude, it is safe to assume that an integrated analysis approach will no longer be optional for future CMB missions, but rather a strict prerequisite. Establishing both the experience and appropriate code required to implement such an approach for future CMB missions is the main long-term scientific motivation for the BeyondPlanck program; current experiments such as *Planck* and *WMAP* provide real-world test-beds that ensure that the BeyondPlanck approach is both realistic and practical.

1.6. The BeyondPlanck program

In this context, we are now ready to formulate the main goal of the BeyondPlanck program:

> BeyondPlanck *aims to implement and apply a single statistically coherent analysis pipeline to Planck and other CMB data sets, processing raw uncalibrated time-ordered data into final astrophysical component maps, angular power spectra, and cosmological parameters within one single code.*

Important secondary goals include

1. to model and propagate instrumental uncertainties from raw time-ordered data into final high-level *Planck* scientific results;
2. to provide a computationally convenient interface to the raw *Planck* data that can be accessed and extended by external users;
3. to develop a framework that allows joint analysis of *Planck* with other data sets; and
4. to prepare for next-generation CMB experiments, in particular those aiming to detect primordial gravitational waves through their imprint on large-scale polarization of the CMB.

The "BeyondPlanck" name serves as a reminder that this work builds directly on several decades of *Planck* efforts and experience, while at the same time highlights the fact that it aims to apply the *Planck* methodology to data sets beyond *Planck*, both archival and future.

Clearly, this is a very ambitious program that will require long-term and dedicated support. The first stage of the program, which is reported in the current suite of papers, has been funded within an EU-based Horizon 2020 action called "Leadership in Enabling and Industrial Technologies" (LEIT), as well as through various individual grants. This funding only covers end-to-end analysis of the *Planck* LFI data, which is smaller in volume than HFI data, and therefore serves as a convenient real-world test case for development purposes, while still representing a very important scientific data set in its own right.

As detailed in the H2020 LEIT contract, the BeyondPlanck program started on March 1st 2018, and ended on November 30th 2020; the total duration of the program is thus strictly limited to two years and nine months. During this period, large amounts of software, products and documentation had to be written from scratch. Indeed, a first fully operational pipeline was completed as late as June 2020. With an effective run-time of six to eight weeks to achieve convergence on our current computer systems, we have been able to complete two full end-to-end data reprocessings since that time. While two full iterations certainly are a technical achievement, they are insufficient for detailed fine-tuning and thus the results are uncharacteristically unpolished relative to a typical data release. They however demonstrate the power of the analysis process itself. Further, the current BeyondPlanck release is not intended to be a static and final analysis solution for one specific data set, i.e., *Planck* LFI, but

rather a common community-wide platform that will allow scientists to explore different data sets both individually and jointly. As such, we expect numerous updates to emerge in the coming months and years, both from BeyondPlanck members and from external researchers, that will gradually refine the current products in a collaborative effort.

2. BeyondPlanck analysis strategy and organization

2.1. End-to-end Bayesian CMB analysis

Recognizing the lessons learned from *Planck* as summarized in Sect. 1.4, the defining design philosophy of BeyondPlanck is tight integration of all steps from raw time-ordered data processing to high-level cosmological parameter estimation. Traditionally, this process has been carried out in a series of weakly connected steps, pipelining independent executables with or without human intervention. Some steps have mostly relied on frequentist statistics, employing forward simulations to propagate uncertainties, while other steps have adopted a Bayesian approach, using the posterior distribution to quantify uncertainties. For instance, traditional mapmaking is a typical example of the former (e.g., Ashdown et al. 2007b), while cosmological parameter estimation is a typical example of the latter (e.g., Lewis & Bridle 2002); for component separation purposes, both approaches have been explored in the literature (e.g., Planck Collaboration Int. XLVI 2016).

BeyondPlanck is the first real-world CMB analysis pipeline to adopt an end-to-end Bayesian approach. This solution was in fact first proposed by Jewell et al. (2004). However, it took more than 15 years of computational and algorithmic developments to actually make it feasible.

Perhaps the single most important advantage of a uniform Bayesian approach is that it allows seamless propagation of uncertainties within a well-established statistical framework. This aspect will become critically important for future experiments, as demonstrated by *Planck*. For most CMB experiments prior to *Planck*, the dominant source of uncertainty was noise; for most CMB experiments after *Planck*, the dominant source of uncertainty will be instrumental systematics, foreground contamination, and the interplay between the two. As a logical consequence of this fact, BeyondPlanck adopts a consistent statistical framework that integrates detailed error propagation as a foundational feature.

The Bayesian approach also has several notable advantages in terms of intuition and transparency. In particular, the most critical step for any Bayesian analysis is the definition of the data model. This may often be described in terms of a handful of equations, and these equations subsequently serve as a road-map for the entire analysis. While the complexity of the numerical implementation may vary from model to model, the posterior distribution itself has a very intuitive and direct interpretation.

At a practical level, integrating the entire pipeline into a single computational code also has significant advantages in terms of net computational speed and resources. Not only are slow disk operations reduced to a minimum by performing all operations within one single code, but more importantly, all intermediate human interactions are eliminated from the process. This both saves significant amounts of human time required for "code shepherding" and file transfers, and it significantly reduces the risk of human errors. Thus human resources are saved that can be better spent on fundamental modelling aspects.

A fourth significant advantage of end-to-end integration is increased transparency of implicit and explicit priors. For a distributed analysis process, it is critically important to communicate all assumptions made in each step to avoid errors, while in an integrated approach internal inconsistencies become much more visible; there are simply fewer opportunities for misunderstandings to propagate undetected throughout an integrated analysis pipeline.

2.2. `Commander`

We adopt `Commander2` (Eriksen et al. 2004, 2008; Seljebotn et al. 2019), a well-established Bayesian CMB Gibbs sampler developed for *Planck*, as the starting point of our pipeline. As demonstrated in Planck Collaboration IV (2020), this code already supports Bayesian multi-resolution component separation, which is precisely the operation that connects low-level mapmaking to high-level cosmological parameter estimation. A main implementational goal for BeyondPlanck is thus to extend this framework to incorporate Bayesian calibration and mapmaking, as well as to connect component separation and cosmological parameter estimation.

We will refer to three different versions of the `Commander` code in the following. `Commander1` refers to the original implementation described by Eriksen et al. (2004, 2008), which at the beginning of the BeyondPlanck project represented the most mature version in terms of foreground spectral parameter fitting. However, a major limitation of that code is a requirement of common angular resolution among all data sets. `Commander2` removes this limitation through explicit beam convolution for each frequency map during component separation, as detailed by Seljebotn et al. (2019), and thereby allows for full resolution analysis of the *Planck* data. Due to the much higher computational cost associated with increased angular resolution, the development of `Commander2` required a re-implementation of the original algebra from scratch, adopting a much more fine-grained parallelization strategy than `Commander1`.

Finally, `Commander3` refers to the time-domain version of the algorithm, as developed in BeyondPlanck, and is a direct generalization and extension of `Commander2` in terms of code implementation. As a result, `Commander2` is no longer an independent code, but we will still refer to it in cases where it might be convenient for pedagogical purposes to distinguish between multi-resolution component separation in the pixel-domain versus the time-domain. All `Commander` source codes are available under a GNU Public Library (GPL) OpenSource license.[2]

2.3. Paper organization

The BeyondPlanck methodology and results are described in a suite of companion papers, as listed in Table 1. The present paper provides a broad overview in terms of motivation, algorithms, and main results. However, it is not intended to be comprehensive, as specific details are instead deferred to the relevant companion papers.

The remaining papers may be divided into four main categories, namely 1) pipeline papers; 2) instrument characterization papers; 3) cosmological and astrophysical results papers; and 4) analysis of external data. The first category of papers provides a comprehensive overview of the current implementation of the BeyondPlanck pipeline, at a level that is hopefully sufficiently detailed to allow external users to understand intuitively its statistical and computational basis, what assumptions it relies on, and what its limitations are. The ultimate goal of these papers is

[2] https://github.com/Cosmoglobe/Commander

Table 1. Overview of BeyondPlanck papers.

Reference	Title
Pipeline	
BeyondPlanck Collaboration (2020)	I. Global Bayesian analysis of the *Planck* Low Frequency Instrument data
Keihänen et al. (2020)	II. CMB mapmaking through Gibbs sampling
Galloway et al. (2020a)	III. Computational infrastructure and `Commander3`
Brilenkov et al. (2020)	IV. Time-ordered data simulations
Gerakakis et al. (2020)	V. Open Science and reproducibility
Instrument characterization	
Ihle et al. (2020)	VI. Noise characterization and modelling
Gjerløw et al. (2020)	VII. Calibration
Galloway et al. (2020b)	VIII. Sidelobe modelling
Svalheim et al. (2020a)	IX. Bandpass and beam leakage modelling
Cosmological and astrophysical results	
Suur-Uski et al. (2020)	X. LFI frequency map posteriors and sample-based error propagation
Colombo et al. (2020)	XI. CMB analysis with end-to-end error propagation: Temperature anisotropies
Paradiso et al. (2020)	XII. CMB analysis with end-to-end error propagation: Likelihood and cosmological parameters
Andersen et al. (2020)	XIII. Intensity foregrounds, degeneracies and priors
Svalheim et al. (2020b)	XIV. Polarized foreground emission between 30 and 70 GHz
Herman et al. (2020)	XV. Limits on polarized anomalous microwave emission
External analysis	
Aurlien et al. (2020)	XVI. Application to simulated *LiteBIRD* observations
Watts et al. (2020)	XVII. Application to *WMAP*
Galeotta et al. (2020)	XVIII. End-to-end validation of BeyondPlanck

that external users should be able to repeat and extend the work that is presented here.

The second category of papers address the various relevant instrumental parameters required to process the raw time-ordered data into sky maps. These include noise characterization, gain estimation, sidelobe corrections, and bandpass and beam mismatch modelling. Each paper aims both to provide an intuitive understanding of the effect in question, and to show how it impacts the final results. These papers also demonstrate how to quantitatively model each instrumental effect, and how to propagate uncertainties into other parameters. Particular emphasis is placed on building intuition regarding leading internal parameter degeneracies, both among the various instrumental parameters and with astrophysical and cosmological parameters.

The third category of papers present the main scientific results in terms of frequency and component maps, as well as angular power spectra and cosmological parameters. Consistency between the BeyondPlanck products and non-*Planck* sets is also considered in this category of papers.

The fourth category includes papers that aim to generalize the BeyondPlanck data model to other data sets. For now, the main emphasis is put on *WMAP* (Watts et al. 2020) and *LiteBIRD* (Aurlien et al. 2020).

We note that, in the spirit of reproducibility and accessibility, a significant emphasis is put on intuition and background throughout the BeyondPlanck papers. The paper suite is intended to be largely self-contained, and detailed knowledge of the *Planck* publication list is not an assumed prerequisite. As such, a substantial amount of review material is included, both in terms of general background material and algorithmic details. The style of the papers is consciously tuned toward Ph.D. students and early postdoctoral fellows, rather than seasoned CMB experts.

3. Parameterizing the microwave sky

As already noted, the single most important component in any Bayesian analysis is the parametric model that is fitted to the data. In our case, this model consists of both astrophysical and instrumental components. In this section we consider the cosmological and astrophysical parameters, before introducing the instrumental parameters in the next section.

3.1. Conventions: Stokes parameters, pixelization, spherical harmonics, and units

In order to characterize each astrophysical component quantitatively, we need to introduce some general notation and conventions. First, each astrophysical component will be described in terms of three Stokes parameters, namely intensity (denoted either I or T) and two linear polarizations (denoted Q and U). We will ignore circular polarization (V) for now, but we note that this may be added in future work.

To discretize the Stokes parameters on the sphere, we adopt the `HEALPix` pixelization[3] (Górski et al. 2005). This pixelization has several highly desirable properties, including equal-area pixels and support for fast spherical harmonics transforms, and is now effectively a standard in modern CMB analysis. The `HEALPix` pixel resolution is controlled through a parameter called $N_{\rm side}$, and the total number of pixels on the sky is $N_{\rm pix} = 12N_{\rm side}^2$. We organize the Stokes parameters into vectors of length $3N_{\rm pix}$, simply by stacking $\{T, Q, U\}$ into a map vector $\boldsymbol{s}(\hat{n})$, where $\hat{n}$ is a unit direction vector.

Unless otherwise noted, we define the Stokes parameters with respect to Galactic coordinates. We adopt the cosmological convention for the polarization angle, γ, in which $\gamma = 0$ for vectors pointing north and increases westward. This is opposite

[3] `http://healpix.jpl.nasa.gov`

to the IAU convention used in most other fields of astronomy, in which γ increases eastward. To convert from one convention to the other, one must multiply Stokes U by -1.

The Stokes polarization parameters Q and U form a spin-2 field, which intuitively may be interpreted as a "headless vector field". In contrast, the intensity T is a spin-0 field, and does not change under rotations. Thus, when rotating Stokes parameters by an angle α, the transformed Stokes parameters are

$$\begin{bmatrix} T' \\ Q' \\ U' \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos 2\alpha & -\sin 2\alpha \\ 0 & \sin 2\alpha & \cos 2\alpha \end{bmatrix} \begin{bmatrix} T \\ Q \\ U \end{bmatrix}. \tag{1}$$

As described by Zaldarriaga & Seljak (1997), the polarization Stokes parameters may be expanded into spherical harmonics through the following relations,

$$T(\hat{n}) = \sum_{\ell=0}^{\ell_{\rm max}} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\hat{n}) \tag{2}$$

$$(Q \pm iU)(\hat{n}) = \sum_{\ell=2}^{\ell_{\rm max}} \sum_{m=-\ell}^{\ell} {}_{\pm 2}a_{\ell m}\, {}_{\pm 2}Y_{\ell m}(\hat{n}), \tag{3}$$

where ${}_k a_{\ell m}$ are called (spin-k) spherical harmonic coefficients. The polarization coefficients are often combined algebraically into E and B coefficients,

$$a^E_{\ell m} = -\frac{1}{2}\left({}_2a_{\ell m} + {}_{-2}a_{\ell m}\right) \tag{4}$$

$$a^B_{\ell m} = \frac{i}{2}\left({}_2a_{\ell m} - {}_{-2}a_{\ell m}\right), \tag{5}$$

which each form a spin-0 field, fully analogous to the intensity T.

From the spherical harmonic coefficients we may compute the observed angular power spectrum as

$$\sigma^{XY}_\ell = \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} \left(a^X_\ell\right)^* a^Y_{\ell m}, \tag{6}$$

where $\{X, Y\} \in \{T, E, B\}$. These quantify the strength of fluctuations at a given multipole ℓ as directly measured from some sky map. In addition, we define the ensemble-averaged power spectrum as

$$C^{XY}_\ell \equiv \left\langle \left(a^X_\ell\right)^* a^Y_{\ell m} \right\rangle = \left\langle \sigma^{XY}_\ell \right\rangle, \tag{7}$$

where brackets indicate an average over statistical realizations. This function is thus independent of the observed sky, and only depends on the model that describes the field in question.

Finally, each sky map $\boldsymbol{s}$ must be quantified in terms of a physical unit. In the following work, we will encounter many different conventions for this, depending on the particular application in question. However, three conventions are more common than others, and we limit our discussion here to these special cases.

The first measure is *surface brightness per solid angle*, which simply measures the amount of energy emitted by some source per surface area, per frequency interval, per sky solid angle. This is often measured in units of $\mathrm{MJy\,sr^{-1}} \equiv 10^{-20}\,\mathrm{W\,m^{-2}\,Hz^{-1}sr^{-1}}$, and it quantifies the specific intensity I_ν of a given source as a function of wavelength, ν.

The second measure we will use is *thermodynamic temperature*. In this case, we identify the intensity with that emitted by a blackbody source with temperature T,

$$I_\nu = B_\nu(T) = \frac{2h\nu^3}{c^2} \frac{1}{\mathrm{e}^{\frac{h\nu}{kT}} - 1}, \tag{8}$$

where h is Planck's constant, c is the speed of light, and k is the Boltzmann constant. This measure is particularly useful for CMB applications, because the CMB is itself a near-perfect blackbody, and a single temperature $T(\hat{n})$ therefore uniquely specifies its intensity at any wavelength at a given position. The unit for thermodynamic temperature is denoted $\mathrm{K_{CMB}}$ or simply K.

Our third and final measure is the *brightness temperature* or *Rayleigh-Jeans temperature*, $T_{\rm RJ}$. This is defined by the the long wavelength limit ($h\nu \ll kT$) of Eq. (8), such that

$$I_\nu = \frac{2\nu^2 k T_{\rm RJ}}{c^2}. \tag{9}$$

While the thermodynamic temperature is convenient to describe the CMB, most astrophysical foreground signals have a non-blackbody nature, and are more naturally quantified in terms of brightness temperature. In particular, while the spectral energy density of many foregrounds can span many tens of orders of magnitude when expressed in $\mathrm{K_{CMB}}$, they are usually limited to a few orders of magnitude when expressed in either $\mathrm{MJy\,sr^{-1}}$ or $\mathrm{K_{RJ}}$. To avoid numerical problems, all astrophysical components are therefore expressed in units of $\mathrm{K_{RJ}}$ internally in `Commander`, and only converted to the respective natural unit before outputting results to disk. Monochromatic conversion between $\mathrm{K_{RJ}}$ and $\mathrm{MJy\,sr^{-1}}$ is performed through Eq. (9), while monochromatic conversion between $\mathrm{K_{RJ}}$ and $\mathrm{K_{CMB}}$ is given by

$$\Delta T_{\rm CMB} = \frac{(\mathrm{e}^x - 1)^2}{x^2 \mathrm{e}^x} T_{\rm RJ}, \tag{10}$$

where $x = h\nu/kT_0$, and $T_0 = 2.7255\,\mathrm{K}$ is the mean CMB temperature (Fixsen 2009). Note that this conversion applies only to small temperature variations around the CMB mean value, $\Delta T \equiv T - T_0$, which is precisely the form of most CMB temperature maps in common use today.

We are now ready to write down parametric models for each of the main astrophysical components that are relevant for the *Planck* frequency range. Each component will be described in terms of a spectral energy density (SED) in brightness tempeerature units, and, in some cases, in terms of an angular power spectrum or some other similar spatial coherence measure.

3.2. Cosmic microwave background anisotropies

We start our survey with the CMB component, which is the scientifically most important one for *Planck*. For this, we first define $\boldsymbol{s}^{\rm CMB}$ to be a $3N_{\rm pix}$ sky vector of CMB Stokes parameters as described above. Second, we assume that the CMB SED may be approximated as a blackbody. As such, its brightness temperature SED is given by Eq. (10),

$$\boldsymbol{s}^{\rm CMB}_{\rm RJ}(\nu) \propto \frac{x^2 \mathrm{e}^x}{(\mathrm{e}^x - 1)^2} \boldsymbol{s}^{\rm CMB}, \tag{11}$$

where $x = h\nu/kT_0$. (Note that we define the effective SED only up to a normalization constant, as we will typically parameterize each component in terms of an amplitude map at a given reference frequency times the SED normalized to unity at the reference; any normalization factor is therefore accounted for in the amplitude coefficient.)

For component separation purposes, this is the *only* assumption we make regarding the CMB. However, for cosmological parameter estimation purposes, we make two important additional assumptions, namely that the CMB temperature fluctuations are both Gaussian distributed and statistically isotropic.

The assumption of Gaussianity determines the conditional probability distribution for the CMB signal,

$$P(\boldsymbol{s} \mid C_\ell) \propto \frac{\mathrm{e}^{-\frac{1}{2}\boldsymbol{s}^t \mathsf{S}^{-1}\boldsymbol{s}}}{\sqrt{|\mathsf{S}|}}, \tag{12}$$

where S is the covariance matrix of the CMB fluctuation field, and we have dropped the "CMB" superscript for convenience. The assumption of statistical isotropy implies that S is fully specified in terms of the angular power spectrum,

$$\mathsf{S}^{XY}_{\ell m,\ell' m'} \equiv \left\langle \left(a^X_\ell\right)^* a^Y_{\ell' m'} \right\rangle = C^{XY}_{\ell m} \delta_{\ell\ell'} \delta_{mm'}. \tag{13}$$

For practical parameter estimation purposes, both of these assumptions have been shown to be excellent approximations to the true CMB sky (see, e.g., Planck Collaboration VII 2020; Planck Collaboration IX 2020, and references therein).

The connection to cosmological parameters, such as the Hubble constant H_0 or the reionization optical depth τ, is made through cosmological Boltzmann codes, such as `CMBfast` (Seljak & Zaldarriaga 1996) or `CAMB` (Lewis et al. 2000). These deterministically calculate the ensemble-averaged CMB power spectrum based on well-understood physics given some specific set of cosmological parameters, ξ. However, this calculation is only straightforward going from ξ to C_ℓ; it is highly nontrivial to go directly from C_ℓ to ξ. Instead, Markov Chain Monte Carlo (MCMC) methods such as `CosmoMC` (Lewis & Bridle 2002) are typically employed to perform the inversion, in which a series of parameter combinations are proposed and rejected or accepted, ultimately resulting in a set of parameter samples that jointly represents the final parameter posterior distribution. As described in Sect. 1.6, the goal of the BeyondPlanck program is to implement a similar MCMC method that eventually accounts for the entire process from raw time-ordered data to final cosmological parameters; in its current form, the final product of the BeyondPlanck pipeline will be the so-called CMB power spectrum likelihood, $\mathcal{L}(C_\ell) \equiv P(\boldsymbol{d} \mid C_\ell)$, and the final parameter estimation process will still be performed using `CosmoMC`.

3.3. Galactic foreground emission

The second most important class of sky emission components consists of diffuse Galactic foregrounds. These all originate from within the Milky Way, and are due to particles (electrons, ions, dust, etc.) associated with various processes such as star formation or supernova explosions. Furthermore, these particles all interact with the same magnetic field, and as a result they produce correlated polarized emission. In this section, we provide a brief survey of each of the main physical emission mechanisms, with a particular focus on parametric models.

3.3.1. Synchrotron emission

At low microwave frequencies, synchrotron emission dominates the radio sky. This emission is mostly due to relativistic electrons ejected from supernova, spiralling in the magnetic field of the Milky Way. CMB observations are typically made at frequencies in the range of tens or hundreds of GHz, and at these frequencies, the synchrotron SED falls rapidly with increasing frequency. Indeed, detailed models and observations both suggest that the effective spectrum may be closely approximated by a power-law at frequencies higher than a few gigahertz, with some evidence for possible curvature. In this work, we therefore follow Kogut (2012), and adopt a general SED model of the form

$$s^{\mathrm{synch}}_{\mathrm{RJ}}(\nu) \propto \left(\frac{\nu}{\nu_{0,\mathrm{s}}}\right)^{\beta + C\ln\nu/\nu_{0,\mathrm{s}}}, \tag{14}$$

where $\nu_{0,\mathrm{s}}$ is a reference frequency, β is a power-law index, and C is a curvature parameter. However, in most cases we set $C = 0$, as the signal-to-noise ratio for this parameter is very low with the limited data set considered in this work.

When the local magnetic field is highly structured, synchrotron emission can be highly polarized, with a theoretical maximum polarization fraction of $p = 75\,\%$. In practice, this value is decreased due to line-of-sight and volume integration effects, and according to *Planck* and *WMAP*, more typical values are $\lesssim 15\,\%$ at high Galactic latitudes, with extreme cases reaching 30–50 % only in a few large-scale supernova remnants that, when projected on the sky, take the form of so-called "Galactic spurs" (Planck Collaboration XXV 2016).

At low frequencies, polarized synchrotron emission is also significantly affected by Faraday rotation (e.g., Beck et al. 2013, and references therein). This effect is caused by circular birefringence, i.e., left- and right-handed circular polarized emission travel at different speeds through a magnetic field embedded in an ionized medium, resulting in a net rotation of the polarization angle of linearly polarized emission. The polarization angle rotation is proportional to the magnetic field strength as well as to the square of the wavelength of the emission. Numerically, the rotation angle is typically a few degrees at 23 GHz at low Galactic latitudes, but reaches hundreds of degrees at 2.3 GHz (Carretti et al. 2019; Fuskeland et al. 2019). Therefore, where relevant, we account for Faraday rotation when comparing our results with low-frequency surveys such as S-PASS (2.3 GHz; Carretti et al. 2019), but we neglect it for higher-frequency surveys such as *Planck* and *WMAP*.

3.3.2. Free-free emission

Free-free emission (or *bremsstrahlung*) arises primarily from free electrons scattering off protons without being captured, and emitting a photon in the process. Since free electrons only exist in appreciable amounts when the temperature of the medium is comparable to the hydrogen binding energy, corresponding to $10^3 - 10^4$ K, free-free emission predominantly traces hot H II regions and, as such, active star forming regions. Free-free emission is particularly important for CMB experiments because it is the only foreground component that is non-negligible at all frequencies between 1 and 1000 GHz, and it is therefore particularly sensitive to degeneracies with respect to both the CMB and other foreground components.

The free-free SED depends primarily on the number of free protons and electrons along the line of sight, which typically is quantified in terms of the *emission measure* (EM), i.e., the integral of the square electron density along the line of sight,

$$\mathrm{EM} \equiv \int_0^\infty n_e^2 \, \mathrm{d}l, \tag{15}$$

where the number densities of free protons and electrons are assumed to be equal. The conventional unit adopted for the EM is pc cm^{-6}, and typical values for the Milky Way range between 0 and 1000 (Planck Collaboration X 2016).

Assuming local thermodynamic equilibrium and first considering an optically thick medium, the free-free SED is determined by a blackbody spectrum given its electron temperature,

T_e, alone. Since the optical depth drops rapidly with increasing frequency, however, free-free emission in astrophysical contexts and at CMB frequencies is optically thin. Hence, the effective SED can be expressed as

$$s^{\rm ff}_{\rm RJ}(\nu) = T_e\,(1 - {\rm e}^{-\tau}). \tag{16}$$

As shown by Dickinson et al. (2003) and Draine (2011), τ may be very well approximated by

$$\tau = 0.05468 \cdot T_e^{-3/2} \cdot \nu_9^{-2} \cdot {\rm EM} \cdot g_{\rm ff}, \tag{17}$$

where

$$g_{\rm ff} = \log\left\{\exp\left[5.960 - \sqrt{3}/\pi \log(\nu_9 \cdot T_4^{-3/2})\right] + {\rm e}\right\} \tag{18}$$

is called the Gaunt factor, and ν_9 and T_4 are the frequency and the electron temperature measured in units of GHz and 10^4 K, respectively.

This SED is a nonlinear function of EM and T_e. A complete free-free model therefore corresponds to a complicated probability distribution with expensive special-purpose sampling algorithms, as for instance employed in Planck Collaboration IX (2016). In this work, we instead adopt a simpler linearized version of Eq. (16) that is only strictly valid in the optically thin case, $\tau \ll 1$, namely

$$s^{\rm ff}_{\rm RJ}(\nu) \propto \frac{g_{\rm ff}(T_e)}{\nu^2}, \tag{19}$$

and we correspondingly quantify the free-free amplitude in terms of the observed signal at a given reference frequency in $\mu{\rm K_{RJ}}$, as opposed to the full nonlinear EM parameter described above.

There is essentially no effective alignment mechanism for thermal electrons in a hot medium, and large-scale free-free emission is therefore expected to be nearly unpolarized. The main exception to this are sharp edges around hot H II regions, which do introduce a preferred direction in the emission geometry. However, even these are only expected to be mildly polarized, and over large angular scales, the net polarization fraction is expected to be well below 1 % (see discussion in Keating et al. 1998). In this paper, we thus assume that free-free emission is completely unpolarized.

3.3.3. Thermal dust emission

The interstellar medium (ISM) is filled not only with hydrogen and electrons, but also with tiny dust grains ranging in diameter from less than a nanometer (i.e., a few atoms across) to roughly a micron (i.e., thousands of atoms across). Dust grains typically condense from stellar outflows and ejecta, and so dust abundance is correlated with star formation. Newly-formed dust is rapidly mixed in the dynamic, turbulent ISM, where it undergoes significant processing. Dust is therefore ubiquitous in the Galaxy, found wherever there is interstellar gas.

It is known from spectroscopic features that dust is made from, at minimum, silicate and carbonaceous materials. However, the precise grain composition is likely to vary with local environment. Dust grains are heated by ambient stellar radiation, and large grains accounting for the bulk of the dust mass equilibriate to a steady-state temperature ranging between 10 and 30 K. This energy is thermally re-emitted with a peak wavelength in the sub-mm frequency range, typically between 1000 and 3000 GHz. Since these grains are inefficient radiators at longer wavelengths, the thermal dust SED falls rapidly at frequencies below the peak, where CMB observations are typically carried out. The varied composition and geometry of ISM dust particles makes the thermal dust SED significantly more complicated to model from first principles, when compared to the free-free emission described above; for recent examples of such modelling efforts, see, e.g., Guillet et al. (2018) and Draine & Hensley (2020).

In practice, simpler fitting formulae are therefore usually adopted for practical analyses, and one particularly popular class of models is the so-called modified blackbody spectrum, which in intensity units reads

$$I^{\rm d}_\nu \propto \tau \nu^{\beta_{\rm d}} B_\nu(T_{\rm d}). \tag{20}$$

This function is simply a blackbody spectrum with temperature $T_{\rm d}$, modulated by a power-law having index $\beta_{\rm d}$. In physical terms, this corresponds to dust having an opacity that scales as $\nu^{\beta}_{\rm d}$, a reasonable approximation for wavelengths longer than $\sim 20\,\mu$m (Hensley & Draine 2020).

The amplitude is, as for free-free emission, given by the optical depth, τ, which depends directly on the surface density of particles along the line of sight. Typical numerical values for these three parameters are $\tau \sim 10^{-6}$, $\beta_{\rm d} \sim 1.6$, and $T_{\rm d} \sim 20$ K. Intuitively speaking, $\beta_{\rm d}$ determines the slope (or first derivative in log-log space) of the SED below 200 GHz, while $T_{\rm d}$ determines the SED peak position, and second derivative at lower frequencies. However, we will model thermal dust emission in terms of brightness temperature, and in these units the effective SED may be written in the form

$$s^{\rm d}_{\rm RJ}(\nu) \propto \frac{\nu^{\beta_{\rm d}+1}}{{\rm e}^{h\nu/kT_{\rm d}} - 1}. \tag{21}$$

Interaction with gas and radiation torques up grains, and they tend to rotate about their axis of greatest moment of inertia, i.e., their short axis. Dust grains having unpaired electrons can develop a non-zero magnetic moment anti-parallel to their angular velocity through the Barnett effect (Dolginov & Mitrofanov 1976). Dissipative processes act to align the rotation axis with the local magnetic field. For a more detailed discussion of grain alignment, see Andersson et al. (2015).

The preferential alignment of the short axes of grains with the local magnetic field leads to significant net polarization from the ensemble of grains. Thermal dust polarization fractions as large as 20 % are found using the high frequency *Planck* polarization measurements (Planck Collaboration XI 2020). We therefore include all three Stokes parameters in our thermal dust model. At the same time, we note that the highest polarization-sensitive *Planck* frequency channel is 353 GHz, and this does not provide sufficient frequency range to allow an independent estimate of the thermal dust temperature in polarization. We therefore assume the same $T_{\rm d}$ for intensity and polarization, while $\beta_{\rm d}$ is allowed to be different.

3.3.4. Spinning dust (or anomalous microwave) emission

Dust grains rotate with rotational kinetic energy of order the thermal energy in the ambient gas. Consequently, sub-nanometer grains can achieve rotational frequencies of tens of GHz. If these grains possess an electric dipole moment, as generally expected for particles of this size (Macià Escatllar & Bromley 2020), this rotation produces emission in the microwave frequency range, as first predicted theoretically by Erickson (1957), and

described quantitatively by Draine & Lazarian (1998). The spinning dust mechanism currently provides the most popular theoretical explanation for so-called "anomalous microwave emission" (AME) observed around 20 GHz in CMB surveys, as first identified and named by Leitch et al. (1997).

In this work, we will adopt a spinning dust model for this component, starting from an SED template, $s_0^{\rm sd}(\nu)$, computed with the SpDust2 code (Ali-Haïmoud et al. 2009; Ali-Haimoud 2010; Silsbee et al. 2011) for environmental parameters typifying the Cold Neutral Medium. This spectrum is intrinsically computed in intensity units, in which it peaks at 30 GHz. After converting to brightness temperature by scaling with ν^{-2}, as given by Eq. (9), the peak shifts to 17.4 GHz, and the overall spectrum is less than 1 % of its peak value at frequencies below 1.3 GHz or above 66 GHz. To fit this SED model to the data, we follow Bennett et al. (2013), and introduce a peak position parameter, $\nu_{\rm p}$, that shifts the spectrum rigidly in $\log\nu$–$\log s$ space,

$$s_{\rm RJ}^{\rm sd}(\nu) \propto \nu^{-2}\, s_0^{\rm sd}\left(\nu \cdot \frac{30.0\,\mathrm{GHz}}{\nu_p}\right) \tag{22}$$

We note, however, that this emission component is associated with large uncertainties, both in terms of the physical mechanism that is actually responsible for the observed emission, and in terms of detailed modelling within the chosen paradigm. In a companion BeyondPlanck paper, Herman et al. (2020), we perform a non-parametric analysis of the observed diffuse AME spectrum in polarization. However, for all main analyses presented in this work, we adopt the spinning dust model in Eq. (22) for the AME component.

Despite sensitive searches in individual objects (Génova-Santos et al. 2015; Génova-Santos et al. 2016) and over large sky areas (Macellari et al. 2011), polarization has not been detected in the AME. In principle, AME could be highly polarized if small spinning grains are efficiently aligned. Theoretical estimates of the alignment efficiency of ultrasmall grains vary widely, with predicted AME polarization fractions ranging from $\lesssim 1\%$ (Hoang et al. 2013) to completely negligible (Draine & Hensley 2016). We perform a detailed study of AME polarization in Herman et al. (2020), but assume it to be unpolarized in all other analysis.

3.3.5. Carbon monoxide emission

In the same way that rotating dust particles can emit radio emission, so can molecules with a non-zero electric dipole moment. One particularly important example of such molecules is carbon monoxide (CO), which resides primarily in dense clouds where it is shielded from destruction by UV radiation. The most common isotopologe of CO is $^{12}C^{16}O$ (abbreviated ^{12}CO), which is typically 10–100 times more abundant than $^{13}C^{16}O$ (abbreviated ^{13}CO) (Szűcs et al. 2014).

For a simple system such as CO, quantum mechanical effects are highly significant. In particular, only very specific rotational states are allowed by quantization of angular momentum. Let us denote the masses of the two atoms by $m_{\rm C}$ and $m_{\rm O}$, respectively, and the corresponding atomic distances from their center of mass by $r_{\rm C}$ and $r_{\rm O}$. We also define $r_{\rm CO} = r_{\rm C} + r_{\rm O}$ to be the effective atom size and $m_{\rm CO} = m_{\rm C}m_{\rm O}/(m_{\rm C} + m_{\rm O})$ its reduced mass.

With this notation, the moment of inertia of the CO molecule is $I = m_{\rm C}r_{\rm C} + m_{\rm O}r_{\rm O}$. The corresponding eigenvalues of the Schrödinger equation are given by

$$E_{\rm rot} = \frac{J(J+1)\hbar^2}{2I}, \tag{23}$$

where $J = 0, 1, \ldots$ is the angular momentum quantum number. Quantum mechanically allowed energy changes are given by $\Delta J = \pm 1$, and each such transition either absorbs or emits a photon with wavelength

$$\nu_0 = \frac{\Delta E_{\rm rot}}{h} = \frac{\hbar J}{2\pi I} = \frac{\hbar J}{2\pi m r_{\rm CO}^2}, \quad J = 1, 2, \ldots \tag{24}$$

For the ^{12}CO J=1←0 transition, one finds $\nu_0 = 115.27$ GHz, while for the ^{13}CO J=1←0 transition, it is $\nu_0 = 110.20$ GHz. Higher-order transitions, such as J=2←1, are very nearly multiples of these frequencies.

The width of CO lines is small compared to the broad *Planck* bandpasses, and so we model the corresponding SED by a delta function at the respective frequency,

$$s_{\rm RJ}^{\rm CO}(\nu) \propto \delta(\nu - \nu_0). \tag{25}$$

We note that specific intensity units are not appropriate for CO emission since all of the energy is being emitted in a narrow spectral range. Instead, CO emission is conventionally quantified in terms of $\mathrm{K\,km\,s^{-1}}$. Because the central frequency is known from theory, emission away from line center can be attributed to radial motion with a definite mapping between frequency and velocity. The line intensity in brightness temperature units is integrated over all velocities, yielding $\mathrm{K\,km\,s^{-1}}$.

CO emission is expected to be only weakly polarized, with a polarization fraction around 1 % in molecular clouds (Greaves et al. 1999). Detecting such low levels of true polarization is challenging with the currently available *Planck* data, primarily due to instrumental temperature-to-polarization leakage. For now, we assume CO line emission to be fully unpolarized, but note that this is likely to change in future analysis.

Finally, we note that although the base CO frequencies do not lie within the *Planck* LFI frequency bands themselves, CO emission is nevertheless indirectly important for LFI because of its prevalence in the HFI channels, and these are in turn critical to model thermal dust emission for LFI.

3.4. Extra-galactic foreground emission

In addition to the Galactic foreground emission mechanisms discussed in Sect. 3.3, several extra-galactic effects are also important for CMB frequencies.

3.4.1. Extra-galactic compact sources

For LFI frequencies, the most important class of extra-galactic components are compact radio sources. All the emission mechanisms listed above operate in external galaxies, but the radio source population is dominated by active galactic nuclei (AGN). Radio emission from AGN is largely synchrotron, and comes from either the galactic nucleus itself or from jets and lobes associated with the nucleus. While the morphology of individual sources may be complicated, few are resolved by most CMB experiments and hence can be treated as "point" sources. Thus, while individual components of an AGN may exhibit polarized microwave emission, the emission from an unresolved source as a whole is rarely strongly polarized; typical polarization fractions are a few percent (Datta et al. 2019).

AGN have a wide range of SEDs, but most AGN spectra at CMB frequencies can be adequately modeled by a simple power law with a spectral index determined primarily by the energy spectrum of the relativistic electrons generating the synchrotron

emission. The spectral indices (in brightness) typically range from 0 to −1, and most fall in a narrower range of −0.5 to −0.7. Hence we adopt a simple power law SED as our model for radio sources, and fit for the amplitude and spectral index of the radio source contribution,

$$s_{\mathrm{RJ}}^{\mathrm{src}}(\nu) \propto \left(\frac{\nu}{\nu_{\mathrm{src}}}\right)^{\alpha-2} \tag{26}$$

Here ν_{src} is a fixed reference frequency, and α is the spectral index defined in intensity units; the conversion between intensity and brightness temperature is proportional to ν^2.

As we move to higher CMB frequencies, or to more sensitive experiments, the counts of extra-galactic sources begin to include dusty galaxies. These objects emit modified blackbody radiation, like Galactic dust, but typically at higher temperatures. Emission from this class of objects is included in the cosmic infrared background discussed below.

Unlike the dusty galaxies, which tend to be clustered, synchrotron-dominated radio sources are quite randomly distributed on the sky, and hence have a flat angular power spectrum. On the other hand, the emission of synchrotron dominated sources is frequently variable, on time scales ranging from days to years. Time variability is not accounted for in the current model, and variable sources are therefore likely to leave residuals in the final maps. For this reason, we will apply a dedicated point source mask during the final CMB parameter estimation, to minimize contamination in the final cosmological parameters.

3.4.2. Sunyaev-Zeldovich effect

Some CMB photons happen to pass through one or more clusters of galaxies on their way through the universe. Such clusters are very hot, some reaching temperatures as high as 10^8 K. At such high temperatures, the inter-cluster medium is highly ionized.

CMB photons have a non-negligible probability of scattering on these free electrons, and when they do, they gain energy from the free electrons. As a result, their spectrum is shifted to slightly higher frequencies compared to the standard blackbody form. This is called the thermal Sunyaev-Zeldovich (tSZ) effect (Sunyaev & Zeldovich 1972), and it is an effective probe of the intergalactic medium in high-redshift clusters.

In this work, we will mostly ignore the tSZ effect, as it has a relatively modest impact on the LFI measurements, due to both their limited sensitivity and angular resolution. In the cases where we do consider it, we adopt the non-relativistic model of the effect, which in brightness temperature units takes the form

$$s_{\mathrm{RJ}}^{\mathrm{SZ}}(\nu) \propto \frac{x^2 \mathrm{e}^x}{(\mathrm{e}^x-1)^2}\left(\frac{x\mathrm{e}^x+1}{\mathrm{e}^x-1}-4\right), \tag{27}$$

where $x = h\nu/kT_0$.

In additional to the thermal SZ effect, non-zero cluster velocities give rise to an additional contribution called the kinetic SZ effect. This does not affect the SED shape of the underlying photons, but simply changes the apparent temperature fluctuation at a given position. For typical cluster velocities of $\lesssim 10^3\,\mathrm{km\,s^{-1}}$, these modifications are however small, at the level of a few μK, and we therefore neglect this effect in the following. Likewise, we also neglect small polarization effects in the thermal SZ case, which are expected to be well below 1 μK in amplitude.

3.4.3. Cosmic infrared background

The last extra-galactic component that will be encountered in this analysis is the cosmic infrared background (CIB). This is the thermal dust emission emitted by many distant galaxies. The CIB may be spatially approximated as a continuous field, similar to the CMB, but with an SED that is defined as an average of a large number of independent thermal dust SEDs, each redshifted according to the distance of the emitting galaxy.

The CIB affects CMB observations in two different ways. First, uncertainties in the CIB monopole translate into uncertainties in the zero-level of each frequency channel. In particular, current models predict a CIB monopole of about 400 μK at 353 GHz, but with a model uncertainty of about 20 %. If left unmitigated, such large uncertainties would translate into massive uncertainties in the thermal dust spectral parameters, β_{d} and T_{d}. In practice, the HFI monopoles are currently determined through cross-correlation with H I (Planck Collaboration III 2020); however, this approach is of course associated with its own uncertainties.

Second, CIB fluctuations dominate over Galactic thermal dust fluctuations near the Galactic poles, where local thermal dust emission is low. In the foreground model employed in the current work, we do not account separately for CIB fluctuations, as we do not have a high enough signal-to-noise ratio to robustly separate them from Galactic emission. The thermal dust emission estimates presented in the following therefore correspond to the sum of Galactic thermal dust emission and CIB fluctuations. Since CIB and thermal dust emission have very similar SEDs, this has only a small effect on other components, most importantly on the CMB. However, it does complicate the physical interpretation of the resulting thermal dust parameter maps and dust parameters derived from them.

3.5. Zodiacal light emission

The last emission component we will consider is zodiacal light emission (ZLE). Similar to both CO and CIB emission, this component is far more important for HFI than LFI frequencies, and its mean amplitude is only about 0.5 μK at 70 GHz (Planck Collaboration X 2016).

The ZLE is emitted by dust particles located within the Solar system, primarily in the asteroid belt between Mars and Jupiter. These grains are heated by solar radiation to a temperature of about 150 K, and this energy is then thermally re-emitted with a thermal dust-like SED with the corresponding temperature. As such, its frequency spectrum is similar to both Galactic thermal dust and CIB fluctuations across CMB bands.

However, unlike Galactic dust and CIB, the ZLE is not stationary on the sky throughout a given survey. As *Planck* moves with the Earth around the Sun throughout a year, the ZLE is observed through different lines of sight. This both allows for, and indeed requires, dynamic modelling of the effect, taking into account the precise location of the satellite as a function of time. For this purpose, we adopt the *COBE*-DIRBE ZLE model (Kelsall et al. 1998), as integrated and re-implemented natively into `Commander` for efficiency purposes.

The *COBE* model treats the ZLE in terms of six distinct constituents (a smooth cloud, three asteroidal dust bands, a circumsolar ring, and a trailing blob), each specified in terms of its own geometry and optical properties. In total, the model has 43 free parameters, and in this work we adopt the parameter set described by Planck Collaboration XIV (2014), which takes into account information from both *COBE*-DIRBE and *Planck*. Future work will aim to estimate these parameters internally in our analysis framework, but this will require time-domain *Planck* HFI and *COBE*-DIRBE observations, and is outside the scope of the present work. In the current work, we adopt the same cor-

rections as described in Planck Collaboration Int. LVII (2020) for the HFI channels, and neglect ZLE for the LFI channels.

3.6. Default sky model

Based on the above survey, and unless specified otherwise, the default BeyondPlanck astrophysical sky model (in brightness temperature units) reads as follows,

$$\boldsymbol{s}_{\rm RJ} = \boldsymbol{a}_{\rm CMB}\,\frac{x^2 e^x}{(e^x-1)^2}\,\frac{(e^{x_0}-1)^2}{x_0^2 e^{x_0}} + \tag{28}$$

$$+\,\boldsymbol{a}_{\rm s}\left(\frac{\nu}{\nu_{0,\rm s}}\right)^{\beta_{\rm s}} + \tag{29}$$

$$+\,\boldsymbol{a}_{\rm ff}\,\frac{g_{\rm ff}(\nu;T_e)}{g_{\rm ff}(\nu_{0,\rm ff};T_e)}\left(\frac{\nu_{0,\rm ff}}{\nu}\right)^2 + \tag{30}$$

$$+\,\boldsymbol{a}_{\rm AME}\left(\frac{\nu_{0,\rm sd}}{\nu}\right)^2\,\frac{s_0^{\rm sd}\left(\nu\cdot\frac{\nu_p}{30.0\,\rm GHz}\right)}{s_0^{\rm sd}\left(\nu_{0,\rm sd}\cdot\frac{\nu_p}{30.0\,\rm GHz}\right)} + \tag{31}$$

$$+\,\boldsymbol{a}_{\rm d}\left(\frac{\nu}{\nu_{0,\rm d}}\right)^{\beta_{\rm d}+1}\,\frac{e^{h\nu_{0,\rm d}/kT_{\rm d}}-1}{e^{h\nu/kT_{\rm d}}-1} + \tag{32}$$

$$+\sum_{j=1}^{N_{\rm src}}\boldsymbol{a}_{\rm src}^j\left(\frac{\nu}{\nu_{0,\rm src}}\right)^{\alpha_{j,\rm src}-2}, \tag{33}$$

where $x = h\nu/kT_0$ and $\nu_{0,i}$ is the reference frequency for component i. Thus, $\boldsymbol{a}_i$ is the amplitude of component i in units of $\mu K_{\rm RJ}$, as observed at a monochromatic frequency $\nu_{0,i}$. The sum in line 33 runs over all sources brighter than some flux threshold as defined by an external source catalog, and both the amplitude and spectral index are fitted individually per source. We adopt the same catalog as Planck Collaboration IV (2020), which is hybrid of the AT20G (Murphy et al. 2010), GB6 (Gregory et al. 1996), NVSS (Condon et al. 1998) and PCCS2 (Planck Collaboration XXVI 2016) catalogs comprising a total of 12 192 individual sources.

Only $\{\boldsymbol{s}_{\rm RJ}, \boldsymbol{a}_{\rm CMB}, \boldsymbol{a}_{\rm s}, \boldsymbol{a}_{\rm d}\}$ are assumed to be polarized in this model, and these comprise 3-component vectors including Stokes T, Q, and U parameters. The remaining amplitudes parameters, $\{\boldsymbol{a}_{\rm ff}, \boldsymbol{a}_{\rm AME}, \boldsymbol{a}_{\rm src}^j\}$, are assumed unpolarized, and have vanishing Stokes Q and U parameters.

For algorithmic reasons, we distinguish between linear and nonlinear parameters. The former group includes $\{\boldsymbol{a}_{\rm CMB}, \boldsymbol{a}_{\rm s}, \boldsymbol{a}_{\rm ff}, \boldsymbol{a}_{\rm AME}, \boldsymbol{a}_{\rm d}, \boldsymbol{a}_{\rm src}\}$, collectively denoted $\boldsymbol{a}$; as described in Sect. 8.3.6, this set of parameters may be estimated jointly and efficiently through a multivariate Gaussian Monte Carlo sampler. In contrast, the nonlinear parameters include $\{\beta_{\rm s}, T_e, \nu_p, \beta_{\rm d}, T_{\rm d}, \beta_{\rm src}\}$, and these must be estimated independently and with computationally far more expensive algorithms; see Sect. 8.3.5 for specific details. In practice, we fit individual compact source amplitudes jointly with the corresponding spectral indices using a general sampling algorithm, since these are much more correlated with these than with any of the diffuse component parameters.

4. Instrument characterization

We now turn to the second half of the parametric model employed in the BeyondPlanck analysis, which describes the instrument used to collect the measurements. So that the BeyondPlanck analysis may freely be used by others, we aim to keep the presentation and notation as general as possible, and only introduce BeyondPlanck and LFI-specific notation where strictly necessary. We start our discussion by first defining an ideal detector response model, and then increase the level of realism step-by-step, until we reach the final instrument model.

4.1. Ideal instrument model

Let us first consider an ideal monochromatic detector observing at frequency ν a stationary sky signal with local Stokes parameters $\{T, Q, U\}$ at Galactic coordinates (l, b) and polarization angle ψ. We also initially assume infinite angular resolution. In this ideal case, the signal recorded by the detector as a function of time t may be written as

$$d(t) = g(t)\,[T + Q\,\cos 2\psi + U\,\sin 2\psi] + n(t), \tag{34}$$

where g is a multiplicative factor called the gain, which converts between physical signal units (which in our case will be $K_{\rm CMB}$) and digitized instrumental detector units (which in our case will be V), and n denotes instrumental noise.

In order to obtain data that may be processed on a computer, it is necessary to discretize the measurements by averaging over some (short) time period, Δt. For most CMB experiments, typical samples rates are between 10 and 200 Hz. A single recorded datum, d_t, thus corresponds to the detector output averaged over a period typically between 0.005 and 0.1 s.

For an ideal detector, the noise may be approximated as Gaussian and uncorrelated in time, and, as such, its variance decreases proportionally to $1/\Delta t$. We define the standard deviation of a single time sample to be σ_0.

A CMB experiment scans the sky according to some scanning strategy, $p(t) = [l(t), b(t), \psi(t)]$, while continuously recording the signal d_t. To describe this behaviour in a convenient notation, we first discretize the sky as described in Sect. 3.1, $\boldsymbol{s} = \boldsymbol{s}_p$, and then re-write Eq. (34) in vector form as follows,

$$\boldsymbol{d} = \mathsf{G}\,\mathsf{P}\boldsymbol{s} + \boldsymbol{n}, \tag{35}$$

where $\boldsymbol{d} = [d_1, d_2, \ldots, d_{n_{\rm TOD}}]^t$ and $\boldsymbol{n} = [n_1, n_2, \ldots, n_{n_{\rm TOD}}]^t$ are time-domain vectors of length $N_{\rm TOD}$, and G is a diagonal $N_{\rm TOD} \times N_{\rm TOD}$ matrix with g_t on the diagonal. The scanning strategy is encoded in an $N_{\rm TOD} \times 3N_{\rm pix}$ matrix that contains $(1, \cos 2\psi, \sin 2\psi)$ in the columns that correspond to pixel p that happens to be observed at time t, and zero elsewhere, i.e.,

$$\mathsf{P} = \begin{bmatrix} 0 & 1 & 0 & \ldots & 0 & \cos 2\psi_1 & 0 & \ldots & 0 & \sin 2\psi_1 & 0 \\ 1 & 0 & 0 & \ldots & \cos 2\psi_2 & 0 & 0 & \ldots & \sin 2\psi_2 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & \ldots & 0 & 0 & \cos 2\psi_1 & \ldots & 0 & 0 & \sin 2\psi_1 \end{bmatrix}. \tag{36}$$

This matrix is called the pointing matrix.[4] Correspondingly, the sky vector consists of the three pixelized Stokes parameter maps stacked into a single vector, $\boldsymbol{s} = [T_1, \ldots, T_{N_{\rm pix}}, Q_1, \ldots, Q_{N_{\rm pix}}, U_1, \ldots, U_{N_{\rm pix}}]^t$.

Equation (35) describes an ideal instrument that cannot be realized in actual hardware. The remainder of this section is therefore dedicated to generalizing this equation to the point that it actually does serve as a useful model for real-world CMB experiments.

[4] Only the nonzero entries need to be stored in the pointing matrix, and memory requirements are therefore manageable.

4.2. Spectral response, bandpass averaging, and unit conversion

The first generalization we will consider is the assumption of monochromaticity. No real detector can measure a single frequency signal, but it is instead sensitive to a range of frequencies. This sensitivity is described by a so-called *bandpass profile* or *spectral transmission*, $\tau(\nu)$, which quantifies how much of the radiation at a given frequency is actually recorded by the detector. We define τ to be normalized to unity when integrated across all frequencies. Adopting brightness temperature units for all quantities (i.e., τ, $\boldsymbol{d}$, and the monochromatic sky signal, $\boldsymbol{s}(\nu)$), the data model in Eq. (35) generalizes to

$$\boldsymbol{d} = \mathsf{GP} \int \boldsymbol{s}(\nu)\tau(\nu)\,\mathrm{d}\nu + \boldsymbol{n}, \tag{37}$$

after taking into account the bandpass effect.[5]

However, most data sets are not provided in terms of brightness temperature units, but more often in either thermodynamic temperature or intensity units. As described in detail in Planck Collaboration IX (2014), in order to convert from unit convention X_i to unit convention X_j, one must multiply with a unit conversion factor that is given by

$$U_{ij} = \frac{\int \tau(\nu) \frac{dI_\nu}{dX_i}\,\mathrm{d}\nu}{\int \tau(\nu) \frac{dI_\nu}{dX_j}\,\mathrm{d}\nu}, \tag{38}$$

where dI_ν/dX_i is the intensity derivative expressed in unit convention X_i. In particular, the conversion factors from brightness temperature to thermodynamic temperature and intensity units are given by

$$U_{\mathrm{K_{RJ}}\rightarrow\mathrm{K_{CMB}}} = \frac{\int \tau(\nu) \frac{2h\nu^2}{c^2}\,\mathrm{d}\nu}{\int \tau(\nu)\, b'_\nu\,\mathrm{d}\nu} \tag{39}$$

$$U_{\mathrm{K_{RJ}}\rightarrow\mathrm{MJy\,sr^{-1}}} = \frac{\int \tau(\nu) \frac{2h\nu^2}{c^2}\,\mathrm{d}\nu}{\int \tau(\nu) \frac{\nu_c}{\nu}\,\mathrm{d}\nu}, \tag{40}$$

where

$$b'_\nu = \left.\frac{\partial B(T,\nu)}{\partial T}\right|_{T=T_0} \tag{41}$$

is the derivative of the blackbody function with respect to temperature, evaluated at the CMB temperature T_0, and ν_c is an arbitrary selected reference frequency for the channel in question. For other conversions, including to $\mathrm{K\,km\,s^{-1}}$ and the SZ y-parameter, we refer the interested reader to Planck Collaboration IX (2014). Taking into account both bandpass integration and unit conversion, the instrument model reads

$$\boldsymbol{d} = U\mathsf{GP} \int \boldsymbol{s}(\nu)\tau(\nu)\,\mathrm{d}\nu + \boldsymbol{n}. \tag{42}$$

We aim to constrain $\boldsymbol{s}$ given $\boldsymbol{d}$. It is therefore important to be able to quickly evaluate the integral and unit conversion factors in Eq. (42). With this in mind, we consider signal component i as defined by the sky model in Sect. 3.6, and write it in the general form $\boldsymbol{s}_i(\nu) = \boldsymbol{a}_i\, f_i(\nu;\nu_0,\beta)$, where $\boldsymbol{a}_i$ is the linear amplitude relative to some reference frequency, $\nu_{0,i}$, and $f_i(\nu;\beta)$ is the frequency scaling from that reference frequency to an arbitrary frequency ν, which depends on some set of spectral parameters β. The total signal measured by detector j may then be written as

$$\boldsymbol{s}^j = \sum_{i=1}^{N_{\mathrm{comp}}} \boldsymbol{a}_i \left[U_j \int f_i(\nu;\beta)\,\tau_j(\nu)\,\mathrm{d}\nu \right] \equiv \sum_{i=1}^{N_{\mathrm{comp}}} \mathsf{M}_i^j\,\boldsymbol{a}_i = \mathsf{M}^j\,\boldsymbol{a}, \tag{43}$$

where M_i^j is called the *mixing matrix*. In order to take into account bandpass integration and unit conversion, the idealized data model in Eq. (35) must be generalized as follows,

$$\boldsymbol{d} = \mathsf{GPM}\boldsymbol{a} + \boldsymbol{n}. \tag{44}$$

It is evident that M depends only on the spectral parameters β and the bandpass τ, but not the amplitudes. Since most signal components are parameterized with limited number of spectral parameters (see Sect. 3), and these parameters are typically also fairly uniform on the sky, it is possible to pre-compute accurate lookup tables for M for each component and detector. In our current code, we adopt (bi-)cubic splines with regular grids for these lookup tables, and the computational cost of performing a full bandpass integral is thus equal to that of a simple polynomial evaluation.

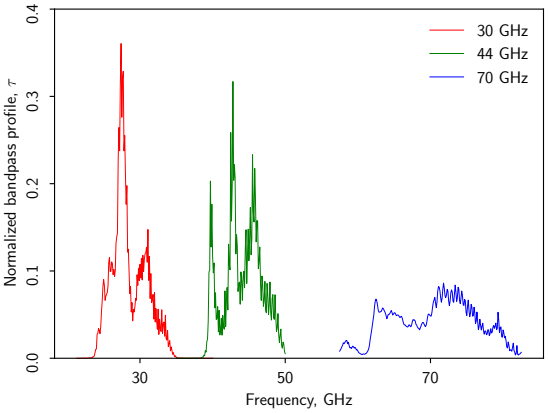

Fig. 2. Detector averaged bandpass profiles, τ, for the three *Planck* LFI frequency channels.

4.2.1. Bandpass uncertainties and corrections

While the bandpass integral described by Eq. (37) may look simple enough at first glance, it does lead to a wide variety of important complications in practice. The most important among these is the fact that the exact shape of the bandpass profile itself is unknown. In particular, it is highly nontrivial to measure τ accurately in a laboratory for a combined multi-component instrument, and it is obviously impossible to do so after commissioning for satellite missions.

As a concrete real-world illustration of this, Fig. 2 shows the laboratory-determined (normalized) bandpass profiles after averaging over all radiometers for a given LFI channel. First, we see that the profiles for both 44 and 70 GHz are truncated, and therefore significant response is likely present outside the measured range. Second, for all three channels we see notable small

[5] Note that many experiments, including *Planck* HFI, defines the bandpass profile in intensity units rather than brightness temperature units, and in this case an additional factor of $2h\nu^2/c^2$ must be included in the integral, as given by Eq. (9); see Planck Collaboration IX (2014) for details.

scale ripples, which are due to standing waves. These may be due to real standing waves within the optical assembly of the LFI instrument itself; but some part of them may also be due to standing waves in the test equipment used to make the measurements. In addition to these two visually obvious effects, there may also be systematic errors in the actual shape, for instance in the form of a smooth slope across the bands, or in the precise position of the peaks within the band.

As described in Sect. 8.3.1, the CMB dipole serves as our primary calibrator for BeyondPlanck, following both *WMAP* and the official *Planck* pipelines. Because the CMB SED very closely follows a blackbody spectrum, which translates into a frequency independent scaling in thermodynamic units, the precise shape of the bandpass is irrelevant for the CMB component. Instead, errors in the bandpass shape effectively translate into incorrectly estimated foreground components, and introduce inaccuracies in the relative foreground SEDs between different frequency channels. In turn, foreground errors can affect the CMB reconstruction.

To account for the uncertainties noted above, we introduce one or more free parameters that can modify the bandpass shape, and allow the data to inform us about, and hence mitigate, potential inaccuracies in the laboratory bandpass measurements. The simplest and most common model we adopt is a simple linear shift, $\Delta_{\rm bp}$, in frequency space,

$$\tau(\nu) = \tau_0(\nu + \Delta_{\rm bp}), \tag{45}$$

where τ_0 is the default laboratory measurement. One value of $\Delta^i_{\rm bp}$ is allowed per radiometer i, but (in most cases) either with the prior that $\sum_i \Delta^i_{\rm bp} = 0$, or that one particular channel is held fixed. Without any priors, the bandpass parameters are fully degenerate with the spectral parameters β of the foreground model, and no stable solution can be found. Various choices of both bandpass models and priors are considered by Svalheim et al. (2020a). In general, we note that the impact of $\Delta_{\rm bp}$ is essentially to scale the amplitude of foregrounds, while leaving the CMB unchanged. At CMB dominated frequency channels, the bandpass shift is therefore non-degenerate with respect to the gain, while at foreground-dominated channels, it is virtually impossible to distinguish between a bandpass error and a gain error.

In addition to this fundamental uncertainty in the bandpass profile for each detector, we note, first, that different detectors within the same frequency band observe different sky signals, and if not properly accounted for, this can create so-called bandpass mismatch errors in co-added frequency maps (see Sect. 7.2.2). Second, as discussed in the next section, the instrumental beam is also intrinsically frequency dependent, with an angular resolution of the main beam that is inversely proportional to the frequency for diffraction-limited observations, as is the case for LFI. In addition, far sidelobes can vary rapidly with frequency through complicated diffraction patterns. Unless properly accounted for, all these effects can potentially compromise final estimates. In BeyondPlanck we account for sidelobes as modelled by the *Planck*team (Planck Collaboration IV 2016), but we do not explore uncertainties in the beam model itself.

4.3. Beam and pixel window convolution

In the same way that no real detector can measure the signal from only a single monochromatic frequency, no real detector can measure the signal from a single point on the sky. Rather, each detector is associated with a so-called "point spread function" (PSF) or "beam", $\boldsymbol{b}(\hat{n})$, that characterizes its instantaneous

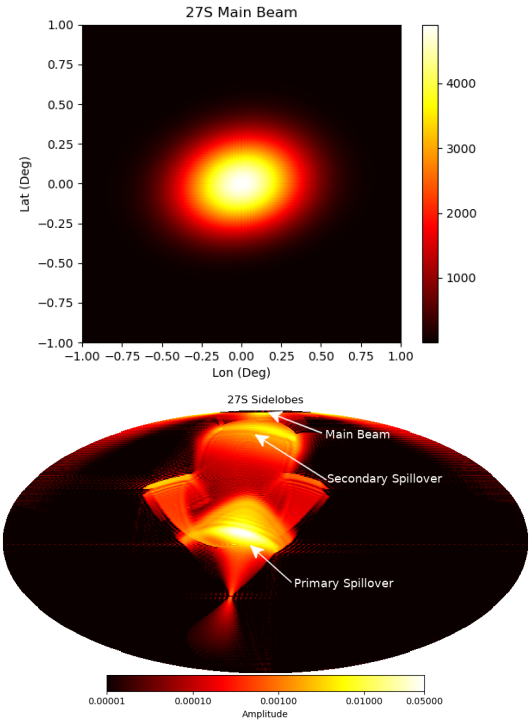

Fig. 3. (*Top*:) Scanning beam (or main beam) of the 30 GHz LFI 27S radiometer in local telescope coordinates, i.e., the instantaneous spatial sensitivity sensitivity to a point source centered at the beam maximum. *Bottom*: Corresponding 4π beam map, oriented such that the main beam is located on the north pole. The main *Planck* far sidelobes are caused by spillover from (i.e., diffraction around) the primary and secondary mirrors. The beams are normalized such that their combined integral over the full sky equals unity.

spatial sensitivity. Conventionally, we define $\boldsymbol{b}(\hat{n})$ to be normalized by setting the full-sky integral equal to unity, and to be oriented such that its maximum value is centered on the north pole.

The recorded value of the sky signal, $s_t^{\rm beam}$, as seen through the beam at time t is then given as the convolution of $\boldsymbol{b}$ and $\boldsymbol{s}$,

$$s_t^{\rm beam} = \int_{4\pi} \boldsymbol{b}_t(\hat{n})\boldsymbol{s}(\hat{n})\,{\rm d}\Omega, \tag{46}$$

where $\boldsymbol{b}_t(\hat{n}) = \mathsf{R}_t(\hat{n}, \hat{n}')\boldsymbol{b}(\hat{n}')$, and R_t is a time-dependent rotation matrix that rotates the beam as given by the scanning strategy of the instrument. Since convolution is a linear operation, we may define a matrix operator, B, such that $\boldsymbol{s}^{\rm beam} = \mathsf{B}\boldsymbol{s}$, and the data model in Eq. (44) may therefore be generalized further into its final form,

$$\boldsymbol{d} = \mathsf{GPBM}\boldsymbol{a} + \boldsymbol{n}, \tag{47}$$

where the position of the operator is defined by noting that the beam only acts on the true sky signal, and not on instrumental effects such as gain or noise.

Noting that modern CMB maps typically comprise up to several hundred million pixels, Eq. (46) is prohibitively expen-

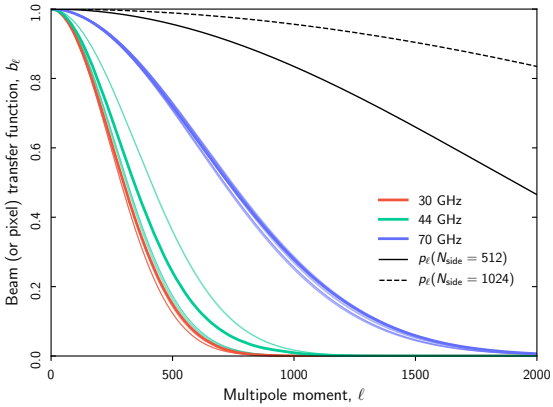

Fig. 4. Azimuthally symmetric and normalized beam transfer functions, b_ℓ for each co-added *Planck* LFI frequency channel (thick colored lines) and for each radiometer (thin colored lines). The former include the effects of non-Gaussian tails, while the latter are strictly Gaussian approximations. Black curves show the `HEALPix` pixel window functions for $N_{\rm side} = 512$ and 1024, respectively.

sive to evaluate directly in pixel space. Instead, we take advantage of the convolution theorem, which states that any convolution in pixel space may be expressed as a multiplication in harmonic space, and vice versa. As first demonstrated by Wandelt & Górski (2001), and later optimized by Prézeau & Reinecke (2010), Eq. (46) may be computed efficiently through reduced Wigner matrices, reducing the cost by a factor of $O(\sqrt{N_{\rm pix}})$ per evaluation for a general $\boldsymbol{b}$.

Another substantial saving can be made if we additionally assume that $\boldsymbol{b}$ is azimuthally symmetric. In that case, the spherical harmonics expansion of $\boldsymbol{b}$ is independent of m, and may be expressed in terms of its Legendre transform, b_ℓ. The full convolution may (by the convolution theorem) in this case be written as[6]

$$s^{\rm beam}(\hat{n}) = \sum_{\ell=0}^{\ell_{\rm max}} \sum_{m=-\ell}^{\ell} b_\ell s_{\ell m} Y_{\ell m}(\hat{n}), \tag{48}$$

where $s_{\ell m}$ are the spherical harmonics coefficients of $\boldsymbol{s}$. Often, b_ℓ is referred to as the beam transfer function.

Note that the bandlimit, $\ell_{\rm max}$, in Eq. (48) should be selected sufficiently large that $b_\ell \approx 0$ as compared to the noise level of the instrument. Conversely, if a too low value of $\ell_{\rm max}$ is adopted for analysis, the most notable artifacts arising from the convolution is ringing around bright point sources, resulting from premature harmonics truncation.

Note also that $s^{\rm beam}(\hat{n})$ in Eq. (48) is written as a function of position rather than time in the above expression, which is only possible in the case of an azimuthally symmetric beam. To obtain the time-dependent signal, one simply reads off the value of $s^{\rm beam}(\hat{n})$ given by the beam center position at time t. In this approximation, a full real-space convolution may be carried out at the cost of only two spherical harmonics transforms.

As discussed in Sect. 3.1, all CMB sky maps are pixelized in order to allow for efficient analysis on a computer. Such pixelization corresponds to an additional smoothing operation of the true sky signal that can be approximated with a top-hat convolution kernel of a given pixel size. For `HEALPix`, the effect of this kernel in harmonic space is described in terms of a pixel window function, p_ℓ, that is provided with the library. Implementationally, it is often convenient to redefine $b_\ell \rightarrow b_\ell p_\ell$ internally in computer codes, as the beam and pixel window affect the signal in the same way, and accounting for the pixel window can therefore usually be done with no additional computational cost compared to beam convolution.

In Euclidean space, the Nyquist theorem assures that any bandwidth limited signal may be reconstructed exactly with at least two samples per bandwidth. No corresponding exact theorem exists on the sphere. Instead, a rough rule of thumb for smooth spherical fields is to allow for at least two or three pixels per beam width. Likewise, no exact multipole bandlimit exists for given a `HEALPix` pixelization; however, numerical experiments suggest that multipoles above $\ell \gtrsim 2.5 N_{\rm side}$ are poorly resolved on the `HEALPix` grid. Combined, these rules of thumb provide useful relationships between a given beam width and the corresponding appropriate values of $N_{\rm side}$ and $\ell_{\rm max}$.

Figure 3 shows the beam of the *Planck* 27S radiometer (Planck Collaboration IV 2014). The bottom panel shows the full 4π beam, while the top panel shows a zoom-in on the north pole. Clearly, this beam pattern is not azimuthally symmetric. However, in this respect it is useful to distinguish between the *main beam*, which is highlighted in the top panel, and the *sidelobes*, which are highlighted in the bottom panel. Furthermore, since convolution is a linear operation, contributions from the main beam and sidelobes may be computed separately.

The sidelobes are caused by optical imperfections, typically by diffraction around the main optical elements. In the case of *Planck*, these are the primary and secondary mirrors (see Fig. 3). As such, the resulting beam structures tend to be highly frequency dependent, and also cover large angular scales. While they clearly cannot be described as azimuthally symmetric in any meaningful way, they are associated with relatively modest bandlimits, $\ell_{\rm max}$, and this leads to acceptable computational costs for treating this component.

The main beam, on the other hand, can often be described reasonably well as azimuthally symmetric, when centered on the north (or south) pole. Of course, the LFI 27S beam shown in the top panel of Fig. 3 exhibits a substantial ellipticity of $\epsilon \approx 1.3$, but this instantaneous beam profile is at least partially symmetrized by averaging due to the scanning strategy. The remaining effects of beam asymmetries may be accounted for, at least in terms of power spectrum bias, by adjusting the transfer function b_ℓ through simulations, as described by, e.g., Mitra et al. (2011).

For simplicity or because of low signal-to-noise, the beam profile is also sometimes approximated in terms of a two-dimensional Gaussian with some full-width-half-maximum (FWHM), or $\sigma_{\rm FWHM}$, in the following expressed in radians. In the Gaussian case, one can derive an explicit expression for the the beam transfer function in the form

$$b_\ell = \mathrm{e}^{-\frac{1}{2}\ell(\ell+1)\frac{\sigma^2_{\rm FWHM}}{8\ln 2}}, \tag{49}$$

where the factor $8\ln 2$ simply accounts for the conversion between the square of the FWHM and the variance for a Gaussian.

Figure 4 compares the azimuthally symmetric beam transfer functions of the three *Planck* LFI channels, co-added over all radiometers, as well as the Gaussian approximations to the individual radiometer beam transfer functions. For reference, we also show the `HEALPix` window transfer functions for $N_{\rm side} = 512$ and 1024, which are the typical pixelizations used for LFI and *WMAP* analysis.

[6] This expression applies to temperature convolution; polarization convolution is notationally slightly more involved, but mathematically fully analogous.

We see that the general azimuthal approximations tend to have slightly heavier tails than the Gaussian approximations, and this is important to account for when estimating the CMB power spectrum, C_ℓ. At the same time, we also see that for applications for which only percent-level accuracy is required, the Gaussian approximations may very well be sufficient. In the following analyses, we will adopt the general azimuthally symmetric approximations for co-added frequency maps, which will be used for component separation and CMB estimation purposes, but Gaussian approximations for radiometer-specific signal modelling during time-domain processing, where the signal-to-noise ratio per sample is low, and sub-percent precision is irrelevant. The reason for the latter approximation is simply that the *Planck* collaboration only provides FWHM estimates for individual radiometers, not full transfer functions.

In the current work, we assume that the transfer functions provided by the *Planck* collaboration are exact, and do not assign dedicated stochastic parameters to them. This is neither a realistic description, nor a testament to the accuracy of the provided products, but only a statement of currently limited human resources; a high-priority task for future work is to implement full support for dynamic beam modelling and error propagation. As presented in this work, however, beam convolution is assumed to be a fully deterministic operation, dependent on officially available beam characterizations alone.

4.4. Gain and analog-to-digital conversion

While the instrument model in Eq. (47) is structurally complete in terms of components, we still need to introduce a few generalizations before we can apply it to our data. The first regards the gain g, simply by reemphasizing that this should be interpreted as a truly time-dependent object, g_t.

To understand why this is the case, it is useful to consider its origin and physical interpretation, and to focus the discussion we will consider the special case of a perfect total-power receiver. The output voltage of such a device is given by

$$P = GkT_{\rm sys}\Delta\nu, \tag{50}$$

where G is a unit-less gain factor, and $\Delta\nu$ is the width of the bandpass. The system temperature is defined as $T_{\rm sys} = T_{\rm ant} + T_{\rm recv}$, where $T_{\rm ant} = T_{\rm CMB} + T_{\rm fg}$ is the antenna temperature, and $T_{\rm recv}$ is the receiver temperature; the latter essentially defines the intrinsic noise level of the receiver.

For a *Planck* LFI 30 GHz radiometer, the bandwidth is 6 GHz, and the receiver temperature is typically 10 K. The antenna temperature is dominated by the CMB temperature, $T_{\rm CMB} = 2.7$ K, as other sky components typically only make up a few mK at most. Assuming, therefore, a system temperature of about 13 K, Eq. (50) predicts that the power measured by this device is $P = 1.1$ pW or $P = -90$ dBm,[7] assuming no amplification ($G = 1$). However, current microwave detectors are typically only able to reliably record power levels larger than $P \gtrsim -30$ dBm. For this reason, the signal level must be actively amplified by a factor of 60 dB or more between the optical assembly and the detector. For *Planck* LFI, such amplification is achieved through the use of high-electron-mobility transistors (HEMTs).

HEMTs provide high gain factors, while adding only very low levels of additional noise to the data. However, they are not perfectly stable in time. Rather, their effective gains exhibit time-dependent drifts with typical overall variations at the $O(10^{-6})$ level, and correlations in time that are often well described by a so-called $1/f$ spectrum (see Sect. 4.5). Unless explicitly accounted for in the model, these time-dependent gain fluctuations can and will bias the derived sky model.

The gain defined by our original instrument model in Eq. (47), denoted G, is in principle the same gain as in Eq. (50), but with two important differences. First, while G is defined as a pure power amplification, and therefore unit-less, G takes into account the end-to-end conversion from a raw sky signal to final recorded data values. As such, G has units of V K^{-1}, in order to be dimensionally correct.

Second, G additionally takes into account the digitization process that converts analog signals to digital bits stored on a computer. This process takes place in a so-called analog-to-digital converter (ADC). An ideal ADC is perfectly linear. Unfortunately, many real-world ADCs exhibit important imperfections, for instance in the form of smooth nonlinear conversion within given signal ranges, or, as for LFI, sharp jumps at specific signal or bit values.

Overall, ADC errors are indistinguishable from gain fluctuations in terms of their direct impact on the recorded data. However, there is one critical difference between the two effects: While gain fluctuations are stochastic and random in time, and do not correlate with the sky signal, ADC errors are perfectly reproducible, and depend directly on the sky signal. Consequently, while the archetypical signature of unmitigated gain fluctuations are coherent stripes or large-scale features in the final sky maps, the corresponding unique signature of unmitigated ADC errors is an asymmetry in the amplitude of the CMB dipole along its positive and negative directions. This effect can be used to characterize and mitigate ADC non-linearity, as done both for *Planck* LFI and HFI (Planck Collaboration II 2020; Planck Collaboration III 2020; Planck Collaboration Int. LVII 2020).

4.5. Instrumental noise

We complete our review of the instrument model by considering the properties of the instrumental noise, $\boldsymbol{n}$. This component may be decomposed into two main contributions, called correlated and white noise,

$$\boldsymbol{n} = \boldsymbol{n}_{\rm corr} + \boldsymbol{n}_{\rm wn}. \tag{51}$$

Both terms may be approximated as Gaussian, but they have different covariances.

The dominant physical source of white noise is Johnson (or thermal) noise, typically excited by thermal electron motions within the electric radiometer circuits. This noise is temperature dependent, and cryogenic cooling is usually required to achieve sufficient sensitivity. The dominant source of the correlated noise term are rapid gain fluctuations modulating the system temperature, $T_{\rm sys}$, as discussed in Sect. 8.3.1.

Based on this decomposition, the standard deviation of the total instrumental noise term for a sample of duration Δt (i.e., σ_0 in Eq. (35)) may be estimated through the so-called radiometer equation,

$$\sigma_0 = T_{\rm sys}\sqrt{\frac{1}{\Delta\nu\Delta t} + \left(\frac{\Delta g}{g}\right)^2}. \tag{52}$$

Here, Δg is the root-mean-square gain variation over Δt, and $\Delta\nu$ is as usual the receiver bandwidth. Intuitively speaking, this

[7] The unit dBm measures power ratios, x, in decibel relative to 1 mW, i.e., $x = 10\,\log_{10}\frac{P}{1\,\mathrm{mW}}$.

equation summarizes the following facts. First, the noise level is proportional to the system temperature, in recognition of the fact that Johnson noise scales with temperature. Second, the white noise term is inversely proportional to the square root of both bandwidth and integration time; this is simply by virtue of collecting more photons, and noting that Gaussian errors add in quadrature. Third and finally, the correlated noise component is proportional to the overall gain fluctuation level. Typical values of σ_0 for the LFI radiometers range between 600 and 1700 μK per sample in temperature units, or between 50 and 200 μV in detector units. If $N_{\rm obs}$ independent observations are made of the same sky pixel p, then the effective noise of the corresponding pixel integrates down roughly as $\sigma_p = \sigma_0/\sqrt{N_{\rm obs}}$.

The different correlation structures of the white and correlated noise terms are most conveniently described in frequency domain through the noise power spectrum density (PSD), $P_n(f) = \left\langle |n_f|^2 \right\rangle$, where n_f are the Fourier coefficients of $\boldsymbol{n}_t$. This PSD is often modelled in terms of a so-called $1/f$ profile, which takes the form

$$P_n(f) = \sigma_0^2 \left[1 + \left(\frac{f}{f_{\rm k}} \right)^{\alpha} \right]. \tag{53}$$

Here, $f_{\rm k}$ is the knee frequency at which the variance of the correlated noise equals that of the white noise, and α is the slope of the spectrum at low frequencies. Typical best-fit values for LFI radiometers are $f_{\rm k} \approx 10$ mHz and $\alpha \approx -1$. However, this model is obviously only approximate; if for no other reasons, the real spectrum has to flatten at low frequencies by energy considerations, whereas the power predicted by this model would approach infinity at low frequencies.

5. Data

The instrument discussion has until this point been general, and applicable to a wide range of different data sets. In this section, we specialize our discussion to one particular combination of data sets, with *Planck* LFI being the primary target of interest. As discussed in Sect. 2, only this data set will be considered in the time-domain, while external data sets will be considered in the form of processed pixelized maps.

We note that the minimal sky model summarized in Sect. 3.6 includes seven distinct astrophysical components, three polarized and four unpolarized. Considering that there are only three LFI frequency channels, we immediately recognize that the LFI data must be augmented with at least four external frequency channels, just in order to make the model minimally constrained. In the default analysis configuration, we therefore include select observations also from *Planck* HFI (Planck Collaboration Int. LVII 2020) and *WMAP* (Bennett et al. 2013), as well as from some ground-based surveys. In this section, we provide a brief overview of these data sets, and refer the interested reader to the respective papers for full details.

The precise combination of data sets used in any particular BeyondPlanck analysis will depend on the goal of the respective application. For instance, the main scientific goal of the current paper is to introduce the concept of Bayesian end-to-end CMB analysis, and provide a first demonstration of this framework as applied to the LFI observations. Consequently, we here only include a minimal set of external observations, allowing LFI to play the dominant role, in particular with respect to CMB constraints. Specifically, in this paper we include only

- *Planck* 857 GHz to constrain thermal dust emission in intensity;
- *Planck* 353 GHz in polarization to constrain polarized thermal dust emission;
- *WMAP* 33, 41, and 61 GHz (called *Ka*, *Q* and *V*-bands, respectively) in intensity at full angular resolution to constrain free-free emission and AME;
- the same *WMAP* channels in polarization to increase the signal-to-noise ratio of polarized synchrotron emission, but only at low angular resolution, where a full noise covariance matrix is available; and
- Haslam 408 MHz (Haslam et al. 1982) to constrain synchrotron emission in intensity.

That is, we include neither intermediate HFI channels nor the *WMAP* *K*-band (23 GHz) channel, because of their higher signal-to-noise ratio relative to the LFI channels. The *WMAP* *W*-band is excluded because of known systematics effects (Bennett et al. 2013), and it does not have particularly unique features with respect to the signal model that are not already covered by other data sets.

We also note that Andersen et al. (2020), Svalheim et al. (2020b), and Herman et al. (2020) focus on general foreground constraints, and these papers therefore also consider additional channels. The ultimate long-term goal of the global Bayesian CMB analysis program in general is of course to integrate as many data sets as possible into a single coherent sky model, and thereby produce the strongest possible constraints on the true astrophysical sky. One leading example of such an effort is the Cosmoglobe[8] project, which specifically aims to combine many state-of-the-art experiments with the ones listed above, including *Planck* HFI (Planck Collaboration III 2020), *COBE*-DIRBE (Hauser et al. 1998) and FIRAS (Mather et al. 1994), PASIPHAE (Tassis et al. 2018), SPIDER (Gualtieri et al. 2018), and many more. The BeyondPlanck methodology presented here represents an ideal statistical framework for performing such global data integration.

5.1. LFI instrument overview

5.1.1. Instrument configuration

We now provide a synthetic description of the LFI instrument configuration, which directly impacts the structure of the LFI data and the potential systematic effects addressed in the BeyondPlanck analysis. For more details on the LFI instrument, its ground calibration and in-flight performance, see Bersanelli et al. (2010), Mennella et al. (2011), and references therein; the overall LFI programme is described by Mandolesi et al. (2010).

The heart of the LFI instrument is an array of 22 differential receivers based on high-electron-mobility transistor (HEMT) low noise amplifiers. The instrument operates in three frequency bands, nominally centred at 30, 44 and 70 GHz, with angular resolutions of about 32′, 28′, and 13′ FWHM, respectively. The front end of the receivers is cooled to 20 K, which dramatically reduces the noise temperature of the HEMT amplifiers and of the overall system. In each receiver, the signal coming from different directions of the sky, intercepted by the telescope as the satellite spins, is compared to a stable internal blackbody reference load at 4 K. It is this differential scheme that allows the LFI to achieve its excellent stability.

Radiation from the sky is coupled to 11 corrugated feed horns, shown in Fig. 6. Each horn is followed by an orthomode transducer (OMT), which splits the incoming radiation into two

[8] http://cosmoglobe.uio.no

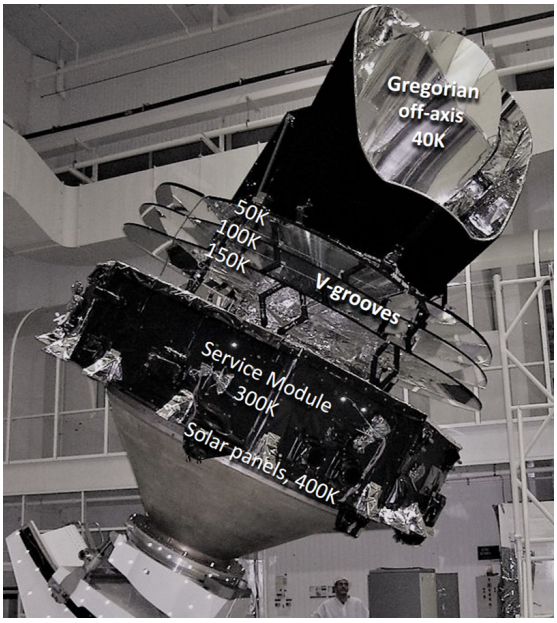

Fig. 5. Flight model of the *Planck* spacecraft. The satellite size is about 4.2×4.2 m, and its mass at launch was 1950 kg. *Planck* was launched on May 14, 2009, and operated for 4.4 years from a Lissajous orbit around the Lagrangian point L2 of the Sun–Earth system. Shown are the approximate temperatures of different critical parts of the satellite during nominal operation in space (see Planck Collaboration II 2011).

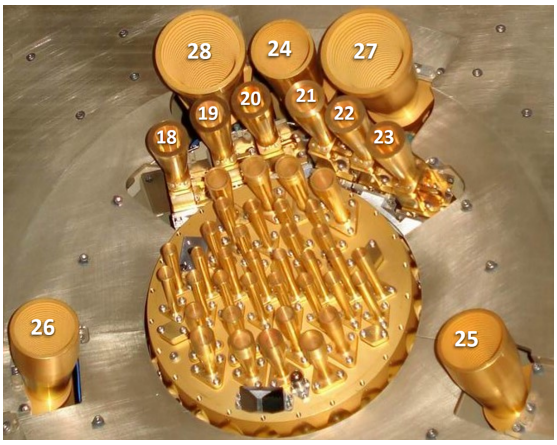

Fig. 6. Top view of the *Planck* focal plane. The central array contains the HFI feed-horns, cooled to 4 K, feeding bolometric detectors cooled to 0.1 K. The LFI horns in the outer part of the array are labelled with numbers; they are cooled to 20 K. The LFI horn numbers 18–23, 24–26, and 27–28 correspond to the 70 GHz, 44 GHz, and 30 GHz channels, respectively.

perpendicular linear polarizations that propagate through two independent differential radiometers; see Fig. 7. The OMT provides exquisite polarization purity, with typical isolation of < -30 dB. Each radiometer pair has a front-end module (FEM), cooled to 20 K, and a back-end module (BEM), operated at 300 K. The FEM is connected to the BEM by four composite wave-guides (two for each radiometer), thermally coupled to the three *Planck* V-groove radiators to minimize parasitic heat transfer to the cold focal plane (see Fig. 5). The cryogenically cooled front-end modules include the first stage HEMT amplifiers and the differencing system, while the back-end modules provide further radio frequency amplification. Detection is made via two square-law detector diodes for each radiometer.

After detection, an analog circuit in the data acquisition electronics is used to adjust the offset to obtain a nearly null DC output voltage, and a programmable gain is applied on-board to match the signal level to the analog-to-digital converter (ADC) input range. After the ADC, data are digitally down-sampled, re-sampled to match beam resolution (> 3 samples per beam), compressed, and assembled into telemetry packets, which are then downlinked to the ground station.

5.1.2. Stabilization

Cryogenic HEMT amplifiers exhibit excellent low-noise performance, but are affected by significant instability in terms of gain and noise-temperature fluctuations, typically modelled in terms of a $1/f$ spectrum as discussed in Sect. 4.5. The LFI system is designed to efficiently reject such fluctuations in the radiometer response. The main differential process responsible for radiometer stabilization takes place in the front-end modules. The signals from the sky and 4 K reference load are injected into a hybrid coupler, which splits the two signals, and redirects them to both of its output ports (see inset of Fig. 7). Then the two mixed signals are amplified by ~30 dB by the two amplifier chains. Thus, any fluctuation in the FEM amplifiers affects both the sky and the reference load components in exactly the same way. After amplification, a second hybrid coupler reconstructs the sky and reference components, which now contain the same fluctuations. Then the signals are transported by the wave-guides in the warm back-end modules, where they are further amplified and detected by the diodes. Finally, when taking the difference between the two diodes, the FEM fluctuations cancel out. This "pseudo-correlation" scheme reduces front-end fluctuations by a factor of $O(10^3)$.

However, instabilities downstream of the FEMs, particularly those originating in the back-end amplifiers and in the detector diodes, would still affect the measurements. For this reason, a further level of stabilization is built into the LFI design. A phase shifter, alternating between 0° and 180° at a frequency of 4096 Hz, is applied in one of the two amplification chains within the front-end modules, as shown in Fig. 7. In this way, the DC output from each diode rapidly alternates the sky and reference signals, with opposite phase in the two detectors. By taking the difference between time-averaged values of sky and reference, any residual fluctuations on time scales longer than $\sim (1/4096)$ s $= 0.244$ ms are removed.

Of course, any non-ideality in the receiver components will introduce some level of residual fluctuations. Further strategies to suppress remaining instabilities and potential systematics introduced by the receiver are described below.

5.1.3. LFI signal model

Based on the above description, the differential power output for each of the four diodes associated with a feedhorn can be written

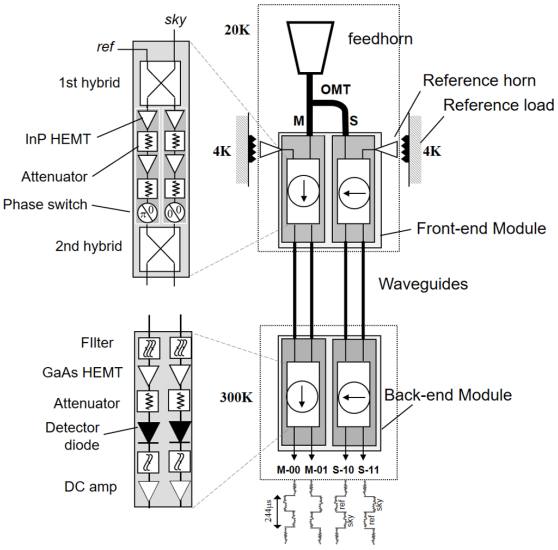

Fig. 7. Schematic of an LFI radiometer chain, consisting of a feedhorn and OMT, and the two associated radiometers, each feeding two diode detectors. The insets show details of a front-end module (top) and of a back end module (bottom).

as the following special case of Eq. (50),

$$P_{\mathrm{out},0}^{\mathrm{diode}} = a\, G_{\mathrm{tot}}\, k\, \Delta\nu \left[T_{\mathrm{sky}} + T_{\mathrm{noise}} - r\,(T_{\mathrm{ref}} + T_{\mathrm{noise}})\right], \tag{54}$$

where G_{tot} is the total gain, k is the Boltzmann constant, $\Delta\nu$ the receiver bandwidth, and a is the diode constant. T_{sky} and T_{ref} are the sky and reference load antenna temperatures at the inputs of the first hybrid and T_{noise} is the receiver noise temperature, averaged over an appropriate integration time. The gain modulation factor, r, is a factor of order unity ($0.8 < r < 1.0$ depending on channel) used to balance in software the temperature offset between the sky and reference load signals. This has the important effect of minimising the residual $1/f$ noise and other non-idealities in the differential data stream. In the DPC analysis (Planck Collaboration II 2020) r was determined using the approximation

$$r = \frac{T_{\mathrm{sky}} + T_{\mathrm{noise}}}{T_{\mathrm{ref}} + T_{\mathrm{noise}}} \approx \frac{V_{\mathrm{sky}}}{V_{\mathrm{ref}}}, \tag{55}$$

and we adopt the same procedure without modification in BeyondPlanck for now. However, we do note that the NPIPE analysis pipeline implements an alternative approach in which V_{ref} is low-pass filtered prior to differencing, and this reduces the amount of high frequency noise in the final maps. Future BeyondPlanck versions can and should implement a similar solution.

Although somewhat counter-intuitive, the sensitivity of the LFI radiometers does not depend significantly on the absolute temperature of the reference load. In fact, to first order, the white noise spectral density at the output of each diode is given by

$$\Delta T_0^{\mathrm{diode}} = \frac{2\,(T_{\mathrm{sky}} + T_{\mathrm{noise}})}{\sqrt{\Delta\nu}}. \tag{56}$$

However, a large imbalance between T_{sky} and T_{ref} would have the effect of amplifying residual fluctuations in the differential signal. For this reason the LFI reference loads are cooled to about 4 K, exploiting the HFI pre-cooling stage.

The above description holds for the ideal case where all front-end amplifiers and phase switches have perfectly balanced properties. In presence of some level of mismatch, the separation of the sky and reference load signals after the second hybrid is not perfect and the outputs are slightly mixed. If the front-end imbalance is small, Eq. (56) may be written as

$$\left(\Delta T^{\mathrm{diode}}\right)^2 \approx \left(\Delta T_0^{\mathrm{diode}}\right)^2 \left(1 \pm \frac{\epsilon_{A_1} - \epsilon_{A_2}}{2} + \alpha\epsilon_{T_{\mathrm{n}}}\right), \tag{57}$$

where $\epsilon_{T_{\mathrm{n}}}$ is the imbalance in front end noise temperature between the two radiometer arms, and ϵ_{A_1} and ϵ_{A_2} are the imbalance in signal attenuation in the two states of the phase switch. Eq. (57) shows that the output is identical for the two diodes apart from the sign of the term $(\epsilon_{A_1} - \epsilon_{A_2})/2$, representing the phase switch amplitude imbalance. For this reason, the LFI scientific data streams are obtained by averaging the voltage outputs from the two diodes in each radiometer,

$$V_{\mathrm{out}}^{\mathrm{rad}} = w_1 V_{\mathrm{out}}^{\mathrm{diode}\,1} + w_2 V_{\mathrm{out}}^{\mathrm{diode}\,2}, \tag{58}$$

where w_1 and w_2 are inverse-variance weights calculated from the data. Thus, the diode-diode anti-correlation is cancelled, and the radiometer white noise becomes

$$\Delta T^{\mathrm{rad}} \approx \frac{\Delta T_0^{\mathrm{diode}}}{\sqrt{2}} \left(1 + \alpha\epsilon_{T_{\mathrm{n}}}\right)^{1/2}. \tag{59}$$

In Eqs. (57) and (59), $\epsilon \ll 1$, while α is a term of order unity defined by a combination of the input signals and noise temperature of the radiometer; for details, see Eq. (8) in Mennella et al. (2011).

In the current BeyondPlanck processing, we follow the LFI DPC procedure for all these steps. Future versions of the framework may also account for these pre-processing steps, and jointly estimate r, α, ϵ_i, and w_i, but this is left for future work, simply due to the strong time limitations of the current project (see Sect. 1.6).

5.1.4. Naming convention

As described in the previous section, LFI has 11 horns and associated OMTs, FEMs and BEMs; 22 radiometers (two for each horn); and a total of 44 detectors (two for each radiometer). For historical reasons, the 11 horns are labelled by numbers from 18 to 28 as shown in Fig. 6.

The radiometers associated with each horn are labelled as "M" or "S" depending on the arm of the OMT they are connected to ("Main" or "Side", as shown in Fig. 7). Each radiometer has two output diodes that are labelled with binary codes "00", "01" (radiometer M) and "10", "11" (radiometer S), so that the four outputs of each radiometer pair can be named with the following sequence; M-00, M-01, S-10, S-11.

As the telescope scans, the observed region of the sky sweeps across the focal plane in the horizontal direction as appearing in Fig. 6. Since the reconstruction of the polarization information requires at least two horns, every pair of horns aligned in the scan direction are oriented such that their linear polarizations are rotated by 45° from each other (with the exception of LFI-24, which is an unpaired 44 GHz horn). Thus, LFI can produce independent polarization measurements from the "horn pairs" 18–23, 19–22, 20–21 (at 70 GHz); 25–26 (at 44 GHz); and 27–28 (at 30 GHz).

5.2. Implementation details

Since the BeyondPlanck project aims to establish an open-source, reproducible and externally extendable analysis framework, it is no longer possible to rely on direct access to the existing LFI-DPC database, which both employs proprietary software and runs on one specific computer. To circumvent this issue, we convert the LFI TOD into a convenient HDF5 format (Galloway et al. 2020a) that may be accessed using publicly available tools. This, however, does lead to some adjustments in the scientific pre-processing pipeline, which now uses this new interface. At the same time, we have converted the scientific pipeline to C++11, and a number of optimizations are applied at the same time, exploiting the new possibilities given by that language.

5.2.1. Unprocessed Level-1 data

The extraction of time-ordered Level-1 data from the LFI-DPC database and the conversion to HDF5 format only need to be performed once in the LFI-DPC environment. We create one file for each LFI horn for each Operational Day, i.e., the time between two consecutive daily telecommunication periods. The extracted file contains sky, reference load and quality flags for each of the diodes of the horn and timing information, including On-Board Time, Spacecraft Event Time (SCET) and Modified Julian Date (MJD). It also contains attitude information that is critical for the analysis; Pointing Period ID (PID); start and end time of each Pointing Period; end time of the maneuver of each Pointing Period; and number of data samples.

To optimize the computational time of Level-2 processing, various deterministic operations are implemented during extraction. For instance, missing data are added back into the time streams and flagged as bad data; this ensures that all the timelines for each frequency are of the same length. Also, planet transits are flagged, and instrumental flags are added to the extracted data.

5.2.2. Pre-processed Level-2 data

In the DPC pipeline, the main pre-processing of the LFI data occurs at the Level-2 stage (see Planck Collaboration II 2020 and references therein). The same is true in the BeyondPlanck framework. First, the data are corrected for the effect of ADC non-linearities and 1 Hz frequency spikes, as outlined in the following sections.

Then the appropriate detector pointing for each sample is computed, based on auxiliary data and beam information. The Level-2 pipeline also calculates the gain modulation factor, r, from the data streams and applies it to minimize $1/f$ noise; see Eq. (55). The outputs from the two detector diodes of each radiometer are then combined with appropriate noise weights, to remove the effect of phase switch mismatch, as given by Eq. (58).

As for Level-1 data, the output is one HDF5 file for each LFI horn for each Operational Day. Each file contains the differentiated signal, detector pointing (θ, ϕ, ψ) and quality flags for each radiometer, time information (On-Board Time, SCET and MJD) and the same attitude information as in the input data.

5.2.3. 1 Hz spike correction

The output signal of the LFI receivers exhibits a set of narrow spikes at 1 Hz and harmonics with different amplitude and shape for each detector. These subtle artifacts are due to a common-mode additive effect caused by interference between scientific and housekeeping data in the analog circuits of the LFI data acquisition electronics. The spikes are present at some level in the output from all detectors, but affect the 44 GHz data most strongly because of the low voltage and high post-detection gain values in that channel. The spikes are nearly identical in sky and reference load samples, and therefore almost completely removed by the LFI differencing scheme. However, a residual effect remains in the differenced data, which needs to be carefully considered in the data processing.

These features are synchronous with the On-Board time, with no measurable change in phase over the entire survey, allowing construction of a piecewise-continuous template by stacking the data for a given detector onto a one second interval. In the DPC analysis the spikes were found to produce negligible effects in the 30 and 70 GHz channels, and were removed only from the 44 GHz time-ordered data via template fitting. We adopt the same procedure without changes.

5.2.4. Analog-to-digital conversion correction

The analog signal from each detector is processed by an analog-to-digital converter (ADC), which ideally provides a digitized output exactly proportional to the applied voltage. If the voltage step sizes between successive binary outputs of the ADC are not constant, then the ADC introduces a nonlinear response that leads to calibration errors. In differential measurements such as those of LFI, small localized distortions due to ADC non-linearity can have a significant impact, since the calibration reconstruction depends on the gradient of the ADC response curve at the point at which the differential measurements are made.

A non-linearity of the ADC produces a variation in the white noise level of a detector which does not correspond to a variation in the input voltage level, as one would expect if the effect were due to a gain shift. This subtle effect was observed in some of the LFI radiometer data for the first time in flight, where drops of a few percent were observed in the voltage white noise but not in the output level over periods of few weeks (Planck Collaboration III 2014). Because of their lower detector voltages, the 44 GHz channels showed the strongest effect, reaching levels of 3 to 5 %. The typical amplitude of the region where the non-linearity occurs is on the order of 1 mV, corresponding to about three bits in the ADC.

The ADC non-linearity effect has been characterised from flight data and removed from the data streams. The correct response curves is reconstructed by tracking how the noise amplitude varies with the apparent detector voltage in the TOD. Under the assumption that the radiometers are stable, the intrinsic white noise is taken to be constant, so any voltage variations are taken to be due to a combination of gain drift and ADC effects. A mathematical model of the effect and the details of the correction method are described in Appendix A of Planck Collaboration III (2014). Again, we adopt these corrections without modification, and leave full posterior sampling of ADC corrections to future work.

5.3. Pixel-domain data

In addition to time-domain LFI data, we consider several external data sets in the pixel domain, as described in the introduction to this section, simply in order to be able to constrain the full astrophysical sky model as defined in Sect. 3.6.

5.3.1. *Planck* HFI data

The first external data set we consider is *Planck* HFI, primarily in order to constrain thermal dust emission in the LFI frequencies. The HFI measurements were taken during the first 29 months of *Planck* observations, from August 2009 until January 2011, at which time the helium coolant was depleted. The HFI instrument includes a total of six frequency bands, centered on 100, 143, 217, 353, 545, and 857 GHz, respectively. The first four channels are polarized, while the latter two are (at least nominally) only sensitive to intensity.

While LFI employs coherent radiometers and HEMTs for signal detection, HFI employs bolometers. One important difference between these two detector types is that while the former records both the phase and the amplitude of the incoming electric field, the latter is sensitive only to the amplitude. In practice, this difference translates into different sensitivity as a function of frequency, as well as different instrumental systematics. Generally speaking, bolometers have lower noise levels than coherent radiometers over relevant CMB frequencies, but they also tend to be more susceptible to various systematic errors. For instance, for the LFI 70 GHz radiometers the noise equivalent temperature[9] is $152\,\mu\mathrm{K_{CMB}\,s^{-1/2}}$ (Planck Collaboration II 2016), while it for the HFI 143 GHz bolometers is $57.5\,\mu\mathrm{K_{CMB}\,s^{-1/2}}$ (Planck Collaboration VII 2016). At the same time, the size of CMB detectors typically scales with wavelength, and it is therefore possible to fit a larger number of high frequency detectors than low-frequency detectors into the same focal plane area. In sum, HFI nominally has more than six times higher sensitivity than LFI with respect to CMB fluctuations, as measured in terms of white noise alone. However, a non-negligible fraction of this sensitivity advantage is lost because of higher sensitivity to cosmic rays, ADC non-linearities, and long-duration bolometer time constants (Planck Collaboration III 2020).

Several different HFI analysis pipelines were developed within the nominal *Planck* collaboration period, as detailed by Planck Collaboration VI (2014), Planck Collaboration VII (2016), and Planck Collaboration III (2020). The two most recent and advanced efforts are summarized in terms of the `SROLL2` (Delouis et al. 2019) and `NPIPE` (Planck Collaboration Int. LVII 2020) pipelines. For BeyondPlanck, we adopt by default the `NPIPE` processing as our HFI data set, which is the most recent among the various available options. However, we note that most analyses here will only consider the highest frequency channels (857 GHz in temperature and 353 GHz in polarization), in order to constrain thermal dust emission, and the precise details of the HFI processing are largely irrelevant for these purposes.

The HFI data are pre-processed as follows before integration into the BeyondPlanck pipeline. First, we note that the HFI frequency channels have angular resolutions ranging between 9.7 arcmin at 100 GHz and 4.4 arcmin at 857 GHz. The natural `HEALPix` pixel resolution for HFI is thus either $N_{\mathrm{side}} = 2048$ or 4096. While our computational codes do support full resolution analysis, such high resolution is computationally wasteful for the purposes of LFI analysis. We therefore smooth the HFI maps to a common angular resolution of 10′ FWHM (which is still smaller than the 14′ beam of the 70 GHz channel), and we re-pixelize each map at $N_{\mathrm{side}} = 1024$. Overall, this reduces both CPU and memory requirements for the component separation phase of the algorithm by about one order of magnitude. Second, we subtract estimates of both zodiacal light and the kinematic CMB quadrupole from each sky map prior to analysis, following Planck Collaboration Int. LVII (2020).

[9] The noise equivalent temperature (NET) represents the noise standard deviation, σ_0, expressed in thermodynamic units of $\mu\mathrm{K_{CMB}}$ with an integration time of $\Delta t = 1$ s.

5.3.2. *Wilkinson Microwave Anisotropy Probe*

Second, we consider observations from the *Wilkinson Microwave Anisotropy Probe* (*WMAP*; Bennett et al. 2013), primarily in order to constrain synchrotron, free-free, and anomalous microwave emission. *WMAP* was funded by the National Aeronautics and Space Administration (NASA), and operated for 9 years between 2001 and 2010. *WMAP* observed the microwave sky in five frequency bands, centered on 23, 33, 41, 61, and 94 GHz, with an angular resolution varying from 53′ at 23 GHz to 13′ at 94 GHz, and with sensitivities that range between 0.8 and 1.6 $\mathrm{mK\,s^{-1/2}}$.

Like LFI, the *WMAP* detectors are based on coherent HEMT technology. However, there are (at least) two critical differences between the practical implementation of the two experiments. First, while the LFI detectors measure the difference between the sky signal in a single direction and that from an internal 4 K reference load, the *WMAP* detection chain is intrinsically differential. That is, each radiometer is coupled to two independent feedhorns that are separated by an angle of 141° on the sky, and each TOD sample is given by the difference between the signals recorded by those two horns. For this reason, each *WMAP* channel is often referred to as a "differencing assembly" (DA), rather than a radiometer. Second, while the basic *Planck* scanning strategy is fixed by its single reaction wheel, supporting smooth rotation only around a single axis, the *WMAP* satellite carried three orthogonal reaction wheels that allow for much more tightly interconnected scanning strategies. In sum, these differences lead to independent instrumental systematics between the two instruments and consequently to different strategies to minimise their impact. The two data sets are thus complementary, and can be used to break each other's internal degeneracies.

As discussed above, we will in this paper only use enough external data to break parameter degeneracies that cannot be resolved by *Planck* LFI alone, thereby leaving enough room to allow this data set to provide the main CMB constraints. Therefore, we include in the following only the *WMAP* channels between 33 and 61 GHz. In intensity, we use the *WMAP* 9-year full-resolution maps with a diagonal noise covariance matrix, while in polarization we use the low-resolution maps with full noise covariance. No pre-processing is applied to any *WMAP* data before integration into the BeyondPlanck pipeline.

5.3.3. Low-frequency surveys

As discussed by Planck Collaboration X (2016), because of the roughly similar shapes of the synchrotron, free-free and AME SEDs between 20 and 70 GHz, *Planck* and *WMAP* are not able to resolve these components on their own. Rather, it is critically important to complement these data with at least one low-frequency survey in order to establish a statistically non-degenerate model.

In BeyondPlanck, we follow Planck Collaboration X (2016), and include the celebrated 408 MHz survey by Haslam et al. (1982). Although this is widely believed to suffer more from instrumental systematic errors than comparable recent surveys, such as S-PASS (Carretti et al. 2019) or C-BASS (King et al. 2014), it also has the distinct advantages of both being publicly available and covering the full sky. This full-sky coverage was achieved by combining observations taken by the Jodrell Bank

MkI 76 m telescope, the Bonn 100 m telescope, and the Parkes 64 m telescope during the 1960's and 1970's. A second advantage is its very low frequency, which allows for a very clean separation of synchrotron emission, with only a minor additional contribution from free-free emission.

We adopt the reprocessed version of the Haslam map that was presented by Remazeilles et al. (2015) for our analyses, and, following Planck Collaboration X (2016), we model the uncertainty of this map with a uniform standard deviation of 0.8 K per pixel, added in quadrature to 1 % of the amplitude in that pixel. Finally, we adopt the monopole and dipole corrections presented by Wehus et al. (2017) to fix the largest angular scales.

Very recently, the S-PASS survey by Carretti et al. (2019) was made publicly available. This provides a detailed map of the southern celestial hemisphere in both intensity and polarization at 2.3 GHz. In principle, this map could play a similar role to the 408 MHz survey for BeyondPlanck. However, its limited sky fraction leads to significant complications. Additionally, the S-PASS polarization map is significantly affected by Faraday rotation (Krachmalnicoff et al. 2018; Fuskeland et al. 2019), which require detailed modelling before inclusion into the main BeyondPlanck pipeline is possible.

6. Bayesian analysis and MCMC sampling

We have now defined an effective parametric model of the astrophysical sky in Sect. 3.6, and an effective instrument model in Eq. (47). We now seek to constrain these models using the data summarized in Sect. 5. Let us for convenience denote the combined set of all free parameters by ω, such that $\omega \equiv \{g, \Delta_{\rm bp}, \boldsymbol{n}_{\rm corr}, \boldsymbol{a}_i, \beta_i, C_\ell, \dots\}$. In BeyondPlanck, we choose to work within the well-established Bayesian framework, and as such, our main goal is to estimate the posterior distribution, $P(\omega \mid \boldsymbol{d})$, where $\boldsymbol{d}$ denotes all available data, both in the form of time-ordered LFI observations and pre-pixelized external sky maps.

Clearly, this distribution involves billions of non-Gaussian and highly correlated parameters. Figure 8 is an informal attempt to visualize some of the main degeneracies of this distribution. Thick arrows indicate particularly strong correlations, while thin arrows indicate weaker ones. This chart is just intended to be a rough illustration, based on our practical experience, rather than a formal posterior exploration, and so it is obviously incomplete. Still, it may serve as useful reminder for new readers about how individual parameters affect other parts of the system. To consider one specific example, the gain has a direct and strong impact on both the CMB and foreground maps by virtue of multiplying the TOD, and this impact goes both ways; if the current CMB or foreground parameters are biased, then the estimated gains will also be biased. The same observations also hold with respect to the correlated noise and bandpasses, although at a lower level. On the other hand, the gains are only weakly dependent on the monopoles or sidelobes. The sidelobes do affect the CMB dipole, however, which is a critically important component for the gain estimation, and so there is a second-order dependency. Similar observations hold for most other parameters; the distribution is tightly integrated, and each parameter affects a wide range of the full model, either directly or indirectly. This integrated nature of the full posterior distribution emphasizes the importance of global end-to-end analysis with full propagation of uncertainties, as implemented in the following.

To start our formal exploration of this full posterior distribution, we write down Bayes' theorem,

$$P(\omega \mid \boldsymbol{d}) = \frac{P(\boldsymbol{d} \mid \omega) P(\omega)}{P(\boldsymbol{d})} \propto \mathcal{L}(\omega) P(\omega), \tag{60}$$

where $P(\boldsymbol{d} \mid \omega) \equiv \mathcal{L}(\omega)$ is called the likelihood; $P(\omega)$ is called the prior; and $P(\boldsymbol{d})$ is a normalization factor usually referred to as the "evidence". By virtue of being independent of ω, the evidence is irrelevant for parameter estimation purposes, and we ignore it in the current work, although we note that it is important for model selection applications.

For a one-, two-, or three-dimensional parametric model, the simplest way to numerically evaluate the posterior distribution is often to compute the right-hand side of Eq. (60) over some grid in ω. However, this approach quickly becomes computationally expensive in higher-dimensional parameter spaces, since the number of grid points grows exponentially with the number of parameters. For models with more than three parameters, it is common practice to resort to Markov Chain Monte Carlo (MCMC) sampling techniques rather than grid techniques. The main advantage of these techniques is that computing resources are mostly spent on exploring the peak of the posterior, which is the region in parameter space that actually matters for final parameter estimates. In contrast, gridding techniques spend most of their time evaluating probability densities that are statistically equivalent to zero. In this section, we will briefly review three particularly important examples of such MCMC sampling techniques, as they play a fundamental role in the BeyondPlanck pipeline.

6.1. Metropolis sampling

By far the most commonly applied, and widely known, MCMC algorithm is the Metropolis sampler (Metropolis et al. 1953). Let ω_i denote the ith sample in a Markov chain,[10] and $T(\omega_{i+1} \mid \omega_i)$ be a stochastic transition probability density for ω_{i+1} that depends on ω_i, but not on earlier states. Assume further that T is symmetric, such that $T(\omega_{i+1} \mid \omega_i) = T(\omega_i \mid \omega_{i+1})$. The most typical example of such a transition rule is a Gaussian distribution with mean equal to ω_i and with some predefined standard deviation (or "step size"), σ.

With these definitions, the Metropolis sampling algorithm can be summarized in terms of the following steps:

1. Initialize the chain at some arbitrary parameter set, ω_0.
2. Draw a random proposal[11] for the next sample based on the transition rule, i.e., $\omega_{i+1} \leftarrow T(\omega_{i+1} \mid \omega_i)$.
3. Compute the acceptance probability, a, defined by

$$a = \min\left(1, \frac{P(\omega_{i+1} \mid \boldsymbol{d})}{P(\omega_i \mid \boldsymbol{d})}\right) \tag{61}$$

4. Draw a random number, η, from a uniform distribution, $U[0,1]$. Accept the proposal if $\eta < a$; otherwise, set $\omega_{i+1} = \omega_i$.
5. Repeat steps 2–4 until convergence.

The critical component in this algorithm is the acceptance rule in Eq. (61). On the one hand, this rule ensures that the chain

[10] A Markov chain is a stochastic sequence of parameter states, $\{\omega_i\}$, in which ω_i only depends on ω_{i-1}, but not earlier states.

[11] The symbol "$\leftarrow$" indicates setting the symbol on the left-hand side equal to a sample drawn from the distribution on the right-hand side.

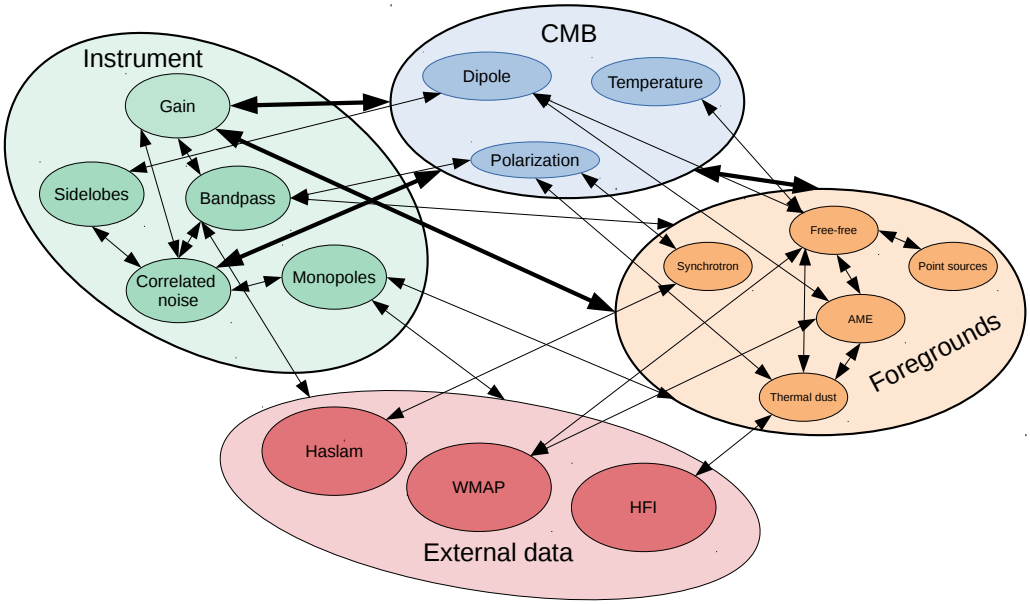

Fig. 8. Schematic overview of the primary parameters and external data sets considered in the current BeyondPlanck analysis and their interdependencies. This chart is intended to visualize the deeply integrated nature of a modern CMB analysis problem; changing any one of these parameter can lead to significant changes in a wide range of other parameters, and tracing these joint uncertainties is critically important for high-precision experiments.

is systematically pushed toward the posterior maximum by always accepting proposals that are more likely than the previous step. In this sense, the Metropolis sampler can be considered a nonlinear optimization algorithm that performs a random walk in the multidimensional parameter space. However, unlike most standard optimization algorithms, the method also does allow samples with lower probability density than the previous state. In particular, by accepting samples with a probability given by the relative posterior ratio of the two samples, one can show that the time spent at a given differential parameter volume is proportional to the underlying distribution density at that state. Thus, the multidimensional histogram of MC samples produced with this algorithm converges to $P(\omega \mid \boldsymbol{d})$ in the limit of an infinite number of samples.

6.2. Metropolis-Hastings sampling

We note that there is no reference to the proposal distribution T in the Metropolis acceptance probability as defined by Eq. (61). This is because we have explicitly assumed that T is symmetric. If we were to choose an asymmetric transition distribution, this equation would no longer hold, as proposals within the heavier tail would be systematically proposed more often than proposals within the lighter tail, and this would overall bias the chain.

For asymmetric transition distributions, we need to replace Eq. (61) with

$$a = \min\left(1, \frac{P(\omega_{i+1} \mid \boldsymbol{d})}{P(\omega_i \mid \boldsymbol{d})} \frac{T(\omega_i \mid \omega_{i+1})}{T(\omega_{i+1} \mid \omega_i)}\right), \tag{62}$$

as shown by Hastings (1970). Without further changes, the algorithm in Sect. 6.1 is then valid for arbitrary distributions T, and the algorithm is in this case called Metropolis-Hastings sampling.

6.3. Gibbs sampling

While the Metropolis and Metropolis-Hastings samplers are prevalent in modern Bayesian analysis applications, they do require a well-tuned proposal distribution T in order to be computationally efficient. If the step size is too small, it takes a prohibitive number of proposals to move from one tail of the distribution to another, whereas if the step size is too large, then all proposals are in effect rejected by the acceptance rate. The latter issue is particularly critical in high-dimensional spaces, and for this reason Metropolis-type samplers are usually only applied to moderately high-dimensional parameter spaces, for instance 20 or 50 dimensions. For millions of dimensions, traditional non-guided Metropolis sampling becomes entirely intractable.

In order to achieve acceptable efficiencies in such cases, one must typically exploit additional information within the transition probability. For instance, the Hamiltonian sampler exploits the derivative of the posterior distribution to establish proposals (e.g, Liu 2008), while the Langevin Monte Carlo algorithm can also incorporate second-order derivatives (Girolami & Calderhead 2011).

Another effective way of improving computing efficiency is to decompose complicated high-dimensional joint distributions into its various conditional distributions, a process that is called Gibbs sampling (Geman & Geman 1984). In this case, one exploits the shape of the posterior distribution itself to make proposals, but only in the form of conditionals. To illustrate the process, let us for the sake of notational simplicity consider a two-

dimensional distribution $P(\alpha,\beta)$. In that case, the Gibbs sampling transition probability takes the form

$$T_{\rm Gibbs}(\alpha_{i+1},\beta_{i+1} \mid \alpha_i,\beta_i) = P(\alpha_{i+1} \mid \beta_i)\,\delta(\beta_{i+1}-\beta_i), \tag{63}$$

where $\delta(x)$ denotes the Dirac delta function, which vanishes for $x \neq 0$, but has a unit integral. The δ function in Eq. (63) ensures that $\beta_{i+1} = \beta_i$, i.e., that β is kept fixed.

This is an asymmetric proposal distribution, and the corresponding acceptance probability is therefore given by inserting Eq. (63) into the Metropolis-Hastings rule in Eq. (62):

$$\begin{aligned}
a &= \frac{P(\alpha_{i+1},\beta_{i+1})}{P(\alpha_i,\beta_i)}\frac{T_{\rm Gibbs}(\omega_i \mid \omega_{i+1})}{T_{\rm Gibbs}(\omega_{i+1} \mid \omega_i)} && (64)\\
&= \frac{P(\alpha_{i+1},\beta_{i+1})}{P(\alpha_i,\beta_i)}\frac{P(\alpha_i \mid \beta_{i+1})\,\delta(\beta_i-\beta_{i+1})}{P(\alpha_{i+1} \mid \beta_i)\,\delta(\beta_{i+1}-\beta_i)} && (65)\\
&= \frac{P(\alpha_{i+1},\beta_i)}{P(\alpha_i,\beta_i)}\frac{P(\alpha_i \mid \beta_i)}{P(\alpha_{i+1} \mid \beta_i)} \qquad \beta_{i+1}=\beta_i && (66)\\
&= \frac{P(\alpha_{i+1} \mid \beta_i)\,P(\beta_i)}{P(\alpha_i \mid \beta_i)\,P(\beta_i)}\frac{P(\alpha_i \mid \beta_i)}{P(\alpha_{i+1} \mid \beta_i)} \qquad P(\alpha,\beta)=P(\alpha \mid \beta)P(\beta) && (67)\\
&= 1, && (68)
\end{aligned}$$

where we have used the definitions of both conditional[12] and marginal[13] distributions; the equations marked in gray indicate which relation is used in a given step. From this calculation, we see that when proposing samples from a conditional distribution within a larger global joint distribution, the Metropolis-Hastings acceptance rate is always unity. Consequently, there is no need to even compute it, and this can save large amounts of computing time for complex distributions. However, one does of course have to propose from the proper conditional distribution for this result to hold.

It is also important to note that only a sub-space of the full distribution is explored within a single Markov step with this algorithm. To explore the full distribution, it is therefore necessary to iterate through all possible conditionals, and allow changes in all dimensions. Note, however, that there are no restrictions in terms of order in which the conditionals are explored. Any combination of sampling steps is valid, as long as all dimensions are explored sufficiently to reach convergence.

The Gibbs sampling algorithm forms the main computational framework of the BeyondPlanck analysis pipeline. However, within this larger framework a large variety of different samplers are employed in order to explore the various conditionals. For convenience, Appendix A provides a summary of the most important samplers, while specific implementation details are deferred to the individual companion papers.

We conclude this section by noting that Gibbs sampling only works well for uncorrelated and weakly degenerate distributions. For strongly degenerate distributions, the number of Gibbs iterations required to explore the full distribution becomes prohibitive, as the algorithm only allows parameter moves parallel to coordinate axes. In such cases, it is usually necessary either to reparametrize the model in terms of less degenerate parameters; or, if possible, sample the degenerate parameters jointly. A commonly used trick in that respect is to exploit the identity $P(\alpha,\beta) = P(\alpha \mid \beta)P(\beta)$, which tells us that a joint sample may be established by first sampling β from its *marginal* distribution, and then α from the corresponding conditional distribution as before. The marginal sampling step ensures the Markov chain correlation length becomes unity. This trick is used in several places in the BeyondPlanck Gibbs chain, for instance for the combination of instrumental gain and correlated noise (Gjerløw et al. 2020), and for the combination of astrophysical component amplitudes and spectral parameters in intensity (Andersen et al. 2020), both of which are internally strongly correlated.

[12] Definition of a conditional distribution: $P(\alpha \mid \beta) \equiv P(\alpha,\beta)/P(\beta)$

[13] Definition of a marginal distribution: $P(\beta) \equiv \int P(\alpha,\beta)\,{\rm d}\alpha$

7. Global model specification

The previous section provides a very general overview of our analysis strategy. In this section, we provide a detailed specification of the parametric BeyondPlanck model that is appropriate for actual implementation and processing.

7.1. Global parametric model

Following the general model introduced in Sects. 3–4, we adopt the following time-ordered data model,

$$d_{j,t} = g_{j,t}\mathsf{P}_{tp,j}\left[\mathsf{B}^{\rm symm}_{pp',j}\sum_c \mathsf{M}_{cj}(\beta_{p'},\Delta^j_{\rm bp})a^c_{p'} + \mathsf{B}^{\rm asymm}_{j,t}\left(\boldsymbol{s}^{\rm orb}_j + \boldsymbol{s}^{\rm fsl}_t\right)\right] + n^{\rm corr}_{j,t} + n^{\rm w}_{j,t}. \tag{69}$$

Here j represents a radiometer label, t indicates a single time sample, p denotes a single pixel on the sky, and c represents one single astrophysical signal component. Further,

- $d_{j,t}$ denotes the measured data value in units of V;
- $g_{j,t}$ denotes the instrumental gain in units of $\mathrm{V\,K^{-1}_{cmb}}$
- $\mathsf{P}_{tp,j}$ is the $N_{\rm TOD} \times 3N_{\rm pix}$ pointing matrix defined in Eq. (36), where ψ is the polarization angle of the respective detector with respect to the local meridian;
- B_j denotes the beam convolution in Eq. (46) in the form of a matrix operator; note that for computational efficiency reasons we only take into account beam asymmetries for the sidelobes and orbital dipole in this paper;
- $\mathsf{M}_{cj}(\beta_p,\Delta_{\rm bp})$ denotes element (c,j) of an $N_{\rm comp}\times N_{\rm comp}$ mixing matrix defined in Eq. (43), describing the amplitude of component c as seen by radiometer j relative to some reference frequency j_0 when assuming some set of bandpass correction parameters $\Delta_{\rm bp}$;
- a^c_p is the amplitude of component c in pixel p, measured at the same reference frequency as the mixing matrix M, and expressed in brightness temperature units;
- $\boldsymbol{s}^{\rm orb}_j$ is the orbital CMB dipole signal in units of $\mathrm{K_{cmb}}$, including relativistic quadrupole corrections;
- $\boldsymbol{s}^{\rm fsl}_j$ denotes the contribution from far sidelobes, also in units of $\mathrm{K_{cmb}}$;
- $n^{\rm corr}_{j,t}$ denotes correlated instrumental noise, as defined by Eqs. (51) and (53); and
- $n^{\rm w}_{j,t}$ is uncorrelated (white) instrumental noise.

For notational convenience, we also define

$$\boldsymbol{s}^{\rm sky}_j = \sum_c \mathsf{M}_{cj}(\beta,\Delta^j_{\rm bp})\boldsymbol{a}^c \tag{70}$$

to be the sky model for detector j without beam convolution, but integrated over the bandpass.

For external data sets, which are defined in terms of prepixelized maps, this model simplifies to

$$d_{j,p} = g_j\mathsf{B}^{\rm symm}_{pp',j}\sum_c \mathsf{M}_{cj}(\beta_{p'},\Delta^j_{\rm bp})a^c_{p'} + n^{\rm w}_{j,p}, \tag{71}$$

which is identical to the `Commander2` data model considered by Seljebotn et al. (2019).

The free parameters in Eq. (69) are $\{\boldsymbol{g}, \Delta_{\rm bp}, \boldsymbol{n}_{\rm corr}, \boldsymbol{a}, \beta\}$. All other quantities are either provided as intrinsic parts of the original data sets (e.g., the pointing matrix, the beam profile, and the orbital dipole), or given as a deterministic function of already available parameters (e.g., the mixing matrix and the far sidelobe component). The only exception to this is the white noise component, which is neither fitted explicitly nor given by prior knowledge, but is simply left as a stochastic uncertainty in the model.

In addition to the parameters defined by Eq. (69), our model includes a handful of parameters that describe the statistical properties of the stochastic random fields included in the model. Specifically, we associate each of the astrophysical component maps $\boldsymbol{a}^c$ with a covariance matrix S^c, which in most cases is assumed to be statistically isotropic. Expanding $a^c_p = \sum_{\ell m} a^c_{\ell m} Y_\ell(p)$ into spherical harmonics, this matrix may then be written as

$$S^c_{\ell m,\ell' m'} \equiv \left\langle a^c_{\ell m} a^{c,*}_{\ell' m'} \right\rangle = C^c_\ell \delta_{\ell\ell'} \delta_{mm'}, \tag{72}$$

where C^c_ℓ denotes the angular power spectrum of component c. (Here we have for notational simplicity assumed that the component in question is unpolarized; the appropriate generalization to polarization is straightforward, and will be discussed in Sect. 8.3.8.) This power spectrum is a stochastic parameter on the same footing as $\boldsymbol{a}$ or β, and may as such be included in the model fitted to the data. Alternatively, the power spectrum may be modelled in terms some smaller set of parameters, ξ, through some deterministic function $C_\ell(\xi)$, in which case ξ is the set of stochastic parameters included in the model. For notational simplicity, we will only include the power spectrum in the various posterior distributions below, but we note that C_ℓ may be replaced with ξ without loss of generality.

Finally, similar considerations hold for the two noise components. First, the white noise component is assumed to be piecewise stationary and Gaussian distributed with vanishing mean and a covariance matrix equal to $\mathsf{N}^{\rm w}_{tt'} = \sigma_0^2 \delta_{tt'}$. In the following, we will assume the stationary period to be given by PIDs, and σ_0 will be fitted independently for each period. Second, the correlated noise component is also assumed to be piece-wise stationary and Gaussian distributed with zero mean, but with a nontrivial covariance structure in time, for instance as given by the $1/f$ model in Eq. (53). With this approximation, the total noise PSD is modelled in terms of a total of three free parameters, namely the white noise level σ_0, a knee frequency $f_{\rm knee}$, and a low-frequency spectral slope α. We denote the spectral noise parameters collectively as ξ_n.

So far, the discussion has been kept general, aiming to fit all necessary parameters into one succinct and computationally convenient framework. However, at this point it is useful to remind ourselves that one of the astrophysical component carries particular importance in this work, namely the CMB. This component is accommodated in Eq. (69) in the form of $\boldsymbol{a} = \boldsymbol{a}^{\rm cmb}$ and $\mathsf{M}^{\rm cmb} = 1$ in thermodynamic temperature units, with an angular CMB power spectrum defined as $C_\ell = \left\langle |a^{\rm cmb}|^2 \right\rangle$. Computing $P(C_\ell \mid \boldsymbol{d})$ (or $P(\xi \mid \boldsymbol{d})$, where ξ represents a set of cosmological parameters) properly marginalized over all relevant astrophysical and instrumental parameters, is the single most important scientific goal of the current algorithm.

In summary, the total set of free stochastic parameters adopted in this work is $\omega \equiv \{\boldsymbol{g}, \Delta_{\rm bp}, \boldsymbol{n}_{\rm corr}, \xi_n, \boldsymbol{a}, \beta, C_\ell\}$, where each symbol collectively represents a larger set of individual parameters, typically depending on radiometer, time, pixel, or component. For notational convenience, we will usually suppress individual indices, unless explicitly required for context. Likewise, we also note that in most cases, each of the parameters and quantities discussed above is associated with its own technicalities, which have been omitted in the above discussion. Such details will be provided in dedicated companion papers, with appropriate references given where appropriate. Finally, a full specification of the astrophysical component model considered in this analysis is provided in Sect. 3.6.

7.2. Deterministic quantities

Before considering the posterior distribution $P(\omega \mid \boldsymbol{d})$, it is useful to introduce some extra notation regarding various quantities that may either be derived deterministically from ancillary information or from other parameters in our model. These quantities are not stochastic variables in their own right within our model, and are as such not associated with independent degrees of freedom, but they are simply computationally convenient auxiliary variables.

7.2.1. Frequency maps and leakage corrections

The first derived quantity we consider are frequency maps, which we will denote $\boldsymbol{m}_\nu$. In our framework, frequency maps are not stochastic parameters, but instead they represent a deterministic compression of the full data set from time-ordered data into sky pixels, conditioning on any parameter or quantity that is not stationary, such as the gain, correlated noise, and the orbital dipole.

In order to construct frequency sky maps, we start by computing the following *residual calibrated TOD* for each detector,

$$r^{(0)}_{j,t} = \frac{d_{j,t} - n^{\rm corr}_{j,t}}{g_{t,j}} - \left(s^{\rm orb}_{j,t} + s^{\rm fsl}_{j,t}\right). \tag{73}$$

According to Eq. (69), $r_{j,t}$ now contains only stationary sky signal and white noise, given the current estimates of all other parameters.

In principle, $r^{(0)}_{j,t}$ could be individually binned into a pixelized map for each radiometer j given the pointing information in P^j_{tp}. Unfortunately, due to the poor cross-linking properties of the *Planck* scanning strategy, it is very difficult to solve for three independent Stokes parameters per pixel based on only information from a single radiometer. In practice, four radiometers are required in order to obtain well-conditioned maps with robust statistical properties. In the following we will mostly consider full-frequency maps, combining all four, six and twelve LFI radiometers into respective 30, 44 and 70 GHz maps.

Unfortunately, combining multiple radiometers into a single pixelized map carries its own complications. Since each radiometer has its own separate bandpass and beam profile, the observed sky will appear slightly different for each radiometer. However, when creating a single joint frequency map, only one single value per pixel is allowed. Any deviation from this mean value will be interpreted within the data model as either correlated or white noise, and consequently be filtered according to ξ_n or down-weighted according to σ_0 during processing, or be split among the various other free parameters, including the CMB map. This typically gives rise to artifacts proportional to the total signal amplitude, but modulated by the scanning strategy of the instrument. These effects are often referred to as bandpass or

beam mismatch contamination, respectively. Informally speaking, this is also often referred to as "temperature-to-polarization leakage," in recognition of the fact that the temperature signal is orders of magnitude brighter than the polarization signal, and therefore even a small bandpass or beam difference can induce a large spurious polarization signal.

Fortunately, with the model described above, which includes a full and explicit model of the astrophysical sky signal as part of its parameter space, it is possible to correct for such leakages. As described by Svalheim et al. (2020a), we adopt a very straightforward approach by simply subtracting a correction from each detector TOD, prior to map binning, of the form

$$\delta s^{\mathrm{leak}}_{j,t} = \mathsf{P}^j_{tp}\mathsf{B}^j_{pp'}\left(s^{\mathrm{sky}}_{jp'} - \left\langle s^{\mathrm{sky}}_{jp'}\right\rangle\right), \tag{74}$$

where s^{sky}_j denotes the sky model as seen by detector j, accounting for separate bandpass profiles, and angle brackets indicate an average over all radiometers included in the map. For computational efficiency reasons, the beam is here approximated as azimuthally symmetric, which allows the average over detector indicated by brackets in the equation to be performed pixel-by-pixel. However, since $\delta\boldsymbol{s}^{\mathrm{leak}}$ is already a difference between two very similar sky models with slightly different bandpasses, the error due to asymmetric beams is a second-order effect, and completely negligible compared to instrumental noise.

In order to correct for bandpass and beam leakage effects, we modify Eq. (73) accordingly,

$$r_{j,t} = \frac{d_{j,t} - n^{\mathrm{corr}}_{j,t}}{g_{t,j}} - \left(s^{\mathrm{orb}}_{j,t} + s^{\mathrm{fsl}}_{j,t} + \delta s^{\mathrm{leak}}_{j,t}\right). \tag{75}$$

After applying this correction, all detector TODs exhibit the same net sky signal, up to the accuracy of the instrument model, which itself is sampled over within the Markov chain. At the same time, the mean signal is not affected by this correction, independent of the accuracy of the instrument model, as $\left\langle \delta s^{\mathrm{leak}}\right\rangle = 0$ when averaged over all detectors.

With calibrated and cleaned TOD ready at hand which contain exclusively equalized signal and white noise for each detector, optimal mapmaking is performed simply by solving the corresponding normal equations pixel-by-pixel (see, e.g., Appendix A.2 or Ashdown et al. 2007a),

$$\left(\sum_{j\in\nu}\mathsf{P}^t_j(\mathsf{N}^{\mathrm{w}}_j)^{-1}\mathsf{P}_j\right)\boldsymbol{m}_\nu = \sum_j \mathsf{P}^t_j(\mathsf{N}^{\mathrm{w}}_j)^{-1}\boldsymbol{d}_j. \tag{76}$$

For our pointing matrix definition and white noise covariance matrix, this equation may for a single pixel be written explicitly as

$$\begin{bmatrix} \sum \frac{1}{\sigma^2_{0,j}} & \sum \frac{\cos 2\psi_{j,t}}{\sigma^2_{0,j}} & \sum \frac{\sin 2\psi_{j,t}}{\sigma^2_{0,j}} \\ \sum \frac{\cos 2\psi_{j,t}}{\sigma^2_{0,j}} & \sum \frac{\cos^2 2\psi_{j,t}}{\sigma^2_{0,j}} & \sum \frac{\cos 2\psi_{j,t}\sin 2\psi_{j,t}}{\sigma^2_{0,j}} \\ \sum \frac{\sin 2\psi_{j,t}}{\sigma^2_{0,j}} & \sum \frac{\sin 2\psi_{j,t}\cos 2\psi_{j,t}}{\sigma^2_{0,j}} & \sum \frac{\sin^2 2\psi_{j,t}}{\sigma^2_{0,j}} \end{bmatrix} \begin{bmatrix} T \\ Q \\ U \end{bmatrix} = \begin{bmatrix} \sum \frac{d_j}{\sigma^2_{0,j}} \\ \sum \frac{d_j \cos 2\psi_{j,t}}{\sigma^2_{0,j}} \\ \sum \frac{d_j \sin 2\psi_{j,t}}{\sigma^2_{0,j}} \end{bmatrix}, \tag{77}$$

where the sums run over both detector j and all time samples t that point toward pixel p. The associated inverse white noise pixel-pixel covariance matrix, $\mathsf{N}^{-1}_{pp'}$, is given simply by the inverse of the matrix on the left-hand side of Eq. (76).

It is important to note that the frequency maps defined by Eq. (76) have a slightly different statistical interpretation than those delivered by earlier CMB analysis pipelines, for instance from the *Planck* DPCs or *WMAP* science team. With our definition, $\boldsymbol{m}_\nu$ represents one possible realization of the frequency sky map *assuming perfect knowledge about the correlated noise, gain, bandpass, leakage effects*, among others; the only unmitigated stochastic quantity is instrumental white noise. The uncertainties due to all those other effects are instead accounted for by the fact that we produce an entire ensemble, $\boldsymbol{m}^i_\nu$, each with different combinations of systematic effects. For full error propagation, it is thus important to analyze the full set of available frequency maps, not just one single realization. In contrast, traditional frequency maps represent an approximation to the overall maximum likelihood solution, and error propagation can only be achieved through analysis of end-to-end simulations.

We conclude this section by emphasizing that $\boldsymbol{s}^{\mathrm{leak}}$ as defined above is not a separate stochastic parameter within our model. It neither increases the total uncertainty in the system, nor does it induce new parameter degeneracies; it is a simple deterministic correction that removes a known bias in co-added frequency maps.

7.2.2. Spurious leakage maps

The correction for spurious leakages from bandpass and beam mismatch defined in Eq. (74) is only exact to the extent that the assumed bandpass and beam profiles are accurate. In order to monitor the efficiency of the leakage correction, it is therefore useful to establish a dedicated goodness-of-fit statistic for this correction. For this purpose, we adopt the "spurious map" approach pioneered by Page et al. (2007), and later adapted within various pipelines, including Planck Collaboration II (2020) and Planck Collaboration Int. LVII (2020).

The central idea underlying this approach is to modify the pointing matrix to allow for a set of additional temperature maps, each corresponding to the difference between the temperature sky as seen by radiometer j and the temperature sky as seen by the mean of the detectors at that frequency. However, to prevent the linear mapmaking equation from becoming degenerate, one can at most include $N_{\mathrm{det}} - 1$ such spurious maps for a configuration involving N_{det} detectors. Thus, we generalize the pointing model for a single observation in terms of the Stokes parameters and spurious maps as follows,

$$s_j = T + Q\cos 2\psi_j + U\sin 2\psi_j + \sum_{i=1}^{N_{\mathrm{det}}-1} S_i\,\delta_{ij}. \tag{78}$$

Given this definition, the mapmaking equation in Eq. (76) generalizes straightforwardly, and for the special case of three detectors, the contribution of a single sample from detector j takes the schematic form

$$\begin{bmatrix} 1 & \cos 2\psi & \sin 2\psi & \delta_{1j} & \delta_{2j} \\ \cos 2\psi & \cos^2 2\psi & \cos 2\psi\sin 2\psi & \cos 2\psi\,\delta_{1j} & \cos 2\psi\,\delta_{2j} \\ \sin 2\psi & \sin 2\psi\cos 2\psi & \sin^2 2\psi & \sin 2\psi\,\delta_{1j} & \cos 2\psi\sin 2\psi\,\delta_{2,j} \\ \delta_{1j} & \cos 2\psi\,\delta_{1i} & \sin 2\psi\,\delta_{1j} & \delta_{1j} & 0 \\ \delta_{2j} & \cos 2\psi\,\delta_{2i} & \sin 2\psi\,\delta_{2j} & 0 & \delta_{2j} \end{bmatrix} \begin{bmatrix} T \\ Q \\ U \\ S_1 \\ S_2 \end{bmatrix} = \begin{bmatrix} d \\ d\cos 2\psi \\ d\sin 2\psi \\ d\,\delta_{1j} \\ d\,\delta_{2j} \end{bmatrix}. \tag{79}$$

For *WMAP*, it is in fact possible to solve this equation pixel-by-pixel, due to the highly interconnected *WMAP* scanning strategy (Page et al. 2007). The resulting Stokes parameter maps solved jointly with S were therefore released as primary mission products (Bennett et al. 2013). Unfortunately, the same is not possible for *Planck* without inducing an unacceptable increase in the overall noise level, as the coupling matrix in Eq. (79) is poorly conditioned over most of the sky. However, the resulting S maps are still very useful for monitoring purposes, and we will in fact use these maps to optimize a small number of bandpass parameters, for which a high level of noise is of no concern; see Svalheim et al. (2020a) and Sects. 4.2 and 8.3.4 for further details.

7.2.3. Orbital dipole

The third derived quantity we will need is the orbital dipole, $s^{\rm orb}_{j,t}$. Including a relativistic quadrupole correction, this has a closed form as given by

$$s^{\rm orb}_{j,t} = \frac{T_{\rm CMB}}{c}\left(\boldsymbol{v}_{\rm sat}\cdot\hat{\boldsymbol{n}}_{j,t} + q(\boldsymbol{v}_{\rm sat}\cdot\hat{\boldsymbol{n}}_{j,t})^2\right), \tag{80}$$

where

$$q = \frac{x(\mathrm{e}^{2x}+1)}{\mathrm{e}^{2x}-1}; \quad x = \frac{h\nu}{2kT_{\rm CMB}} \tag{81}$$

is the frequency dependency of the relativistic quadrupole term. The CMB temperature is in our analysis fixed to $T_{\rm CMB} = 2.7255\,\mathrm{K}$, following Fixsen (2009). Finally, c is the speed of light, h is Planck's constant, k is Boltzmann's constant, $\boldsymbol{v}_{\rm sat}$ is the satellite velocity, and $\hat{\boldsymbol{n}}_{j,t}$ is the pointing vector of detector j at time t. The satellite velocity is known with an absolute precision better than $1\,\mathrm{cm\,s^{-1}}$ (Planck Collaboration Int. LVII 2020). An efficient convolution algorithm for this component that takes into account the full 4π beam is described by Galloway et al. (2020b).

It is important to note the critical role of this particular signal term. Depending only on the velocity of the satellite (which is known to exceedingly high precision) and the CMB temperature (which is known to a precision of 0.02 %; Fixsen 2009), it provides the best absolute calibration source in microwave astronomy, if not all of astronomy. For BeyondPlanck, as for both *Planck* and *WMAP*, this signal is therefore used to determine the overall absolute calibration of the entire data set.

7.2.4. Far sidelobe corrections

The last derived quantity we will need at this stage is the far sidelobe correction, $\boldsymbol{s}^{\rm fsl}$, as defined in Sect. 4.3. As shown by Planck Collaboration IV (2016), the *Planck* LFI optics have several significant sidelobes at large angles from the optical axis. The most important is due to spillover around the main reflector, and located about 85° from the main beam. The second most important is due to spillover around the secondary reflector, and located about 20° from the main beam. To account for these, we convolve the parametric sky model with the (near-)4π beam profile, B, of each radiometer (regions closer than 5° from the main beam are excluded),

$$s^{\rm fsl}_{j,t} = \int_{4\pi} [\mathsf{R}(\Omega_t)\mathsf{B}(\Omega)]\, s^{\rm sky}_j(\Omega)\,\mathrm{d}\Omega, \tag{82}$$

where $\mathsf{R}(\Omega_t)$ is a rotation matrix that rotates the beam as specified by the satellite pointing at time t. To evaluate this integral, we employ an algorithm that is algebraically equivalent to the `conviqt` approach described by Prézeau & Reinecke (2010), but implemented in terms of spin harmonics, as described by Galloway et al. (2020b).

We stress, however, that uncertainties in the far-sidelobe model are not yet accounted for, and this represents a significant model uncertainty in the current analysis. Generalizing the parametric model in Eq. (69) to allow for new beam-related degrees of freedom is an important goal for future algorithm development.

8. The BeyondPlanck Gibbs sampler

8.1. Global posterior distribution

Given the global parametric model defined in Sect. 7.1, and the ancillary quantities summarized in Sect. 7.2, we are now finally ready to consider the full global BeyondPlanck posterior distribution, $P(\omega \mid \boldsymbol{d})$, and describe the computational algorithms required to map it out. In practice, this entails writing down explicit expressions for the likelihood and priors in Eq. (60), as well as specifying an explicit Gibbs chain that is able to explore the posterior distribution efficiently.

Starting with the likelihood, $\mathcal{L}(\omega)$, we first note that the data model defined in Eqs. (69)–(71) is given as a linear sum of various components, all of which are specified precisely in terms of our free parameters ω. This applies even to the correlated noise component, $\boldsymbol{n}^{\rm corr}$, which for the purposes of the likelihood is fully equivalent to any of the other physical components. As such, we may symbolically write $\boldsymbol{d} = \boldsymbol{s}^{\rm tot}(\omega) + \boldsymbol{n}^{\rm w}$, where $\boldsymbol{s}^{\rm tot}(\omega)$ is the sum of all model components in Eq. (69), whether they have a cosmological, astrophysical or instrumental origin. With this notation, we immediately see that

$$P(\boldsymbol{d} \mid \omega) \propto P(\boldsymbol{n}^{\rm w} \mid \omega) \propto \exp\left(-\frac{1}{2}\left(\frac{\boldsymbol{d}-\boldsymbol{s}^{\rm tot}(\omega)}{\sigma_0}\right)^2\right), \tag{83}$$

since $\boldsymbol{n}^{\rm w} = \boldsymbol{d} - \boldsymbol{s}^{\rm tot}(\omega)$, $P(\boldsymbol{n}^{\rm w}) \propto N(0,\sigma^2)$, and $\boldsymbol{s}^{\rm tot}$ is deterministically given by ω.

Next, the prior $P(\omega)$ should encapsulate all our prior knowledge about any of the model parameters. For instance, we may use this term to introduce information regarding the instrumental gain from temperature measurements of the 4 K load onboard the *Planck* satellite during the calibration stage; or we can use it to impose prior knowledge regarding the CIB zero-level amplitude at each frequency during component separation; or we may introduce a prior on the Hubble constant during cosmological parameter estimation; or we may use it to regularize posterior volume effects through the application of a Jeffreys ignorance prior (Jeffreys 1946). A detailed breakdown of the priors used in this particular analysis will be presented in association with the respective steps.

8.2. Overview of Gibbs chain

As already discussed, the posterior distribution defined by Eq. (60) involves millions of tightly correlated and non-Gaussian parameters, and it is clearly unfeasible to optimize or sample from it directly. We therefore resort to the Gibbs sampling algorithm described in Sect. 6.3: We compute a Markov chain of correlated samples by initializing on some arbitrary parameter combination, ω_0, and then iteratively sample from each conditional distribution from the full distribution. In practice, most runs are initialized on the outcome of an earlier analysis, in order to save burn-in time.

The BeyondPlanck Gibbs chain may be written schematically as follows,

$$\boldsymbol{g} \leftarrow P(\boldsymbol{g} \mid \boldsymbol{d}, \xi_n, \Delta_{\rm bp}, \boldsymbol{a}, \beta, C_\ell) \tag{84}$$
$$\boldsymbol{n}_{\rm corr} \leftarrow P(\boldsymbol{n}_{\rm corr} \mid \boldsymbol{d}, \boldsymbol{g}, \xi_n, \Delta_{\rm bp}, \boldsymbol{a}, \beta, C_\ell) \tag{85}$$
$$\xi_n \leftarrow P(\xi_n \mid \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\rm corr}, \Delta_{\rm bp}, \boldsymbol{a}, \beta, C_\ell) \tag{86}$$
$$\Delta_{\rm bp} \leftarrow P(\Delta_{\rm bp} \mid \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\rm corr}, \xi_n, \boldsymbol{a}, \beta, C_\ell) \tag{87}$$
$$\beta \leftarrow P(\beta \mid \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\rm corr}, \xi_n, \Delta_{\rm bp}, C_\ell) \tag{88}$$
$$\boldsymbol{a} \leftarrow P(\boldsymbol{a} \mid \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\rm corr}, \xi_n, \Delta_{\rm bp}, \beta, C_\ell) \tag{89}$$
$$C_\ell \leftarrow P(C_\ell \mid \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\rm corr}, \xi_n, \Delta_{\rm bp}, \boldsymbol{a}, \beta), \tag{90}$$

where the conditional variables have been vertically aligned for clarity only. As usual, the symbol ← means setting the variable on the left-hand side equal to a sample from the distribution on the right-hand side. For convenience, in the following we also define the notation "$\omega \setminus \xi$" to imply the set of parameters in ω except ξ.

Note that the first conditional in this Gibbs chain, $P(\boldsymbol{g} \mid \boldsymbol{d}, \ldots)$ represents a marginal distribution with respect to $\boldsymbol{n}_{\rm corr}$. As such, $\boldsymbol{g}$ and $\boldsymbol{n}_{\rm corr}$ are in effect sampled jointly in the BeyondPlanck Gibbs chain (Gjerløw et al. 2020; Ihle et al. 2020), using the properties discussed in Sect. 6.3. The reason for this choice is that these two parameters are particularly strongly degenerate, and joint sampling therefore leads to a much shorter overall correlation length than strict Gibbs sampling. This outweighs by far the somewhat higher computational cost per iteration that is required for sampling the gain from its marginal distribution.

The same trick is applied when sampling astrophysical component parameters, $\boldsymbol{a}$ and β in the case of intensity maps. In this case, we first sample β marginalized over $\boldsymbol{a}$, and then $\boldsymbol{a}$ conditionally on β (Andersen et al. 2020). Since $\boldsymbol{a}$ is a set of linear parameters, the integral over $\boldsymbol{a}$ may be computed analytically, as first exploited for CMB component separation purposes in the `Miramare` code (Stompor et al. 2009; Stivoli et al. 2010). For polarization, we still sample β conditionally on $\boldsymbol{a}$, as described by Svalheim et al. (2020b), because the low-resolution *WMAP* data with full covariance matrix prohibits smoothing to a common angular resolution, as needed for the marginal sampling approach.

We will now describe each of these distributions in turn, with the main goal being to build intuition regarding each distribution. For specific implementational details we refer the interested reader to companion papers.

At this point, we note that if a joint maximum likelihood estimate is required as opposed to a sample set, the same methodology applies as described below, with the exception that one should then maximize each conditional, rather than sample from it. The algorithm then becomes equivalent to a (slow but computationally convenient) steepest descent nonlinear optimizer. In our codes, we have implemented support for both modes of operation.

8.3. Specification of conditional sampling steps

8.3.1. Gain and calibration sampling

We start our review of the various Gibbs sampling steps with the gain, g_t. In this paper, we only summarize the main central steps, and we refer the interested reader to Gjerløw et al. (2020) for full algorithmic details.

The gain is among the most critical parameters in our model in terms of the overall resulting data quality, and even relative errors at the $O(10^{-4})$ level are highly significant. At the same time, it is also one of the parameters we have the least prior information about, as it is entirely specific for each individual instrument. To estimate the gain robustly, we therefore exploit the following observations: First, we note that the orbital CMB dipole (see Sect. 7.2.3) depends only the satellite velocity, which is known to a precision of 10^{-6} (Godard et al. 2009), and the CMB monopole value, which is known to a precision of 0.02 % (Fixsen 2009). The orbital dipole therefore by far provides the most robust constraints on the mean calibration.[14] However, since the Earth's orbital velocity is 30 km s^{-1} and the CMB monopole is 2.7255 K, the absolute amplitude of the orbital dipole is only 270 μK, which is small compared to typical signal and noise variations. As a result, the orbital dipole is not strong enough to directly determine the gain alone on short time scales.

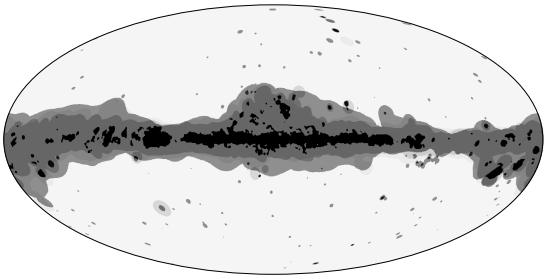

Fig. 9. Processing masks used for low-level TOD processing. The gray regions show the main processing mask used for gain and correlated noise sampling (with 30 GHz shown in lightest gray, and 70 GHz in darkest gray), which remove 27, 19, and 23 % of the sky for 30, 44 and 70 GHz, respectively. The black regions show the bandpass correction sampling mask, which only removes 4.7 % of the sky.

In contrast, the amplitude of the solar CMB dipole is 3 mK, about ten times brighter than the orbital dipole. Of course, the true solar CMB dipole parameters are unknown, and must be estimated jointly with everything else; but we do know that all detectors observe the *same* solar dipole. We also know that its frequency spectrum is given by a perfect blackbody with temperature $T_{\rm CMB}$. Together, these two facts provide a strong handle on relative gain variations, both between detectors and in time.

First, we note that the fundamental data model in Eq. (69) may for each detector be written in the form

$$d_t = g_t s_t^{\rm tot} + n_t^{\rm corr} + n_t^{\rm w} \tag{91}$$

where

$$s_t^{\rm tot} = \mathsf{P}_{tp}\left[\mathsf{B}_{pp'}^{\rm symm} s_{p'}^{\rm sky} + \mathsf{B}_{pp'}^{\rm asymm}\left(s_t^{\rm orb} + s_t^{\rm fsl}\right)\right] \tag{92}$$

is the total sum of all true astrophysical signals entering through the optical assembly of the instrument. Noting that both the gain and correlated noise *a priori* are unknown quantities with no and weak priors, respectively, it is clear from the structure of Eq. (91) that these two parameters are highly degenerate: significant variations in $\boldsymbol{g}$ can be accounted for by adjusting $\boldsymbol{n}_{\rm corr}$ with only a small penalty in terms of total goodness-of-fit through the noise power spectrum, ξ.

Recognizing the importance of this degeneracy, and the resulting sensitivity to potential modelling errors, we decompose the full time-dependent gain function into three components, and sample each of these with a special-purpose sampler. Specifically, we write the full gain for detector i in the form $g_t^i = g_0 + \Delta g^i + \delta g_t^i$, where g_0 is the gain averaged both over all detectors within a given frequency map and over time; Δg^i is the mean gain differential for detector i averaged over time, with the additional requirement that $\sum_i \Delta g^i = 0$; and δg_t^i represents the time dependence in gain, with the additional requirement that $\sum_t g_t^i = 0$ for each i. In addition, when sampling the these gain

[14] The term "calibration" refers in this paper to the time average of the gain.

parameters, we marginalize over the correlated noise component, as discussed in Sect. 8.3.1, in order to minimize the Gibbs chain correlation length. In total, the data model used for gain sampling therefore reads

$$d_t^i = (g_0 + \Delta g^i + \delta g_t^i) s_t^{\rm tot} + n_t^{\rm tot}, \tag{93}$$

where $n_t^{\rm tot} = n_t^{\rm corr} + n_t^{\rm wn}$ is the total noise contribution with a full covariance matrix given by ξ^n.

Formally speaking, the statistically optimal sampling algorithm for any of the three gain parameters is given by correlating the full sky signal with the observed data. In effect, this direct approach was adopted by the LFI DPC pipeline (Planck Collaboration II 2020). A significant disadvantage of this direct approach, however, is sensitivity to foreground and bandpass mismatch errors. Instead, we adopt the following three-step approach, which is structurally similar to the NPIPE algorithm (Planck Collaboration Int. LVII 2020).

First, we sample g_0 with the orbital CMB dipole alone as calibration source, based on the residual

$$r_t^i \equiv d_t^i - \hat{g}(s_t^{\rm tot} - s^{\rm orb}) - (\Delta\hat{g}^i + \delta\hat{g}_t^i) s_t^{\rm orb} = g_0 s_t^{\rm orb} + n_t^{\rm tot}, \tag{94}$$

where the symbol "^" denotes the respective estimate from the previous Gibbs iteration. Noting that $n^{\rm tot}$ is a Gaussian field with covariance N and $\boldsymbol{s}^{\rm orb}$ is a deterministic quantity, the sampling algorithm for g_0 is given by that of a univariate Gaussian distribution as described in Appendix A.1,

$$g_0 = \frac{\sum_i \boldsymbol{s}_i^{{\rm orb},t} \mathsf{N}_i^{-1} \boldsymbol{r}_i^{\rm orb}}{\sum_i \boldsymbol{s}_i^{{\rm orb},t} \mathsf{N}_i^{-1} \boldsymbol{s}_i^{\rm orb}} + \frac{\eta}{\sqrt{\sum_i \boldsymbol{s}_i^{{\rm orb},t} \mathsf{N}_i^{-1} \boldsymbol{s}_i^{\rm orb}}}, \tag{95}$$

where the sum runs over all detectors in the frequency map, and $\eta \sim N(0,1)$. We note that this expression represents a formal violation of the Gibbs chain, since g_0 in reality affects both $\boldsymbol{s}^{\rm tot}$ and $\boldsymbol{s}^{\rm orb}$. By fitting only to $\boldsymbol{s}^{\rm orb}$ we effectively gain robustness with respect to modelling errors at the cost of increased statistical uncertainty.

Note that the noise covariance matrices in Eq. (95) include both correlated and white noise contributions; this is algebraically equivalent to marginalizing over $\boldsymbol{n}_{\rm corr}$ as described above. In constrast, in a classic Gibbs sampling algorithm we would subtract $\boldsymbol{n}_{\rm corr}$ from r^i in Eq. (96), and then only include the white noise component in N_i. This, however, would lead to a much longer Markov correlation length, since the joint uncertainties between g_0 and $\boldsymbol{n}_{\rm corr}$ then would have to be explored by sampling, rather than algebraically.

Second, we sample the detector dependent calibration factors, Δg_i, based on the residual

$$r_t^i \equiv d_t^i - (g_0 + \delta\hat{g}_t^i) s_t^{\rm tot} = \Delta g_i s_t^{\rm tot} + n_t^{\rm tot}, \tag{96}$$

for each detector, which now includes contributions from both the solar CMB dipole and astrophysical foregrounds, and therefore supports a significantly higher signal-to-noise ratio than the orbital dipole alone. At the same time, we impose the additional linear constraint that

$$\sum \Delta g_i = 0, \tag{97}$$

such that possible contamination in this step does not affect the absolute mean calibration of the full frequency channel. The total system may be solved using the Lagrange multiplier technique (e.g., Bertsekas 1996) with a Lagrangian of the form

$$\mathcal{L}(\Delta g_i, \lambda) = \sum_i \left(\boldsymbol{r}^i - \Delta g_i \boldsymbol{s}^{\rm tot}\right)^t \mathsf{N}_i^{-1} \left(\boldsymbol{r}^i - \Delta g_i \boldsymbol{s}^{\rm tot}\right) + \lambda \sum_i \Delta g_i, \tag{98}$$

where λ is a Lagrange multiplier. The maximum posterior solution is found by solving the linear equations resulting from setting $\partial\mathcal{L}/\partial x_i = 0$ with $x_i = \{\Delta g_i, \lambda\}$.

Third and finally, the time-dependent gain fluctuations are sampled starting from the residual

$$r_t^i \equiv d_t^i - (g_0 + \Delta g_i)\, s_t^{\rm tot} = \delta g_i s_t^{\rm tot} + n_t^{\rm tot}, \tag{99}$$

where δg_t^i is assumed to be constant within each PID period, but free to vary between consecutive PIDs. Ideally, δg_t^i should be sampled from a multivariate Gaussian distribution, taking into account their known $1/f$ power dependency. In the current implementation, however, we adopt a similar solution as the LFI DPC pipeline (Planck Collaboration II 2020). That is, we first estimate δg_t^i independently for each time segment, and then smooth the resulting gain estimates with a moving boxcar average. In contrast, NPIPE does not smooth the estimates (Planck Collaboration Int. LVII 2020).

To prevent foreground modelling errors from affecting the various gain estimates, we apply the processing masks indicated in gray in Fig. 9 in each of the above equations. Any sample that falls within the masked region is omitted from the corresponding inner product, and does not contribute to the overall estimate. The same applies to any sample that may be flagged by the instrument database. Removing individual samples, however, does introduce a slight computational complication because of the $\mathsf{N}_{\rm tot} = \mathsf{N}_{\rm corr} + \mathsf{N}_{\rm wn}$ operator, which denotes a dense noise covariance matrix that includes both correlated and white noise. Application of this operator at full temporal TOD resolution is computationally expensive. However, we note that since the gain is defined only by a single value per PID, small-scale fluctuations can be averaged down with minimal loss of information in all the above equations. We therefore down-sample each time-ordered data object to 1 Hz before evaluating the above equations, and this reduces the overall computational cost for gain sampling by almost two orders of magnitude; see Gjerløw et al. (2020); Ihle et al. (2020) for further details.

Finally, we note that several discrete events took place during the *Planck* observations, for instance maneuvers or cooler maintenance, and many of these led to sharp jumps in the LFI gain (Planck Collaboration II 2020). To avoid smoothing across such steps, we boxcar average independently over each continuous PID range between two events. A total of 23 events are included in this manner. However, we do note that this represents a significant unmodelled uncertainty in the current BeyondPlanck analysis: The presence of unknown jumps can contaminate the final resulting polarization maps at significant level. Indeed, as discussed in Sect. 9.5, unmodelled gain jumps are viable candidates for explaining one of the main residual systematic effect in the 44 GHz channel, namely strong stripes visible in the Southern Galactic hemisphere. Automatically identifying jumps as an intrinsic step of the sampling algorithm is an important task for future work.

8.3.2. Correlated noise sampling

Since the gain is sampled from a marginal distribution with respect to correlated noise, not a conditional distribution, it is essential to sample the correlated noise immediately following the gain; otherwise the Gibbs chain would end up in an internally inconsistent state. However, as far as the actual sampling algorithm for the correlated noise is concerned, this is a normal conditional with respect to the gain, akin to any other standard Gibbs step, and was first described in a CMB setting by Wehus et al. (2012).

The same algorithm has now also, for the first time, been used to solve the CMB mapmaking problem by Keihänen et al. (2020).

To derive the appropriate sampling equation for $\boldsymbol{n}^{\rm corr}$, we return to the full data model in Eq. (69), and note that it may be written on the form

$$r_t \equiv d_t - g_t s_t^{\rm tot} = n_t^{\rm corr} + n_t^{\rm w}, \tag{100}$$

where $s_t^{\rm tot}$ is defined in Eq. (92). As discussed in Sect. 7.1, $\boldsymbol{n}^{\rm corr}$ is assumed to be Gaussian distributed with zero mean and covariance $\mathsf{N}^{\rm corr}$, while the white noise term is uncorrelated Gaussian with variance σ_0^2. Eq. (100) therefore also describes a correlated Gaussian distribution, and the sampling equation is in this case given by Eq. (A.10) with a template matrix $\mathsf{T} = 1$, a signal covariance matrix $\mathsf{S} = \mathsf{N}^{\rm corr}$, a noise covariance matrix $\mathsf{N} = \mathsf{N}^{\rm w}$, and data $\boldsymbol{d} = \boldsymbol{r}$.

Let us first consider the ideal case of a single PID with no missing data due to either instrument flags or processing mask. In that case, Eq. (A.10) can be solved very conveniently in the Fourier domain, and the appropriate sampling equation for the kth Fourier mode reads

$$n_k = \frac{r_k + \eta_1\sigma_0^2/\sqrt{P_k^{\rm corr}} + \eta_2\sigma_0}{1 + \sigma_0^2/P_k^{\rm corr}}. \tag{101}$$

For this case, we note that the computational cost is equivalent to two Fourier transforms of the full time-ordered data.

As usual, the first term in the numerator of Eq. (101) is simply a Wiener filtered version of the residual, $\boldsymbol{r}$. As such, it represents a biased estimate of $\boldsymbol{n}^{\rm corr}$, with a noise suppression factor given by the relative inverse signal-to-noise ratio, $P_k^{\rm corr}/\sigma_0^2$. The two last terms are stochastic random fluctuations that ensure that the resulting sample has the appropriate covariance structure.

Equation (101) only applies to data with no missing time samples, as the Fourier transforms require continuous inputs. In practice, however, we drop all samples that are removed by either the instrument flags or by the processing mask shown in Fig. 9. In this case, the optimal solution is therefore given by Eq. (A.10), where rows and columns corresponding to masked samples are set to zero in N. The resulting equation is therefore solved efficiently by a Conjugate Gradient (CG) technique, as described by Keihänen et al. (2020). As reported by Galloway et al. (2020a), and summarized in Table 2, this particular step accounts for about 40 % of the total computational cost of the BeyondPlanck Gibbs sampler, and it is as such by far the most expensive single component in the entire analysis.

In the current framework, correlated noise estimation plays the role of the traditional CMB mapmaking problem with correlated noise in a traditional pipeline. In this respect, it is worth noting that the correlated noise sample is constructed based on the signal-subtracted data, r, alone. Under the assumption of a perfect signal model, inaccuracies in the correlated noise model can therefore not introduce any signal bias. Using the analogy of traditional destriping codes (e.g., Maino et al. 1999; Keihänen et al. 2004, 2005, 2010), the signal subtraction plays the same role in the Gibbs sampling approach as the projection operator $\mathsf{Z} = \mathsf{I} - \mathsf{P}(\mathsf{P}^t\mathsf{N}^{-1}\mathsf{P})^{-1}\mathsf{P}^t\mathsf{N}^{-1}$ does for destriping, shielding any stationary signal from the noise filter. The main fundamental difference between the two approaches lies in the fact that while the traditional destriper only exploits information from a single frequency channel at any given time, the Gibbs sampling approach simultaneously exploits information from all frequencies to construct a joint signal model, which then is used to shield the signal during correlated noise estimation. The Gibbs sampling approach is thus mathematically equivalent to destriping all frequencies at once. The effect of this global correlated noise estimation will become evident later, in the form of lower correlated noise residuals in the joint approach.

Second, it is important to note that the correlated noise solution resulting from Eq. (101) is moderately robust against model errors, whether they are due to foreground modelling errors or inaccuracies in the bandpass or beam profile. The reason is simply that Eq. (101) is a Wiener filter, and therefore has greatly suppressed power in any frequency mode for which $P_k^{\rm corr} \ll \sigma_0^2$. Intuitively, this means that any feature that cannot be readily identified in the raw time-ordered data as compared with σ_0, will only be weakly affected by the correlated noise component. Minor errors in the signal model, beam or bandpass profiles are therefore mostly negligible.

There are, however, two important exceptions to this general rule. First, some point sources, such as Tau A or the Galactic center, are sufficiently bright that uncertainties in the beam or foreground model can be large compared to the white noise level. If so, the resulting errors will typically translate into bright stripes passing through the respective source, extending along the scanning path of the satellite. To avoid this, it is critically important to mask all bright sources as part of the processing mask, and replace those regions with a proper constrained realization as described above.

The second important exception is the CMB dipole. This signal is both bright, with a peak-to-peak amplitude of about 3 mK, and concentrated at a very low frequency that corresponds to the satellite spin rate of 1/60 Hz. This is typically comparable to (or lower than) the correlated noise knee frequencies (Planck Collaboration II 2020). Furthermore, the ring-based *Planck* scanning strategy provides minimal modulation of the dipole signal on frequencies beyond the spin frequency. The combination of these facts leads to a strong degeneracy between the CMB dipole parameters, the time-dependent gain, and the correlated noise. Indeed, experience shows that *Planck* is, for all practical purposes, unable to break this degeneracy through statistical power alone. Instead, various strong priors are typically imposed to regularize these degeneracies. For instance, the LFI DPC processing impose the requirement that $\boldsymbol{m}_{\rm D} \cdot \boldsymbol{m} = 0$, where $\boldsymbol{m}_{\rm D}$ is a map of the CMB dipole and $\boldsymbol{m}$ is the sky map; this effectively leaves the full instrumental noise component aligned with the CMB dipole in the final sky map (Planck Collaboration V 2016). Additionally, the LFI pipeline makes no explicit corrections for bandpass mismatch during gain calibration. For the HFI 2018 DPC processing, the dominant assumption is that the gain is fully independent of time, and the only source of apparent gain fluctuations are ADC non-linearities (Planck Collaboration III 2020). For `NPIPE`, two important assumptions are that polarized foregrounds at frequencies between 44 and 217 GHz may be fully modelled in terms of the signal observed by 30 and 353 GHz, and that CMB polarization may be ignored during calibration (Planck Collaboration Int. LVII 2020). Obviously, none of these assumptions are formally correct, and they will necessarily lead to systematic biases at some level.

In BeyondPlanck, we adopt a different approach to the problem, by actually exploiting information beyond *Planck*. Specifically, as described in Sect. 5, we will in the following perform a joint analysis of *WMAP* and *Planck* observations, and thereby take advantage of information in one experiment to break degeneracies in the other. Most notably, the *WMAP* scanning strategy covers 70 % of the sky every hour, as compared to less than 1 % per hour for *Planck*. This strategy is thus obviously better suited for measuring the very largest angular scales on the sky,

despite higher white noise. On the other hand, the differential structure of the *WMAP* differencing assemblies leads to particularly large uncertainties for some specific modes, including $E_{\ell=5}$ and $B_{\ell=3}$ (Jarosik et al. 2011). In BeyondPlanck we therefore choose to combine *Planck* and *WMAP* data while taking into account the full covariance information of each experiment, and thereby optimally leverage the individual strengths of each experiment. Still, we emphasize the importance of visually inspecting binned sky maps of $\boldsymbol{n}_{\rm corr}$ for dipole-like residuals, which is the archetypical signature of calibration errors; such residuals may occur if the assumed signal model is inadequate for the data set in question.

8.3.3. Noise PSD sampling

The third conditional distribution, $P(\xi^n \mid \boldsymbol{d}, \omega \setminus \xi^n)$, in the BeyondPlanck Gibbs chain describes the noise power spectrum density parameters, P_k and σ_0, collectively denoted ξ^n. In the following, we will make the assumptions that ξ^n is constant within each PID and uncorrelated between PIDs. Being closely connected to the previous sampling step, the following procedure was also first presented for CMB applications by Wehus et al. (2012).

To sample from $P(\xi^n \mid \boldsymbol{d}, \omega \setminus \xi^n)$, we recall that $\boldsymbol{n}_{\rm corr} \sim N(0, \mathsf{N}_{\rm corr})$. Therefore,

$$P(\xi \mid \boldsymbol{d}, \omega \setminus \xi^n) \propto P(\xi \mid \boldsymbol{n}_{\rm corr}) \tag{102}$$

$$\propto \frac{\mathrm{e}^{-\frac{1}{2}\boldsymbol{n}_{\rm corr}^t \mathsf{N}_{\rm corr}^{-1} \boldsymbol{n}_{\rm corr}}}{\sqrt{|\mathsf{N}_{\rm corr}|}}, \tag{103}$$

where $\mathsf{N}_{\rm corr} = \mathsf{N}_{\rm corr}(\xi^n)$. To sample from this distribution, we could for instance run a Metropolis sampler over ξ^n, using Eq. (103) to define the acceptance probability. However, at this stage we introduce an approximation to the exact solution, trading a small amount of statistical optimality for increased robustness to modelling errors and minimal parameter degeneracies. Specifically, we decouple the white noise variance from the correlated noise model simply by defining

$$\sigma_0^2 \equiv \frac{\mathrm{Var}(r_t - r_{t-1})}{2}, \tag{104}$$

where we define

$$r_t \equiv d_t - s_t^{\rm tot} - n_t^{\rm corr} \tag{105}$$

to be the residual time stream after subtracting both the current total sky signal and correlated noise estimates. Thus, we take the variance of the difference between any two neighboring residual samples to be our white noise variance. On the one hand, this represents the single most robust estimate of the total white noise level one can form from a finite data set. On the other hand, it is of course only an approximation to the true white noise level, since the correlated noise component may also happen to include a flat and non-negligible power spectrum density at the highest frequency mode. This situation typically arises more often for bolometers (as for instance employed by the *Planck* HFI detectors) than for coherent detectors (as employed by the *Planck* LFI detectors and considered here), but the principle is the same both cases.

Thus, we define any flat component of the correlated noise to be part of the white noise, and the correlated noise is consequently defined as the difference between the total noise and the white noise. For error propagation into other parameters in the model, only the sum of the two components is significant. This split is thus essentially just a computational trick that eliminates internal degeneracies between the two noise components, and maximizes the relative contribution of the white noise component. This has two main numerical advantages. First, noting that white noise marginalization is performed algebraically, while correlated noise marginalization is done through sampling, a high relative white noise fraction leads to a shorter overall Markov chain correlation length for all steps in the algorithm. Second, by fixing the white noise level, we break degeneracies within the ξ^n parameters, which otherwise lead a very long correlation length between σ_0, α, and $f_{\rm knee}$, making making convergence assessment difficult.

Given this definition of the white noise variance, the correlated noise level may now be sampled from Eq. (103) by fixing $N^{\rm w}$. Specifically, as discussed by Ihle et al. (2020), the conditional posterior may be written in Fourier space as

$$-\ln P(\xi^n) = \sum_{\nu>0}\left[\frac{|n_\nu^{\rm corr}|^2}{\sigma_0^2\left(\frac{\nu}{\nu_{\rm knee}}\right)^\alpha} + \ln\sigma_0^2\left(\frac{\nu}{\nu_{\rm knee}}\right)^\alpha\right], \tag{106}$$

up to an irrelevant constant, where $n_\nu^{\rm corr}$ are the Fourier coefficients of the correlated noise estimate, $\boldsymbol{n}_{\rm corr}$. We sample from this distribution with a simple inversion sampler (see Appendix A.3), iteratively Gibbs sampling over α and $\nu_{\rm knee}$. Masking and inpainting is handled by the $\boldsymbol{n}_{\rm corr}$ sampling step described in Sect. 8.3.2. Additionally, only frequencies below a fixed upper frequency limit (defined by twice the DPC estimate of $f_{\rm knee}$) is included when sampling α and $\nu_{\rm knee}$, to avoid potential signal leakage and modelling errors from contaminating the correlated noise estimates; for further details, see Ihle et al. (2020).

Before concluding this section, we note once again that uncertainties due to modelling errors as such are not meaningfully accounted for in a traditional Bayesian analysis. This point is important for LFI noise estimation, because Ihle et al. (2020) find that the $1/f$ noise model is not an adequate statistical description for the LFI 30 and 44 GHz frequency channels. In particular, they identify significant noise excess at intermediate temporal frequencies, between 0.1 and 1 Hz, which correspond to angular scales between 6 and 60° on the sky given the *Planck* scanning strategy. As such, this excess noise is highly relevant for large-scale polarization estimation with the 30 and 44 GHz channels. Generalizing the $1/f$ noise model to account for this excess noise is a top priority issue for future work. On the other hand, we do find that the $1/f$ model appears to be good description of the noise at 70 GHz, and by far most of the LFI-specific CMB science has been derived from this particular channel (e.g., Planck Collaboration V 2020).

8.3.4. Bandpass sampling

Next, we consider the bandpass correction conditional distribution, $P(\Delta_{\rm bp} \mid \boldsymbol{d}, \omega \setminus \Delta_{\rm bp})$, and in the following we will consider the most basic form of bandpass correction, namely a linear shift as defined by Eq. (45); see Svalheim et al. (2020a) for further details.

Similar to the gain case, we find it useful to decompose the full bandpass shift for detector j as follows,

$$\Delta_{\rm bp}^j = \bar{\Delta}_{\rm bp} + \delta_{\rm bp}^j. \tag{107}$$

Here the first term is the average over all radiometers within a given frequency channel and the second term is constrained by $\sum_j \delta_{\rm bp}^j = 0$. The motivation for this decomposition is that the two

terms impact the data in qualitatively different ways. The average bandpass shift, $\bar{\Delta}_{\rm bp}$, change the overall effective frequency of the full frequency channel, and is as such highly degenerate with the foreground SED parameters; a given bandpass frequency shift may often be counteracted by adjusting the values of the synchrotron or thermal dust spectral indices. This mean bandpass shift does not in general change the polarization properties of the resulting frequency map. The relative bandpass corrections, however, have a strong impact in terms of polarization through temperature-to-polarization leakage, as discussed in Sect. 4.2 and by Svalheim et al. (2020a).

For this reason, we have implemented two different sampling algorithms for these parameters. First, the mean bandpass correction is sampled with the full time-domain residual on the form

$$\boldsymbol{r}_j = \boldsymbol{d}_j - \boldsymbol{n}_j^{\rm corr} - \mathsf{G}_j\mathsf{P}_j\mathsf{B}_j^{\rm asymm}\left(\boldsymbol{s}_j^{\rm orb} + \boldsymbol{s}_j^{\rm fsl}\right) \tag{108}$$

$$= \mathsf{G}_j\mathsf{P}_j\mathsf{B}_j^{\rm symm}\sum_c \mathsf{M}_{cj}(\beta, \Delta_{\rm bp}^j)\boldsymbol{a}^c + \boldsymbol{n}_j^{\rm w}. \tag{109}$$

Clearly, this residual is highly nonlinear in $\Delta_{\rm bp}$, and no analytic distribution or sampler exist. We therefore once again resort to the Metropolis sampler described in Sect. 6.1. Specifically, we propose small variations to the current mean bandpass shift (while keeping the relative differences between radiometers fixed); we compute the resulting mixing matrices M and sky maps for the new proposals; and we finally then apply the Metropolis acceptance rule as given by the resulting χ^2. Only samples within the small processing mask in Fig. 9 are included in the χ^2. Since mixing matrix updates are computationally expensive, bandpass corrections are among of the most difficult parameters to explore within the entire model. However, as discussed by Svalheim et al. (2020a), the degeneracies between CMB, free-free, AME and $\bar{\Delta}_{\rm bp}$ are too strong to support a robust determination of $\bar{\Delta}_{\rm bp}$ when including only LFI and *WMAP* data. In the final BeyondPlanck production runs, we therefore adopt priors based on the *Planck* 2015 analysis (Planck Collaboration X 2016), which used HFI data to break these degeneracies. In practice, we only apply an overall mean correction of 0.3 GHz to the 30 GHz channel, and no mean corrections to the 44 and 70 GHz channels. In future analyses also including the full HFI set, these priors will obviously be removed.

For the relative bandpass corrections, $\delta_{\rm bp}$, we adopt an alternative approach that is specifically tuned to increase robustness in the final polarization maps. Specifically, after proposing changes to each of the detector-specific bandpasses (under the constraint that their sum vanishes), we compute the resulting *IQUS* map that was defined in Eq. (79) for both the old and new parameter values. Next, we define a special purpose χ^2 of the form

$$\chi^2 = \sum_{j=1}^{N_{\rm det}-1}\sum_p \left(\frac{S_j(p)}{\sigma_j(p)}\right)^2, \tag{110}$$

where $S_j(p)$ is the spurious map corresponding to radiometer j in pixel p, and $\sigma_j(p)$ is the associated uncertainty resulting from the *IQUS* solution. This χ^2 defines the Metropolis acceptance probability as follows,

$$a = \min\left(1, \mathrm{e}^{-\frac{1}{2}(\chi^2_{\rm prop} - \chi^2_{\rm i-1})}\right), \tag{111}$$

where $\chi^2_{\rm prop}$ and $\chi^2_{\rm i-1}$ are the χ^2's of the proposed and previous parameter states, respectively.

Overall, this approach builds on the same fundamental ideas as the *IQUS* approach pioneered by *WMAP* (Page et al. 2007), but using vastly fewer free parameters: Rather than fitting one free parameter per pixel, this algorithm introduces only one additional free parameter per radiometer. To achieve acceptable precision, it instead uses the current foreground model to predict the resulting corrections in each pixel. Thus, while the direct *IQUS* method is not applicable for *Planck* due to its poorly interconnected scanning strategy, our approach implements the same basic idea but without excessively increasing the overall white noise level of the final maps. For further discussion of the method, we refer the interested reader to Svalheim et al. (2020a).

8.3.5. Diffuse component spectral parameter sampling

The fifth conditional distribution in the BeyondPlanck Gibbs chain concerns the foreground SED parameters, $P(\beta \mid \boldsymbol{d}, \omega \setminus \beta)$. Noting that the linear amplitudes $\boldsymbol{a}$ and spectral parameters β are in general highly degenerate for high signal-to-noise models, we employ the same computational trick for intensity sampling as for the gain and correlated noise, and sample these jointly. In practice, this is achieved by first sampling β from the marginal probability distribution with respect to $\boldsymbol{a}$, and then $\boldsymbol{a}$ conditionally on β. For specific details regarding the following algorithm, we refer the interested reader to Andersen et al. (2020). For polarization, we employ a standard Metropolis sampler that is conditional on the foreground amplitudes; see Svalheim et al. (2020b) for details.

For CMB component separation applications, the two-step marginal sampling approach was first described by Stompor et al. (2009) and later implemented in the `Miramare` code by Stivoli et al. (2010). To see how their methodology connects with our notation, as defined by Eq. (69), we can write the relevant residual in the following form,

$$\boldsymbol{r}_j = \left(\boldsymbol{d}_j - \boldsymbol{n}_j^{\rm corr}\right)/g_j - \left(\boldsymbol{s}_j^{\rm orb} + \boldsymbol{s}_j^{\rm sl} + \boldsymbol{s}_j^{\rm mono}\right)$$

$$= g_j\mathsf{P}_j\mathsf{B}_j\boldsymbol{s}_j^{\rm sky}(\beta) + \boldsymbol{n}_j^{\rm w}. \tag{112}$$

The left-hand side in this equation is identical to the residual in Eq. (73), which is the input to the binned mapmaker defined by Eq. (76). Under the assumption of azimuthally symmetric beams,[15] B_j, this expression may therefore be rewritten in terms of binned sky maps on the form

$$\boldsymbol{m}_\nu = \mathsf{A}_\nu(\beta)\boldsymbol{a} + \boldsymbol{n}_\nu^{\rm w}, \tag{113}$$

where $\mathsf{A}(\beta) \equiv \mathsf{B}_\nu\mathsf{M}_\nu(\beta)$ is an effective mixing matrix that accounts for both beam convolution and astrophysical component SEDs. Given this expression, the marginal log-posterior for β then reads (Stompor et al. 2009)

$$P(\beta \mid \boldsymbol{m}) \propto \exp\left(\sum_\nu \left(\mathsf{A}_\nu^t\mathsf{N}_\nu^{-1}\boldsymbol{m}_\nu\right)^t \left(\mathsf{A}_\nu^t\mathsf{N}_\nu^{-1}\mathsf{A}_\nu\right)^{-1}\left(\mathsf{A}_\nu^t\mathsf{N}_\nu^{-1}\boldsymbol{m}_\nu\right)\right). \tag{114}$$

However, the derivation of this expression relies on an assumption of identical beam responses across all frequency channels, and it is therefore necessary to smooth all input maps to a common angular resolution before evaluating this expression. We therefore use this expression only for intensity sampling, coupled to a tuned Metropolis sampler.

[15] In the current BeyondPlanck implementation, we assume azimuthally symmetric beams for all component separation steps, following all previous CMB analysis pipelines.

For polarization, we employ a likelihood given by the original residual defined by Eq. (113),

$$-2\ln P(\beta \mid \boldsymbol{m}, \boldsymbol{a}) = \sum_{\nu} \left(\frac{\boldsymbol{m}_{\nu} - \mathsf{A}_{\nu}(\beta)\boldsymbol{a}}{\sigma_{\nu}(p)} \right)^2 \tag{115}$$

where $\sigma_{\nu}(p)$ is the standard deviation map of channel ν.

When estimating the spectral index of synchrotron emission, we partition the sky into four large disjoint regions, and sample one constant value of β_{s} per region, while still allowing for smooth transitioning between regions. Sky partitioning allows us both to tune the signal-to-noise ratio per fitted parameter, and also to reduce the overall computational cost. All other free spectral parameters are fitted using a single constant value across the full sky. For both temperature and polarization, we employ tuned Metropolis samplers to explore the posterior distribution (Andersen et al. 2020; Svalheim et al. 2020b).

Finally, we note that even with low-dimensional spectral parameter models, it is useful to impose additional priors on β to stabilize the fits. Specifically, we consider two types of priors in the following. First, in order to be able to pre-compute efficient mixing matrix lookup tables for each parameter, we impose a hard uniform prior on each parameter as discussed in Sect. 8.3.8. Second, we impose informative Gaussian priors on β, with parameters informed from the literature; see Andersen et al. (2020) and Svalheim et al. (2020b) for further details.

8.3.6. Diffuse component amplitude sampling

Since we sample β from a marginal distribution with respect to $\boldsymbol{a}$ for the intensity case, we must also sample $P(\boldsymbol{a} \mid \boldsymbol{d}, \omega \backslash \boldsymbol{a})$ directly following β. The relevant data model for $\boldsymbol{a}$ is (similar to β) given by Eq. (113), but this time interpreted as a function of $\boldsymbol{a}$ instead of β. As applied to CMB estimation, this model was first introduced into the CMB literature by Jewell et al. (2004); Wandelt et al. (2004); Eriksen et al. (2004), and later generalized to joint CMB power spectrum estimation and astrophysical component separation by Eriksen et al. (2008). With the uniformized notation defined above, the same formalism applies both to CMB and diffuse astrophysical foregrounds, just with different parametric forms for the mixing matrices, M, signal covariance matrices, S, and optional priors.

Noting that $\boldsymbol{n}_{\nu}^{\mathrm{w}}$ represents Gaussian white noise and $\sum_{\nu} \mathsf{B}_{\nu}\mathsf{M}_{\nu}$ is a deterministic linear operation given $\omega \setminus \boldsymbol{a}$, the appropriate sampling equation for $\boldsymbol{a}$ is yet again given by the multivariate Gaussian sampler in Eq. (A.10) with a template matrix $\mathsf{T} = \sum_{\nu} \mathsf{B}_{\nu}\mathsf{M}_{\nu}$, i.e.,

$$\left(\mathsf{S}^{-1} + \sum_{\nu} \mathsf{M}_{\nu}^{t}\mathsf{B}_{\nu}^{t}\mathsf{N}_{\nu}^{-1}\mathsf{B}_{\nu}\mathsf{M}_{\nu}\right)\boldsymbol{a} = \sum_{\nu} \mathsf{M}_{\nu}^{t}\mathsf{B}_{\nu}^{t}\mathsf{N}_{\nu}^{-1}\boldsymbol{m}_{\nu} + \mathsf{S}^{-1}\mu + \sum_{\nu} \mathsf{M}_{\nu}^{t}\mathsf{B}_{\nu}^{t}\mathsf{N}_{\nu}^{-1/2}\eta_{\nu} + \mathsf{S}^{-1/2}\eta_{0}. \tag{116}$$

Here we have included the signal covariance matrix, $\mathsf{S} = \mathsf{S}(C_{\ell})$, which is a prior that depends on the angular power spectrum of the respective component. If no spatial prior is desired, S^{-1} may simply be set to zero.

Equation (116) arguably represents the single most challenging step in the entire BeyondPlanck analysis pipeline in terms of computational complexity. Fortunately, an efficient iterative solver was recently developed by Seljebotn et al. (2019) for precisely this equation, and this algorithm forms the computational engine of `Commander2` (see Sect. 2.2). The main new idea in that work is the use of a pseudo-inverse preconditioner coupled to a Conjugate Gradient (CG) solver that easily supports multi-resolution observations, as required for Eq. (116). For specific details, we refer the interested reader to Seljebotn et al. (2019).

Computationally speaking, the main complicating factor associated with Eq. (116) is the application of an analysis mask. For CMB likelihood estimation purposes, it is necessary to exclude pixels with particularly bright astrophysical foregrounds by setting $\mathsf{N}_{\nu}^{-1} = 0$, in order not to contaminate the resulting CMB map. Unfortunately, this makes the coefficient matrix on the left-hand side of Eq. (116) poorly conditioned, and the resulting CG search expensive. At the same time, we are also scientifically interested in the properties of astrophysical foregrounds inside the Galactic mask, and simply discarding all this useful information is clearly undesirable.

Rather than directly applying a processing mask, we therefore instead choose to solve Eq. (116) twice. First, within the main Gibbs loop (as defined in Sect. 8.2) we solve Eq. (116) imposing neither a spatial prior on the CMB component, nor an analysis mask. In this configuration the CG search converges typically within $O(10^2)$ iterations, which corresponds to a computational cost that is smaller than the TOD processing steps by one order of magnitude (Galloway et al. 2020a). The resulting CMB sky map samples correspond to prior-free, full-sky CMB maps, similar to those produced by classic component separation algorithms; see, e.g., Planck Collaboration IX (2016) and Planck Collaboration IV (2020).

However, in order to produce the clean full-sky CMB map and power spectrum samples that are required for high-resolution CMB likelihood estimation purposes (see Sect. 8.3.8 and Colombo et al. 2020), we additionally solve Eq. (116) with S^{-1} and a mask, but *condition* on all non-CMB parameters. Statistically speaking, this is equivalent to writing the full joint posterior distribution in Eq. (60) in the form

$$P(\boldsymbol{a}^{\mathrm{CMB}}, \omega \setminus \boldsymbol{a}^{\mathrm{CMB}} \mid \boldsymbol{d}) = P(\boldsymbol{a}^{\mathrm{CMB}} \mid \boldsymbol{d}, \omega \setminus \boldsymbol{a}^{\mathrm{CMB}}) P(\omega \setminus \boldsymbol{a}^{\mathrm{CMB}} \mid \boldsymbol{d}), \tag{117}$$

and using the first main Gibbs loop to draw samples from the second factor on the right-hand side, and the second solution of Eq. (116) to sample from the first factor.

Formally speaking, we note that this approach is only approximate, since C_{ℓ} should in principle also be conditioned upon in the second factor in Eq. (117). The penalty of not doing so is slightly more noise in the non-CMB parameters, since the prior-free CMB sky map sample is less smooth than it is with the prior. However, the practical benefits gained by separating the TOD processing steps from the CMB likelihood estimation step more than outweighs a small increase in statistical uncertainties for several reasons: 1) it greatly reduces overall computational costs for the joint Gibbs chain; 2) it allows CMB estimation from individual frequency channels or channel combinations; and 3) it allows rapid exploration of different analysis masks and/or cosmological models without having to rerun the costly TOD processing steps. Thus, this split plays the same role in the BeyondPlanck pipeline as the split between mapmaking and likelihood estimation does in a traditional CMB analysis pipeline.

We employ a similar trick also for low-resolution likelihood analysis, and re-sample CMB multipoles below $\ell \leq 64$, while conditioning on all higher multipole CMB modes and other parameters. In this case, we do not impose the C_{ℓ} prior term, but rather set $\mathsf{S}^{-1} = 0$ as in the original analysis. This allows us to generate tens of thousands of low-resolution samples at a greatly

reduced computational cost, and derive a well-converged brute-force low-ℓ likelihood from a relatively limited number of full-scale samples. For further details, see Sect. 9.5, Colombo et al. (2020), and Paradiso et al. (2020).

For two of the astrophysical foregrounds, namely free-free emission and AME, we use informative priors to stabilize the model (Andersen et al. 2020). For free-free emission, we adopt the *Planck* 2015 model (Planck Collaboration X 2016) as a spatial template for the prior mean, while the AME prior is based on the *Planck* HFI 857 GHz map, but with a free scaling factor, under the assumption that the AME surface brightness correlates strongly with thermal dust emission (Planck Collaboration X 2016). In both cases, the signal covariance matrices are empirically tuned to allow sufficient variations to statistically fit the data, while at the same time not introducing too many unconstrained degrees-of-freedom.[16]

8.3.7. Compact source sampling

The two previous sections described sampling algorithms for diffuse components (such as CMB, synchrotron or thermal dust emission) in terms of their amplitude and SED parameters. These algorithms are strongly tuned toward global modelling in terms of spherical harmonics expansions through the use of computationally efficient spherical harmonics transforms. However, as discussed in Sect. 3.4.1, a multitude of compact objects also scatters the sky, and some of these are extremely bright. Formally speaking, these may of course also be described in terms of a spherical harmonics decomposition, since the instrumental beam ensures that they are indeed bandwidth limited in the observed data. However, in practice this would require an extremely high bandwidth limit for the diffuse components, and this is therefore impractical because of the high associated computational costs.

Instead, we follow Planck Collaboration IV (2020), and individually model the brightest compact sources based on a pre-existing catalog of object candidates. Each source candidate is mathematically modelled spatially as a delta function convolved with the instrumental beam evaluated at the source location, and with a power-law SED given by an amplitude, $\boldsymbol{a}^{\rm src}$, and a spectral index, α. For *Planck* frequencies, we take into account the full asymmetric beam profiles as evaluated with `FEBeCOP` (Mitra et al. 2011), while for non-*Planck* frequency maps, we adopt azimuthally symmetric beams.

The conditional posterior for the ith compact object is given by subtracting all diffuse components and all other compact objects from the map-based data model in Eq. (113), such that the effective residual at frequency ν reads

$$\boldsymbol{r}_i = \boldsymbol{m}_\nu - \sum_{c\neq i} \mathsf{B}_\nu \mathsf{M}^c_\nu \boldsymbol{a}_c, \tag{118}$$

where c runs both over all diffuse components and all compact objects except the i'th source. The likelihood then takes the form

$$-2\ln P(a_i,\alpha_i \mid \boldsymbol{m},\omega\setminus\{a_i,\alpha_i\}) = \sum_\nu \left(\frac{\boldsymbol{m}_\nu - U_\nu a_i \left(\frac{\nu}{\nu_{\rm ptsrc}}\right)^{\alpha-2}_i \boldsymbol{t}^i_\nu}{\sigma_\nu(p)}\right)^2, \tag{119}$$

where $\nu_{\rm ptsrc}$ is the reference frequency adopted for the point source component, $\boldsymbol{t}^i_\nu$ is the spatial (pre-computed) beam template for the current source, and U_ν is the unit conversion factor for frequency ν. (As usual, bandpass integration is suppressed in the notation for readability, but is of course taken into account in the actual calculations, as described in Sect. 4.2.)

In addition, we impose a Gaussian prior on the spectral index of $P(\alpha) = N(-0.1, 0.3)^2$, motivated by Bennett et al. (2013), and a positivity prior on the amplitude, $a_i \geq 0$.

The full conditional posterior is sampled using a Metropolis sampler for (a_i, α_i), running 100 MCMC steps for each source, while completing 3 full scans through the full source set per full Gibbs cycle. This step represents a relatively minor computational cost, due to extensive pre-computation of both effective beam and bandpass profiles.

8.3.8. C_ℓ and cosmological parameter sampling

The final conditional distribution in the BeyondPlanck Gibbs chain concerns the angular power spectrum, C_ℓ, of each component, possibly as parameterized in terms of a smaller number of general parameters. In the following, we will actually apply this only to the angular CMB power spectrum, but we note that the formalism applies without changes to any other statistically isotropic component, for instance the CIB.

Before we start the discussion, we remind the reader that, as mentioned in Sect. 8.3.6, we apply three different different sampling steps for the CMB amplitude map:

1. full-resolution solution of Eq. (116) with no spatial CMB prior, $\mathsf{S}^{-1}_{\rm CMB} = 0$; the resulting samples are primarily used for CMB prior-free component separation and deriving unbiased frequency maps, but not directly for cosmological parameter estimation;
2. low-resolution solution of Eq. (116) with no spatial CMB prior,[17] $\mathsf{S}^{-1}_{\rm CMB} = 0$, but only including multipoles $\ell \leq 64$, and conditioning on all other parameters; typically, 50 low-resolution samples are drawn based on each high-resolution sample. These samples form the basis for the low-ℓ temperature-plus-polarization CMB likelihood described below.
3. full-resolution solution of Eq. (116) with a spatial CMB prior, $\mathsf{S}^{-1}_{\rm CMB} \neq 0$, where C_ℓ is sampled with an inverse Wishart sampler as summarized below. The resulting samples form the basis for our high-ℓ temperature likelihood.

In practice, the first step is run together with the full Gibbs analysis, including both TOD and component separation steps, while the other two are performed by re-running the code after the main run has been completed. From the point of view of CMB estimation alone, the primary purpose of the main Gibbs run is thus to derive an ensemble of frequency maps and corresponding astrophysical sky models, that later can be re-sampled with respect to CMB parameters.

Low-resolution temperature-plus-polarization likelihood
From step 2 above, we typically have a sample set of $O(10^4)$ CMB-only samples, each corresponding to one possible combination of TOD, foreground and high-ℓ CMB parameters. Clearly, the information contained in this sample set may be combined into an effective CMB likelihood in many different

[16] Note that S plays a fully analoguous role in a multi-variate Gaussian prior as the usual standard deviation in a univariate Gaussian prior, and can be used to adjust the strength of the prior.

[17] In practice, we do formally apply a prior also in this case, but with a sufficiently large numerical value that $\mathsf{S}^{-1}_{\rm CMB} \approx 0$.

ways, each with its own algorithmic advantages and disadvantages. For instance, they could form the basis of a highly robust cross-spectrum estimator, by analysing two halves of the data set at a time, and cross-correlating the resulting CMB map; for a recent example of such cross-spectrum approach applied to the *Planck* data, see, e.g., Planck Collaboration V (2020).

However, since our main goal of this paper is to introduce the end-to-end Bayesian approach as such, rather than deriving the most sensitive CMB likelihood possible, we prefer in this paper to stay as close as possible to the exact Bayesian solution. And, practically speaking, that corresponds most closely to a Gaussian multivariate distribution on the form,

$$P(C_\ell \mid \hat{\mathbf{s}}_{\rm CMB}) \propto \frac{\mathrm{e}^{-\frac{1}{2}\hat{\mathbf{s}}^t_{\rm CMB}(\mathsf{S}(C_\ell)+\mathsf{N})^{-1}\hat{\mathbf{s}}_{\rm CMB}}}{\sqrt{|\mathsf{S}(C_\ell)+\mathsf{N}|}}, \tag{120}$$

where $\hat{\mathbf{s}}_{\rm CMB}$ represents a CMB-plus-noise map and N is its corresponding effective noise covariance map.[18] Since we at this point have access to a full ensemble of low-resolution CMB samples that span the full allowed posterior volume, we may estimate these quantities as

$$\hat{\mathbf{s}}_{\rm CMB} = \left\langle \mathbf{s}^i_{\rm CMB} \right\rangle \tag{121}$$

$$\mathsf{N} = \left\langle (\mathbf{s}^i_{\rm CMB} - \hat{\mathbf{s}}_{\rm CMB})(\mathbf{s}^i_{\rm CMB} - \hat{\mathbf{s}}_{\rm CMB})^t \right\rangle, \tag{122}$$

where brackets indicate average over the sample set. In the limit of an infinite number of samples, these quantities will converge to the Gaussian approximation of the full pixel-based CMB posterior.

This approach is conceptually very similar to that adopted by both the *Planck* LFI DPC Planck Collaboration V (2020) and the *WMAP* science team Hinshaw et al. (2013) for low-ℓ likelihood estimation, both of which rely on brute-force likelihood estimation according to Eq. (120). However, there is one critically important difference: with our approach, all sources of uncertainty that are sampled over in the Gibbs chain with ω are seamlessly propagated to the CMB likelihood, including gain and bandpass uncertainties; foreground uncertainties; correlated noise etc. For the traditional approaches, typically only correlated noise and overall calibration is accounted for in the covariance matrix.

An important question regarding the practicality of Eq. (120) is how many samples are required for convergence. As discussed by Sellentin & Heavens (2016), an absolute minimum criterion for a sampled $n \times n$ covariance matrix simply to be invertible is that $N_{\rm samp} > n$. However, this is by no means sufficient to obtain a *robust* estimate, and, more typically, numerical experiments indicate that many times this is required for matrices of moderate size and relatively weak correlations; the precise value, however, is something that must be tested on a case-by-case matrix.

In any case, since we have a relatively limited number of samples available, it is of great interest to compress the relevant information in $\hat{\mathbf{s}}_{\rm CMB}$ into as few spatial modes as possible, while still retaining the lion's share of its full information content. With this in mind, we note that the main scientific target for low-ℓ likelihood estimation for *Planck* is the reionization optical depth, τ. In this case, τ typically only depends on the first 6 or 8 multipoles, because of the limited sensitivity of the instrument (Planck Collaboration V 2020). As such, a first natural compression is to retain only modes with $\ell \leq 8$, which corresponds to a total of $3(\ell_{\rm max}+1)^2 \approx 240$ modes. However, many of these modes fall within a typical analysis mask (Colombo et al. 2020), and therefore carry no statistical weight in the final answer.

One particularly convenient method of isolating the actually useful modes is through Karhunen-Loève compression, as discussed by Tegmark et al. (1997) and Gjerløw et al. (2015). This approach essentially corresponds to retaining the eigen-vectors of $\mathsf{S}+\mathsf{N}$ with the highest eigenvalues, where S is evaluated for a typical model of interest. Adopting the notation of Gjerløw et al. (2015), we organize the eigenmodes with eigenvalues higher than some user-specified threshold row-by-row into a projection operator, P, and apply this to the CMB samples derived above. The compressed data and covariance matrix then reads

$$\tilde{\mathbf{s}}_{\rm CMB} = \mathsf{P}\hat{\mathbf{s}}_{\rm CMB} \tag{123}$$

$$\tilde{\mathsf{N}} = \mathsf{P}\mathsf{N}\mathsf{P}^t \tag{124}$$

$$\tilde{\mathsf{S}} = \mathsf{P}\mathsf{S}\mathsf{P}^t. \tag{125}$$

Adopting a multipole threshold of $\ell_{\rm max} = 8$ and a signal-to-noise threshold of 10^{-6} typically leaves around 170 spatial modes in the full data set, for which we that convergence is typically reached with about 30 000 fast samples, corresponding to 600 full samples including all systematic effects; see Sect. 9.5.3 and Colombo et al. (2020). The computational cost of a single likelihood evaluation is also correspondingly reduced because of this compression, and only takes a few hundredths of a second.

High-resolution Blackwell-Rao estimator The above estimator can only be employed at low angular resolution because of its strong dependence on the size of the covariance matrix. For high angular resolution analysis, we use another well-established solution, namely the Blackwell-Rao estimator (Chu et al. 2005), which works very well for high signal-to-noise data. In practice, we only use this for temperature analysis in the current paper, since the signal-to-noise ratio for high-ℓ polarization is very low with only LFI and *WMAP* data. However, we keep the following presentation general, such that it can be used for both temperature and polarization analysis for other experiments.

To derive the appropriate sampling algorithm for $P(C_\ell \mid \boldsymbol{d}, \omega \setminus C_\ell)$ from first principles, we first note that $P(C_\ell \mid \boldsymbol{d}, \omega \setminus C_\ell) = P(C_\ell \mid \boldsymbol{a}^{\rm CMB})$; if the true CMB map, $\boldsymbol{s}_{\rm CMB}$, is perfectly known, then no further knowledge regarding the measured data can possibly provide more useful information about the angular CMB power spectrum, C_ℓ. Second, as discussed in Sect. 3.2, we assume that the CMB fluctuation field is isotropic and Gaussian distributed, and its probability distribution is therefore given by Eq. (12). Noting that individual $a_{\ell m}$'s are statistically independent by the assumption of isotropy, we can therefore write

$$P(C_\ell \mid \boldsymbol{a}^{\rm CMB}) \propto P(\boldsymbol{a}^{\rm CMB} \mid C_\ell) P(C_\ell) \tag{126}$$

$$= \prod_{m=-\ell}^{\ell} \frac{\mathrm{e}^{-\frac{1}{2}\boldsymbol{a}^\dagger_{\ell m}\mathsf{C}_\ell^{-1}\boldsymbol{a}_{\ell m}}}{\sqrt{|\mathsf{C}_\ell|}} P(\mathsf{C}_\ell) \tag{127}$$

$$= \frac{\mathrm{e}^{-\frac{2\ell+1}{2}\mathrm{tr}(\sigma_\ell \mathsf{C}_\ell^{-1})}}{|\mathsf{C}_\ell|^{\frac{2\ell+1}{2}}} P(\mathsf{C}_\ell), \tag{128}$$

[18] We note that this expression does not correspond to the exact Bayesian solution, strictly speaking, because the true uncertainty of a given pixel may be non-Gaussian due to the presence of both foregrounds and TOD corrections. To account for this, cosmological parameters should ideally be sampled within the full-resolution Gibbs chain, for instance using the algorithms proposed by Racine et al. (2016); this, however, is left for future work, and we adopt a Gaussian approximation for now.

where $\boldsymbol{a}_{\ell m} = \{a^T_{\ell m}, a^E_{\ell m}, a^B_{\ell m}\}$ and

$$\mathsf{C}_\ell \equiv \begin{bmatrix} C^{TT}_\ell & C^{TE}_\ell & C^{TB}_\ell \\ C^{TE}_\ell & C^{EE}_\ell & C^{EB}_\ell \\ C^{TB}_\ell & C^{EB}_\ell & C^{BB}_\ell \end{bmatrix}; \tag{129}$$

$$\sigma_\ell \equiv \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} \begin{bmatrix} (a^T_{\ell m})^* a^T_{\ell m} & (a^T_{\ell m})^* a^E_{\ell m} & (a^T_{\ell m})^* a^B_{\ell m} \\ (a^E_{\ell m})^* a^T_{\ell m} & (a^E_{\ell m})^* a^E_{\ell m} & (a^E_{\ell m})^* a^B_{\ell m} \\ (a^B_{\ell m})^* a^T_{\ell m} & (a^B_{\ell m})^* a^E_{\ell m} & (a^B_{\ell m})^* a^B_{\ell m} \end{bmatrix}. \tag{130}$$

We typically adopt uniform priors on C_ℓ (although for a discussion of non-uniform priors, see Wandelt et al. 2004 and Larson et al. 2007), and the distribution in Eq. (128) is then known as the inverse Wishart distribution, which has a very simple sampling algorithm (Larson et al. 2007):

1. Draw $2\ell - n$ Gaussian random vectors, η_i, from the empirical covariance matrix $(2\ell+1)\sigma_\ell$, each of length n, where n is the dimension of C_ℓ;
2. Compute the outer product of these vectors, $\rho_\ell = \sum_{i=1}^{2\ell-n} \eta_i \eta_i^t$;
3. Set $\mathsf{C}_\ell = \sigma_\ell / \rho_\ell$.

Note that if C is block-diagonal, as for instance is the case if $C^{TB}_\ell = C^{EB}_\ell = 0$, then this algorithm should be applied separately block-by-block. Also, if binning is desired, for instance to increase the effective signal-to-noise ratio of a given power spectrum coefficient, this is most conveniently done in terms of $D_\ell = C_\ell \ell(\ell+1)/2\pi$; for details, see Larson et al. (2007).

In the current work, which applies this only to the CMB temperature power spectrum, we only bin multipoles above $\ell \geq 600$, where we apply uniform binning of $\Delta\ell = 50$. To improve the Monte Carlo mixing in the low signal regime, we interleave the straight inverse Wishart sampler with the joint $(a_{\ell m}, C_\ell)$ re-scaling MCMC sampler introduced by Jewell et al. (2009), where the proposal rule is defined as

$$a^{\rm new}_{\ell m} = \sqrt{\frac{C^{\rm new}_\ell}{C^{\rm old}_\ell}}\, a^{\rm old}_{\ell m}. \tag{131}$$

Essentially, this proposal corresponds to a move along the diagonal of the joint $(a_{\ell m}, C_\ell)$ space, and is efficient wherever these two parameters are highly degenerate, which is the case in the low signal-to-noise regime. However, we do note that for final cosmological parameter constraints, we will only use the LFI-based likelihood presented here up to $\ell < 600$, and rather use the HFI-based *Planck* 2018 likelihood at higher multipoles, where appropriate.

The above algorithm describes an efficient Gibbs-based approach to CMB power spectrum sampling, as originally suggested by Wandelt et al. (2004). The product from this procedure is a set of joint samples $(s_{\rm CMB}, C_\ell)_i$. However, the algorithm does not specify how to constrain cosmological parameters from these samples. Indeed, many different approaches may be adopted for this purpose, each making different assumptions and choices with regard to computational cost and robustness to systematic errors. Some approaches presented in the literature include

- *the Blackwell-Rao estimator* (Chu et al. 2005): Direct averaging over σ_ℓ samples given the analytic smoothing kernel in Eq. (128). Exact, but converges slowly in low signal-to-noise regime. Used by *WMAP* low-ℓ TT likelihood (Hinshaw et al. 2013).
- *the Gaussianized Blackwell-Rao estimator* (Rudjord et al. 2009): Multivariate Gaussian approximation to the above, following a Gaussian change-of-variable defined by univariate marginal distribution. Converges much faster than direct Blackwell-Rao estimator, and is highly accurate for typical masks. Used by *Planck* low-ℓ TT likelihood (e.g., Planck Collaboration V 2020).
- *joint Metropolis-Hastings sampling* of $\{\boldsymbol{a}_{\ell m}, \mathsf{C}_\ell\}$ (Jewell et al. 2009; Racine et al. 2016): Efficient in both low and high signal-to-noise regimes; may be applied to both C_ℓ and cosmological parameter estimation.

The first two of these methods define a CMB power spectrum likelihood function, $\mathcal{L}(\mathsf{C}_\ell)$, which then must be coupled to a cosmological parameter estimation code. We employ the widely employed `CosmoMC` (Lewis & Bridle 2002) code for this purpose, as detailed in Paradiso et al. (2020). In contrast, when applied to cosmological parameter estimation, the third method requires a means to convert between cosmological parameters and angular power spectra, such as `CAMB` (Lewis et al. 2000). In this paper, we adopt the Gaussianized Blackwell-Rao estimator as our default solution, and leave the full integrated MCMC sampling approach for future work.

8.4. Computational requirements and optimization

The end-to-end algorithm summarized in the last few sections represents a significant computational challenge, both in terms of fundamental hardware requirements and in terms of software optimization. In this section we briefly review some critical computational features implemented in the current code, while in-depth presentations are provided by Galloway et al. (2020a) and Gerakakis et al. (2020). In addition, we highly recommend the interested reader to consult the source code.[19] At the same time, we emphasize that these codes are most definitely works in progress, and still undergo rapid development. Nearly every single component and function have room for further improvement and optimization. However, it is our hope and intention that by providing all codes to the general community under an open-source license, new collaborations, efforts and ideas will emerge, and this will leading to more mature, efficient and generally applicable code.

With these caveats in mind, Table 2 summarize the overall computational cost of the current implementation, both in terms of initialization and cost per sample. These benchmarks were obtained by running the pipeline on a single compute node with 72 Intel Xeon E7-8870 2.10 GHz cores and 1.5 TB of RAM. All time related costs are provided in units of wall-time, and must therefore be multiplied with 72 to convert to CPU time.

Overall, the computational complexity of the BeyondPlanck Gibbs sampler is determined by three fundamentally different types of operations. First, the low-level analysis is dominated by TOD memory management. Second, the high-level amplitude sampling step is dominated by spherical harmonic transforms. Third, the spectral index sampling step is dominated by map-based operations, typically either spherical harmonic transforms or χ^2 evaluations. Efficient parallelization of each of these three types of operations is therefore the critical design driver for the current implementation. We now briefly review how the BeyondPlanck pipeline optimizes each of these aspects, and refer the interested reader to Galloway et al. (2020a) for further details.

[19] The BeyondPlanck software is available under a GNU Public Library (GPL) open-source license at https://github.com/Cosmoglobe/Commander.

Table 2. Computational resources required for end-to-end BeyondPlanck processing. All times correspond to wall-clock duration while running on a single 72-core node. All times are also approximate, and vary by $\lesssim 5\,\%$ from sample to sample; the values reported correspond to one specific but arbitrarily selected sample. We also note that the TOD processing stage is memory-bus limited, and the reported run-times are therefore significantly correlated between various operations.

Item	30 GHz	44 GHz	70 GHz	Sum
Data volume				
Uncompressed data volume	761 GB	1 633 GB	5 522 GB	7 915 GB
Compressed data volume/RAM requirements	86 GB	178 GB	597 GB	**861 GB**
Processing time (cost per run)				
TOD initialization/IO time	176 sec	288 sec	753 sec	1217 sec
Other initialization				663 sec
Total initialization				**1880 sec**
Gibbs sampling steps (cost per sample)				
Data decompression	36 sec	105 sec	252 sec	393 sec
TOD projection (P operation)	33 sec	49 sec	248 sec	330 sec
Sidelobe evaluation ($\boldsymbol{s}_{\rm sl}$)	58 sec	85 sec	337 sec	480 sec
Orbital dipole ($\boldsymbol{s}_{\rm orb}$)	45 sec	61 sec	343 sec	449 sec
Gain sampling (g)	13 sec	10 sec	71 sec	94 sec
Correlated noise sampling ($\boldsymbol{n}_{\rm corr}$)	355 sec	390 sec	2393 sec	3138 sec
TOD binning (P^t operation)	22 sec	34 sec	442 sec	498 sec
Loss due to poor load-balancing	62 sec	305 sec	135 sec	502 sec
Sum of other TOD steps	32 sec	135 sec	139 sec	306 sec
TOD processing cost per sample	656 sec	1074 sec	4666 sec	**6396 sec**
Amplitude sampling, $P(\boldsymbol{a} \mid \boldsymbol{d}, \omega \setminus \boldsymbol{a})$				527 sec
Spectral index sampling, $P(\beta \mid \boldsymbol{d}, \omega \setminus \beta)$				1080 sec
Other steps				149 sec
Total cost per sample				**8168 sec**

8.4.1. Low-level optimization

Starting with the low-level TOD-oriented operations, we first note in Table 2 that the full data volume of four years of *Planck* LFI observations is 8 TB. This number includes all science and housekeeping data. A single read of the full data set from spinning disks on a typical intermediate-sized high-performance computing (HPC) cluster therefore requires a few hours of wall time, assuming $O(1\,\mathrm{GB\,s^{-1}})$ read speed. While acceptable as a one-time initialization cost, integrating such expenses into the Gibbs loop clearly leads to impractical run times. A first requirement for efficient end-to-end TOD processing is thus that the entire data set may be stored in RAM. Likewise, noting that the memory bus from the RAM to the CPU is relatively slow compared to CPU operations, a corollary requirement is that the overall memory footprint should be aggressively minimized.

With these observations in mind, we first choose to read only those parts of the data that are strictly required for the analysis in question; all unnecessary housekeeping data are omitted. For each *Planck* LFI radiometer the only retained quantities therefore include 1) differenced detector voltages, $\boldsymbol{d}_t$ (one float per sample); 2) pointing, P_t (three double precision values per sample); and 3) flags, f_t (one integer per sample). Nominally, a total of 32 bytes/sample/radiometer are required to store the TOD information.

However, as detailed by Galloway et al. (2020a), because the pointing and flags are both very smooth functions of time, they lend themselves to highly efficient compression. We exploit this by transforming and discretizing each relevant quantity into integers; taking the difference between consecutive samples to minimize their dynamic range; and finally Huffman compressing (Huffman 1952) the resulting time streams, i.e., we assign bit patterns of variable lengths to each integer according to their relative frequency. The average number of bits per sample is thus reduced by a factor of 5–6. These compressed TOD arrays are then stored in memory PID-by-PID, and only decompressed when needed. The total data volume is in this way reduced from 8 TB to 861 GB, which fits into the RAM of a single modern compute node. The decompression cost accounts for about 5 % of the total analysis wall time, which we consider well worth the memory savings. However, as discussed by Galloway et al. (2020a), this compression does have notable implications in terms of the overall Gibbs sampling structure, as the full decompressed TOD set can never be stored in memory at once, nor is it possible to store multiple copies of the TOD. Accordingly, careful relative ordering of the various Gibbs sampling steps is necessary. In practice, four full scans are made through the entire TOD within each Gibbs iteration, where each scan corresponds to sampling one global TOD-related parameter, namely three three gain components (see Sect. 8.3.1) and the bandpass correction parameter; none of these can be sampled simultaneously without breaking the Gibbs chain.

Next, the low-level parallelization scheme for TOD processing is organized according to PIDs, such that each computing core processes a distinct subset of PIDs. Load balancing is achieved by first measuring the effective computing time for each PID, and then distributing them according to cost in a round-robin manner among the various computing cores.

Inspecting the costs of individual steps in Table 2, we see that that dominant TOD operation is associated with sampling $\boldsymbol{n}_{\rm corr}$, which makes intuitive sense: While most operations scale linearly in the number of samples, $O(N_{\rm tod})$, the correlated noise step requires two Fourier transforms, and therefore scales as $O(N_{\rm tod} \log N_{\rm tod})$. To optimize this step, we first of all employ the `FFTW` library (Frigo & Johnson 2005) for all FFT operations. Second, we note that the speed required for a single FFT transform depends sensitively and highly non-linearly on $N_{\rm tod}$. Values of $N_{\rm tod}$ that happen to factorize into particularly favorable combinations of primes may happen to be, say, three to five times faster than neighboring values. We exploit this by first measuring the

time required per FFT for every length between 1 and 10^6, and construct a table of optimal lengths, with at least one value per 100th sample. At read time, we then truncate the length of each PID until it equals the closest lower optimal length. As such, we lose on average one second of data per PID, corresponding to about 0.03 % of the total data volume, while gaining a factor of three or more in overall TOD processing time.

After the FFT-based operations, the dominant TOD operations are the sidelobe and orbital dipole evaluations, as well as the pointing projections operators, P and P^t. Here it is worth noting that the TOD analysis is currently memory-bus limited. That is, the cost is associated simply with transferring data from RAM into the CPU. As such, the specific algorithmic details of each step are largely irrelevant, and the important factor is simply the total data volume. To improve the performance of these steps, the best approach would be to run across multiple nodes, which thereby increase the number of memory buses available. On the other hand, this also leads to lower performance for the CPU dominated operations, and most notably the spherical harmonics transforms. A future optimal solution should implement a better tuned parallelization strategy where SHTs are parallelized within nodes, while TOD operations are parallelized across nodes; this is left for future development.

Next, the two TOD projection operators warrant a few comments. First, we recall that P converts a map into a time stream. This represents a computational challenge in itself, because each core then needs access to all pixels in the map. However, actually storing the full map per core would require substantial amounts of memory. To solve this, we exploit a MPI-3 shared memory feature, and only store one copy of the map per compute node, rather than one per core. However, we do observe that the memory access speed associated with these shared-memory arrays is typically five times slower than for local arrays, and further optimizations are therefore possible.

In contrast, the P^t operation co-adds samples in a time-stream into a map. In terms of practical code, this is a more complex operation than P, since all cores need to update the values stored in each sky map pixel, not only read them. This can easily lead to race conditions in which different cores simulatenously write to the same parts of memory, resulting in corrupt data, and a direct shared array approach is therefore impractical. At the same time, allocating a full sky map per core is not an option due to the same memory constraints discussed above. As a compromise, we instead first scan the full pointing stored by each core, and accumulate a list of all locally observed pixels. Due to the sparse *Planck* scanning strategy, this typically amounts to only 5–10 % of all pixels for each core. Allocating and maintaining a sub-map of this limited size is acceptable in terms of total memory footprint. Co-addition over cores is then achieved using a combination of shared arrays within each computing node, and a single `MPI_ALLREDUCE` operation between nodes. Clearly, further optimization is very likely possible also with respect to this operation.

8.4.2. High-level parallelization and optimization

Next, we consider optimization of the high-level routines, and in particular of the amplitude and spectral index sampling steps. These are largely overlapping in terms of essential low-level routines, and so we will also discuss them jointly.

The single most important computational routine involved in these operations is the spherical harmonics transform, needed both for solving the Wiener filter defined by Eq. (116) and for smoothing maps to a common angular resolution as required for Eq. (114). Indeed, the importance of this operation is so critically important that we base our entire map parallelization strategy of our codes around it. With this in mind, we adopt the `libsharp2` (Reinecke & Seljebotn 2013) spherical harmonics library for all harmonic transforms, which is the most efficient library for this purpose available today. This library is based on a deep parallelization level in both pixel and harmonic space, distributing both constant-latitude rings and constant-m harmonics across different cores. We adopt these parallelization conventions without modification.

The second most important operation involved in these operations is multiplication with the mixing matrix, $\mathsf{M}(\beta;\Delta_{\rm bp})$. As described in Sect. 4.2, this expression involves integration of an ideal parametric SED with the bandpass of each instrumental detector. It also varies from pixel-to-pixel, depending on the local properties of the spectral parameters, β. For this reason, we precompute the full mixing matrix prior to each full amplitude sampling step, pixel-by-pixel. Taking advantage of the `libsharp` parallelization scheme, which distributes rings across all available cores, the memory requirements for this is fairly limited. Furthermore, employing the spline-based library discussed in Sect. 4.2, the actual evaluation of this matrix only carries a cost equal to a polynomial evaluation per pixel. However, it is important to note that actually changing the bandpass correction parameters, $\Delta_{\rm bp}$, requires a full re-evaluation of the underlying splines, as well as all higher-level mixing matrices, and this particular operation is therefore very computationally intensive. As a result, it is done as infrequently as possible.

Finally, as described above, many of the various sampling steps are carried out with a standard Metropolis sampler. Although conceptionally and implementationally straightforward, this sampler does have the drawback of requiring specific tuning of the step size to be efficient. For most of these samplers, we therefore typically run a short tuning chain during the first iteration, if the computational cost of the sampler is limited (which, for instance, is the case for the point source sampler), or insert a pre-calculated proposal matrix into the run through a parameter file (which, for instance, is the case for the bandpass correction sampler). Such tuning is essential to achieve acceptable mixing for the overall chain.

9. Results

We are now finally ready to present the main results resulting from applying the algorithms summarized in Sects. 7–8 to the data combination described in Sect. 5. For the analysis shown here, we produce a total of six independent Monte Carlo Markov chains of samples drawn from the posterior distribution $P(\omega|\boldsymbol{d})$, as described in Sect. 8.2. Each chain has different length but includes at least 225 samples, and we discard the first 25 samples for burn-in. For applications where comparison between chains is useful, we thus retain 200 samples from each chain, for a total of 1200 accepted samples, while for applications where sample order does not matter, we use the full set of accepted samples. With a computational cost of 2.3 wall hours/sample (see Table 2), this set took about three weeks of continuous run time to produce on six nodes, for a total computational cost of 220 000 CPU hours. Although not directly comparable, it is still interesting to note that the production of the *Planck* FFP8 simulation set required a total of 25 million CPU hours, and the cost of constructing only a single component of a single Monte Carlo realization of the 70 GHz channel cost 9360 CPU-hours (Planck Collaboration XII 2016). The full analysis shown in the following thus carries a total computational cost that is equivalent to

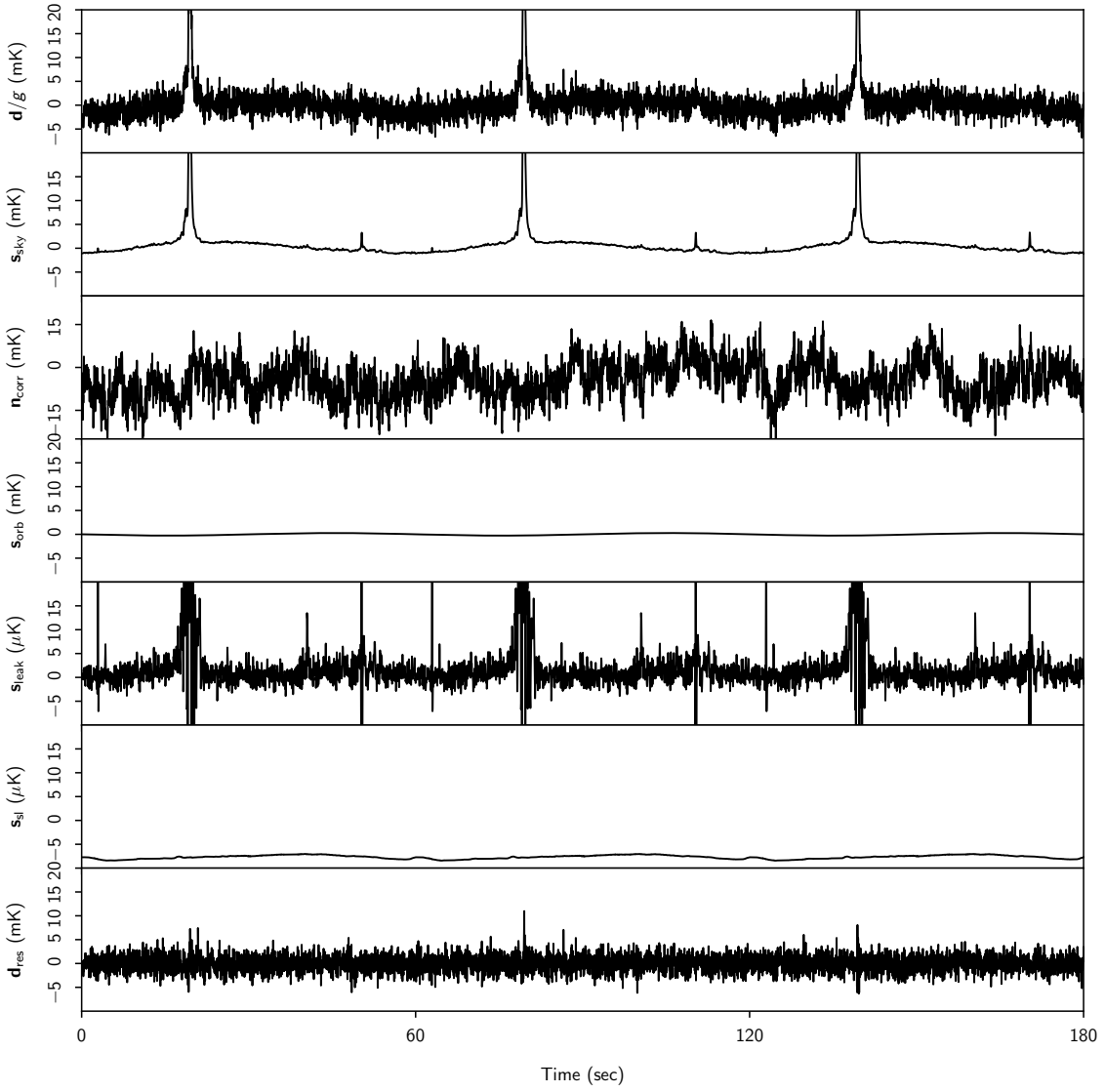

Fig. 10. Time-ordered data segment for the 30 GHz LFI 27M radiometer. From top to bottom, the panels show 1) raw calibrated TOD, $\boldsymbol{d}/g$; 2) sky signal, $\boldsymbol{s}_{\rm sky}$; 3) calibrated correlated noise, $\boldsymbol{n}_{\rm corr}/g$; 4) orbital CMB dipole signal, $\boldsymbol{s}_{\rm orb}$; 5) leakage mismatch correction, $\boldsymbol{s}_{\rm leak}$; 6) sidelobe correction, $\boldsymbol{s}_{\rm sl}$; and 7) residual TOD, $\boldsymbol{d}_{\rm res} = (\boldsymbol{d} - \boldsymbol{n}_{\rm corr})/g - \boldsymbol{s}_{\rm sky} - \boldsymbol{s}_{\rm orb} - \boldsymbol{s}_{\rm leak} - \boldsymbol{s}_{\rm sl}$. Note that the units differ between panels by a factor of 1000, such that each panel spans either a range of 30 mK or 30 μK depending on the respective signal amplitude.

$O(10)$ *Planck* FFP8 70 GHz simulations. This clearly demonstrates the computational feasibility of the Bayesian end-to-end approach, and the algorithms shown here do not require the use of a massive super-computer center to be useful. At the same time, it is also clear that future work should concentrate on increasing the concurrency of the current implementation through better parallelization schemes, such that the wall time can be reduced to hours or days, as opposed to weeks and months, when more resources are available.

The rest of the section is organized as follows. We first review the various low-level instrumental parameters in Sect. 9.1. In Sect. 9.2 we consider the impact of the various time-domain corrections in the pixel domain, as well as the resulting frequency maps after taking these corrections into account. We discuss goodness-of-fit in Sect. 9.3, before presenting astrophysical sky model parameters in Sect. 9.4 and CMB posterior distributions in Sect. 9.5. We note that this presentation is not intended to be comprehensive, but rather aims to provide a broad overview of the full process, and build intuition regarding the interpretation and interplay between the various components. Individual results are discussed in detail in the companion papers listed in Table 1.

9.1. Instrumental parameters

We start our review by inspecting the data and model at the lowest level, and for this purpose we select the 30 GHz channel as a reference test case, for which the foreground contamination is the largest, and therefore the calibration challenge the hardest. The top panel of Fig. 10 shows a 3 minute chunk of the 30 GHz LFI 27M TOD, in which the only pre-processing steps are differencing with the 4 K load signal and ADC corrections (see Sect. 5 for details). In addition, for comparison purposes, we scale the entire TOD segment by the inverse gain, converting the signal from volts to μK; we recall, however, that the gain model is constant over this whole section of data, and this does therefore not affect the visual structure, but only the plotting scale. Each panel in Fig. 10 is plotted with a range of either 30 mK or 30 μK, to allow direct visual comparison within groups of similar quantities.

In this plot, we can readily identify by eye many of the features discussed in Sect. 7. First, we see a slow sinusoidal oscillation with a peak-to-peak amplitude of about 3 mK; this is the signature of the solar CMB dipole modulated by the *Planck* scanning strategy. The second most obvious feature is a sharp peak at $t \approx 15$ sec (and periodically repeating with 60 sec intervals), as well as a smaller peak at $t \approx 45$ sec. These correspond to pointings where the satellite points toward the Galactic plane, either near the center or near the anti-center. Beyond that, the only other visually significant fluctuations are due to instrumental noise.

The second panel shows the total sky signal model projected into time-domain, $\mathsf{P}\boldsymbol{s}^{\rm sky}$. Here we can see the same signal features noted above, both the solar dipole and the Galactic plane crossings, only with much higher signal-to-noise, as the total sky model is derived from the full data set, not just the 30 GHz LFI 27M radiometer. Indeed, here we also note some smaller fluctuations superimposed on the dipole signal, and these are primarily higher-order CMB temperature fluctuations. Comparing the amplitude of these fluctuations with the noise level seen in the top panel provides some intuition regarding the amount of data coaddition that is required in order to measure true CMB temperature variations, and the corresponding requirements regarding instrument stability.

The third panel shows the (calibrated) correlated noise component, $\boldsymbol{n}^{\rm corr}/g$; the calibration is performed here only in order to be able to compare its magnitude with the signal components in units of μK. Overall, we see that the correlated noise spans roughly the same range as the sky signal component, but exhibits a distinct $1/f$ type correlation structure in time, and it does not depend on the pointing. In combination, these two facts allow separation of correlated noise from true sky signal.

The fourth panel shows the orbital CMB dipole, caused by the satellite's motion around the Sun, $\boldsymbol{s}^{\rm orb}$. Although its amplitude appears visually small in this plot compared to the signal and correlated noise terms in the second and third panels, it is important that recall that there are no free parameters or uncontrolled uncertainties associated with this particular component. It is therefore possible to correlate its apparent amplitude with the raw data over the entire mission, and thereby derive one absolute calibration value.

The fifth panel shows the bandpass and beam leakage correction, $\boldsymbol{s}^{\rm leak}$, computed as the difference between the sky signal as observed by the 30 GHz LFI 27M radiometer and the mean over all 30 GHz radiometers. Note that the scale of this panel is a factor of one thousand smaller than those of the previous panels. In this plot, we see the impact of the bandpass mismatch in the form of broad spikes centered on the Galactic plane. The magnitude of this effect can reach several tens of μK at low Galactic latitudes, but is almost negligible at high Galactic latitudes. The rapid fluctuations are due to beam mismatch between the various

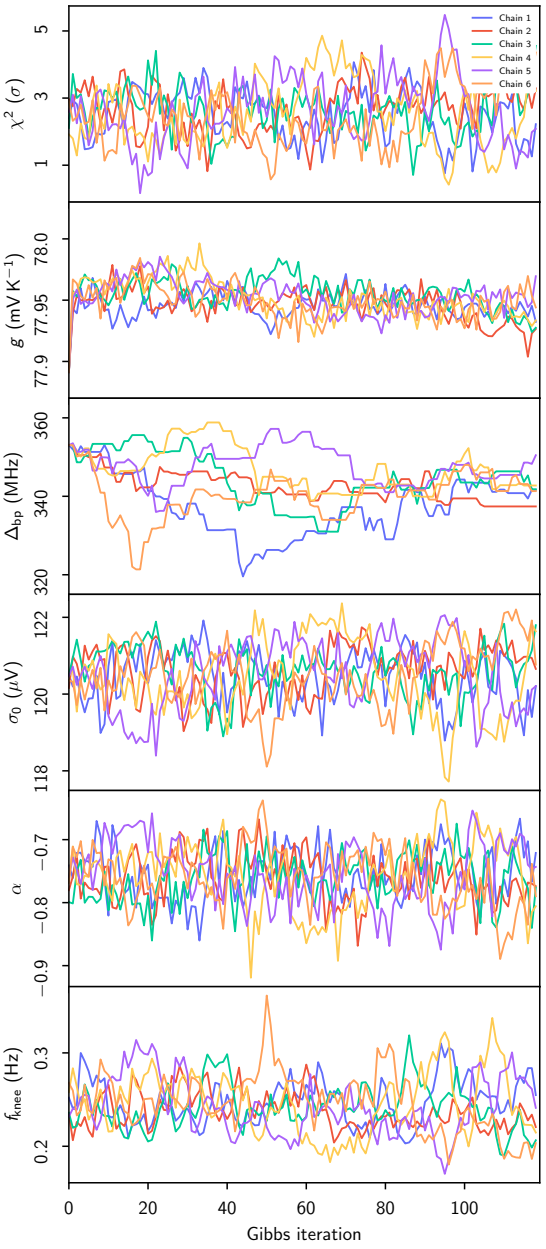

Fig. 11. Example of TOD parameters Gibbs chain for the 30 GHz LFI 27M radiometer. From top to bottom, the panels show normalized reduced χ^2 for a single Pointing Period; gain for the same PID in units of mV K^{-1}; bandpass correction in MHz; white noise level, σ_0; correlated noise slope, α; and correlated noise knee frequency, $f_{\rm knee}$. The six different colored curves correspond to six independent Gibbs chains.

30 GHz detectors, and correspond to CMB fluctuations observed through slightly different beam solid angles. The standard deviation of this component is on the order of 1 μK.

The sixth panel shows the contribution from the far sidelobes, $s^{\rm sl}$. Overall, the two dominant structures in this function are a slow dipole modulation, caused by the solar and orbital CMB dipoles, and a weak imprint of the Galactic plane crossings at about 60 and 120 sec. For this particular combination of radiometer and PID, the overall magnitude of the sidelobe correction is rather limited, although in other cases it may be an order of magnitude brighter than what is seen here.

The bottom panel in Fig. 10 shows the residual resulting from subtracting each of the model terms in Eq. (69) from the raw data. Ideally, this should be uncorrelated stationary Gaussian noise with a standard deviation equal to σ_0. This hypothesis will be tested explicitly in Sect. 9.3, where we compute the χ^2 per PID for each radiometer throughout the mission. In practice, however, correlated features may appear in this residual, and these translate into a bias in the sky map. In this particular example, we see a notable residual near the Galactic plane crossings, which eventually should be masked in higher-level analyses.

Figure 10 represents one single Gibbs sample in the full chain. In contrast, Fig. 11 shows samples from all six Gibbs chains for the instrumental parameters for the same PID, but this time plotted as a function of Gibbs iteration. For perfect Markov chain mixing, these should all scatter around a well-defined mean value with a short correlation length.

The top panel shows the normalized reduced χ^2 for a single PID (arbitrarily selected to be PID = 5000), as defined by

$$\chi^2 \equiv \frac{\sum_{t=1}^{N_{\rm tod}} \left(\frac{d_t - s_t^{\rm tot}}{\sigma_0}\right)^2 - N_{\rm tod}}{\sqrt{2N_{\rm tod}}}. \tag{132}$$

Recalling that the χ^2 distribution with n degrees of freedom converges towards a Gaussian with mean equal to n and variance equal to $2n$, this quantity should be approximately distributed as $N(0,1)$ for ideal data, with deviations measured in units of σ. We adopt this χ^2 as a convenient goodness-of-fit measure.

Regarding the χ^2 values plotted in the top panel of Fig. 11, we see that the mean value is $\chi^2 \approx 2.5$, which thus indicates that the Gaussian white noise null hypothesis is rejected by more than 2σ for this particular PID. This is an early indication that something is indeed wrong with the noise model, and we will return to this observation below.

The second panel shows the gain g for the same PID. In this case, the Markov correlation length appears to be perhaps 20 samples by eye, while the burn-in phase is very short. Considering that we have only 1200 full Gibbs samples available, this implies that the number of independent gain samples per PID is rather limited, and probably around 50. However, this is still sufficient to provide a useful estimate of the true underlying distribution, and the marginal posterior may be summarized as $g = 77.95 \pm 0.02$ mV K^{-1}. Even when adopting a conservative estimate of only 30 independent samples, the sampling uncertainty accounts for less than 20 % of the quoted uncertainty. For further discussion of the gain posteriors, see Gjerløw et al. (2020).

The third panel shows the bandpass shift, $\Delta_{\rm bp}$, for the 30 GHz LFI 27M radiometer. As already noted in Sect. 4.2, this parameter is the single most difficult quantity to estimate in the entire framework, because of the highly non-Gaussian and global nature of its impact; virtually *all* stochastic variables in the entire model depend on the instrumental bandpass in one form or another, and changes in this parameter therefore take a substantial amount of time to propagate throughout the model. Furthermore, the sampling algorithm used for this parameter is a basic Metropolis sampler, simply because of a lack of better alternatives. The result is a long correlation length of about 50 samples, resulting in perhaps as few as 20 uncorrelated samples. Still, even with this crude sampler, we do see that the six chains mix reasonably well, and it is possible to establish a useful estimate for the marginal posterior, which in this case may be summarized as $\Delta_{\rm bp} = 342 \pm 6$ MHz. However, in this case the sampling uncertainty accounts for at least 30 % of the error bar. For further discussion of the bandpass posteriors, see Svalheim et al. (2020a).

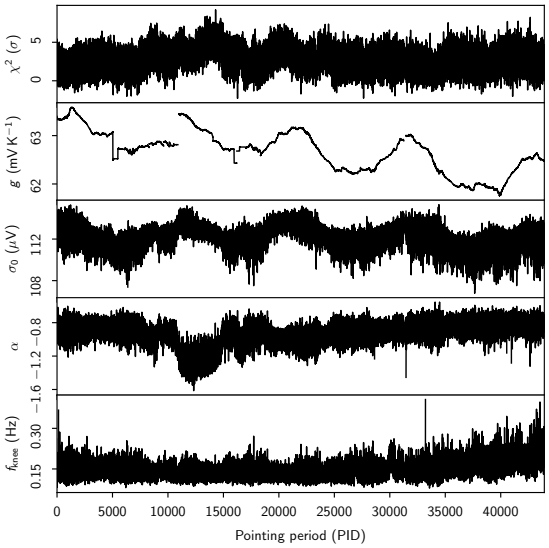

Fig. 12. A single Gibbs sample for the 30 GHz LFI 28M radiometer as a function of PID. From top to bottom the panels show 1) normalized reduced χ^2; 2) gain, g; 3) white noise level, σ_0; 4) correlated noise slope, α; and 5) correlated noise knee frequency, $f_{\rm knee}$.

The three last panels show the three noise PSD parameters, σ_0, α and $f_{\rm knee}$, for the same radiometer. These are much better constrained, because of the large number of samples within each PID. However, there are still obvious internal correlations between the three parameters, such that a low value of σ_0 leads to high values of α and $f_{\rm knee}$, which makes intuitive sense. It is also worth noting that for other parameters in the model, only the actual noise PSD matters, as defined by the combination of these parameters. For further discussion of the noise posteriors, see Ihle et al. (2020).

Figure 12 shows corresponding values as a function of PID for one single Gibbs sample, this time for the 30 GHz LFI 28M radiometer. The top panel shows the normalized reduced χ^2, as defined by Eq. (132). As discussed above, this function should ideally be independent between PIDs, and distributed according to $N(0,1)$. This plot is therefore a powerful monitor for identifying unmitigated and non-stationary instrumental systematic effects in a given radiometer. We see that the χ^2 excess noted above for a single 27M PID was not a fluke, but rather a typical example: As discussed by Ihle et al. (2020), all 30 GHz and 44 GHz radiometers exhibit a χ^2 excess per PID at the 2–3 σ

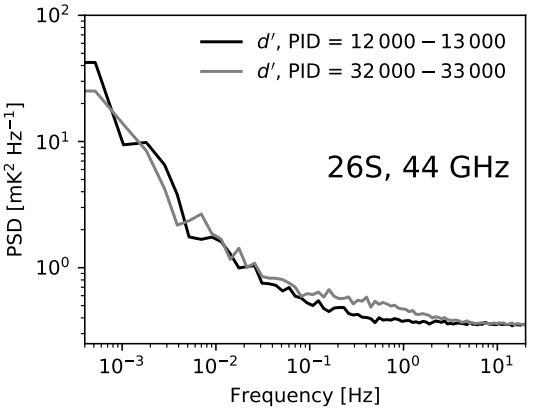

Fig. 13. Comparison of the mean PSD of $\boldsymbol{d} - \boldsymbol{s}^{\rm sky}$ for the 44 GHz LFI 26S radiometer, each of which are averaged over two PID ranges. Note the excess seen in the frequency range between 0.1 and 1 Hz for PIDs between 32 000 and 33 000; this noise contribution cannot be described by a simple $1/f$ model.

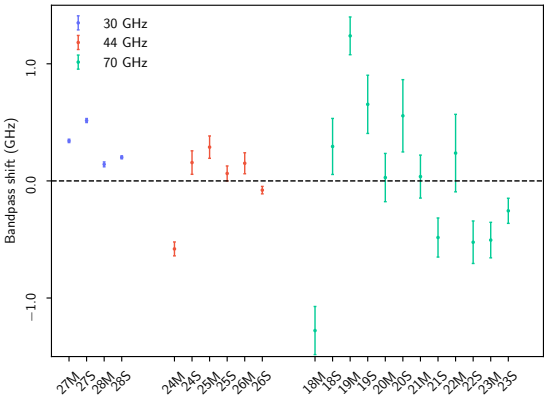

Fig. 14. Estimated bandpass corrections for each LFI radiometer. Error bars indicate $\pm 3\sigma$ uncertainties. Note these parameters are constrained to have vanishing mean within each frequency, and are as such strongly correlated.

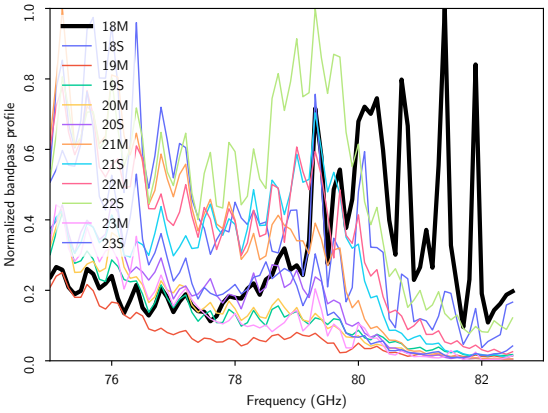

Fig. 15. High frequency tail of individual ground-measured 70 GHz radiometer bandpasses. The 70 GHz LFI 18M radiometer, for which a large negative bandpass correction parameter is derived in the BeyondPlanck Gibbs sampler, is shown as a thick black curve.

level, strongly suggesting a break-down of the assumed noise model for these channels.

The second panel shows the gain for the same 28M radiometer as a function of PID. Here we see clear evidence of a systematic oscillation with a period of one year, and a maximum variation of about 1–2 % throughout the mission. The oscillatory behaviour is primarily due to variations in the incoming solar radiation during the year, effectively changing the heating of the instrument depending on its precise orientation with respect to the Sun. In addition, we also see evidence for a handful of discrete jumps; these are caused by known events, such as satellite maneuvers or cooler maintenance, as summarized in Planck Collaboration I (2016). We note that this list of discrete jumps is currently inserted by hand into the analysis, and no automatic searches are performed. Undetected jumps may therefore contaminate higher-level analyses, and, as already noted, this issue may represent one of the most important outstanding questions with regards to the 44 GHz channel in the current analysis.

The three bottom panels of Fig. 12 show corresponding plots of the three noise parameters as a function of PID. The same features as observed in the gain are seen also here, although with lower signal-to-noise ratio. Overall, it is visually obvious that the noise properties of this channel are not stationary throughout the mission, but rather vary significantly in time. In particular, the white noise level varies by 3–4 % throughout the mission, and mirrors the gain variations seen above. For the slope, α, the most noteworthy feature are overall steeper values between PIDs 11 000 and 15 000; as shown by Ihle et al. (2020), these can be traced to changes in the thermal environment of the satellite using house-keeping data.

Returning to the systematic χ^2 excess noted above, Fig. 13 compares the power spectral density (PSD) of $\boldsymbol{d} - \boldsymbol{s}^{\rm sky}$ for the 44 GHz LFI 26S radiometer, as averaged over two different PID ranges. Ideally, this should follow a $1/f$ noise profile, and for PIDs between 12 000 and 13 000 this appears to be a reasonable assumption. However, later in the mission there is a clear excess noise contribution at intermediate frequencies, between 0.1 and 1 Hz, that clearly cannot be described as $1/f$ noise. Ihle et al. (2020) shows that this particular radiometer abruptly changes behaviour around PID 20 000, after which this term becomes dominant. For this reason, we choose to exclude the 26S radiometer entirely from the BeyondPlanck analysis during the second half of the mission. Furthermore, to avoid temperature-to-polarization leakage we also exclude the 26M radiometer. In total, this reduces the total available data volume at 44 GHz by 17 %, and the resulting BeyondPlanck 44 GHz frequency map is therefore also correspondingly more noisy than the corresponding DPC and NPIPE maps. While this data cut should, at least in theory, result in a more accurate noise characterization for the BeyondPlanck maps, we do emphasize that all other 30 and 44 GHz radiometers also exhibit similar noise excesses, although at quantitiatively lower levels than 26S. In order to actually establish a statistically acceptable noise description of these channels, the assumed $1/f$ noise model will need to be generalized to account for excess noise at intermedate frequencies. This is clearly a top priority for next-generation processing

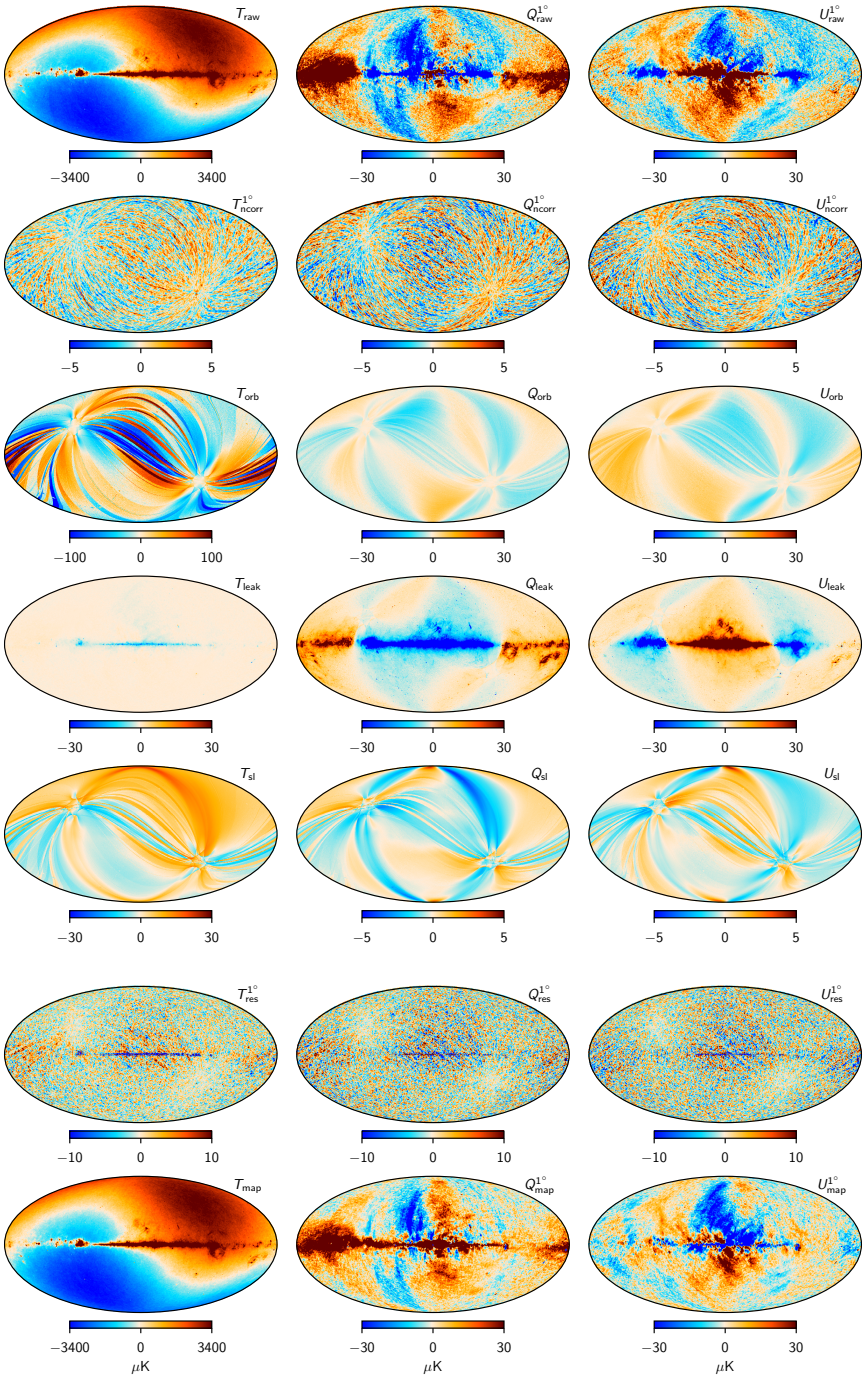

Fig. 16. Comparison between individual individual TOD correction terms and sky maps for the 30 GHz channel for a single Gibbs sample, projected into sky maps. Columns show Stokes T, Q, and U parameters. Rows show, from top to bottom, 1) the raw TOD projected into a sky map; 2) correlated noise, $\boldsymbol{n}_{\rm corr}$; 3) the orbital dipole, $\boldsymbol{s}_{\rm orb}$; 4) bandpass and beam mismatch leakage; and 5) sidelobe corrections, $\boldsymbol{s}_{\rm sl}$. The bottom row shows the final sky map obtained by subtracting all these correction terms from the raw TOD, while the second to last row shows the residual map obtained when also subtracting a model of the sky, $\boldsymbol{s}_{\rm sky}$, from the TOD prior to map binning. Some components, as indicated in the respective labels, have been smoothed to an angular resolution of 1° FWHM. All units are $\mu{\rm K}_{\rm CMB}$.

We now return to our review of the instrumental parameter posterior distributions. In that respect, we note that some low-level parameters are global in nature, and do not map into PID space. The most prominent example of this is the set of bandpass corrections, $\Delta_{\rm bp}$, which are summarized in Fig. 14 in terms of the marginal posterior mean and standard deviation for each radiometer. As described by Svalheim et al. (2020a), the overall shift within each frequency is fixed by comparing the estimates reported by Planck Collaboration X (2016) against the current data set. Doing so, we find that an overall frequency shift of 0.3 GHz is indeed required for the 30 GHz channel, in agreement with Planck Collaboration X (2016), whereas both the 44 and 70 GHz are compatible with no shift. In contrast, the relative distribution of bandpass corrections within a frequency is determined through minimizing the so-called spurious S-maps discussed in Sect. 7.2.2, and this is only mildly dependent on the details of the foreground model. While the absolute bandpass correction for a given channel is directly dependent on the foreground model, the polarization leakage only depends on it to second order, and may therefore be estimated much more directly from a single data set. At the same time, we do note that the uncertainties in the bandpass parameters depend sensitively on the amplitude of the foregrounds. As a result, the relative bandpass uncertainties are small at 30 GHz, where the foregrounds are bright, while they are large at 70 GHz, where the foregrounds are weak.

One visually striking result seen in this plot is the fact that the first 44 and 70 GHz radiometer (24M and 18M, respectively) have a much larger (and negative) correction than other radiometers. To provide some intuition regarding this result, Fig. 15 shows a comparison between the individual 70 GHz radiometer bandpass profiles as measured on ground, zooming in on their high frequency tails. The thick solid black curve shows the 18M radiometer. Here we see that this particular radiometer exhibits large oscillations near the high frequency cutoff, with an amplitude of order unity. No other radiometer shows nearly as strong fluctuations. One possible hypothesis for this result is that spurious standing waves may have been induced by the test equipment itself. Irrespective of its origin, this effect serves as a useful example of the utility of global analysis, by highlighting a previously unnoticed instrumental effect in terms of the final marginal posterior distribution. At the same time, this result also underlines the importance of choosing a physically meaningful parametric model; in this particular case, a parametric model implementing smooth apodization of the tails may be far better motivated than a simple shift model, as adopted in the current analysis. Such models should be explored further in future work.

Next, we consider the spatial structure of each of the various TOD model terms in pixel space, and Fig. 16 shows each of the TOD objects binned into a 3-component Stokes IQU sky map for one arbitrarily selected sample (specifically, sample number 101 from the first chain).

The top panel shows the raw TOD binned into a sky map, and provides intuition regarding the overall quality of the data before applying any corrections. Indeed, for the temperature component it is very difficult to spot major artifacts of any kind; the most notable feature is a few correlated noise stripes in the lower left quadrant. For polarization, the dominant effect is the alternating sign along the Galactic plane as a function of longitude, which is due to bandpass mismatch.

The second panel shows the correlated noise component. The most notable features in this map are coherent stripes along the satellite scanning path. It should also be noted that this component is the one that is the least constrained from a-priori considerations among all TOD components, and therefore acts as a "trash can" for possible unmodelled errors; this is the first place one expects to see residuals from modelling errors. And, indeed, we do see hints of artifacts in these maps, perhaps most notably in the Stokes Q parameter: There is a broad bright region near the South Ecliptic Pole. Comparing this with the other panels in the same figure, this feature appears qualitatively similar to all three of the orbital, sidelobe, and leakage corrections. The common denominator among these three terms is their direct dependency on the assumed beam model. As reported by Planck Collaboration II (2020), about 1 % of the *Planck* 30 GHz solid angle has not been accounted for by the current state-of-the-art GRASP model for the LFI 30 GHz channel, and this missing power could at least in principle account for some of the excess structure seen in $\boldsymbol{n}^{\rm corr}$.

The third row shows the orbital dipole. For a single PID, this signal is defined by a perfect dipole along the scanning ring with an amplitude of about 270 μK, convolved with the 4π LFI beam. However, when the same ring is observed six months apart, the phase of the signal is reversed, and the total sum is then both small and difficult to predict. Also, although the intrinsic signal is entirely unpolarized, convolution with far-sidelobes algebraically couples this model to the polarization sector as well.

The fourth row shows the bandpass and beam leakage correction. This effect is clearly the strongest among all polarization corrections, with amplitudes exceeding many tens of μK in the Galactic plane, while still being almost entirely negligible in temperature. Morphologically speaking, the archetypal signature of bandpass mismatch is a variable sign along the Galactic plane, tracing the specific orientation of the detector polarization angles as the different detectors observe at slightly different effective frequencies. At high latitudes, this map is dominated by temperature-to-polarization leakage resulting from different radiometers observing the signal model with different beam FWHMs; large angular scales are dominated by CMB dipole leakage, while small angular scales are dominated by foregrounds and CMB temperature fluctuations.

The fifth row shows the impact of sidelobe pickup. In temperature, the two dominant features are, first, a large-scale pattern broadly aligned with the solar CMB dipole resulting from interactions with the intermediate sidelobes, and, second, individual rings created by the far sidelobes hitting the Galactic plane. The same features are also seen in polarization, but now a more complicated pattern arises due to the additional modulation by the relative orientation of the polarization angles at any given time.

The second to last row shows the TOD residuals binned into a sky map. For most of the sky, this is consistent with white noise, but clear residuals are seen in the Galactic plane, reflecting the structures seen in the bottom panel of Fig. 10. This indicates that the adopted foreground and/or instrument model is not statistically adequate in these very bright regions of the sky. Subsequent CMB-oriented analyses will clearly need to mask these regions prior to power spectrum or parameter estimation.

The last row shows the final co-added frequency map. To assess the overall impact of the various correction terms, this may be directly compared with the raw map shown in the top row.

A complete survey of similar maps for all three LFI frequencies is provided by Suur-Uski et al. (2020), and for the most part these are qualitatively similar to those presented here. However, there is one special case that warrants a dedicated discussion, namely the correlated noise map for the 44 GHz channel. This is shown in Fig. 17, after smoothing to an effective angular reso-

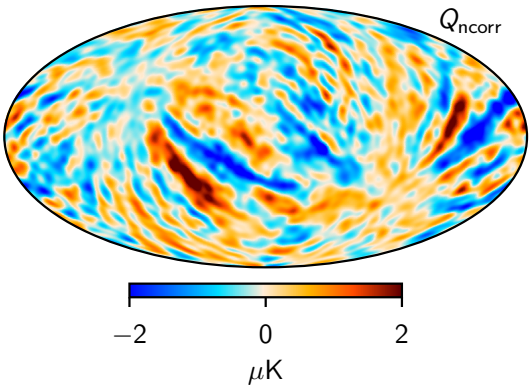

Fig. 17. Correlated noise map, $\boldsymbol{n}^{\rm corr}$, for the 44 GHz frequency channel, smoothed to an effective angular resolution of 5° FWHM. Reproduced from Suur-Uski et al. (2020).

lution of 5° FWHM. The most striking features in this map are strong stripes extending from the right-most edge through most of the Southern Galactic hemisphere. The origin of these stripes has not been robustly identified, and this is an area of intense on-going research. However, we have found that by reducing the gain smoothing window (see Sect. 8.3.1 and Gjerløw et al. 2020) by a factor of two, making the gain model more flexible, these stripes vanish from $\boldsymbol{n}^{\rm corr}$ (Gjerløw et al. 2020). However, that also introduces large gain fluctuations in periods of low CMB Solar dipole amplitude, and this is not an acceptable solution for production analysis. Still, based on preliminary work, it seems likely that these strong 44 GHz stripes are associated with shortcomings in the current gain model, and this will require further work to resolve. Furthermore, excess χ^2 values in the Southern Galactic hemisphere represents the biggest challenge with respect to full-sky CMB polarization reconstruction for BeyondPlanck, and this is thus clearly a top priority issue. In the current analysis, we only use data covering the Northern Galactic hemisphere to derive final low-ℓ CMB polarization constraints.

9.2. Frequency maps

We now turn our attention to co-added frequency maps, as solved for deterministically through Eq. (76). For many users, these represent the most convenient form of the BeyondPlanck products, and we provide these maps both in the form of individual samples, each corresponding to one possible realization of all modelled systematic effects, and as more traditional posterior mean and standard deviation maps,

$$\hat{\boldsymbol{m}}_\nu = \left\langle \boldsymbol{m}_\nu^i \right\rangle \tag{133}$$

$$\sigma_\nu(p) = \sqrt{\left\langle \left(m_\nu^i(p) - \hat{m}_\nu(p)\right)^2 \right\rangle}, \tag{134}$$

where brackets indicate averaging over Monte Carlo samples. Note that σ_ν, as defined here, only accounts for systematic uncertainties per pixel, not white noise uncertainties as defined by the diagonal of the inverse coupling matrix in Eq. (76), $\sigma_\nu^{\rm wn}(p)$. To obtain the full uncertainty, these two terms must be added in quadrature,

$$\sigma_\nu^{\rm tot}(p) = \sqrt{\sigma_\nu(p)^2 + \sigma_\nu^{\rm wn}(p)^2}. \tag{135}$$

We stress, however, that analysis of these posterior mean maps is likely to be sub-optimal for most scientific applications, and will not exploit the full power of the BeyondPlanck framework. Instead, we highly recommend users to analyze the full ensemble of individual posterior samples; that is by far the most robust and statistically correct method for propagating BeyondPlanck uncertainties into any higher-level analysis.

With these caveats in mind, Fig. 18 shows the posterior mean maps $\hat{\boldsymbol{m}}_\nu$ for each frequency and each Stokes parameter. The polarization maps have been smoothed to an angular resolution of 1° FWHM to reduce noise. Note that the BeyondPlanck temperature maps retain the CMB dipole, similar to `NPIPE` Planck Collaboration Int. LVII (2020), but contrary to the *Planck* 2018 and *WMAP* frequency maps. Leaving this component in the maps ensures that the full information content of the data is available for subsequent component separation and calibration applications.

Figure 19 shows the corresponding posterior standard deviation maps, σ_ν, as defined above, after smoothing to a common angular resolution of 2° FWHM. These maps summarize the combined effect of the various systematic corrections made to the frequency maps, and are are such morphologically rich. The most striking features include:

1. dipole variations in the 30 GHz intensity map, reflecting uncertainties in the absolute calibration of this channel;
2. excess variance for rings aligned with the Galactic plane, reflecting the higher uncertainties in the time-variable gain resulting from the processing mask;
3. excess variance along the Galactic plane, reflecting the higher uncertainties here due to gain and bandpass variations;

all of which are super-imposed on the general *Planck* scanning pattern, which itself reflects correlated noise variations. We also note that the upper limit of the temperature color scale is only 2.1 μK, which indicates that these variations are much lower than the intrinsic variance of the CMB temperature fluctuations, which is about 50 muK on these angular scales, and minor details in the systematic model are therefore unlikely to affect final cosmological results. In contrast, the standard deviation of the polarization maps at high Galactic latitudes is typically about 0.5 μK, which is of the same order of magnitude as the expected polarization imprint from cosmic reionization.

Next, Fig. 20 shows the corresponding white noise standard deviation maps. These maps are fully specified by the detector white noise level σ_0, the time-variable gains g_t, and the number of observations per pixel. They do not account for any coupling to astrophysical foregrounds or instrumental effects beyond those listed above, and, as a result, these maps are structurally much cleaner than the systematic uncertainty maps.

The above figures concern the average frequency map posterior distribution. It is also interesting to consider variations between individual Monte Carlo samples, as they often more easily convey spatial correlations between individual modes. Figure 21 is an example of this, showing the difference between two frequency map samples, smoothed to a common angular resolution of 7° FWHM. Here we clearly see correlated noise stripes along the *Planck* scan direction in all three frequency channels, but significantly more pronounced in the 30 GHz channel than in the other two frequencies. We also see fluctuations along the Galactic plane, which are dominated by uncertainties in the bandpass correction parameters, $\Delta_{\rm bp}$. Clearly, modelling such correlated fluctuations in terms of a single standard deviation per pixel is unlikely to be adequate for any high-precision analysis, and,

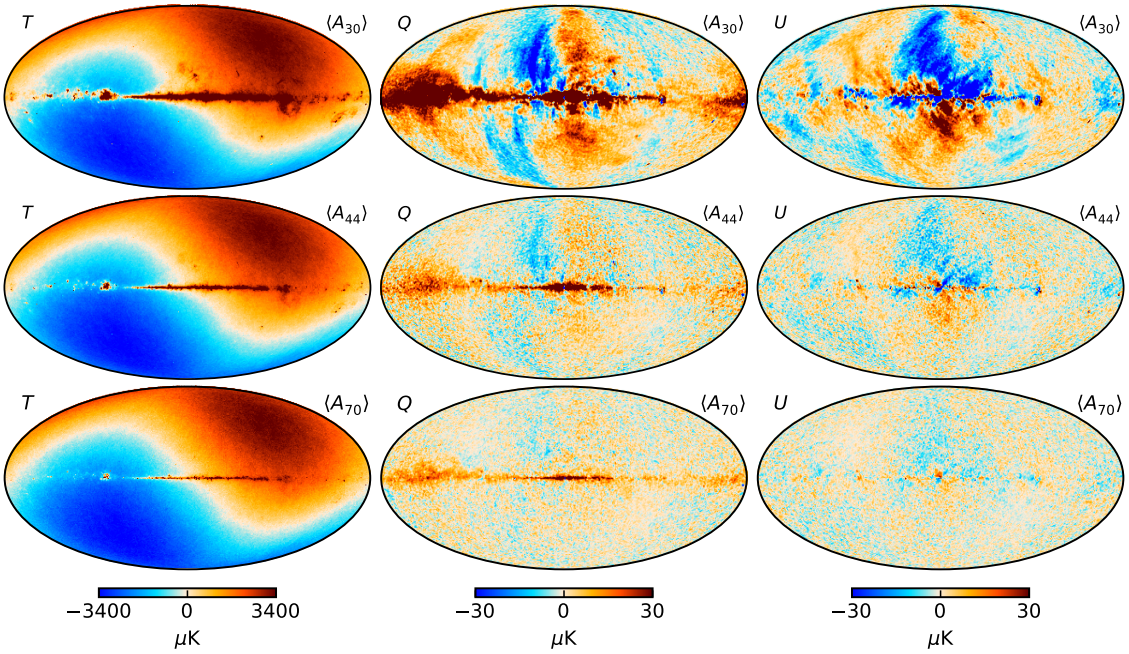

Fig. 18. Posterior mean maps for each LFI frequency. Rows show, from top to bottom, the 30, 44 and 70 GHz frequency channels, while columns show, from left to right, the temperature and Stokes Q and U parameters. The polarization maps have been smoothed to a common angular resolution of 1° FWHM to visually reduce the noise level.

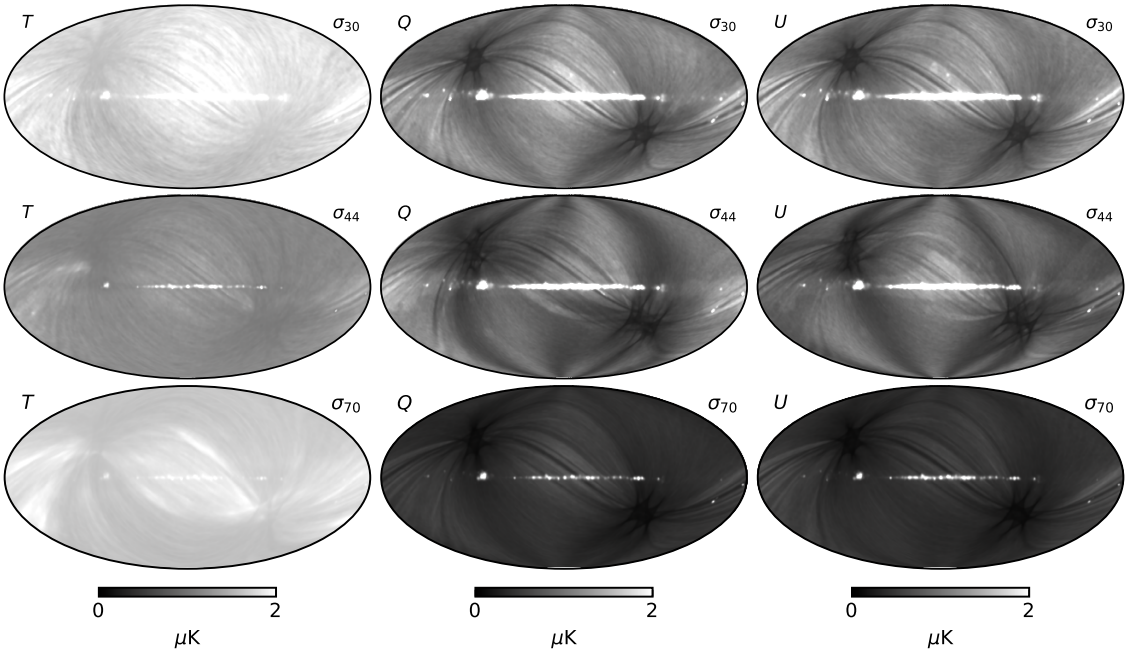

Fig. 19. Posterior standard deviation maps for each LFI frequency. Rows show, from top to bottom, the 30, 44 and 70 GHz frequency channels, while columns show, from left to right, the temperature and Stokes Q and U parameters. Note that these maps do not include uncertainty from instrumental white noise, but only variations from the TOD-oriented parameters included in the data model in Eq. (69).

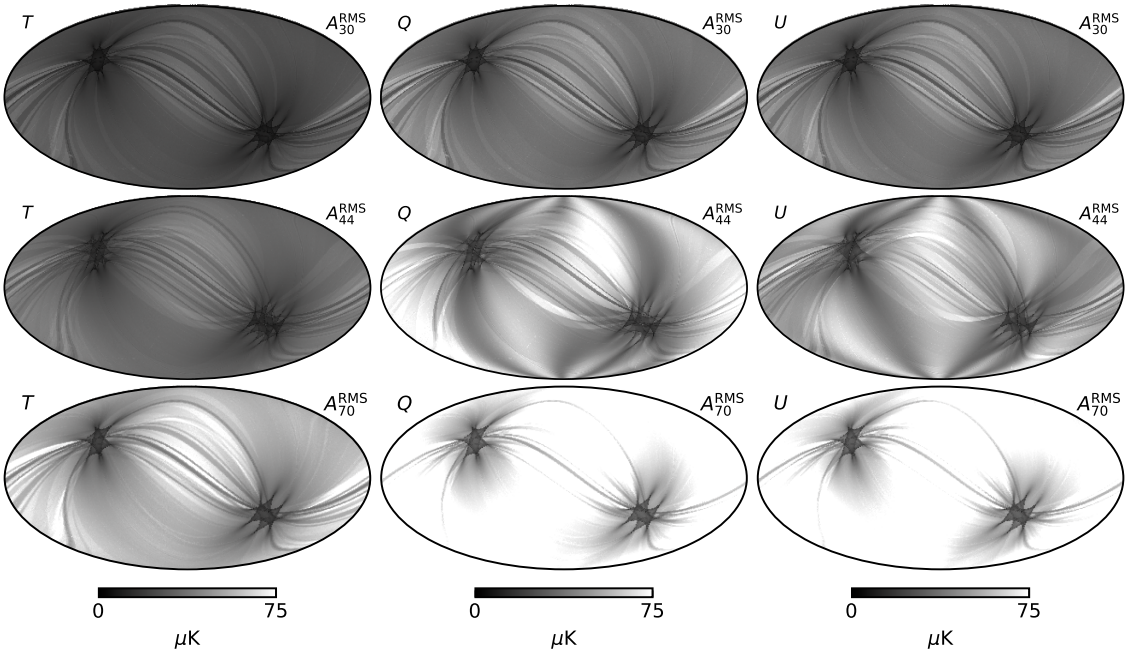

Fig. 20. White noise standard deviation maps for a single arbitrarily selected sample. Rows show, from top to bottom, the 30, 44 and 70 GHz frequency channels, while columns show, from left to right, the temperature and Stokes Q and U parameters. Note that the 70 GHz maps are scaled by a factor of 2, to account for the fact that this map is pixelized at $N_{\mathrm{side}} = 1024$, while the two lower frequencies are pixelized at $N_{\mathrm{side}} = 512$.

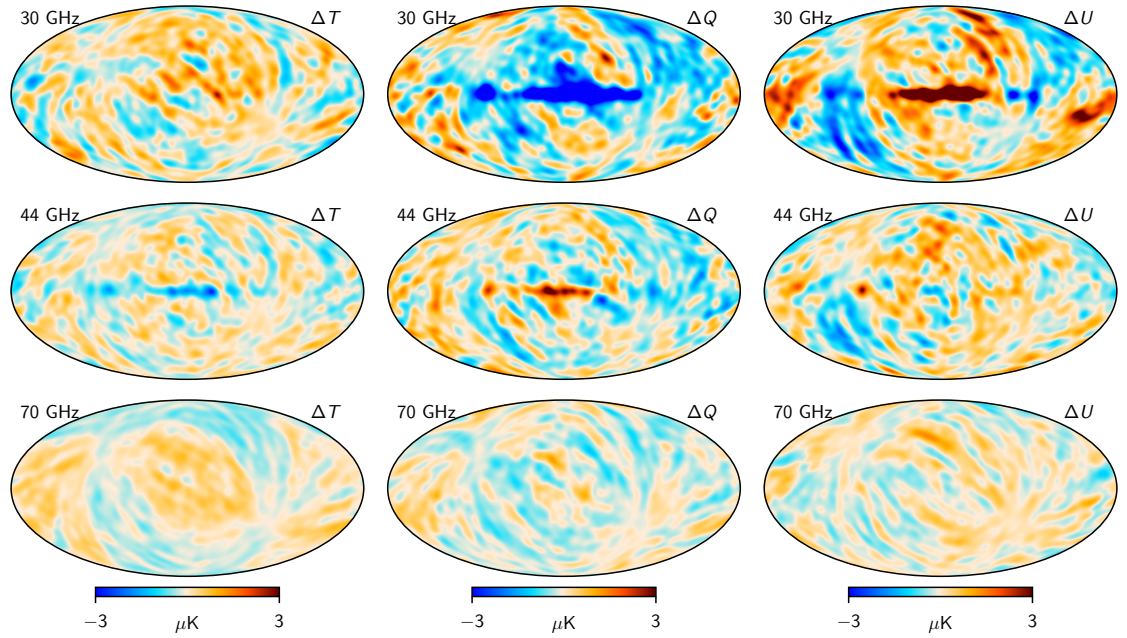

Fig. 21. Difference maps between two frequency map samples, smoothed to a common angular resolution of 7° FWHM. Rows show, from top to bottom, the 30, 44 and 70 GHz frequency channels, while columns show, from left to right, the temperature and Stokes Q and U parameters.

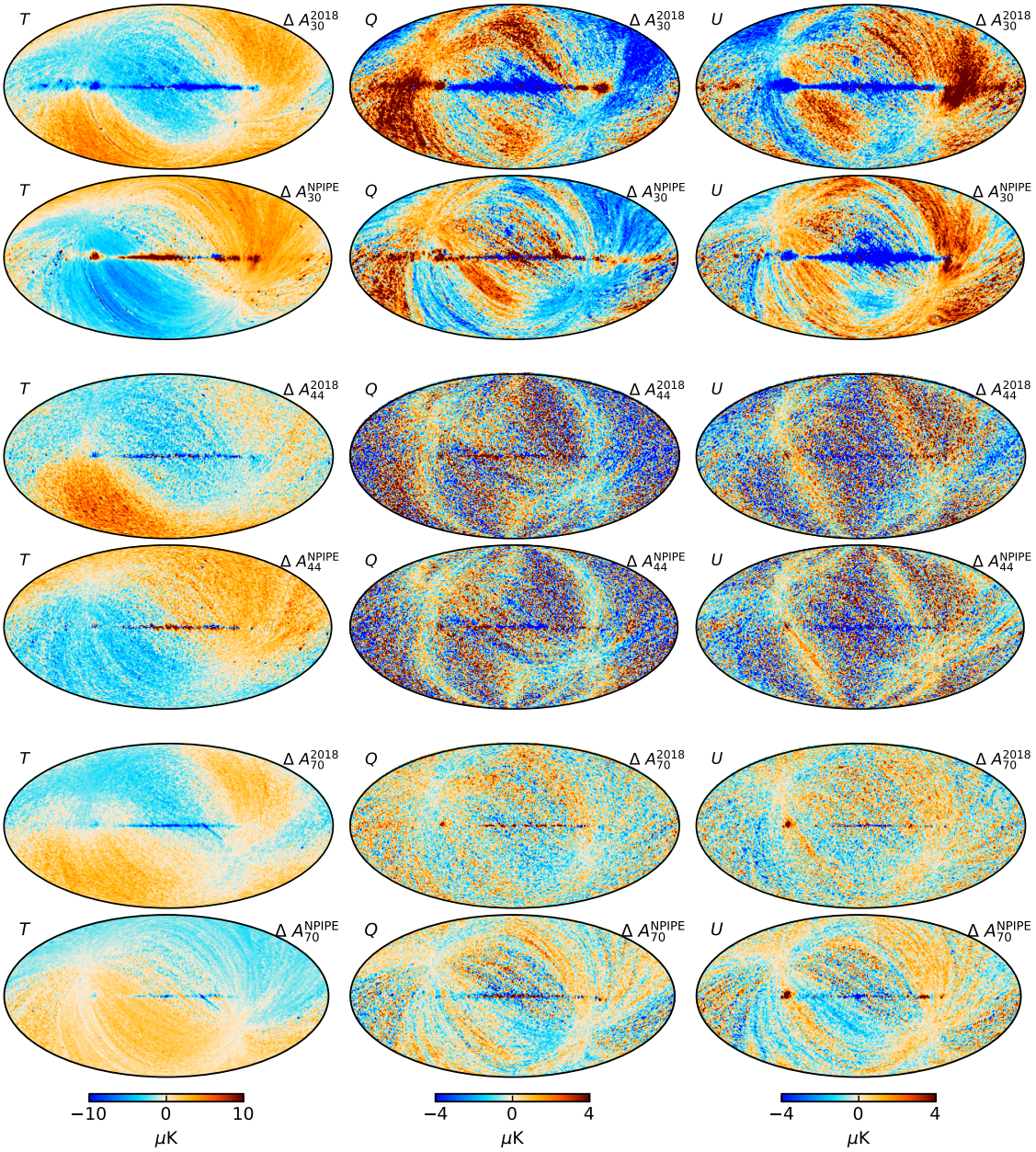

Fig. 22. Differences between BeyondPlanck and 2018 or NPIPE frequency maps, smoothed to a common angular resolution of 2° FWHM. Columns show Stokes T, Q and U parameters, respectively, while rows show pair-wise differences with respect to the pipeline indicated in the panel labels. A constant offset has been removed from the temperature maps, while all other modes are retained. The 2018 maps have been scaled by their respective beam normalization prior to subtraction.

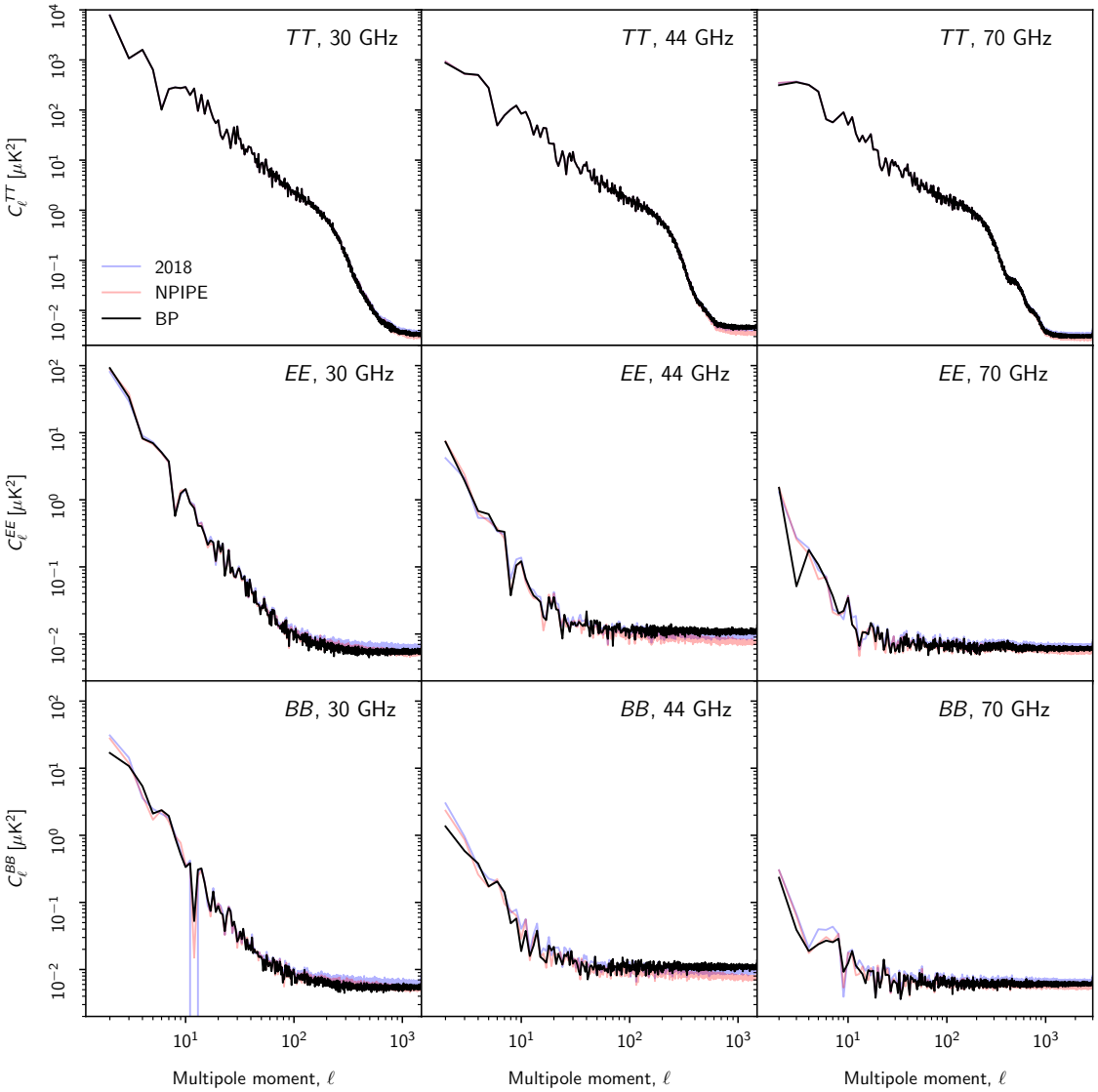

Fig. 23. Comparison between angular auto-spectra computed from the BeyondPlanck (*black*), *Planck* 2018 (*blue*), and NPIPE (*red*) full-frequency maps. Rows show different frequencies, while columns show *TT*, *EE*, and *BB* spectra. All spectra have been estimated with PolSpice using the *Planck* 2018 common component separation confidence mask (Planck Collaboration IV 2020).

again, we strongly recommend analysis of the full map ensemble when using the BeyondPlanck products for external analysis.

Figure 22 shows differences between the BeyondPlanck frequency maps and those presented in the *Planck* 2018 and NPIPE data releases. To ensure that this comparison is well defined, the 2018 maps have been scaled by the uncorrected beam efficiencies, and the best-fit *Planck* 2018 solar dipole has been added to each map, before computing the differences. Overall, we see that the BeyondPlanck maps agree with the other two pipelines to $\lesssim 10\,\mu$K in temperature, and to $\lesssim 4\,\mu$K in polarization. In temperature, we see that the main difference between NPIPE and BeyondPlanck is an overall dipole, while differences with respect to the 2018 maps show greater morphological differences. The sign of the NPIPE dipole differences changes with frequency. This result is consistent with the original characterization of the NPIPE maps derived through multi-frequency component separation in Planck Collaboration Int. LVII (2020); that paper reports a relative calibration difference between the 44 and 70 GHz channel of 0.31 %, which corresponds to $10\,\mu$K in the map-domain. Overall, in temperature BeyondPlanck is thus morphologically similar to NPIPE, but it im-

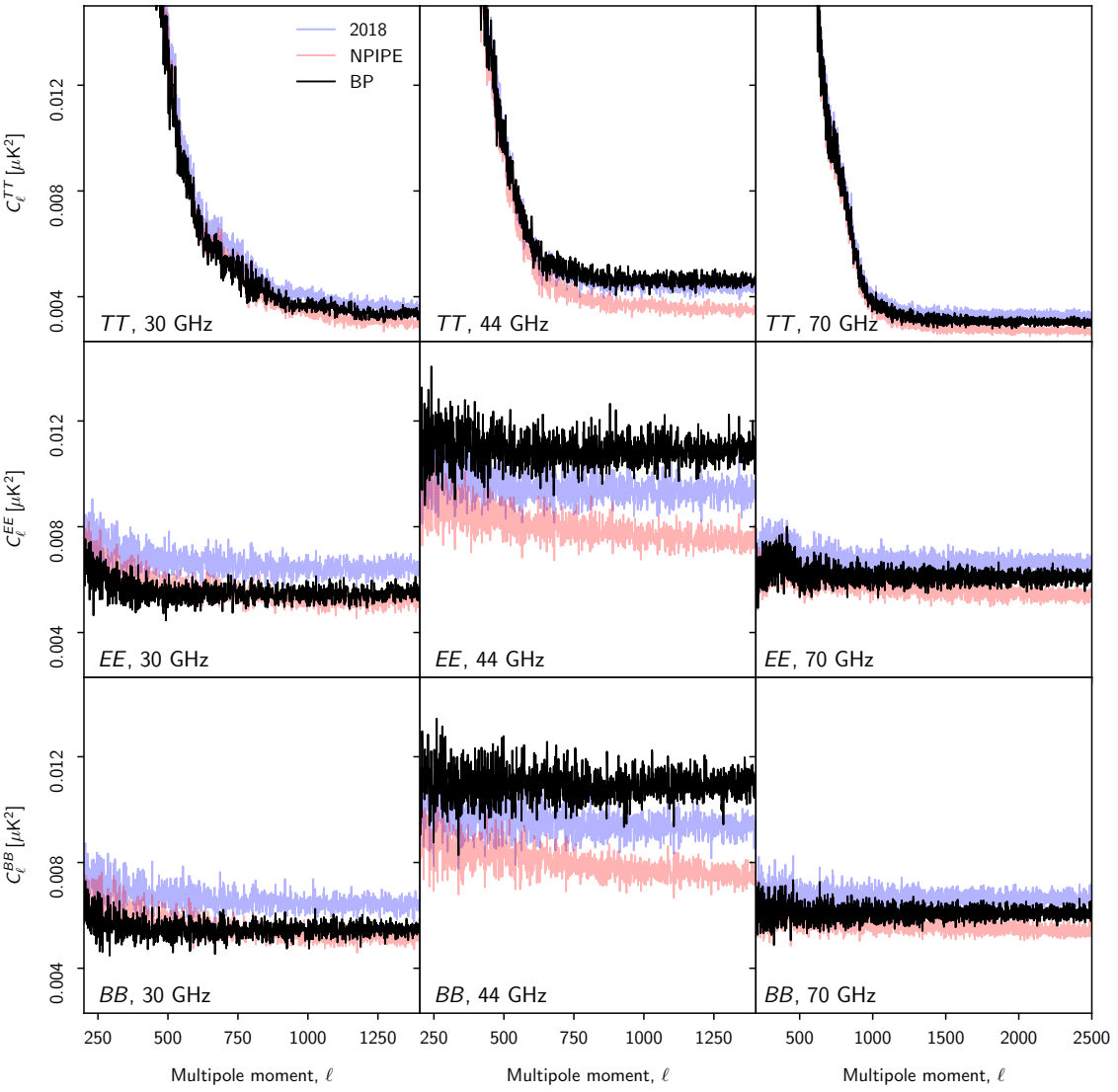

Fig. 24. Same as Fig. 23, but zooming in on the noise-dominated high-ℓ multipole range.

proves a previously reported relative calibration uncertainty between the various channels by performing joint analysis.

In polarization, the dominant large-scale structures appear to be dominated by effectively different offset determinations per PID, which may originate from different gain or correlated noise solutions. It is worth noting that the overall morphology of these difference maps is structurally similar between frequencies, and that the apparent amplitude of the differences falls with frequency. This strongly suggests that different foreground modelling plays a crucial role. In this respect, two observations are particularly noteworthy: First, while both the *Planck* 2018 and NPIPE pipelines incorporate component separation as an external input as defined by the *Planck* 2015 data release (Planck Collaboration X 2016), BeyondPlanck performs a joint fit of both astrophysical foregrounds and instrumental parameters. Second, both the LFI DPC and the NPIPE pipeline consider only *Planck* observations alone, while BeyondPlanck also exploits *WMAP* information to establish the sky model, which is particularly important to break scanning-induced degeneracies in polarization.

Regarding the 44 GHz channel, two main features stand out in the difference maps in Fig. 22. First, these difference maps obviously exhibit a much higher white-noise level than the corresponding 30 or 70 GHz maps. This is because we exclude data from both the 44 GHz LFI 26M and 26S radiometers for the second half of the mission, as discussed above. Second, we see the same coherent stripes extending through the Southern Galactic hemisphere as seen in the correlated noise map in Fig 17, with clearly different amplitudes in the 2018 and NPIPE differences.

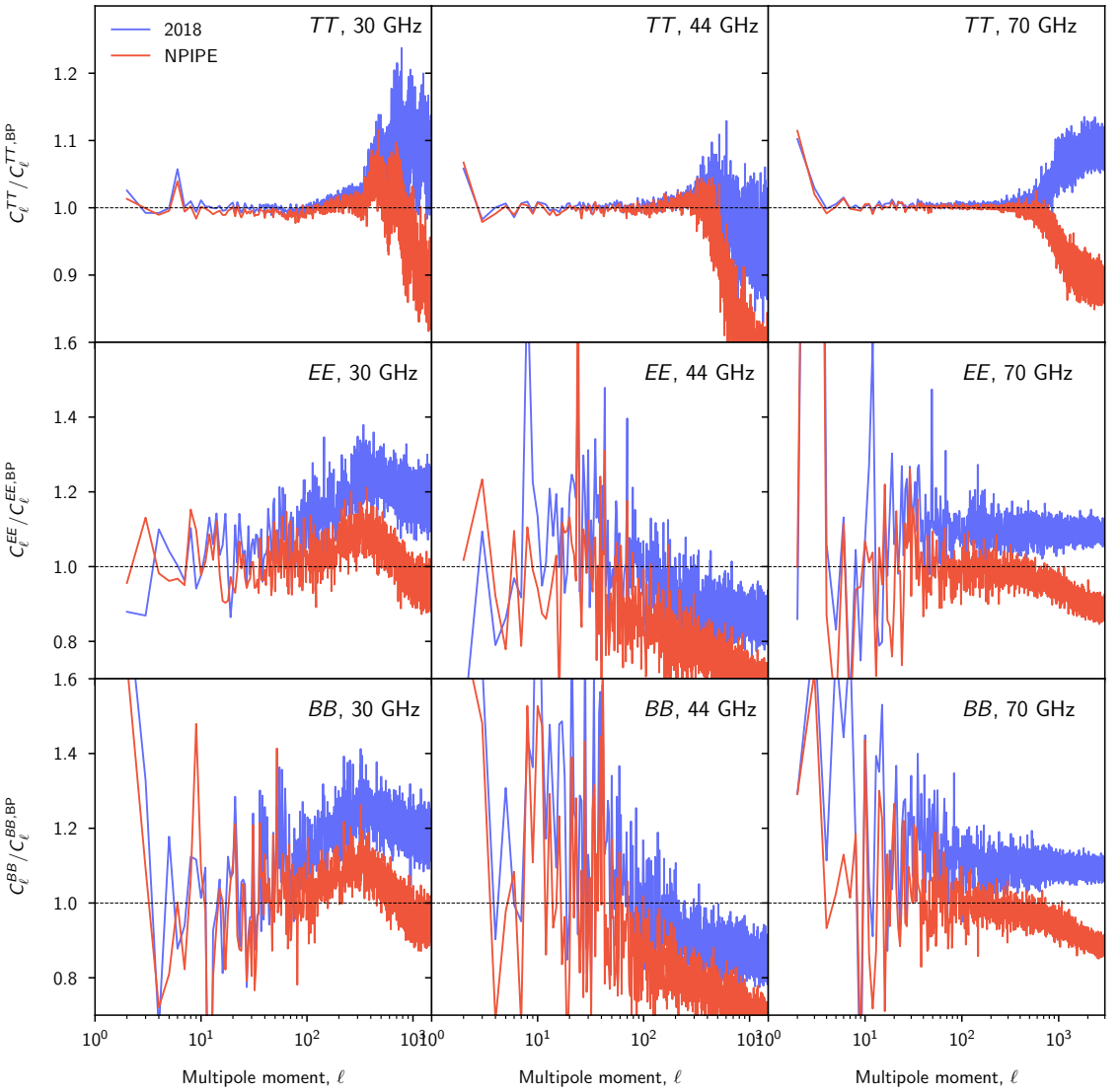

Fig. 25. Ratios between the angular auto-spectra shown in Fig. 23, adopting the BeyondPlanck spectra as reference. *Planck* 2018 results are shown as blue lines, while NPIPE results are shown as red lines. Values larger than unity imply that the respective map has more power than the corresponding BeyondPlanck spectrum.

This shows that at least two out of the three different processing pipelines struggle with this particular problem, whatever its true origin may be.

Finally, and as already noted, the dominant high-latitude structures in the 70 GHz residual maps are morphologically very similar to the 30 GHz differences. This suggesting a foreground-related common mode error in one or more of these maps.

We now turn our attention to the angular power spectrum properties of the BeyondPlanck frequency maps. Figure 23 shows auto-correlation spectra as computed with `PolSpice` (Chon et al. 2004) outside the *Planck* 2018 common component separation confidence mask (Planck Collaboration IV 2020), which accepts a sky fraction of 80 %. All these spectra are clearly signal-dominated at large angular scales (as seen by the rapidly decreasing parts of the spectra at low ℓ's), and noise-dominated at small angular scales (as seen by the flat parts of the spectra at high ℓ's); note that the "signal" in these maps includes both CMB and astrophysical foregrounds. Overall, the three pipelines agree well at the level of precision supported by the logarithmic scale used here; the most striking differences appear to be variations in the high-ℓ plateau, suggesting notably different noise properties between the three different pipelines.

We therefore zoom in on the high-ℓ parts of the spectra in Fig. 24. Here the differences become much more clear, and easy

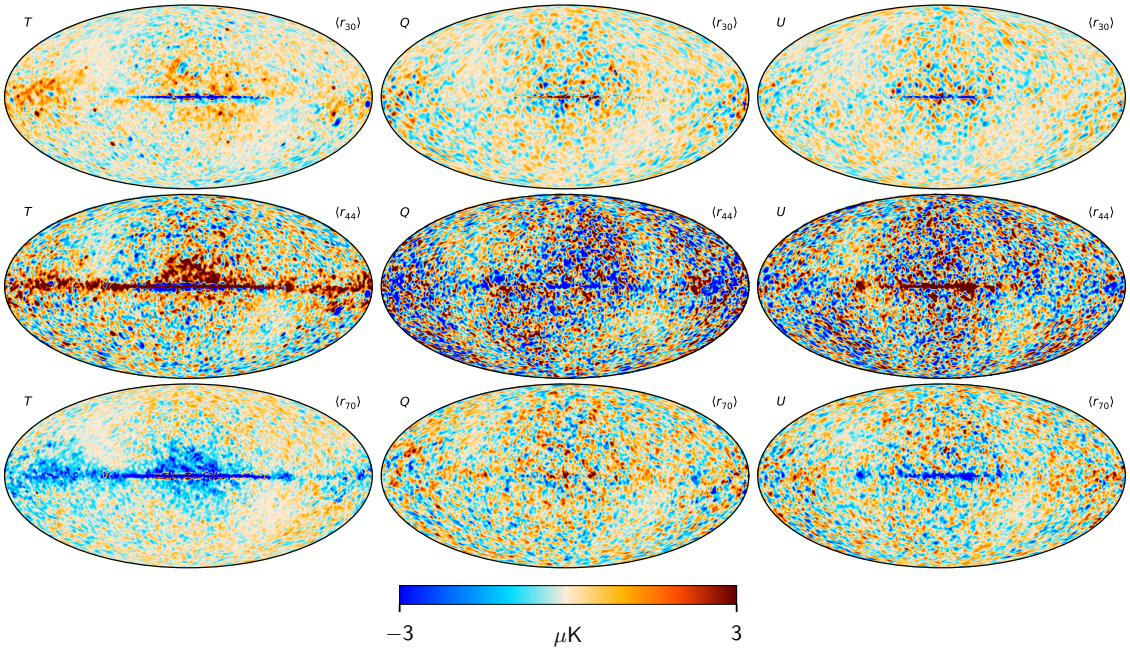

Fig. 26. Posterior mean total data-minus-model residual maps $\boldsymbol{d}_\nu - \boldsymbol{s}_\nu$ for BeyondPlanck LFI 30 (*top*), 44 (*middle*), and 70 GHz (*bottom*). All maps are smoothed to a common angular resolution of 2° FWHM.

to interpret. And in general we note two different trends. First, we note that the overall noise levels of the BeyondPlanck 30 and 70 GHz maps are slightly lower than in the *Planck* 2018 maps, but also higher than NPIPE, although the latter holds less true for 30 GHz than for 70 GHz. The BeyondPlanck 44 GHz map clearly has higher noise than either of the other two, as discussed above. Second, we also note that the BeyondPlanck spectra are notably flatter than the other two pipelines, and in particular than NPIPE, which shows a clearly decreasing trend toward high multipoles.

These differences are further elucidated in Fig. 25, which simply shows the power spectrum ratios between *Planck* 2018 and NPIPE, respectively, and BeyondPlanck. Again, we see that the three codes generally agree to well within 1 % in TT in the signal-dominated regimes of the spectra, but diverge in the noise-dominated regimes. Indeed, at the highest multipoles for 30 and 70 GHz NPIPE typically exhibits 10–15 % less white noise than BeyondPlanck, while BeyondPlanck exhibits 10 % less noise than *Planck* 2018. As discussed in Planck Collaboration Int. LVII (2020), NPIPE achieves lower noise than *Planck* 2018 primarily through three changes. First, NPIPE exploits the so-called repointing periods in the *Planck* scanning strategy, i.e., the periods during which the spin axis of the satellite moves, which account for about 8 % of the total data volume. Second, NPIPE smoothes the reference load LFI data prior to TOD differencing, as described in Sect. 5.1.3, and this results in a similar noise reduction. Third, NPIPE includes data from the so-called "ninth survey" at the end of the *Planck* mission, which accounts for about 3 % of the total data volume. In contrast, BeyondPlanck currently uses the repointing data, but neither smooths the reference load (essentially only because of limited time for implementation and analysis), nor includes the ninth survey. The reason for the latter is that we find that the TOD χ^2 statistics during this part of the mission show greater variation from PID to PID, suggesting less stability of the instrument. To be cautious, these data are therefore also omitted for now, similar to the horn 26 data, but may be included later.

These effects explain the different white noise levels. However, they do not (necessarily) explain the different slopes of the spectra, which instead indicate that the level of correlated noise is significantly lower in the BeyondPlanck maps as compared to the other two pipelines. The main reason for this is as follows: While *Planck* 2018 and NPIPE both destripe each frequency map independently, BeyondPlanck effectively performs joint correlated noise estimation using all available frequencies at once, as described in Sect. 8.3.2. This happens when conditioning on the current sky model during the correlated noise estimation phase, as opposed to applying the destriping projection operator Z independently to each channel. Thereby, the 30 GHz channel is in effect helped by the 70 GHz channel to separate true CMB fluctuations from its correlated noise, while the 70 GHz channel is helped by the 30 GHz channel to separate synchrotron and free-free emission from its correlated noise. And both 30 and 70 GHz are helped by both *WMAP* and HFI to separate thermal and spinning dust from correlated noise. Of course, this also means that the correlated noise component are correlated *between* frequency channels, and it is therefore imperative to actually use the Monte Carlo samples themselves to propagate uncertainties faithfully throughout the system.[20]

[20] It should of course be noted that the traditional pipelines also exhibit a correlated noise component between different frequencies, simply because they use the same foreground sky model to estimate bandpass corrections at different frequencies. This, however, is very difficult to

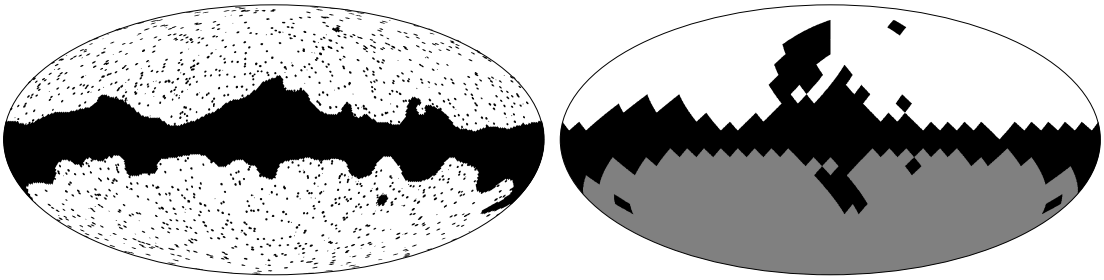

Fig. 27. Temperature (*left*) and polarization (*right*) confidence masks used for BeyondPlanck CMB analysis. In the right panel, the white-plus-gray region defines the BeyondPlanck "full-sky" mask with an accepted sky fraction of 74 %, while the white-only region defines the default BeyondPlanck mask with an accepted sky fraction of 36 %. The sky fraction of the temperature mask is $f_{\rm sky} = 0.69$.

9.3. Residual maps and masking

Having characterized the frequency maps, we will soon turn our attention to the astrophysical and cosmological products as defined by the sky model in Sect. 3.6. However, before doing so, we consider in this section the overall goodness-of-fit of the sky model, potentially unmodelled systematics, and masking required for CMB analysis.

The most detailed and informative summary of goodness-of-fit of a given sky model is provided through pixel-by-pixel residual maps, $\boldsymbol{r}_\nu = \boldsymbol{m}_\nu - \boldsymbol{s}_\nu$. These are shown in Fig. 26 for the LFI channels; for a discussion of all residuals included in the current analysis, we refer the interested reader to Andersen et al. (2020) (for temperature analysis) and Svalheim et al. (2020b) (for polarization analysis).

Starting with the temperature residual maps, we first of all see an imprint of the Galactic plane. This is expected, since the adopted foreground model is quite minimal, and the fits rely on strong priors from external observations; although the current data combination obviously has the statistical strength to identify model errors, it is for instance not strong enough to robustly distinguish between synchrotron spectral index variations from spatial variations in the free-free electron temperature. In particular, the morphology seen in the temperature residuals in Fig. 26 matches well that of Galactic dust emission, which suggests a mild deficiency in the AME model used for the current processing. Further data from experiments like C-BASS (Jew et al. 2019) and QUIJOTE (Génova-Santos et al. 2016) is obviously needed to refine these models, and as a result, a Galactic mask needs to be imposed before performing high-precision CMB analysis. However, we do note that the magnitude of these residuals is modest, with typical peak-to-peak amplitudes smaller than 3 μK at Galactic latitudes higher than a few degrees.

At high Galactic latitudes, we see clear point source residuals in the 30 GHz channel. We note that many radio sources are intrinsically variable, and such variability is not accounted for in the current model.

Turning our attention to the polarization maps, we note that both the 30 and 70 GHz residual maps appear quite clean, and are dominated by white noise, although some slight hints of large-scale systematics may be seen. We recall that similar structures were seen in the sample-to-sample difference maps shown in Fig. 21, and these uncertainties are thus at least partially accounted for in the model; whether they are fully accounted for, will only be clear after a full χ^2 analysis, which will be presented later.

However, while the 30 and 70 GHz channels appear clean, the 44 GHz channel exhibits some stronger artifacts, with a morphology that was already pointed out in Sect. 9.2; there are some PID ranges with incorrect large-scale corrections that result in stripes seen near the right edge of the map. Although the nature of these stripes is still not understood, we note that their presence did not become visually clear until after the completion of the current BeyondPlanck processing, and only with these results in hand do we now have the necessary tools to track them down. Given that limited progress on understanding the problems regarding the 44 GHz channel was made during the official *Planck* analysis period, we consider this identification an important, if preliminary, success of the current methodology, and a demonstration of usefulness for identifying and isolating low-level systematic effects.

These effects all correspond to unmodelled systematics in the current BeyondPlanck processing, and must be expected to contaminate the final high-level results at some level. At the same time, their actual impact may be small for any given specific application. In general, each higher-order analysis should therefore be accompanied with an appropriately defined goodness-of-fit assessment, typically involving χ^2 calculations that account for the full uncertainties as described by the Markov Chain ensemble.

However, some effects are more striking than others, and require special attention. The most prominent example of such is the Galactic plane, which obviously needs to be masked for precision CMB applications. For temperature, we construct a mask using both the residual map information shown here; combined with pair-wise difference maps evaluated between different algorithms (following Planck Collaboration IV 2020); and finally a point source mask. The full procedure is summarized in Colombo et al. (2020). For polarization, we adopt the product of the *Planck* 2018 common confidence mask and the 9-year *WMAP* polarization analysis mask as our baseline to remove the Galactic plane. However, in recognition of the residuals seen in the 44 GHz channel, covering much of the Southern Galactic hemisphere, we also define a special mask that leaves unmasked only the northern Galactic hemisphere. We will refer to these two masks as the "full-sky" and default masks, respectively, which have accepted sky fractions of 74 and 36 %. These are shown in Fig. 27. Both masks will be considered for CMB analysis, but as we will see, the full-sky mask results in large χ^2 excesses that effectively prohibits robust CMB inference.

both quantify or propagate, because of the substantial cost of including full component separation within a forward simulation pipeline.

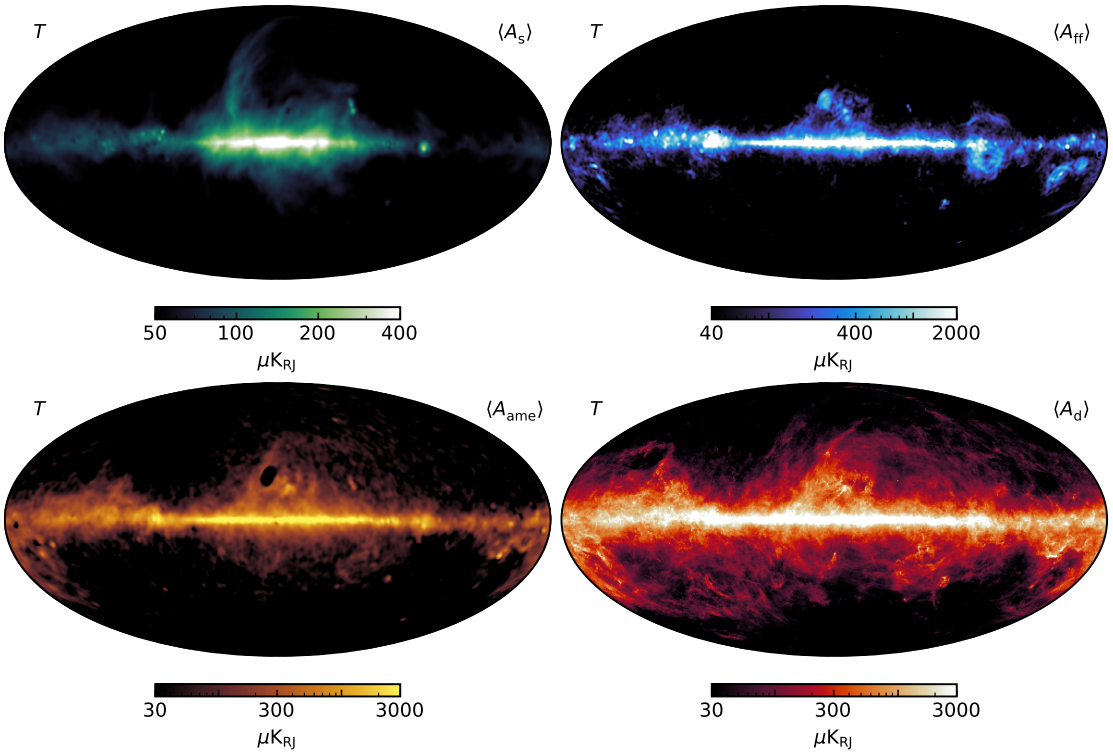

Fig. 28. Posterior mean maps of the aplitude of each of the four intensity foreground components included in the BeyondPlanck analysis. (*Top left:*) Synchrotron amplitude, evaluated at 30 GHz and smoothed to 2° FWHM resolution. (*Top right:*) Free-free amplitude, evaluated at 40 GHz and smoothed to 30′ FWHM resolution. (*Bottom left:*) AME amplitude, evaluated at 22 GHz and smoothed to 2° FWHM resolution. (*Bottom right:*) Thermal dust amplitude, evaluated at 545 GHz and smoothed to 10′ FWHM resolution. Note that the color bars vary between panels. See Andersen et al. (2020) for further discussion of these maps.

9.4. Astrophysical component posteriors

We now turn our attention to the astrophysical component posteriors. However, before presenting the results, we recall that a main design feature of the current analysis was to let the LFI data play the main role in the CMB reconstruction. In practice, this means that neither the CMB-dominated HFI frequencies, nor the *WMAP K*-band observations, are included in the analysis. As a result, we note that the derived foreground posterior constraints shown here are significantly weaker than those presented by the *Planck* team in Planck Collaboration X (2016), Planck Collaboration IV (2020), and Planck Collaboration Int. LVII (2020). Full joint analysis of all data sets is left for future work.

With that caveat in mind, Fig. 28 shows the posterior mean maps for each of the four modelled temperature foregrounds, namely synchrotron, free-free, AME, and thermal dust emission. As discussed by Andersen et al. (2020), these are consistent with earlier results of the same type (Planck Collaboration X 2016), but with notably higher uncertainties, because of the more limited data set employed here.

Similarly, Fig. 29 shows the posterior mean amplitude for polarized synchrotron emission, and Fig. 30 summarizes the posterior mean (left panel) and standard deviation (right panel) for the power-law index of polarized synchrotron emission. In this case, it is worth pointing out that the *Planck* team never published a joint polarized synchrotron solution that included both *Planck* and *WMAP* observations, for the simple reason that these data sets could never made to agree statistically to a satisfactory degree when analyzed separately; when attempting to fit a single synchrotron spectral index across both data sets, the resulting constraints were clearly nonphysical, and led to large χ^2 excesses.

Thus, the BeyondPlanck analysis represents the first reduction of the *Planck* LFI data set for which a joint foreground polarization analysis with *WMAP* yields statistically meaningful results. However, as shown by Svalheim et al. (2020b), even the combination of the two data sets does not constrain the spectral index very strongly, and for this reason we choose to fit only a small number of independent spectral indices across the sky. Specifically, we partition the sky into four disjoint regions, corresponding to the Galactic Center (GC), the Galactic Plane (GP), the North Galactic Spur (NGS), and High Galactic Latitudes (HGL), and treat each region separately. Adopting Planck Collaboration X (2016) as a reference, we enforce a Gaussian prior of $\beta_s \sim N(-3.1, 0.1^2)$. Finally, each spectral index sample is smoothed with a Gaussian beam of 10° FWHM to avoid edge effects.

For the GP and NGS regions, which both have significant signal-to-noise ratio with respect to polarized synchrotron emission and low systematic effects, we fit β_s using the full posterior

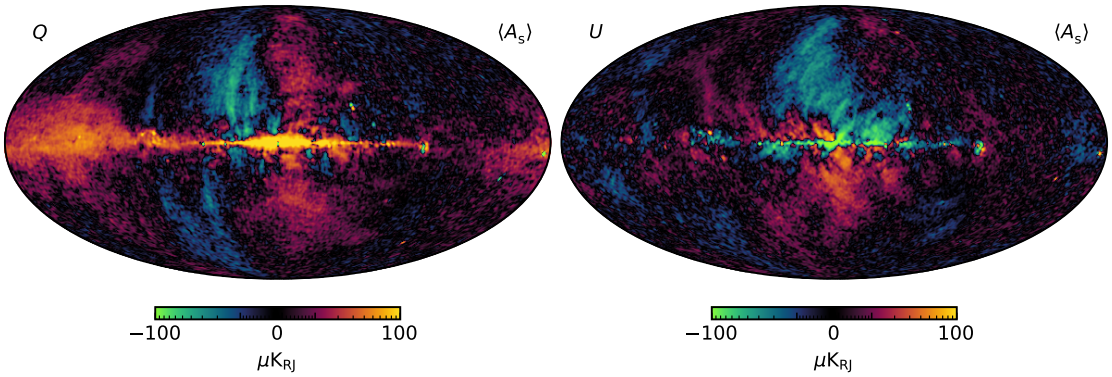

Fig. 29. Posterior mean maps of polarized synchrotron amplitude derived from BeyondPlanck, evaluated at 30 GHz and smoothed to an angular resolution of 1° FWHM. The two columns show Stokes Q and U parameters, respectively; see Svalheim et al. (2020b) for further discussion of these maps.

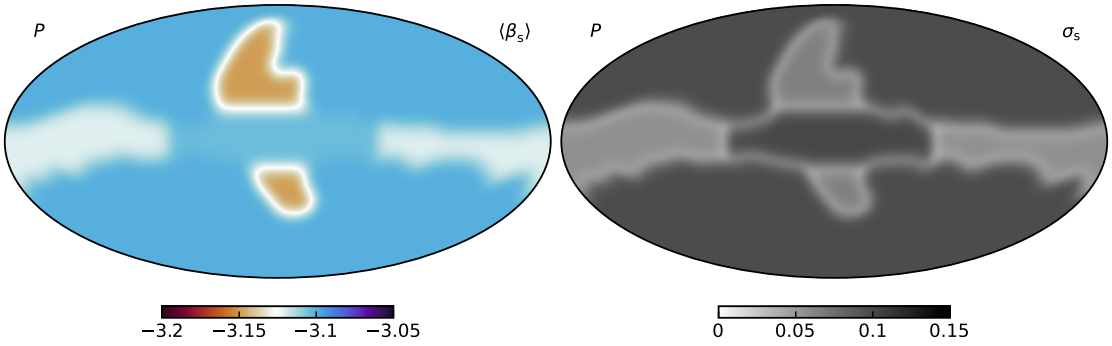

Fig. 30. Posterior mean and standard deviation maps for the spectral index of polarized synchrotron emission, β_s. Note that β_s is fitted in terms of four disjoint regions, each with a constant value but smoothed with a 10° FWHM Gaussian beam to avoid edge effects. The effect of this smoothing is seen in both the mean and standard deviation maps.

distribution as described in Sect. 8.3.5. However, for the HGL region, in which the effective synchrotron signal-to-noise ratio is very low, we simply marginalize over the prior, and exclude the likelihood term. The reason for this is simply that unconstrained degeneracies with other parameters, such as the gain, tend to bias β_s toward high values ($\beta_s^{\rm HGL} \approx -2.5$; see Svalheim et al. 2020b) when fitted freely.

We also do the same for the GC region, for which temperature-to-polarization leakage and bandpass effects are particularly important, and the synchrotron signal may also be biased by Faraday rotation. When fitting this region freely, we find an effective spectral index of $\beta_s^{\rm GC} \approx -4$, which is also clearly unphysical. Rather than letting these unmodelled systematic effects feed into the other components, we marginalize over the physically motivated prior.

This leaves us with two main regions usable for scientific interpretation, and these may be seen as blue regions in the standard deviation map in Fig. 30. Specifically, we find $\beta_s^{\rm GP} = -3.14 \pm 0.05$ and $\beta_s^{\rm NGS} = -3.19 \pm 0.05$, respectively (Svalheim et al. 2020b). On the one hand, we note that these values are broadly consistent with previous temperature-only constraints, such as those reported by Planck Collaboration X (2016), who found $\beta_s = -3.1$. On the other hand, our results show no compelling evidence for a significant spectral steepening from low to high Galactic latitudes, as for instance reported by (Kogut 2012) and Fuskeland et al. (2014, 2019). Rather, our results are qualitatively more similar to those derived using *WMAP* polarization data alone by Dunkley et al. (2009), who found a difference of only $\Delta\beta_s = 0.08$ between low and high Galactic latitudes. In this respect, it is worth noting that the low Galactic latitudes are particularly sensitive to both systematic and astrophysical modelling errors, both in temperature and polarization. For a full discussion of these results, we refer the interested reader to Svalheim et al. (2020b).

9.5. CMB posteriors

Finally, we arrive at the main scientific target application of the paper, the CMB posteriors. We start with a discussion of the CMB dipole in Sect. 9.5.1, before presenting the CMB fluctuation maps in Sect. 9.5.2. The BeyondPlanck CMB low-ℓ power spectrum and likelihood are discussed in Sect. 9.5.3, and the high-ℓ power spectrum and likelihood are discussed in Sect. 9.5.4. Finally, cosmological parameters are presented in Sect. 9.5.5.

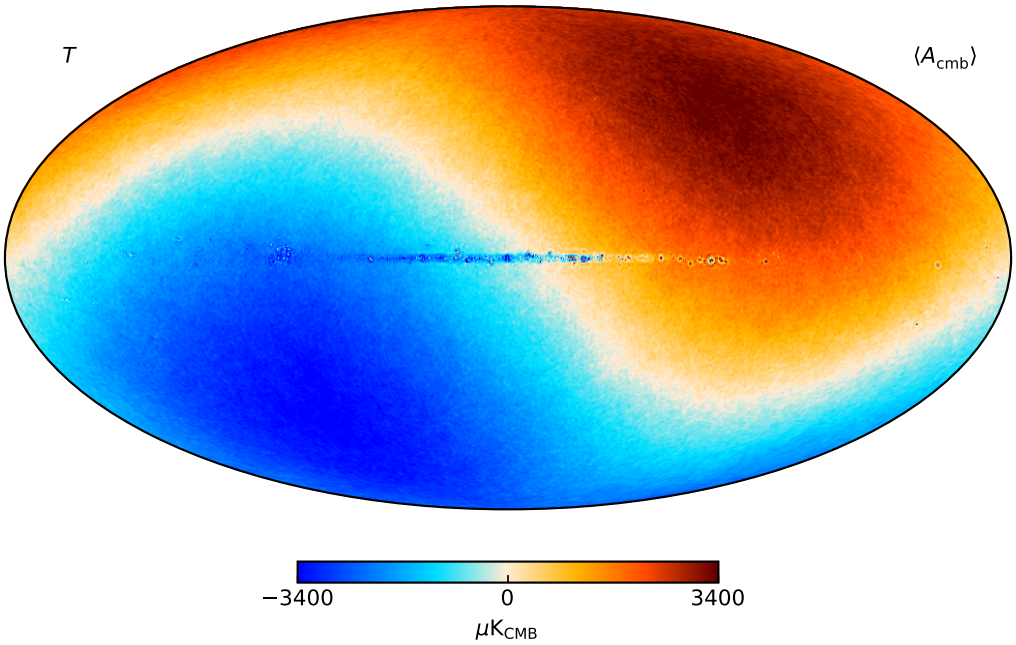

Fig. 31. Posterior mean CMB BeyondPlanck temperature map, smoothed to an angular resolution of 14′ FWHM.

Table 3. Comparison of Solar dipole measurements from *COBE*, *WMAP*, and *Planck*.

		Galactic coordinates		
Experiment	Amplitude [μK_{CMB}]	l [deg]	b [deg]	Reference
COBE[a,b]	3358 ± 23	264.31 ± 0.16	48.05 ± 0.09	Lineweaver et al. (1996)
WMAP[c]	3355 ± 8	263.99 ± 0.14	48.26 ± 0.03	Hinshaw et al. (2009)
LFI 2015[b]	3365.5 ± 3.0	264.01 ± 0.05	48.26 ± 0.02	Planck Collaboration II (2016)
HFI 2015[d]	3364.29 ± 1.1	263.914 ± 0.013	48.265 ± 0.002	Planck Collaboration VIII (2016)
LFI 2018[b]	3364.4 ± 3.1	263.998 ± 0.051	48.265 ± 0.015	Planck Collaboration II (2020)
HFI 2018[d]	3362.08 ± 0.99	264.021 ± 0.011	48.253 ± 0.005	Planck Collaboration III (2020)
NPIPE[a,c]	3366.6 ± 2.6	263.986 ± 0.035	48.247 ± 0.023	Planck Collaboration Int. LVII (2020)
BeyondPlanck[e]	**3359.5 ± 1.9**	**263.97 ± 0.09**	**48.30 ± 0.03**	Section 9.5

[a] Statistical and systematic uncertainty estimates are added in quadrature.
[b] Computed with a naive dipole estimator that does not account for higher-order CMB fluctuations.
[c] Computed with a Wiener-filter estimator that estimates, and marginalizes over, higher-order CMB fluctuations jointly with the dipole.
[d] Higher-order fluctuations as estimated by subtracting a dipole-adjusted CMB-fluctuation map from frequency maps prior to dipole evaluation.
[e] Estimated with a sky fraction of 68 %. Error bars include only statistical uncertainties, as defined by the global BeyondPlanck posterior framework, and they thus account for instrumental noise, gain fluctuations, parametric foreground variations etc. However, they do not account for prior or model selection uncertainties; see Sect. 9.5.1 for a discussion of these priors.

9.5.1. The CMB solar dipole

In the BeyondPlanck framework, the CMB dipole is in principle estimated on completely the same footing as any other mode in the CMB sky, and is represented in terms of three spherical harmonic coefficients in $\boldsymbol{s}_{\rm CMB}$. No special-purpose component separation algorithms are applied to derive the CMB dipole, nor does any individual frequency play a more important role than others, except for as dictated by the relative level of instrumental noise in each channel.

However, as discussed by Ihle et al. (2020), Gjerløw et al. (2020), and Suur-Uski et al. (2020), this apparent algorithmic simplicity does not imply that robust CMB dipole estimation is by any means *easy* in the BeyondPlanck procedure. Indeed, the CMB dipole is quite possibly the single most difficult parameter to estimate in the entire model, simply because it both affects, and relies on, a wide range of other parameters in the model. Some of the most important degeneracies are the following:

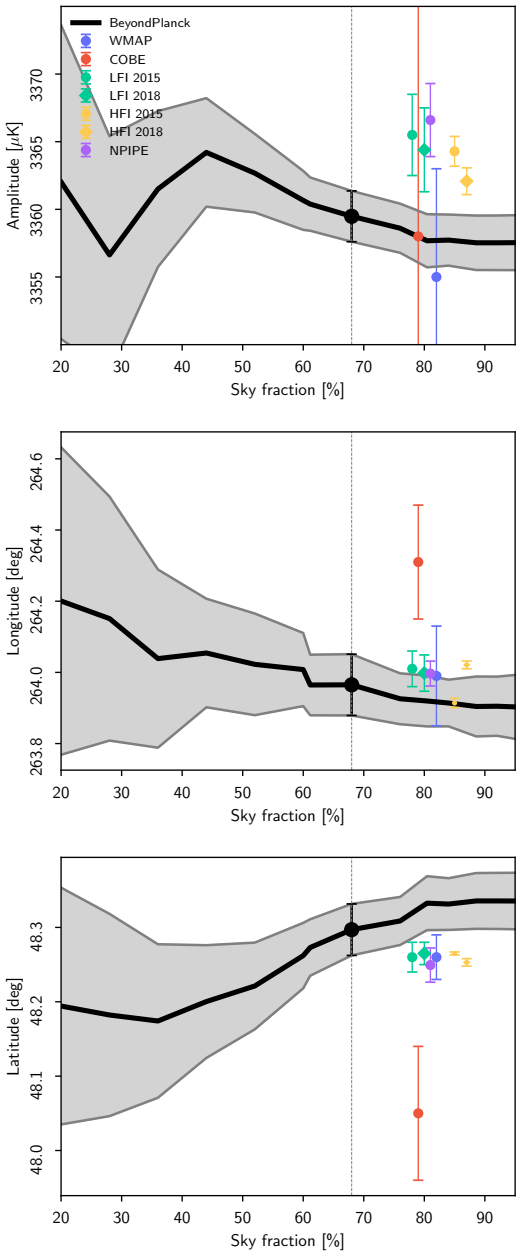

Fig. 32. CMB dipole parameters as a function of sky fraction.Gray bands indicate 68 % posterior confidence regions.

1. *Absolute calibration* – g_0 is the single most important parameter for a robust solar dipole determination, as it directly scales the amplitude of the CMB map. This parameter is itself constrained from the orbital dipole, which is both relatively weak in terms of absolute amplitude, and for significant parts of the mission it is aligned with, and thereby obscured by, the Galactic plane.
2. *Astrophysical foregrounds* – foregrounds have in general both non-zero dipole and higher-order moments with unknown parameters, and must be estimated jointly with the CMB dipole. However, considering that the current data set includes five astrophysical components, each with a free value in each pixel, and we only have eight different frequency maps, the full system is quite poorly constrained; it is therefore possible to add a significant dipole to the CMB map, and subsequently subtract appropriately scaled dipoles from each of the foreground maps, with only a minimal penalty in terms of the overall χ^2. In practice, we see particularly strong degeneracies between the CMB, AME and free-free components, when exploring the full system without priors, leading to massive and obviously nonphysical marginal uncertainties.
3. *Correlated noise* – $n_{\rm corr}$ is only weakly constrained through its PSD parameters, and is therefore able to account for a wide range of modelling errors, including calibration errors. In particular, incorrectly estimated gains will leave a spurious dipole-like residual in the time-ordered data. Since this spurious residual clearly is detector-dependent, it will typically be interpreted by the algorithm as correlated noise, and thereby excite a dipolar structure in $\boldsymbol{n}_{\rm corr}$.
4. *Large-scale CMB quadrupole, foreground and bandpass corrections* – while the CMB polarization quadrupole is predicted by current ΛCDM models to have a very small quadrupole, with a variance of typically less than 0.05 μK^2, there is nothing in the current parametric model that explicitly enforces this. This particular mode therefore opens up a particularly problematic degeneracy for *Planck* through coupling with the gain and bandpass shift as follows: An error in the absolute gain leads to an apparently wrong orbital dipole. However, this can be countered by adding a polarized CMB quadrupole, which has the same SED and nearly the same spin harmonics as the orbital dipole, due to the *Planck* scanning strategy that observes along nearly perfect great circles.[21] Errors in the total polarized sky signal as observed at each frequency can finally be countered by adjusting the combination of relative gains, polarized foreground signals, and bandpass corrections between radiometers, leaving the total χ^2 nearly unchanged.

During the initial test phase of the BeyondPlanck pipeline, the Markov chain was allowed to explore these degeneracies freely, in order to understand their nature, and these runs resulted in a full marginal uncertainty on the dipole amplitude of more than 40 μK, as compared to 3 μK reported by *Planck* LFI for the 70 GHz channel alone (Planck Collaboration II 2020), or 1 μK as reported by HFI (Planck Collaboration III 2020). Although this value by itself could be considered acceptable, given the limited cosmological importance of the CMB dipole, it was also strikingly obvious that all component maps were fully compromised by the poorly constrained calibration, ultimately leading to obviously nonphysical astrophysical component maps.

To break these degeneracies, we instead impose the following effective priors in the analysis pipeline, as already discussed in Sect. 8.3 and by Gjerløw et al. (2020):

1. We estimate the absolute calibration, g_0, using only the orbital dipole as a calibrator; see Sect. 8.3.1. This significantly

[21] This particular degeneracy does not exist for *WMAP*, because of its more complex scanning strategy.

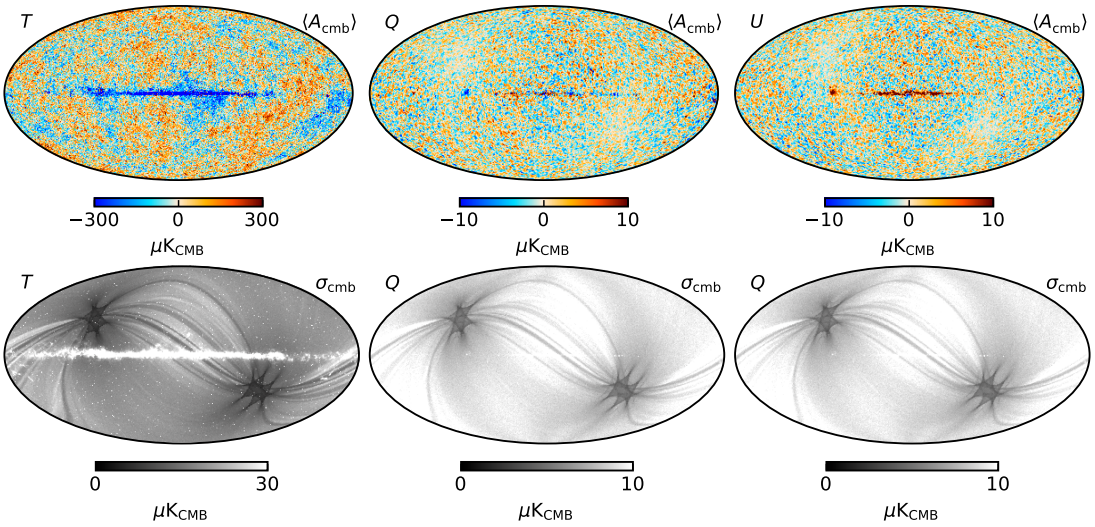

Fig. 33. BeyondPlanck posterior mean (*top row*) and standard deviation (*bottom row*) CMB fluctuation maps. Columns show, from left to right, temperature and Stokes Q and U parameters, respectively. The temperature maps are smoothed to 14′ FWHM resolution, while the polarization maps are smoothed to 1° FWHM.

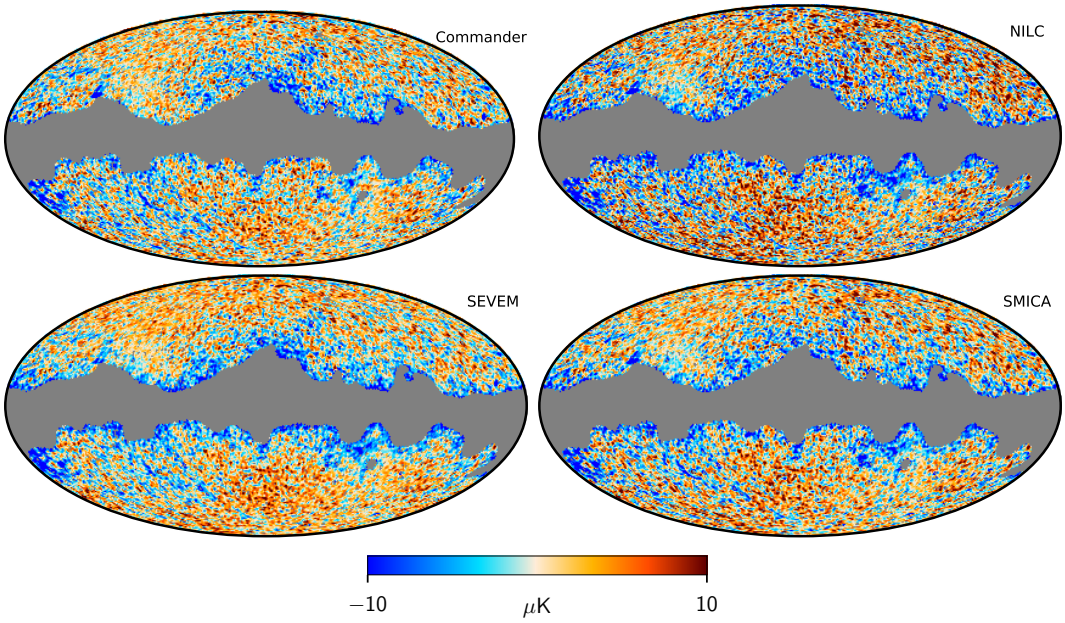

Fig. 34. Difference maps between the BeyondPlanck CMB temperature map and those derived from the full *Planck* 2018 data set (Planck Collaboration IV 2020). From left to right and from top to bottom, the various panels show differences with respect to `Commander`, `NILC`, `SEVEM`, and `SMICA`. All maps are smoothed to a common angular resolution of 1° FWHM.

reduces the degeneracy between the foregrounds and the absolute calibration.

2. We enforce active spatial priors on the free-free and AME amplitude maps, as discussed by Andersen et al. (2020). Specifically, we use (an appropriately scaled version of) the *Planck* 857 GHz map as a spatial prior for AME (Planck Collaboration Int. LVII 2020), and the *Planck* 2015 free-free map (Planck Collaboration X 2016) as a prior for free-free emission. In effect, we thus incorporate external information

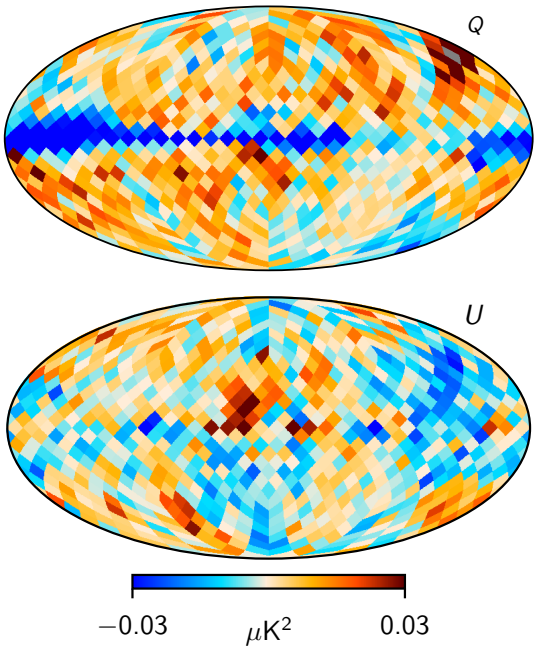

Fig. 35. Single column of the low-resolution CMB noise covariance matrix, constructed as the mean outer-product of Monte Carlo samples. The column corresponds to Stokes Q pixel number 100, which is marked in gray, and located in the top right quadrant near the 'Q' label. Note that non-zero correlations are detected nearly across the full sky in both Stokes parameters.

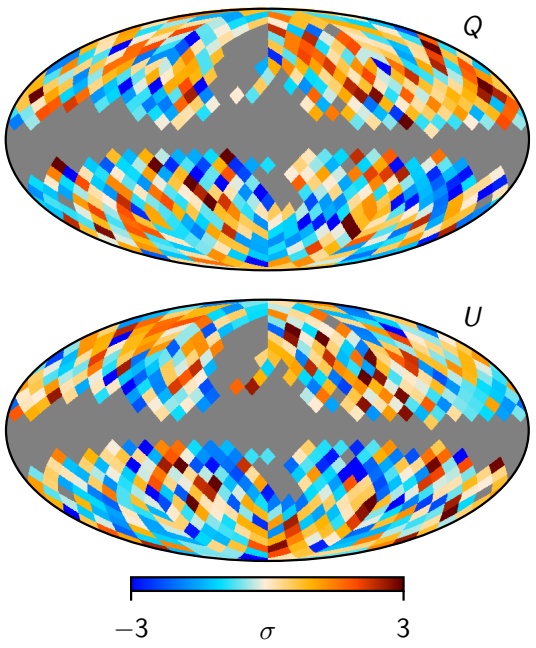

Fig. 36. BeyondPlanck low-resolution and "whitened" CMB polarization map, as defined by $\mathsf{N}_{\mathrm{CMB}}^{-1/2}\boldsymbol{s}_{\mathrm{CMB}}$ at a `HEALPix` resolution of $N_{\mathrm{side}} = 8$ and masked with the BeyondPlanck "full-sky" mask. Top and bottom panel shows Stokes Q and U parameters, respectively, and the color scales span $\pm 3\sigma$.

from *Planck* HFI to help stabilize the model, but without introducing data that directly constrains the CMB component.

3. We set the CMB polarization quadrupole to zero during the relative gain calibration phase. This is less strict than the LFI DPC and `NPIPE` procedures, both of which set the entire CMB polarization signal to zero in this step (Planck Collaboration II 2020; Planck Collaboration Int. LVII 2020). As discussed by Gjerløw et al. (2020), the strong coupling between the gain and the CMB polarization is associated with the quadrupole alone, and not higher-order modes. As a formal justification of this prior, one may argue that this corresponds to imposing a ΛCDM prior to that particular mode. In principle, this does bias the resulting posterior distribution; however, as long as ΛCDM is a reasonable approximation to the true sky, then the effect is negligibly small, because of the very small predicted amplitude of the CMB quadrupole. Furthermore, the true CMB polarization quadrupole is likely to be even smaller than that predicted by ΛCDM, since the CMB temperature quadrupole is already known be low at the 1 % level, and these two are correlated through the TE power spectrum. For additional verification, we will in the following estimate cosmological parameters both with and without the EE $\ell = 2$ mode included, and we find no significant difference.

Figure 31 shows the marginal CMB temperature fluctuation posterior mean map as derived in BeyondPlanck, given both the data, model and priors described above. This map is massively dominated by the CMB solar dipole, with only a small imprint of the Galactic plane being visible in the very center. At high latitudes, CMB temperature fluctuations may be seen as tiny ripples superimposed on the dipole.

Because of the small but non-negligible Galactic plane, we must impose an analysis mask before estimating final dipole parameters. For this purpose, we use the Wiener filter estimator described by Thommesen et al. (2020), which in-paints the Galactic mask with a constrained realization prior to parameter estimation; this is necessary in order to account for, and marginalize over, coupling to higher-order CMB fluctuations. This method was also adopted for the dipole estimates presented in Planck Collaboration Int. LVII (2020), although we introduce one significant difference to that analysis: In the current analysis we estimate the magnitude of systematic uncertainties directly from the BeyondPlanck Gibbs samples, as opposed to putting in it by hand. Specifically, instead of producing 9000 constrained realizations from a single maximum likelihood map, as done in Thommesen et al. (2020) and Planck Collaboration Int. LVII (2020), we now produce 100 constrained realizations from each of the 900 available Gibbs samples. Since each of these realizations have different gain, correlated noise, and foreground residuals, the full ensemble therefore now accounts for these uncertainties automatically. The only additional term we put by hand into to the error budget is a contribution of 0.7 μK from the CMB monopole uncertainty itself (Fixsen 2009).

Using this methodology, we estimate the CMB dipole parameters over a series of Galactic masks, ranging in sky fraction from 20 to 95 %. The results from these calculations are shown in Fig. 32. Overall, we see that the posterior distributions are quite

stable with respect to sky fraction. Furthermore, we note that the uncertainties do not decrease after $f_{\rm sky} \approx 0.7$, as they would if the full error budget could be described in terms of white noise and sky fraction. Rather, the weight of the additional sky coverage is effectively reduced when marginalizing over the various systematic contributions, as desired. We adopt a sky fraction of $f_{\rm sky} = 0.77$ to define our final dipole estimates, corresponding to the sky fraction used for the main CMB temperature analysis. The resulting values are plotted as black points in Fig. 32, and tabulated together with previous estimates in Table 3.

Several points are worth noting regarding these results. First, we see that the reported best-fit BeyondPlanck dipole amplitude is 3359.5 ± 1.9 μK, which is lower than the latest LFI 2018 estimate of 3364.4 ± 3.1 μK, which in turn is lower than the NPIPE estimate of 3366.6 ± 2.6 μK. In fact, it is even lower than the latest HFI estimate of 3362.08 ± 0.99 μK. On the other hand, it is higher than the *WMAP* estimate of 3355 ± 8 μK. Algorithmically speaking, this makes intuitively sense, considering that *WMAP* observations actually are used actively in the BeyondPlanck analysis, while it is not in the *Planck*-only analyses. At the same time, we note that the BeyondPlanck estimate is statistically consistent with any one of these estimates, as measured in terms of their error bars; the biggest outlier is NPIPE, which still only represents a 2.7σ variation.

Regarding the directional parameters, two observations are worth pointing out. First, we see that the BeyondPlanck uncertainties are substantially larger than any of the previous *Planck*-dominated results. Here it is worth recalling again that no additional systematic error contributions are added by hand to the BeyondPlanck directional uncertainties, and the reported values are thus the direct result of degeneracies within the model itself. Perhaps the biggest algorithmic difference in this respect is the fact that the current algorithm explicitly marginalizes over the full foreground model, while most other approaches condition on external constraints. The second observation is that the BeyondPlanck latitude is higher than any of the previous results. The statistical significance of this difference is modest, only about 1–2 σ, but compared with the remarkable internal agreement between *Planck* and *WMAP*, it is still noteworthy. In this respect, we recall that we are currently using the *Planck* 2015 free-free map as an informative prior in the current processing, and CMB and and free-free emission are known to be strongly correlated for the current data set; see Andersen et al. (2020). Performing a joint analysis of LFI, HFI, and *WMAP* without an external free-free prior might be informative regarding this point.

9.5.2. CMB maps

Next, we consider the CMB fluctuation maps, as shown in Fig. 33. The top row shows the posterior mean (after subtracting the dipole from the temperature component), while the bottom row shows the posterior standard deviation. We see a narrow Galactic plane imprint in both temperature and polarization, and both in the mean and standard deviation maps. Fortunately, at least at a visual level, the obviously offending features are well covered by the analysis masks shown in Fig. 27.

Figure 34 shows difference maps between the BeyondPlanck temperature map and the four official *Planck* 2018 foreground-cleaned CMB maps (Commander, NILC, SEVEM, and SMICA; see Planck Collaboration IV 2020). A best-fit monopole and dipole have been removed outside the *Planck* 2018 common confidence mask in each case, and all maps are smoothed to a common resolution of 1° FWHM. Several notable feature may be seen in these difference maps. First of all, we notice a significant amount of random noise at high Galactic latitudes, which is due to the fact that the BeyondPlanck CMB map does not include HFI data.

At low Galactic latitudes, the main difference is a blue Galactic plane. From the top left panel in Fig. 33, we see that this is indeed coming from the BeyondPlanck map, indicating that the current BeyondPlanck signal model over-subtracts foregrounds in the Galactic plane. At the same time, by comparing this figure to Fig. 7 in Planck Collaboration IV (2020) (which shows pairwise differences between each of the four *Planck* 2018 CMB maps), we see that the absolute internal differences within the *Planck* 2018 maps are of comparable order-of-magnitude as the difference between BeyondPlanck and *Planck* 2018, although generally covering a slightly smaller sky fraction.

9.5.3. Low-ℓ CMB likelihood

Our main scientific goal with the BeyondPlanck pipeline is to constrain the CMB power spectrum and cosmological parameters through end-to-end analysis. Both of these operations are most conveniently facilitated through the CMB power spectrum likelihood, $\mathcal{L}(C_\ell)$. As discussed in Sect. 8.3.8, it is most convenient to split this function into two components, one low-dimensional and low-ℓ component that employs dense matrix operations, and one high-dimensional and high-ℓ component that employs faster operations; a similar split is used by both *WMAP* (Hinshaw et al. 2013) and *Planck* (Planck Collaboration V 2020).

As described in Sect. 8.3.8, we employ a brute-force Gaussian pixel-based likelihood estimator at low-ℓ, as has been a standard procedure since *COBE*-DMR,

$$P(C_\ell \mid \hat{\mathbf{s}}_{\rm CMB}) \propto \frac{\mathrm{e}^{-\frac{1}{2}\hat{\mathbf{s}}^t_{\rm CMB}(\mathsf{S}(C_\ell)+\mathsf{N})^{-1}\hat{\mathbf{s}}_{\rm CMB}}}{\sqrt{|\mathsf{S}(C_\ell)+\mathsf{N}|}}. \tag{136}$$

The only noteworthy variation is that we employ an optimized basis set that reduce the number of basis vectors in $\hat{s}_{\rm CMB}$ from $N_{\rm pix}$ to the number of modes with a relative signal-to-noise ratio larger than 10^{-6} over $2 \leq \ell \leq 8$; see Gjerløw et al. (2015) for details. We choose an upper truncation of $\ell \leq 8$ as a compromise between minimizing the dimensionality of the noise covariance matrix and retaining as much constraining power with respect to τ as possible. As shown by Colombo et al. (2020), all main results shown in the following are robust against variations in this cut-off.

We construct the inputs to the low-ℓ likelihood as follows:

1. For each of the 900 full-resolution Gibbs samples discussed in the introduction to this section, we draw 50 new samples from $P(\mathbf{s}^{\ell\leq 64}_{\rm CMB} \mid \boldsymbol{d}, \omega \setminus \mathbf{s}^{\ell\leq 64}_{\rm CMB})$, resulting in a total of 45 000 low-ℓ CMB map samples. Note that each of these samples represents a complete sample from the full posterior; the only difference is that this particular sub-volume of ω is sampled 50 times more densely than for other parameters.
2. Next, each sample is individually downgraded to $N_{\rm side} = 8$. In temperature, we additionally smooth to 20° FWHM before downgrading to avoid sub-pixel effects from high signal-to-noise high-ℓ fluctuations, while for polarization we simply boxcar average over each $N_{\rm side} = 8$ pixel.
3. We then compute the posterior mean CMB map, $\hat{\mathbf{s}}_{\rm CMB}$ and covariance matrix $\mathsf{N}_{\rm CMB}$ simply by averaging over samples, as given by Eq. (122). Additionally, we add 1 μK of random Gaussian regularization noise to the temperature component, both in $\hat{\mathbf{s}}_{\rm CMB}$ and $\mathsf{N}_{\rm CMB}$, to make the noise covariance matrix non-singular, and we set all off-diagonal temperature-

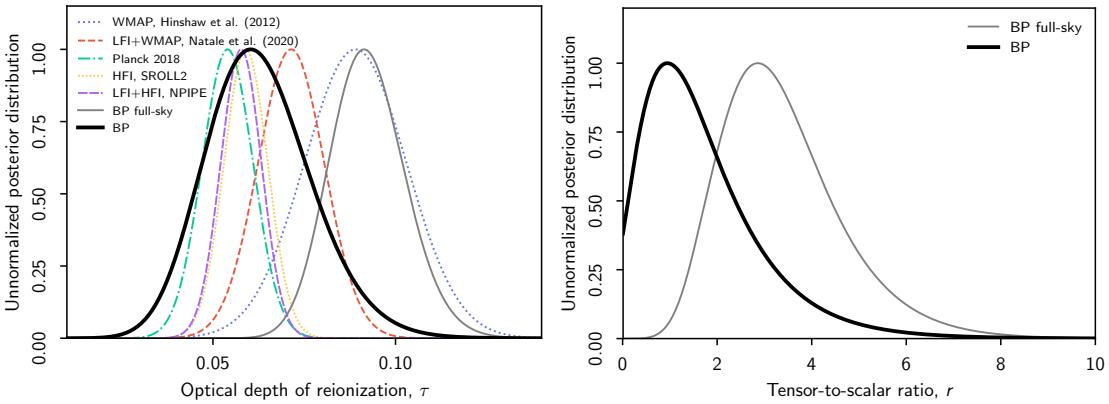

Fig. 37. (*Left panel*:) Comparison of (un-normalized) marginal posterior distributions of the reionization optical depth from *Planck* 2018 (green; Planck Collaboration VI 2020), 9-year *WMAP* (red; Hinshaw et al. 2013), *WMAP Ka–V* and LFI 70 GHz (orange; Natale et al. 2020); and BeyondPlanck *EE*-only using multipoles $\ell = 2$–8 (gray/black; Paradiso et al. 2020). The thin gray line shows BeyondPlanck constraints derived using the full-sky mask ($f_{\rm sky} = 0.74$), while the solid black line shows constraints using the northern mask only ($f_{\rm sky} = 0.36$). (*Right panel*:) Corresponding marginal BeyondPlanck tensor-to-scalar ratio posteriors derived using *BB* multipoles between $\ell = 2$–8, including either the Northern Galactic hemisphere (*black*) or the full sky (*gray*).

Table 4. Summary of cosmological parameters dominated by large-scale polarization, and goodness-of-fit statistics. Note that the BeyondPlanck full-sky case has an unacceptable goodness-of-fit, and should not be used for cosmological analysis. The main science result from the current analysis are summarized in the top two entries, and are evaluated with a small polarization sky fraction. For completeness, the third row shows results evaluated with nearly full-sky data, but these are strongly contamined by systematic errors, as indicated by the high χ^2 value.

Analysis Name	Data Sets	$f_{\rm sky}^{\rm pol}$	τ	$r_{95\%}^{BB}$	χ^2 PTE	Reference
BeyondPlanck, $\ell = 2$–8	LFI, *WMAP Ka–V*	0.36	$0.060^{+0.015}_{-0.013}$	< 4.3	0.16	Paradiso et al. (2020)
BeyondPlanck, $\ell = 3$–8	LFI, *WMAP Ka–V*	0.36	$0.061^{+0.015}_{-0.014}$	< 5.4	0.16	Paradiso et al. (2020)
BeyondPlanck, $\ell = 2$–8, full-sky	LFI, *WMAP Ka–V*	0.74	$0.091^{+0.010}_{-0.098}$	$2.9^{+1.3}_{-1.0}$	$5 \cdot 10^{-4}$	Paradiso et al. (2020)
WMAP 9-yr	*WMAP Ka–V*	0.76	0.089 ± 0.014			Hinshaw et al. (2013)
Natale et al.	LFI 70, *WMAP Ka–V*	0.54	0.071 ± 0.009			Natale et al. (2020)
Planck 2018	HFI 100×143	0.50	0.051 ± 0.009	< 0.41		Planck Collaboration V (2020)
SROLL2	HFI 100×143	0.50	0.059 ± 0.006			Pagano et al. (2020)
NPIPE (Commander CMB)	LFI+HFI	0.50	0.058 ± 0.006	< 0.16		Tristram et al. (2020)

polarization elements in $\mathsf{N}_{\rm CMB}$ to zero, noting that temperature is massively signal-dominated.

4. We apply the Galactic analysis mask (as defined in Fig. 27) by removing the relevant rows and columns from both $\hat{\boldsymbol{s}}_{\rm CMB}$ and $\mathsf{N}_{\rm CMB}$.
5. Finally, we eigenvalue decompose the $\mathsf{S}(C_\ell) + \mathsf{N}_{\rm CMB}$ matrix, where C_ℓ is defined by the best-fit *Planck* 2018 ΛCDM spectrum for $2 \leq \ell \leq 8$, and define the projection operator P as the set of all columns corresponding to an eigenvalue, ϵ_i, larger than $10^{-6} \max(\epsilon_i)$. This multipole range includes all values for which the LFI data has a significant signal-to-noise ratio, while the eigenvalue threshold is chosen to avoid issues with numerical precision.

To build useful intuition regarding $\hat{\boldsymbol{s}}_{\rm CMB}$ and $\mathsf{N}_{\rm CMB}$, we first show an single column of $\mathsf{N}_{\rm CMB}$ prior to eigen-mode compression in Fig. 35; this column corresponds to pixel 100 in the Stokes Q map, which is marked by gray in the figure (located near the Q label). Several interesting features may be seen here. First, we see that there is a broad stripe (or region) extending from the pixel in question in both Q and U. This stripe corresponds to correlated noise and gain fluctuations modulated by the *Planck* scanning strategy. Second, we see that the Galactic plane is negatively correlated with the pixel in question in Q, and positively correlated in U; these correlations result from variations in both gain and bandpass, coupled to the foreground components. Third, we also see that there are significant correlations at further distances in the form of broad extended regions; these are at least partially due to the *WMAP* horn separation of 141°, which in effect couples pixels across the entire sky. The complexity of this map illustrates a significant advantage of the Monte Carlo sampling approach: Constructing this full matrix analytically would be extremely difficult, which is why it has never been done in the CMB literature until now. To date, only correlated noise, absolute calibration, and linear template corrections have been accounted for in this matrix; see, e.g., Hinshaw et al. (2013) and Planck Collaboration V (2020).

Figure 36 shows the "whitened" CMB polarization map, as defined by $\mathsf{N}_{\rm CMB}^{-1/2} \boldsymbol{s}_{\rm CMB}$. This essentially measures the local signal-to-noise ratio in each pixel (taking into account the full covariance matrix structure), and should be consistent with Gaussian random noise with vanishing mean and unit standard deviation for a strongly noise-dominated map. Overall, this map

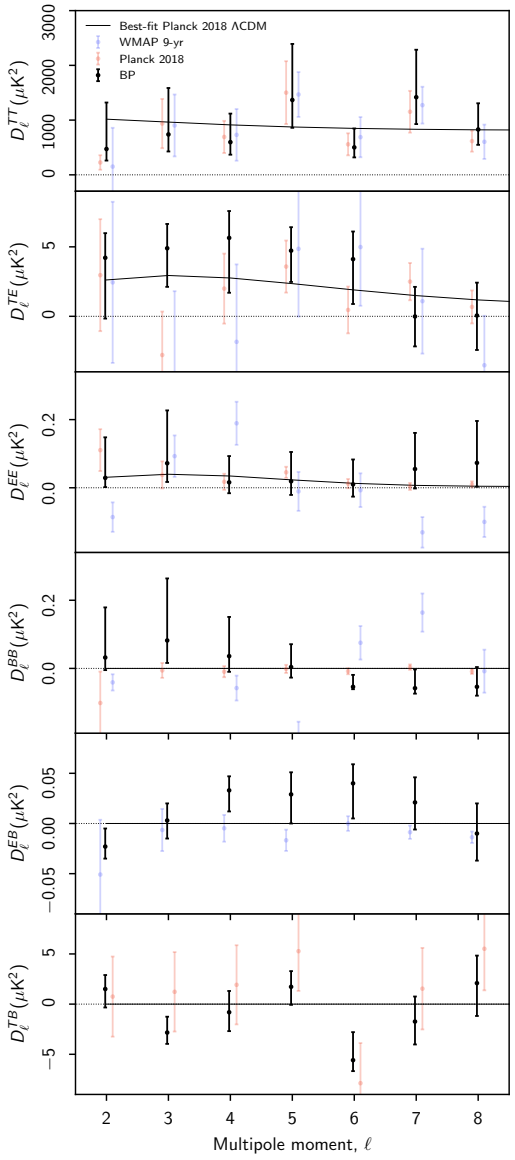

Fig. 38. Comparison between low-ℓ angular CMB power spectra, as derived by the *Planck* collaboration using both LFI and HFI data (blue points; Planck Collaboration V 2020); by the *WMAP* team using just *WMAP* data (red points; Hinshaw et al. 2013); and by BeyondPlanck using both LFI and *WMAP* data (black points; Colombo et al. 2020). Thin black lines indicate the *Planck* 2018 best-fit ΛCDM spectrum (Planck Collaboration VI 2020). The BeyondPlanck data points are evaluated by conditionally slicing the posterior distribution ℓ-by-ℓ with respect to the best-fit ΛCDM model, by holding all other multipoles fixed at the reference spectrum while mapping out $P(C_\ell|\boldsymbol{d})$, to visualize the posterior structure around the peak.

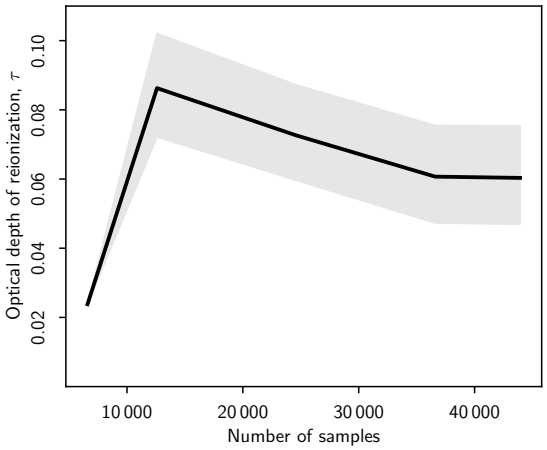

Fig. 39. Convergence of constraints of the reionization optical depth as a function of the number of posterior samples used to construct the CMB mean map and covariance matrix. The solid black line shows the posterior mode for τ, and the gray region shows the corresponding 68 % confidence region.

does appear largely noise dominated, but there are also some intruiging coherent large-scale features that could indicate excess signal. Also, there is fortunately no compelling evidence of residual foreground contamination in this map, and the chosen mask appears to perform well.

As a first quality assesment, we start by slicing the marginal posterior distribution in Eq. (136) with respect to the two main low-ℓ polarization parameters, τ and r. Specifically, we adopt the best-fit *Planck* 2018 ΛCDM model (Planck Collaboration V 2020) to define a reference model, but vary in turn τ and r. Each model spectrum, C_ℓ, is computed with `CAMB` (Lewis et al. 2000). When constraining τ we vary only the *EE* spectrum, while when constraining r we vary only the *BB* spectrum. In all these calculations, we keep $A_\mathrm{s}\mathrm{e}^{-2\tau}$ fixed at the best-fit *Planck* value (Planck Collaboration V 2020) to break a strong degeneracy with the amplitude of scalar perturbations, A_s, essentially adopting the high-ℓ HFI information as a effective prior on the overall amplitude of the spectrum. These calculations are repeated for both the full-sky and the Northern Galactic hemisphere masks as defined above.

To assess the internal consistency of the resulting models, we evaluate the χ^2 of that model with respect to the data in question,

$$\chi^2 = \hat{\boldsymbol{s}}^t_\mathrm{CMB}\left(\mathsf{S}(C_\ell^\mathrm{bf}) + \mathsf{N}_\mathrm{CMB}\right)^{-1}\hat{\boldsymbol{s}}_\mathrm{CMB}. \tag{137}$$

For a Gaussian and isotropic random field, this quantity should be distributed according to a $\chi^2_{n_\mathrm{dof}}$ distribution, where n_dof is the number of degrees of freedom, which in our case is equal to the number of basis vectors in $\hat{\boldsymbol{s}}_\mathrm{CMB}$. The signal covariance matrix is in each case evaluated using the best-fit value of τ derived from the parameter exploration described above. The results are summarized in Fig. 37 and Table 4.

Starting with the full-sky case, we first note that the χ^2 probability-to-exceed (PTE) for this region is $5\cdot10^{-4}$, indicating a poor fit. Furthermore, the best-fit cosmological parameters are $\tau = 0.091^{+0.010}_{-0.098}$ and $r = 2.9^{+1.3}_{-1.0}$, both of which are in strong tension with previous results from *Planck* and other experiments (e.g., BICEP2/Keck Array and Planck Collaborations 2015; Planck Collaboration V 2020; Tristram et al. 2020).

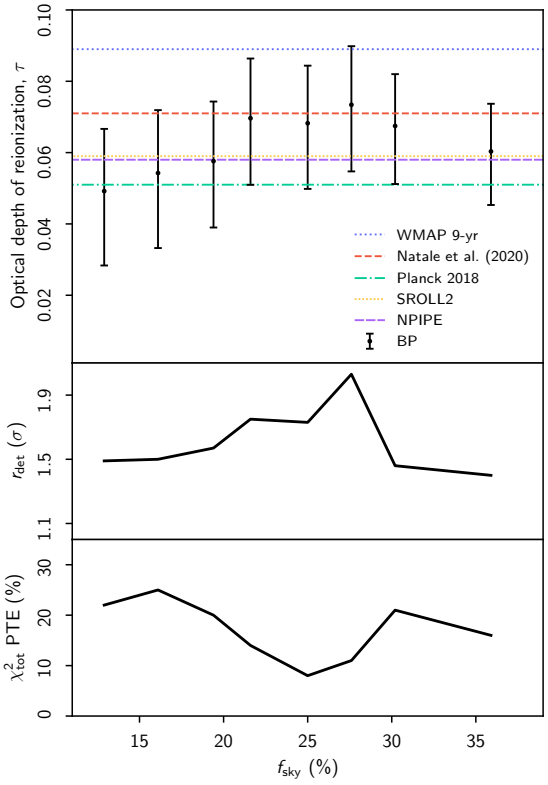

Fig. 40. Estimates of τ (*top panel*), r (*middle panel*), and χ^2 (*bottom panel*) as a function of sky fraction, $f_{\rm sky}$, using only low-ℓ polarization information. The tensor-to-scalar ratio is given as a detection level in units of σ, as measured by the likelihood ratio relative to a model with vanishing B-mode signal.

Thus, even with clearly nonphysical cosmological parameters, this data set fails a basic goodness-of-fit test, and this can only be explained by a break-down of some fundamental assumption, namely either Gaussianity, isotropy, or the basic ΛCDM framework itself. Of course, at this point, we already know about at least one plausible candidate to explain this violation of Gaussianity and statistical isotropy, namely the strong stripes in the 44 GHz channel that are spatially mostly confined to the Southern Galactic hemisphere. Because of this χ^2 excess, we conclude that the BeyondPlanck processing is not yet sufficiently mature to allow analysis of full-sky data.

We therefore turn our attention to the Northern hemisphere alone, for which no strong stripes are seen in the 44 GHz channel. In this case, we see that the χ^2 PTE is 16 %, which is statistically fully acceptable. Furthermore, the best-fit tensor-to-scalar ratio is consistent with zero within a statistical significance of $1.4\,\sigma$, and with an upper 95 % confidence limit of $r < 4.3$. The fact that the peak location of this distribution is slightly positive could be related to the breakdown of the $1/f$ noise model for the 30 and 44 GHz channels reported by Ihle et al. (2020), but the impact of this effect must clearly be small compared to the much larger statistical uncertainties. The best-fit value of the optical depth of reionization is $\tau = 0.060^{+0.015}_{-0.013}$, which is in excellent agreement with previous results. Of course, the uncertainty of this measurement is relatively large, since the accepted sky fraction is small ($f_{\rm sky} = 0.36$).

To visualize the low-ℓ CMB angular power spectrum, we once again employ Eq. (120), and simply compute slices through $\mathcal{L}(C_\ell)$ along each dimension, while fixing all other elements at the best-fit *Planck* 2018 ΛCDM spectrum.[22] For each slice, we report the conditional posterior maximum value as a point estimate, and an asymmetric 68 % confidence interval as its uncertainty. Figure 38 shows a comparison of the resulting BeyondPlanck low-ℓ power spectra evaluated from the Northern Galactic hemisphere with those published by *WMAP* (Hinshaw et al. 2013) and *Planck* 2018 (Planck Collaboration V 2020); the latter two are simply reproduced from tabulated values without reprocessing.

Before concluding this section, we consider two internal consistency tests. First, Fig. 39 shows constraints on τ as a function of number of samples used to generate the low-ℓ likelihood inputs. As discussed by Sellentin & Heavens (2016), it is not sufficient that $\mathsf{N}_{\rm CMB}$ itself converges to some specified precision to obtain robust results, but also that $\mathsf{N}^{-1}_{\rm CMB}$ and $|\mathsf{N}_{\rm CMB}|$ reach a corresponding precision. We see that about 35 000 samples are required to achieve a precision that results in a Monte Carlo uncertainty that is significantly smaller than the posterior width. Regarding the detailed behaviour of this function, we note that the sharp increase around 10 000 samples happens when the final covariance matrix first becomes well-conditioned, and there are no longer any algebraically degenerate modes. However, at this point only a sub-space of the systematic parameter volume has been probed by the Monte Carlo Markov chains, and the corresponding predicted variance is therefore too low, resulting in a positive bias in τ. Only around 35 000 has the full distribution become stationary, and no further improvements are made with more samples.

In the second consistency test we plot τ, r and the χ^2 PTE as a function of $f_{\rm sky}$, adopting the same suite of masks as Planck Collaboration Int. LVII (2020), but after multiplying each with our Northern Galactic mask. The results from these calculations are summarized in Fig. 40. We see that all values of τ are consistent with the *Planck* HFI results to better than $1\,\sigma$. For the tensor-to-scalar ratio, we find that the overall detection level varies between 1.4 and $2.0\,\sigma$, and this variation is also reflected in the χ^2, for which the PTE varies between 8 and 25 %. Overall, all results are statistically consistent for all sky fractions.

9.5.4. High-ℓ CMB likelihood

We now turn our attention to the high-ℓ CMB power spectrum and likelihood. In this respect, we note that the LFI and *WMAP* polarization measurements have very low CMB signal-to-noise ratios at multipoles $\ell \gtrsim 10$, and we therefore restrict ourselves in the following to temperature alone. This has two main advantages. First, the cost of producing a single high-ℓ constrained CMB sample is reduced by a factor of three, since we no longer need to perform polarized spherical harmonics transforms. Second, this also allows us to employ the so-called Blackwell-Rao estimator (Chu et al. 2005), which is by far the preferred estimator for sampling-based likelihood estimators due to its high level of precision combined with excellent computational speed.

[22] A more common convention is to fix other spectra at the joint maximum likelihood solution. However, this sometimes leads to pathological likelihood shapes for strongly noise-dominated modes, and the resulting spectra are in such cases poor representations of the likelihood shape that is actually relevant for parameter estimation.

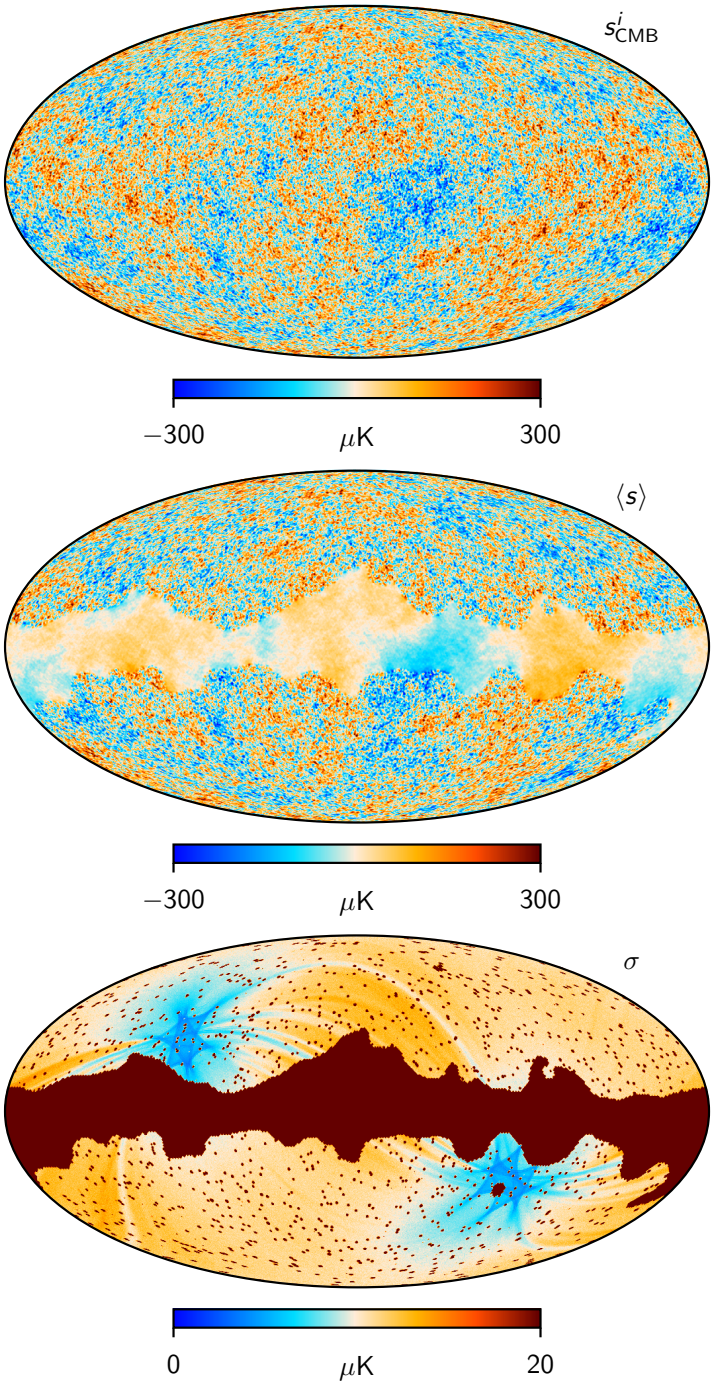

Fig. 41. Full-resolution CMB temperature constrained realization maps. (*Top*:) Single constrained realization, $\boldsymbol{s}^i$, drawn from $P(\boldsymbol{s} \mid \boldsymbol{d}, C_\ell, \ldots)$. (*Middle*:) Posterior mean map, $\langle \boldsymbol{s} \rangle$, as evaluated from the ensemble of constrained CMB realizations; note that the small-scale signal amplitude inside the mask decreases smoothly to zero with increasing distance from the edge of the mask. (*Bottom*:) CMB posterior standard deviation map, as evaluated pixel-by-pixel from the ensemble of constrained CMB realizations. This map is dominated by instrumental noise outside the mask, and by random fluctuations informed by the assumptions of isotropy and Gaussianity inside the mask.

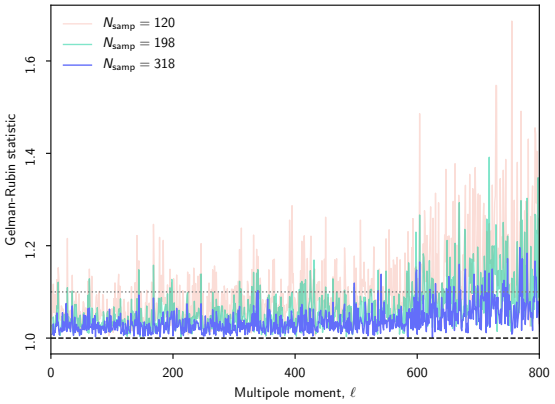

Fig. 42. Gelman-Rubin convergence statistic for the BeyondPlanck TT angular power spectrum, as evaluated from eight σ_ℓ chains with each N samples. A value lower than 1.1 (dotted gray line) typically indicates acceptable convergence.

However, it does require a substantial signal-to-noise ratio per multipole in order to converge, which is not the case for current polarization data. In practice we employ the Gaussianized Blackwell-Rao estimator, as presented by Rudjord et al. (2009) and used by the *Planck* 2015 and 2018 likelihoods (Planck Collaboration XI 2016; Planck Collaboration V 2020), in order to reduce the number of samples required for convergence at high multipoles.

The inputs to this likelihood estimator are generated as follows. For each full-resolution Gibbs sample[23] discussed in the introduction to this section, we draw a new sample from $P(\boldsymbol{s}_{\rm CMB} \mid \boldsymbol{d}, \omega \setminus \boldsymbol{s}_{\rm CMB})$, as we did for the low-$\ell$ estimator, but with the following important differences:

1. The re-sampling step is performed at full angular resolution.
2. The noise level of each input frequency map is set to infinity for each pixel that is excluded by the BeyondPlanck temperature mask, giving these zero weight in the fit.
3. We condition on a CMB power spectrum prior, $\mathsf{S}(C_\ell)$, when solving the Wiener filter equation, which results in a sample, $\boldsymbol{s}^i$, that has the Galactic plane in-painted with a Gaussian constrained realization.
4. The power spectrum, C_ℓ^i, is Gibbs sampled over between each sky sample, $\boldsymbol{s}^i$, such that $\{\boldsymbol{s}^i, C_\ell^i\}$ explore the full corresponding posterior distribution. Multipoles higher than $\ell > 800$ are fixed at the best-fit *Planck* ΛCDM power spectrum, while all multipoles below $\ell \leq 800$ are sampled ℓ-by-ℓ.

For each sky sample, we compute the observed power spectrum, σ_ℓ^i, as defined by Eq. (6), which serves as the actual input to the Blackwell-Rao estimator. Note that this procedure is very nearly identical to that described by Chu et al. (2005), and later adopted by both *WMAP* (Hinshaw et al. 2013) and *Planck* (Planck Collaboration V 2020), but with one fundamental difference: While all previous analyses drew samples from one fixed set of frequency sky maps, we now marginalize over a whole ensemble of frequency sky maps. Thus, this is the first time that low-level instrumental systematic errors are also propagated through the

[23] At the time of writing, only 400 high-resolution samples have been produced after burn-in. The wall-time cost per sample is 5 hours.

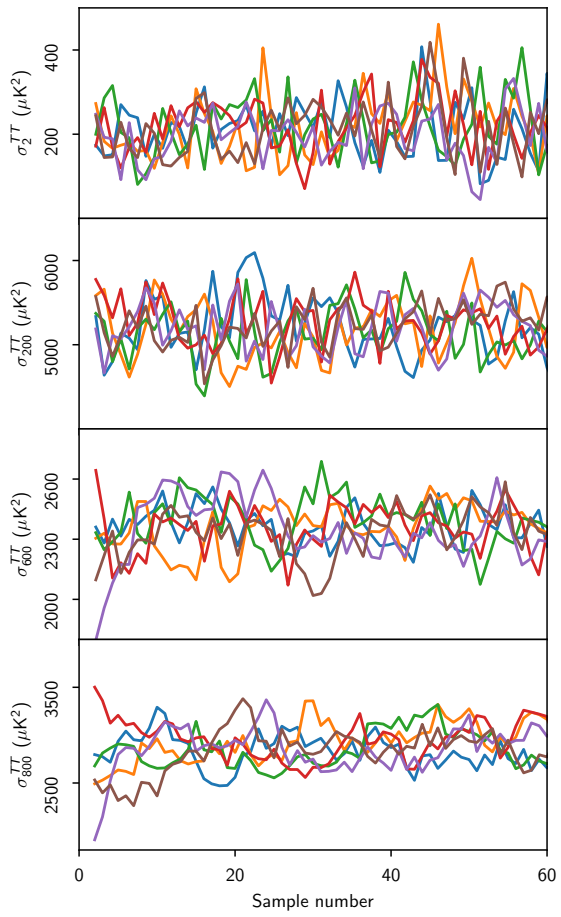

Fig. 43. Trace plots for four representative angular power spectrum multipole coefficients, σ_ℓ^{TT}. From top to bottom, each of the four panels show $\ell = 2$, 200, 600, and 800. Each curve shows one independent Gibbs chain.

Blackwell-Rao estimator into high-level cosmological parameters, in addition to instrumental noise, sky cut, and foreground uncertainties (Planck Collaboration V 2020).

To provide some useful visual intuition, Fig. 41 summarizes the properties of the constrained realizations that feed into the Blackwell-Rao estimator. The top panel shows a single CMB sky map sample from the full posterior distribution. This represents one perfect full-sky CMB map that is *consistent* with the underlying data. Note that it is impossible to see the analysis mask in this plot; these pixels are in effect replaced with a Gaussian random realizations with a power spectrum given by C_ℓ^i and phases that are constrained by the high-latitude information. The middle panel shows the corresponding posterior mean map. At high Galactic latitudes, this map is virtually identical to the single sample (only with very slightly lower noise), while inside the mask it is much smoother. Only modes that can be meaningfully estimated from high latitudes under the assumption of statistical isotropy have non-zero value. Finally, the bottom panel show the posterior standard deviation map. At high latitudes, this is dom-

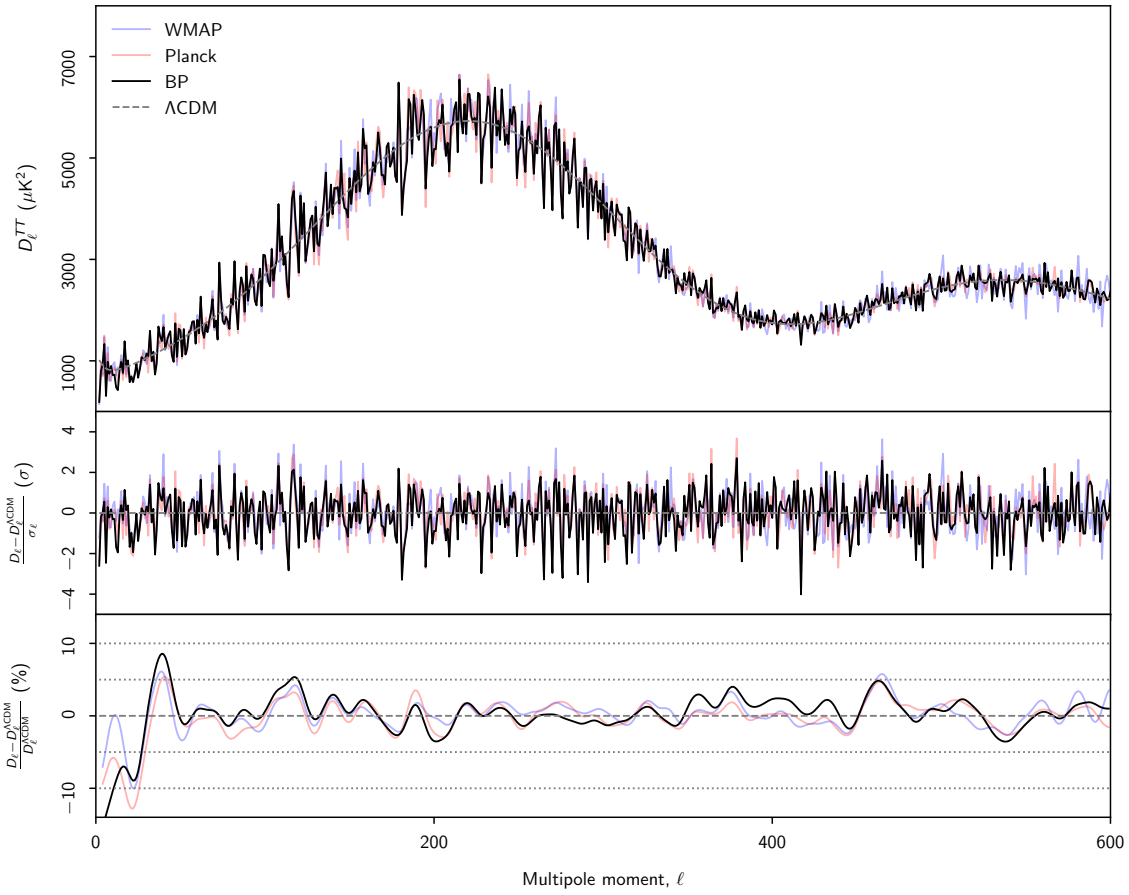

Fig. 44. (*Top*:) Angular CMB temperature power spectrum, D_ℓ^{TT}, as derived by BeyondPlanck (black), *Planck* (red), and *WMAP* (blue). The best-fit *Planck* 2018 ΛCDM power spectrum is shown in dashed gray. (*Middle*:) Residual power spectrum relative to ΛCDM, measured relative to full quoted error bars, $r_\ell \equiv (D_\ell - D_\ell^{\Lambda CDM})/\sigma_\ell$. For pipelines that report asymmetric error bars, σ_ℓ is taken to be the average of the upper and lower error bar. (*Bottom*:) Fractional difference with respect to the *Planck* ΛCDM spectrum. In this panel, each curve has been boxcar averaged with a window of $\Delta\ell = 100$ to suppress random fluctuations.

inated by instrumental noise and systematic effects, while inside the mask it is dominated by CMB sample variance.

The intuitive interpretation of the Blackwell-Rao estimator is now very simple: Assume first that we happened to know the exact CMB sky, without noise, foregrounds or systematic effects. In that case, the CMB power spectrum likelihood is given analytically by the inverse Wishart distribution shown in Eq. (12). However, in reality, we do of course not know the true sky perfectly, and we therefore have to marginalize over all possible CMB skies that are consistent with all observed data. In other words, we have to average the Gaussian likelihood over the ensemble of $\boldsymbol{s}^i$ samples—and that is precisely the definition of the Blackwell-Rao estimator.

As shown by Wandelt et al. (2004), the Blackwell-Rao estimator is guaranteed to converge to the true, exact CMB likelihood in the limit of an infinite number of samples. Of course, in practice we do not have an infinite number of samples, and a common way of assessing convergence is through the so-called Gelman-Rubin R statistic (Gelman & Rubin 1992), which compares the sample variance as measured *within* each chain to the variance *between* chains. With the appropriate scaling factors, R should typically be lower than 1.1 for acceptable convergence. Figure 42 shows the R statistic for the σ_ℓ^i ensemble described above, which suggests that acceptable convergence is achieved up to $\ell \approx 600$–700 with the current sample set.

These numbers are, however, only general rules-of-thumb, and should always be combined with human visual inspection of the actual chains as well. A selection of four representative multipoles are therefore shown in Fig. 43, in which each color represents one independent Gibbs chain. Here we see that $\ell = 2$ and 200 mix very well, and the Markov chain correlation length is essentially zero. As a result, robust convergence will be achieved with a relatively low number of samples. At $\ell = 600$, the correlation length is notably longer, but a good estimate of the full distribution may still be derived from the sample set shown. At $\ell = 800$, it is even longer, and the mixing is about to break down.

We conclude from these results that the current sample set has converged well below $\ell \lesssim 700$. Still, to be conservative we

Table 5. Comparison of basic 6-parameter ΛCDM model parameters as derived by BeyondPlanck (Paradiso et al. 2020), *Planck* 2018 (Planck Collaboration VI 2020), and *WMAP* (Hinshaw et al. 2013). The second column shows results for BeyondPlanck only, using only TT multipoles below $\ell \leq 600$ and polarization below $\ell \leq 8$. The third column shows similar results when also adding TT multipoles between $600 < \ell \leq 2500$ from *Planck* 2018. For *Planck*, we show results from the `Plik` pipeline using the TT+TE+EE+lowE+lensing data combination, while for *WMAP* we show results based on C^{-1}-weighted 9-year *WMAP*-only data. Note that *Planck* and *WMAP* adopt slightly different conventions for some parameters, and we report both where applicable. Columns marked with "Δ" show differences with respect to BeyondPlanck-only results, as measured in units of σ.

	BeyondPlanck		*Planck* 2018		*WMAP*	
Parameter	$\ell \leq 600$	+*Planck* $\ell > 600$	Estimate	$\Delta\,(\sigma)$	Estimate	$\Delta\,(\sigma)$
$\Omega_b h^2$	0.02226 ± 0.00088	0.02230 ± 0.00022	0.02237 ± 0.00015	−0.1	0.02243 ± 0.00050	−0.2
$\Omega_c h^2$	0.115 ± 0.016	0.1227 ± 0.0025	0.1200 ± 0.0012	−0.3	0.1147 ± 0.0051	0
Ω_Λ	⋯	⋯	⋯	⋯	0.721 ± 0.025	⋯
$100\theta_{MC}$	1.0402 ± 0.0048	1.04064 ± 0.00048	1.04092 ± 0.00031	−0.2	⋯	⋯
τ	0.067 ± 0.016	0.074 ± 0.015	0.054 ± 0.007	0.8	0.089 ± 0.0014	−1.4
$10^9 \Delta^2_{\mathcal{R}}$	⋯	⋯	⋯	⋯	2.41 ± 0.10	⋯
$\ln(10^{10} A_s)$	3.035 ± 0.079	3.087 ± 0.029	3.044 ± 0.014	−0.1	⋯	⋯
n_s	0.962 ± 0.019	0.9632 ± 0.0060	0.9649 ± 0.0042	−0.1	0.972 ± 0.013	−0.5

only include multipoles between $9 \leq \ell \leq 600$ in the final BeyondPlanck high-ℓ TT likelihood for cosmological parameter estimation. The resulting power spectrum is shown in Fig. 44, and compared with those presented by *Planck* (Planck Collaboration V 2020) and *WMAP* (Hinshaw et al. 2013). For reference, the gray dashed line shows the best-fit *Planck* 2018 ΛCDM spectrum. The middle panel shows shows the difference of each measured spectrum with respect to the model spectrum in units of each pipeline's respective error bars, while the bottom panel shows the corresponding fractional difference with respect to the best-fit *Planck* 2018 ΛCDM spectrum in units of percent. At $\ell \lesssim 500$, where these data sets are all signal-dominated, the three spectra follow each other almost ℓ-by-ℓ, while at higher multipoles, where *WMAP* becomes noise-dominated, larger variations are seen within multipoles. Overall, the agreement between the three estimates is very good, both as measured by fractional differences and in units of σ.

9.5.5. Cosmological parameters

Finally, we are ready to present cosmological parameters from the BeyondPlanck analysis pipeline. In the following, we use `CosmoMC` (Lewis & Bridle 2002) to explore different cosmological models, coupled to a likelihood of the following form,

$$\ln \mathcal{L}_{\rm BP}(C_\ell) = \ln \mathcal{L}^{\ell=2-8}_{\rm low-}{}_{\ell}(C_\ell) \quad (138)$$
$$+ \ln \mathcal{L}^{\ell=9-600}_{\rm BR}(C_\ell) \quad (139)$$
$$+ \ln \mathcal{L}^{\ell=601-2500}_{Planck}(C_\ell) \quad \text{(optional)}, \quad (140)$$

where the likelihood in line 138 is given by Eq. (120); the likelihood in line 139 is given by Eq. (7) in Rudjord et al. (2009) (which, intuitively, is also given by averaging Eq. (128) over the ensemble of available Gibbs samples); and the likelihood in line 140 is defined by the Gaussian TT-only *Planck* 2018 likelihood (Planck Collaboration V 2020).

The low-ℓ likelihood in Eq. (138) is defined directly in terms of a full multi-variate pixel-based Gaussian distribution with mean and covariance tuned using the posterior samples. Similarly, the Blackwell-Rao estimator in Eq. (139) is defined simply by averaging the inverse Wishart distribution (as defined in Eq. (128)) over all available Gibbs samples. Intuitively, each Gibbs sample represents one possible full-sky and noiseless CMB realization that is consistent with all observed data, and the the Blackwell-Rao estimator is thus simply equivalent to averaging the appropriate distribution for cosmic variance (i.e., the inverse Wishart distribution) over all possible CMB realizations that are consistent with the measurements. Finally, the high-ℓ *Planck* likelihood is simply defined as a multi-variate Gaussian distribution in terms of angular power spectra. We estimate parameters both with and without the *Planck* high-ℓ likelihood.

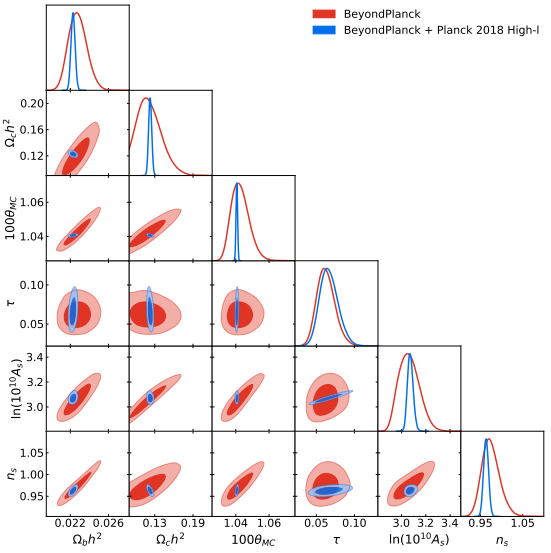

Fig. 45. Marginal cosmological parameter posterior distributions for a basic 6-parameter ΛCDM model as estimated from the BeyondPlanck-only CMB likelihood using TT multipoles up to $\ell \leq 600$ and polarization multipoles up to $\ell \leq 8$ (blue distributions), and also when including $\ell > 600$ temperature information from *Planck* 2018 (red distributions).

For a complete discussion regarding the parameter constraints that is derived from this likelihood, and comparisons with previously published results, we refer the interested reader to Paradiso et al. (2020). Here we only show one single case, namely that corresponding to the basic six-parameter ΛCDM model. These are summarized in terms of posterior means and

standard deviations in Table 5, and in terms of 1- and 2-dimensional marginal distributions in Fig. 45.

Overall, we see that the agreement between BeyondPlanck, *Planck* 2018 and *WMAP* is very good. For *Planck* 2018, which is statistically dominated by HFI measurements, the parameter showing the biggest difference is τ, with a positive shift of $0.8\,\sigma$. We note however that both SROLL2 (Pagano et al. 2020) and NPIPE (Tristram et al. 2020) report slightly higher values of τ than *Planck* 2018; as measured relative to these, the BeyondPlanck excess is only $0.4\,\sigma$.

From Table 5, we also see that the BeyondPlanck uncertainties are generally larger than those of *Planck* 2018, even when combined with high-ℓ information from *Planck*. This is most likely due to the fact that BeyondPlanck does not exploit HFI measurements below $\ell \leq 600$, and therefore observes notably larger uncertainties between the first and third accoustic peaks in the CMB power spectrum. In addition to these differences in raw data volume, it is also important to note that the BeyondPlanck framework marginalizes over a significantly larger model of nuisance parameters than either of the two previous pipelines, and these also contribute to larger uncertainties. For a detailed breakdown of statistical and systematic uncertainties, and their impact on cosmological parameters, we refer the interested reader to Colombo et al. (2020); Paradiso et al. (2020).

10. Reproducibility and Open Science

As discussed in Sect. 1, the main long-term scientific goal and motivation of the BeyondPlanck program is to establish an end-to-end analysis framework for CMB observations that, we hope, will be useful for the general community. This framework is designed to be sufficiently flexible to allow analysis of different and complementary experiments, and thereby exploit the strengths of one instrument to break degeneracies in another. A concrete example of such synergies has already been demonstrated in the current paper, where information contained in the *WMAP* observations is used to support the calibration and component separation of LFI, and, as a result, we are now, for the first time, able to fit for the spectral index of polarized synchrotron emission across the two experiments.

For this project to succeed, substantial efforts have been spent within the BeyondPlanck program on the issue of *reproducibility*. These efforts are summarized by Gerakakis et al. (2020), both in terms of the internal process itself and some lessons learned, and also in terms of the final practical solutions that have been implemented. Here we provide a brief summary of the main points.

10.1. Reproducibility

For the BeyondPlanck framework to be useful for other experiments it must be *reproducible*: Researchers outside of the current collaboration must be able to repeat our analysis, before improving and extending it. To support this, we have focused on four main items:

1. *Documented open-source code* – the full Commander[24] source code, as well as various pre- and post-processing tools,[25] are made publicly available in a GitHub repository under a GPL license, and may be freely downloaded and extended within the general restriction of that license. Preliminary documentation is provided,[26] although it is under continuous development, as is the source code itself.
2. *Cmake compilation* – easy compilation is supported through the Cmake environment; required external libraries are automatically downloaded and compiled.
3. *Data downloader* – a Python-based tool is provided that automatically downloads all BeyondPlanck input data to a user-specified directory, together with the parameter files that are needed to run the code.
4. *Community-based support environment* – we have established a web-based discussion forum[27] dedicated to end-to-end analysis where interested parties may share experiences and discuss issues. All participation in this forum is of course voluntary, with no expectations of either commitments or guarantees from any participant, but the hope is that the forum will grow into a useful and active discussion platform for anything from bugs and code development issues to high-level scientific questions.

In addition, all main results (both full chain files and selected post-processed posterior mean and standard deviation maps) are available from the BeyondPlanck homepage,[28] and eventually through the *Planck* Legacy Archive.[29] For further details regarding the reproducibility aspects of the work, we refer the interested reader to Gerakakis et al. (2020).

10.2. Software

A second requirement for the BeyondPlanck framework to be useful for other users is that the software is computationally efficient so that it can be run on generally available hardware, and also that the source code is extendable without expert knowledge. Regarding the former point, we note that great emphasis has been put on minimizing the required computational resources throughout the implementation. This appears to be at least partially successful, as summarized in Sect. 8.4 and by Galloway et al. (2020a): The full BeyondPlanck analysis, as presented here, has a computational cost of 220 000 CPU hours, which is roughly equivalent to the cost of producing $O(10)$ end-to-end *Planck* FFP8 70 GHz realizations using the traditional pipeline (Planck Collaboration XII 2016). Furthermore, by compressing the TOD inputs the memory footprint of the LFI data set has been reduced by about an order of magnitude (see Table 2 and Galloway et al. 2020a), and now requires only about 1.5 TB of RAM to run. Computers with this amount of memory and clock cycles are now widely available, and a full *Planck* LFI analysis therefore no longer requires the use of expensive supercomputers – although they will of course be beneficial when available.

Regarding the software itself, the current main code base is written in object-oriented Fortran 2003. Clearly, this may represent a significant hurdle for many users, as most astrophysics students today are typically more exposed to languages like Python or C than Fortran. This choice of language is primarily historical, and due to the fact that a large part of the legacy code base was originally written in Fortran, most notably HEALPix (Górski et al. 2005) and Commander (Eriksen et al. 2004, 2008). However, a second important motivation for adopting Fortran is that it remains one of the fastest languages even today in

[24] http://beyondplanck.science
[25] https://github.com/cosmoglobe/c3pp
[26] https://docs.beyondplanck.science
[27] https://forums.beyondplanck.science
[28] http://beyondplanck.science
[29] https://pla.esac.esa.int/

terms of computational speed and memory management. As far as readability and extendability goes, the code has been designed with a strong focus on object-orientation, and we believe that adding support for several types of new sub-classes is relatively straight-forward. This includes classes for new signal components; noise or beam representations; or TOD models. On the other hand, modifying the underlying memory management, component separation infrastructure, or parallelization paradigm, is likely to be difficult without expert knowledge. A guide to the current software is provided by Galloway et al. (2020a). As two real-world demonstrations of the extendability of the framework, we present preliminary applications to both *LiteBIRD* and *WMAP* in two companion papers (Aurlien et al. 2020; Watts et al. 2020).

As useful as we hope the current version will be, we do believe that developing a massively parallel version of `Commander` in Python would be a useful, interesting and intellectually challenging task, and we would encourage (and support!) work in this direction. For reference, the current `Commander` Fortran source code spans 45 000 lines,[30] which can likely be reduced by a significant factor if written in a less verbose language; porting this to Python would obviously be a major undertaking, but certainly feasible for even just a small team of talented and motivated researchers.

11. Conclusions, summary and outlook

The *Planck* project represents a landmark achievement in international cosmology, mapping out the primary temperature fluctuations in the CMB to a precision determined by astrophysical constraints. This achievement was made possible by the dedication and long-term contributions from ESA and NASA; from tens of national funding agencies; and many hundreds of scientists and engineers working for more than two decades. At the end of the mission, a massive amount of knowledge and expertise regarding optimal analysis of CMB experiments had been generated within the collaboration, as clearly demonstrated through more than 150 refereed scientific publications.

A central goal of the BeyondPlanck project was to translate a significant part of this aggregated experience into a practical computer code that can analyse *Planck* data from end-to-end, and to make this code available to the community in general. Due to limited resources and time, BeyondPlanck only considered the *Planck* LFI data in the time domain, although some preliminary work has also been done on *WMAP* and simulated *LiteBIRD* observations.

Regrettably, an application for continued funding for integrating HFI and *LiteBIRD* into the same framework was recently rejected. The referee noted that *the advantages of the proposed methodology over other available methodologies (. . .) are not sufficiently explained*, and *the use of the stakeholder knowledge, such as instrument modellers and astrophysicists, is not sufficiently considered.* These are very reasonable comments, and a busy Principal Investigator of a given experiment who reads this paper may ask very similar questions: "Why should I care about Bayesian statistics? What's in it for me?" To answer these questions, we make the following points:

1. *Faithful error propagation*: BeyondPlanck implements global end-to-end Bayesian CMB analysis framework. The single most important advantage of this is faithful propagation of uncertainties from raw TOD to final cosmological parameters. Instrumental and astrophysical systematic errors are propagated to the final CMB likelihood on the same footing as any other nuisance parameter. While already important for *Planck*, this issue will become absolutely critical for future planned high-precision *B*-mode experiments, such as *LiteBIRD* or *PICO*.
2. *Breaking degeneracies and saving costs by exploiting synergistic observations*: Combining data from complementary sources is essential to break fundamental degeneracies within a given experiment. For instance, both *Planck* and *WMAP* have degenerate polarization modes that they cannot measure well on their own, due to peculiarities in their respective scanning strategies—but there are no degenerate modes in the combined data set. In general, however, the usefulness of joint analysis with external data is often limited by systematic errors. The BeyondPlanck framework addresses this by providing a common platform for performing joint low-level analysis of different experiments. Also noting that the lion's share of the analysis cost of any real-world CMB experiment is associated with understanding degeneracies and systematic errors, we believe that a global approach will lead to better and cheaper science for each experiment.
3. *Fewer human errors*: Tight analysis integration also leads to many important practical advantages, including less room for human errors or miscommunication; greater transparency of both explicit and implicit priors; better optimization of computing resources; and significantly reduced end-to-end wall-clock time by eliminating intermediate human interaction.
4. *"Faster, better and cheaper" through open-source science*: True inter-experiment global analysis will clearly not succeed without active contributions and support from a large part of the general community. For this reason, we make our source codes publicly available under a GPL open-source license to ensure long-term stability of the currently released software. It also means that future improvements must be released under a similarly generous license, in recognition of the fact that this project is intended to be collaborative, open, and inclusive. The use of stakeholder knowledge is critically important—and we hope that many stakeholders will indeed be interested in actively contributing to the program, ultimately leading to "faster, better, and cheaper" science for everyone.

As discussed above, the BeyondPlanck program has primarily focused on the *Planck* LFI data. The reasons for doing so were three-fold. First and foremost, many BeyondPlanck collaborators have been working with the LFI data for one or two decades, and the aggregated experience with this data set within the collaboration implied a low start-up cost; results could be produced quickly. Second, the full LFI data volume is fairly limited in size, comprising less than 1 TB after compression, which is good for fast debugging and testing. Third, the LFI instrument is based on HEMT radiometers, which generally both have a relatively high noise contribution and low systematic errors per sample. The combination of these three points made LFI a natural starting point for the work.

However, now that the computational framework already exists, it will require substantially less effort to generalize it to other and complementary data sets. This work has already started for *LiteBIRD*, *SPIDER*, and *WMAP*, but we welcome initiatives targeting any other experiment as well. In this respect, it may be useful to distinguish between four types of experiments, each with their own set of algorithmic complexities.

[30] Interestingly, only about 6000 lines are directly associated with TOD processing, while 14 000 lines are directly associated with component separation; the rest is spent on general data infrastructure and tools.

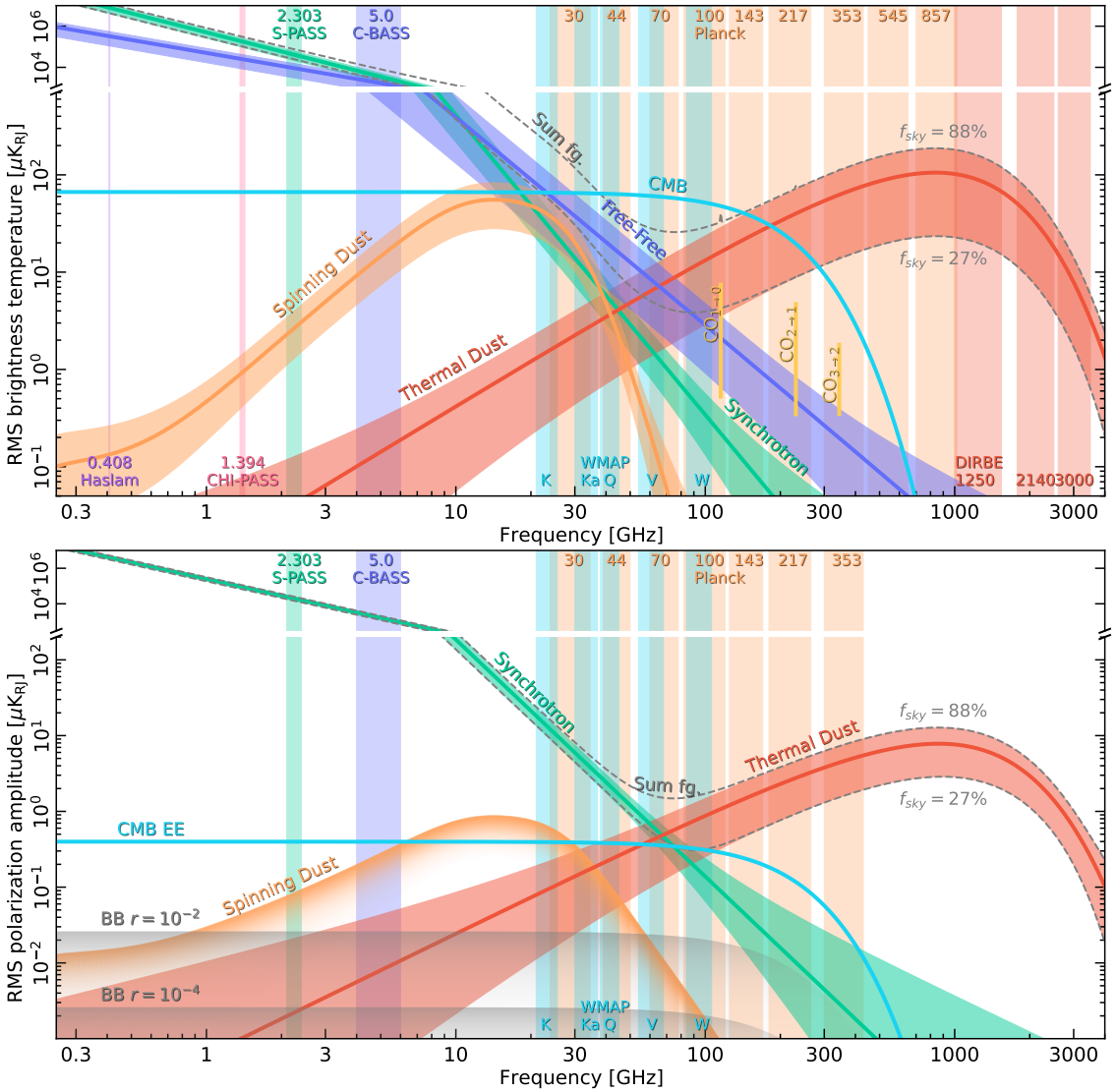

Fig. 46. Brightness temperature (*top panel*) and polarization amplitude (*bottom panel*) RMS as a function of frequency and astrophysical component, and as derived from the BeyondPlanck and *Planck* sky models. Vertical bands indicate the frequency ranges of various experiment. All components have been evaluated at a common angular resolution of 1° FWHM. The widths of each confidence region correspond to the spread when evaluating the RMS over two different masks with sky fractions of 88 and 27 %, respectively. The cyan curve shows the level of CMB fluctuations evaluated for the best-fit *Planck* ΛCDM spectrum. For polarization, the spinning dust component (orange curve) indicates an upper limit as presented by Herman et al. (2020), not a detection. In the current BeyondPlanck analysis, only the three LFI channels are modelled in the time-domain. A long-term future goal is to include all publicly available and relevant data (for instance *WMAP* and *Planck* HFI) into this model; preferrably in the form of time-ordered data, but if this is not technically or financially possible, then at least in the form of pre-processed sky maps. This work will be organized within the Cosmoglobe project.

First, many radio, microwave and sub-millimeter experiments may be modelled within nearly the same sky and instrument model as BeyondPlanck. Examples include C-BASS, QUIET and QUIJOTE, all of which simply provide additional signal-to-noise and/or frequency coverage, as far as the underlying algorithms are concerned. For these, analysis within the BeyondPlanck framework may turn out to amount simply to writing one or more TOD processing modules (for instance using the current LFI module as a template) to take into account the various instrument-specific systematic effects of the experiment in question. These experiments should be, relatively speaking, the easiest to integrate into the current framework.

Other experiments may build on the same sky model and component separation procedures as BeyondPlanck, but require a different mapmaking algorithm. One prominent example of this is *WMAP*, which is differential in nature, and therefore requires a different Conjugate Gradient mapmaking algorithm to translate cleaned TOD into pixelized maps; this work is already on-going (Herman et al. 2020). Experiments of this type should also be relatively straighforward to integrate.

A third class of experiments are those that can use the same type of sky models, but requires a significantly different instrumental model. The most prominent example of such are TES bolometer-based instruments. These often have both higher signal-to-noise ratios and systematic errors per sample, and therefore require a richer set of systematics corrections. They also typically have a significant multiplicative transfer function, which means that unbiased maps cannot be produced simply by introducing additive TOD corrections, as is done in the current implementation. Instead, they will also require a dedicated Conjugate Gradient mapmaker to take into account the multiplicative effects. Examples of potentially relevant experiments include for instance BICEP2, CLASS, SPIDER, and *LiteBIRD*. Integrating these will thus be more challenging than HEMT-based experiments like LFI or *WMAP*, but it should certainly be feasible, and the scientific rewards will be massive.

The fourth and final group of experiments are those that either produce massive amounts of time-ordered data, or very high-resolution data. Important examples are ACT , SPT, Simons Observatory, and CMB-S4. These will all require a fundamental redesign of the existing code base, simply to handle the massive amounts of memory and network communication efficiently. Additionally, experiments that observe only a fraction of the sky, but at high angular resolution, cannot employ the spherical harmonics basis that we currently use for component separation without introducing large degeneracies and singular modes; all spatial modes need to be constrained by at least one experiment for the current implementation to work properly. Developing a new version of the Bayesian framework that can handle higher levels of parallelization, and also use more general basis sets, is thus an important goal for future work.

Returning to the specific scientific results derived by the BeyondPlanck project, we note that cosmological constraints derived from LFI and *WMAP* alone will never be competitive in terms of overall uncertainties as compared to an HFI-based analysis. Nevertheless, many interesting results have been established during the course of the project. Some of the most noteworthy among these are the following:

1. We have, at least partially, succeeded in integrating the LFI 44 GHz channel into a statistically viable low-ℓ CMB likelihood. In the process, we have identified two important breakdowns of the current 44 GHz instrument model, namely a limited range of scanning rings for which the gain model appears to break down, and a general short-coming of a simple $1/f$ model to describe correlated noise in both the 30 and 44 GHz channels. Understanding the nature of these systematic errors, and mitigating them, is an important goal for the immediate future (and may be possible even before the current suite of papers goes into press!). For now, however, only the Northern Galactic hemisphere may be used for cosmological low-ℓ polarization analysis.
2. We have for the first time constructed a full, dense, low-resolution CMB covariance matrix that accounts for marginalization over a wide range of important systematic time-ordered effects, including gain, bandpass, and foreground corrections, in addition to the usual correlated noise. This results in a low-ℓ polarization likelihood that yields results consistent with the latest HFI analyses, and a best-fit value of the reionization optical depth of $\tau = 0.060^{+0.015}_{-0.013}$. The associated χ^2 goodness-of-fit statistics are statistically acceptable, although there might be weak hints of excess power, possibly due to the break-down of the $1/f$ noise model.
3. We have produced a statistically consistent and joint estimate of the CMB dipole using both *Planck* and *WMAP* data. The best-fit dipole amplitude of $3359.5 \pm 1.9\,\mu$K is consistent with all published results, including the latest HFI-based measurements, and the quoted error estimate is derived strictly within the well-defined Bayesian statistical framework.
4. We are for the first time able to fit a physically meaningful spectral index of polarized synchrotron emission using both *WMAP* and *Planck*. This is the direct result of performing a truly joint analysis with LFI and *WMAP* as described above, using information from one experiment to break degeneracies within the other.

Before concluding, we reemphasize that this is not the end. While the BeyondPlanck project itself contractually ends on November 30th, 2020, the work will in general continue with various alternative funding sources, and, we hope, also with the help of a continuously growing community of supporting collaborators and experiments. Figure 46 shows a compilation of the current BeyondPlanck sky model and data sets in both temperature (top panel) and polarization (bottom panel) together with selected external products. The long-term goal of this work is to populate this plot with all available experimental data, and thereby gradually refine the sky model. The ERC-funded Cosmoglobe project aims to coordinate these efforts, and will serve as a stable platform for all parties interested in global Bayesian CMB analysis. Cosmoglobe will also serve as the long-term home for all BeyondPlanck material and products, long after the current BeyondPlanck web portal vanishes.

Finally, we end with an important *caveat emptor*, and emphasize that `Commander` is very much a work-in-progress—and it will remain so for all foreseeable future. Essentially every single step in the pipeline can and will be replaced by smarter and more capable sampling algorithms; there are still known bugs and memory leaks in the code (Galloway et al. 2020a); the user-interface could most certainly be made more intuitive; and so on. This is an unavoidable side-effect of being at the cutting edge of algorithmic research, where new ideas are continuously being explored, implemented and tested. However, at the same time, it is also our belief that the current platform is now finally sufficiently mature to allow external users and developers to use it productively for their own analyses, and to extend it as they see fit. In other words, we believe that now is the right time for Bayesian end-to-end CMB analysis to go OpenSource, and we invite all interested parties to participate in this work.

Acknowledgements. We thank Prof. Pedro Ferreira for useful suggestions, comments and discussions, and Dr. Diana Mjaschkova-Pascual for administrative support. We also thank the entire *Planck* and *WMAP* teams for invaluable support and discussions, and for their dedicated efforts through several decades without which this work would not be possible. The current work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement numbers 776282 (COMPET-4; BeyondPlanck), 772253 (ERC; bits2cosmology), and 819478 (ERC; Cosmoglobe). In addition, the collaboration acknowledges support from ESA; ASI and INAF (Italy); NASA and DoE (USA); Tekes, Academy of Finland (grant no. 295113), CSC, and Magnus Ehrnrooth foundation (Finland); RCN (Norway; grant nos. 263011, 274990); and PRACE (EU).

References

Ali-Haimoud, Y. 2010, SpDust/SpDust.2: Code to Calculate Spinning Dust Spectra
Ali-Haïmoud, Y., Hirata, C. M., & Dickinson, C. 2009, MNRAS, 395, 1055
Alvarez, M. A., Komatsu, E., Doré, O., & Shapiro, P. R. 2006, ApJ, 647, 840
Andersen et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Andersson, B. G., Lazarian, A., & Vaillancourt, J. E. 2015, ARA&A, 53, 501
Ashdown, M. A. J., Baccigalupi, C., Balbi, A., et al. 2007a, A&A, 471, 361
Ashdown, M. A. J., Baccigalupi, C., Balbi, A., et al. 2007b, A&A, 467, 761
Aurlien et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Beck, R., Anderson, J., Heald, G., et al. 2013, Astronomische Nachrichten, 334, 548
Bennett, C. L., Larson, D., Weiland, J. L., et al. 2013, ApJS, 208, 20
Bersanelli, M., Mandolesi, N., Butler, R. C., et al. 2010, A&A, 520, A4
Bertsekas, D. P. 1996, Constrained Optimization and Lagrange Multiplier Methods (Optimization and Neural Computation Series), 1st edn. (Athena Scientific)
BeyondPlanck Collaboration. 2020, A&A, in preparation [arXiv:201x.xxxxx]
BICEP2/Keck Array and Planck Collaborations. 2015, Phys. Rev. Lett., 114, 101301
Brilenkov et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Carretti, E., Haverkorn, M., Staveley-Smith, L., et al. 2019, MNRAS, 489, 2330
Chon, G., Challinor, A., Prunet, S., Hivon, E., & Szapudi, I. 2004, MNRAS, 350, 914
Chu, M., Eriksen, H. K., Knox, L., et al. 2005, Phys. Rev. D, 71, 103002
Colombo et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Condon, J. J., Cotton, W. D., Greisen, E. W., et al. 1998, AJ, 115, 1693
Datta, R., Aiola, S., Choi, S. K., et al. 2019, MNRAS, 486, 5239
de Bernardis, P., Ade, P. A. R., Bock, J. J., et al. 2000, Nature, 404, 955
Delouis, J. M., Pagano, L., Mottet, S., Puget, J. L., & Vibert, L. 2019, A&A, 629, A38
Dickinson, C., Davies, R. D., & Davis, R. J. 2003, MNRAS, 341, 369
Dolginov, A. Z. & Mitrofanov, I. G. 1976, Ap&SS, 43, 291
Draine, B. T. 2011, Physics of the Interstellar and Intergalactic Medium (Princeton University Press)
Draine, B. T. & Hensley, B. S. 2016, The Astrophysical Journal, 831, 59
Draine, B. T. & Hensley, B. S. 2020, arXiv e-prints, arXiv:2009.11314
Draine, B. T. & Lazarian, A. 1998, ApJ, 494, L19
Dunkley, J., Komatsu, E., Nolta, M. R., et al. 2009, ApJS, 180, 306
Erickson, W. C. 1957, ApJ, 126, 480
Eriksen, H. K., Jewell, J. B., Dickinson, C., et al. 2008, ApJ, 676, 10
Eriksen, H. K., O'Dwyer, I. J., Jewell, J. B., et al. 2004, ApJS, 155, 227
Fixsen, D. J. 2009, ApJ, 707, 916
Frigo, M. & Johnson, S. G. 2005, Proceedings of the IEEE, 93, 216, special issue on "Program Generation, Optimization, and Platform Adaptation"
Fuskeland, U., Andersen, K. J., Aurlien, R., et al. 2019, arXiv e-prints, arXiv:1909.05923
Fuskeland, U., Wehus, I. K., Eriksen, H. K., & Næss, S. K. 2014, ApJ, 790, 104
Galeotta et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Galloway et al. 2020a, A&A, in preparation [arXiv:201x.xxxxx]
Galloway et al. 2020b, A&A, in preparation [arXiv:201x.xxxxx]
Gelman, A. & Rubin, D. B. 1992, Statist. Sci., 7, 457
Geman, S. & Geman, D. 1984, IEEE Trans. Pattern Anal. Mach. Intell., 6, 721
Génova-Santos, R., Rubiño-Martín, J. A., Rebolo, R., et al. 2015, MNRAS, 452, 4169
Gerakakis et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Girolami, M. & Calderhead, B. 2011, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73, 123
Gjerløw, E., Colombo, L. P. L., Eriksen, H. K., et al. 2015, ApJS, 221, 5
Gjerløw et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Godard, B., Croon, M., Budnik, F., & Morley, T. 2009, in Proceedings of the 21st International Symposium on Space Flight Dynamics (ISSFD), Toulouse
Górski, K. M., Hivon, E., Banday, A. J., et al. 2005, ApJ, 622, 759
Greaves, J. S., Holland, W. S., Friberg, P., & Dent, W. R. F. 1999, The Astrophysical Journal, 512, L139
Gregory, P. C., Scott, W. K., Douglas, K., & Condon, J. J. 1996, ApJS, 103, 427
Gualtieri, R., Filippini, J. P., Ade, P. A. R., et al. 2018, Journal of Low Temperature Physics, 193, 1112
Guillet, V., Fanciullo, L., Verstraete, L., et al. 2018, A&A, 610, A16
Gunn, J. E. & Peterson, B. A. 1965, ApJ, 142, 1633
Génova-Santos, R., Rubiño-Martín, J. A., Peláez-Santos, A., et al. 2016, Monthly Notices of the Royal Astronomical Society, 464, 4107
Haslam, C. G. T., Salter, C. J., Stoffel, H., & Wilson, W. E. 1982, A&AS, 47, 1
Hastings, W. K. 1970, Biometrika, 57, 97
Hauser, M. G., Arendt, R. G., Kelsall, T., et al. 1998, ApJ, 508, 25
Hensley, B. S. & Draine, B. T. 2020, ApJ, 895, 38
Herman et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Hinshaw, G., Larson, D., Komatsu, E., et al. 2013, ApJS, 208, 19
Hinshaw, G., Weiland, J. L., Hill, R. S., et al. 2009, ApJS, 180, 225
Hoang, T., Lazarian, A., & Martin, P. G. 2013, ApJ, 779, 152
Hu, W. & Dodelson, S. 2002, ARA&A, 40, 171
Huffman, D. A. 1952, Proceedings of the IRE, 40, 1098
Ihle et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Jarosik, N., Bennett, C. L., Dunkley, J., et al. 2011, ApJS, 192, 14
Jeffreys, H. 1946, Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences, 186, 453
Jew, L., Taylor, A. C., Jones, M. E., et al. 2019, Monthly Notices of the Royal Astronomical Society, 490, 2958
Jewell, J., Levin, S., & Anderson, C. H. 2004, ApJ, 609, 1
Jewell, J. B., Eriksen, H. K., Wandelt, B. D., et al. 2009, ApJ, 697, 258
Kamionkowski, M. & Kovetz, E. D. 2016, ARA&A, 54, 227
Keating, B., Timbie, P., Polnarev, A., & Steinberger, J. 1998, ApJ, 495, 580
Keihänen, E., Keskitalo, R., Kurki-Suonio, H., Poutanen, T., & Sirviö, A. 2010, A&A, 510, A57
Keihänen, E., Kurki-Suonio, H., & Poutanen, T. 2005, MNRAS, 360, 390
Keihänen, E., Kurki-Suonio, H., Poutanen, T., Maino, D., & Burigana, C. 2004, A&A, 428, 287
Keihänen et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Kelsall, T., Weiland, J. L., Franz, B. A., et al. 1998, ApJ, 508, 44
King, O. G., Jones, M. E., Blackhurst, E. J., et al. 2014, MNRAS, 438, 2426
Kogut, A. 2012, ApJ, 753, 110
Kogut, A., Spergel, D. N., Barnes, C., et al. 2003, ApJS, 148, 161
Komatsu, E., Dunkley, J., Nolta, M. R., et al. 2009, ApJS, 180, 330
Kovac, J. M., Leitch, E. M., Pryke, C., et al. 2002, Nature, 420, 772
Krachmalnicoff, N., Carretti, E., Baccigalupi, C., et al. 2018, A&A, 618, A166
Larson, D., Dunkley, J., Hinshaw, G., et al. 2011, ApJS, 192, 16
Larson, D. L., Eriksen, H. K., Wandelt, B. D., et al. 2007, ApJ, 656, 653
Leach, S. M., Cardoso, J., Baccigalupi, C., et al. 2008, A&A, 491, 597
Leitch, E. M., Readhead, A. C. S., Pearson, T. J., & Myers, S. T. 1997, ApJ, 486, L23
Lewis, A. & Bridle, S. 2002, Phys. Rev. D, 66, 103511
Lewis, A., Challinor, A., & Lasenby, A. 2000, ApJ, 538, 473
Lineweaver, C. H., Tenorio, L., Smoot, G. F., et al. 1996, ApJ, 470, 38
Liu, J. S. 2008, Monte Carlo Strategies in Scientific Computing (Springer Publishing Company, Incorporated)
Loeb, A. & Barkana, R. 2001, Annual Review of Astronomy and Astrophysics, 39, 19
Macellari, N., Pierpaoli, E., Dickinson, C., & Vaillancourt, J. E. 2011, MNRAS, 418, 888
Macià Escatllar, A. & Bromley, S. T. 2020, A&A, 634, A77
Maino, D., Burigana, C., Maltoni, M., et al. 1999, A&AS, 140, 383
Mandolesi, N., Bersanelli, M., Butler, R. C., et al. 2010, A&A, 520, A3
Mather, J. C., Cheng, E. S., Cottingham, D. A., et al. 1994, ApJ, 420, 439
Mennella, A., Butler, R. C., Curto, A., et al. 2011, A&A, 536, A3
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. 1953, J. Chem. Phys., 21, 1087
Mitra, S., Rocha, G., Górski, K. M., et al. 2011, ApJS, 193, 5
Murphy, T., Sadler, E. M., Ekers, R. D., et al. 2010, MNRAS, 402, 2403
Natale, U., Pagano, L., Lattanzi, M., et al. 2020, arXiv e-prints, arXiv:2005.05600
Pagano, L., Delouis, J. M., Mottet, S., Puget, J. L., & Vibert, L. 2020, A&A, 635, A99
Page, L., Hinshaw, G., Komatsu, E., et al. 2007, ApJS, 170, 335
Paradiso et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Penzias, A. A. & Wilson, R. W. 1965, ApJ, 142, 419
Planck Collaboration. 2005, ESA publication ESA-SCI(2005)/01 [arXiv:astro-ph/0604069]
Planck Collaboration II. 2011, A&A, 536, A2
Planck Collaboration I. 2014, A&A, 571, A1
Planck Collaboration III. 2014, A&A, 571, A3
Planck Collaboration IV. 2014, A&A, 571, A4
Planck Collaboration VI. 2014, A&A, 571, A6
Planck Collaboration IX. 2014, A&A, 571, A9
Planck Collaboration XI. 2014, A&A, 571, A11
Planck Collaboration XIV. 2014, A&A, 571, A14
Planck Collaboration XV. 2014, A&A, 571, A15
Planck Collaboration XVI. 2014, A&A, 571, A16
Planck Collaboration XVII. 2014, A&A, 571, A17
Planck Collaboration I. 2016, A&A, 594, A1
Planck Collaboration II. 2016, A&A, 594, A2
Planck Collaboration IV. 2016, A&A, 594, A4
Planck Collaboration V. 2016, A&A, 594, A5
Planck Collaboration VII. 2016, A&A, 594, A7
Planck Collaboration VIII. 2016, A&A, 594, A8
Planck Collaboration IX. 2016, A&A, 594, A9
Planck Collaboration X. 2016, A&A, 594, A10
Planck Collaboration XI. 2016, A&A, 594, A11
Planck Collaboration XII. 2016, A&A, 594, A12
Planck Collaboration XXV. 2016, A&A, 594, A25

Planck Collaboration XXVI. 2016, A&A, 594, A26
Planck Collaboration I. 2020, A&A, 641, A1
Planck Collaboration II. 2020, A&A, 641, A2
Planck Collaboration III. 2020, A&A, 641, A3
Planck Collaboration IV. 2020, A&A, 641, A4
Planck Collaboration V. 2020, A&A, 641, A5
Planck Collaboration VI. 2020, A&A, 641, A6
Planck Collaboration VII. 2020, A&A, 641, A7
Planck Collaboration IX. 2020, A&A, 641, A9
Planck Collaboration XI. 2020, A&A, 641, A11
Planck Collaboration Int. XLVI. 2016, A&A, 596, A107
Planck Collaboration Int. XLVII. 2016, A&A, 596, A108
Planck Collaboration Int. XLVIII. 2016, A&A, 596, A109
Planck Collaboration Int. LVII. 2020, A&A, in press [arXiv:2007.04997]
Prézeau, G. & Reinecke, M. 2010, ApJS, 190, 267
Racine, B., Jewell, J. B., Eriksen, H. K., & Wehus, I. K. 2016, ApJ, 820, 31
Reinecke, M. & Seljebotn, D. S. 2013, A&A, 554, A112
Remazeilles, M., Dickinson, C., Banday, A. J., Bigot-Sazy, M.-A., & Ghosh, T. 2015, MNRAS, 451, 4311
Rudjord, Ø., Groeneboom, N. E., Eriksen, H. K., et al. 2009, ApJ, 692, 1669
Seljak, U. & Zaldarriaga, M. 1996, ApJ, 469, 437
Seljebotn, D. S., Bærland, T., Eriksen, H. K., Mardal, K. A., & Wehus, I. K. 2019, A&A, 627, A98
Sellentin, E. & Heavens, A. F. 2016, MNRAS, 456, L132
Shewchuk, J. R. 1994, An Introduction to the Conjugate Gradient Method Without the Agonizing Pain
Silsbee, K., Ali-Haïmoud, Y., & Hirata, C. M. 2011, MNRAS, 411, 2750
Smoot, G. F., Bennett, C. L., Kogut, A., et al. 1992, ApJ, 396, L1
Stivoli, F., Grain, J., Leach, S. M., et al. 2010, MNRAS, 408, 2319
Stompor, R., Leach, S., Stivoli, F., & Baccigalupi, C. 2009, MNRAS, 392, 216
Sugai, H., Ade, P. A. R., Akiba, Y., et al. 2020, Journal of Low Temperature Physics, 199, 1107
Sunyaev, R. A. & Zeldovich, Y. B. 1972, Comments on Astrophysics and Space Physics, 4, 173
Suur-Uski et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Svalheim et al. 2020a, A&A, in preparation [arXiv:201x.xxxxx]
Svalheim et al. 2020b, A&A, in preparation [arXiv:201x.xxxxx]
Szűcs, L., Glover, S. C. O., & Klessen, R. S. 2014, MNRAS, 445, 4055
Tassis, K., Ramaprakash, A. N., Readhead, A. C. S., et al. 2018, arXiv e-prints, arXiv:1810.05652
Tegmark, M., Taylor, A. N., & Heavens, A. F. 1997, ApJ, 480, 22
Thommesen, H., Andersen, K. J., Aurlien, R., et al. 2020, arXiv e-prints, arXiv:2007.06250
Tristram, M., Banday, A. J., Górski, K. M., et al. 2020, arXiv e-prints, arXiv:2010.01139
Wandelt, B. D. & Górski, K. M. 2001, Phys. Rev. D, 63, 123002
Wandelt, B. D., Larson, D. L., & Lakshminarayanan, A. 2004, Phys. Rev. D, 70, 083511
Wang, X., Tegmark, M., Jain, B., & Zaldarriaga, M. 2003, Phys. Rev. D, 68, 123001
Watts et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Wehus, I. K., Fuskeland, U., Eriksen, H. K., et al. 2017, A&A, 597, A131
Wehus, I. K., Næss, S. K., & Eriksen, H. K. 2012, ApJS, 199, 15
Zaldarriaga, M. & Seljak, U. 1997, Phys. Rev. D, 55, 1830

Appendix A: Review of frequently used textbook sampling algorithms

As described in Sect. 6.3, the BeyondPlanck pipeline is designed in the form of a Gibbs sampler in which each parameter is sampled conditionally on all other parameters. Each parameter must therefore be associated with a specific sampling algorithm that samples from the correct distribution. In this appendix, we therefore review some of the most common sampling techniques that are used in the BeyondPlanck framework, while noting that all of this is textbook material; this is just provided for reference purposes. In all cases below, we assume that samplers for both the uniform distribution, $U[0,1]$, and the standard univariate normal (Gaussian) distribution, $N(0,1)$, are already available through some numerical library; we use routines provided in `HEALPix`.

Appendix A.1: Univariate and low-dimensional Gaussian sampling

Perhaps the single most common distribution in any Bayesian pipeline is the univariate Gaussian distribution $N(\mu,\sigma^2)$ with mean μ and standard deviation σ. A sample x from this distribution can be trivially generated by

$$x = \mu + \sigma\eta, \tag{A.1}$$

where $\eta \sim N(0,1)$ is a standard normal variate. Note that $\langle x\rangle = \mu$, because $\langle\eta\rangle = 0$ and $\left\langle(x-\mu)^2\right\rangle = \sigma^2$ because $\left\langle\eta^2\right\rangle = 1$.

A sample from a multi-variate normal distribution $N(\bar{\mu},\mathsf{C})$ with mean vector $\bar{\mu}$ and covariance matrix C may be produced in a fully analogous manner,

$$\boldsymbol{x} = \bar{\mu} + \mathsf{C}^{\frac{1}{2}}\bar{\eta}, \tag{A.2}$$

where now $\bar{\eta}$ is a vector of independent $N(0,1)$ variates, and $\mathsf{C}^{\frac{1}{2}}$ denotes some matrix for which $\mathsf{C} = \mathsf{C}^{\frac{1}{2}}(\mathsf{C}^{\frac{1}{2}})^t$. The two most typical examples are the Cholesky decomposition ($\mathsf{C} = \mathsf{L}\mathsf{L}^t$ for positive definite matrices, where $\mathsf{C}^{\frac{1}{2}} = \mathsf{L}$) and singular-value decomposition ($\mathsf{C} = \mathsf{V}\Sigma\mathsf{V}^t$ for singular matrices, where $\mathsf{C}^{\frac{1}{2}} = \mathsf{V}\Sigma^{\frac{1}{2}}\mathsf{V}^t$). A notable advantage regarding the latter is that it is symmetric, and therefore less bug-prone than the Cholesky factor; on the other hand, it is slightly more computationally expensive.

Appendix A.2: High-dimensional Gaussian sampling

It is important to note that evaluating a "square root" of a matrix, whether it is through Cholesky or eigen-vector decomposition, is an $O(n^3)$ operation, where n is the dimension of the matrix. As such, the direct approach is only computationally practical for relatively low-dimensional distributions, and just with a few thousand elements or less. For distributions with millions of correlated variables, the above prescription is entirely impractical. In the following, we therefore describe a widely used method to sample from high-dimensional distributions, effectively by inverting the covariance matrix iteratively by Conjugate Gradients.

Again, let $\boldsymbol{x}$ be a random Gaussian field of n elements with an $n\times n$ covariance matrix S, i.e., $\boldsymbol{x} \sim N(\boldsymbol{0},\mathsf{S})$. Further, to put the notation into a familiar context, we assume we have some observations $\boldsymbol{d}$ that can be modeled as

$$\boldsymbol{d} = \mathsf{T}\boldsymbol{x} + \boldsymbol{n}, \tag{A.3}$$

where $\boldsymbol{n}$ is a stochastic noise vector of size n_d (which in general is different from n) which is drawn from a Gaussian distribution with zero mean and covariance N, and T is a matrix of size $n_d\times n$, which effectively translates $\boldsymbol{x}$ into the vector space of $\boldsymbol{d}$. In other words, we assume that the data may be modelled as a linear combination of $\boldsymbol{x}$ plus a well-defined noise contribution.

Note that this assumption about $\boldsymbol{d}$ does not preclude the cases where we have observations that can be written as $\boldsymbol{d} = \mathsf{T}\boldsymbol{x}+\boldsymbol{n}+\boldsymbol{b}$, where $\boldsymbol{b}$ is known and independent of $\boldsymbol{x}$ - in this case, we are free to redefine $\boldsymbol{d}$: $\boldsymbol{d}' \rightarrow \boldsymbol{d}-\boldsymbol{b}$, in which case our assumption in Eq. (A.3) would be met for $\boldsymbol{d}'$.

In general, T will not depend on $\boldsymbol{x}$. In the context of the Gibbs framework of this paper, however, T typically *will* depend on other quantities that we do sample, but which we assume to be known with respect to the current conditional of the Gibbs chain.

Our goal is then to draw a sample from $P(\boldsymbol{x}\mid\boldsymbol{d},\mathsf{T},\mathsf{S},\mathsf{N})$, the posterior of $\boldsymbol{x}$, given $\boldsymbol{d}$ and the other quantities, denoted $P(\boldsymbol{x}\mid\boldsymbol{d})$ as a shorthand. Using Bayes' theorem, we can write this as

$$P(\boldsymbol{x}\mid\boldsymbol{d}) \propto P(\boldsymbol{d}\mid\boldsymbol{x})P(\boldsymbol{x}). \tag{A.4}$$

Here $P(\boldsymbol{x})$ is a prior for x, which we assume takes the form $\mathcal{N}(\boldsymbol{0},\mathsf{S})$, whereas the likelihood term, $P(\boldsymbol{d}\mid\boldsymbol{x})$, is simply given by a Gaussian distribution with covariance N and mean $\mathsf{T}\boldsymbol{x}$. This gives (neglecting the pre-factors of the exponentials, as they are independent of $\boldsymbol{x}$ and end up as normalization constants)

$$\begin{aligned}-2\ln P(\boldsymbol{x}\mid\boldsymbol{d}) &= \boldsymbol{x}^t\mathsf{S}^{-1}\boldsymbol{x} + (\boldsymbol{d}-\mathsf{T}\boldsymbol{x})^t\mathsf{N}^{-1}(\boldsymbol{d}-\mathsf{T}\boldsymbol{x})\\ &= \boldsymbol{x}^t\mathsf{S}^{-1}\boldsymbol{x} + \boldsymbol{d}^t\mathsf{N}^{-1}\boldsymbol{d} + \boldsymbol{x}^t\mathsf{T}^t\mathsf{N}^{-1}\mathsf{T}\boldsymbol{x} -\\ &\qquad \boldsymbol{d}^t\mathsf{N}^{-1}\mathsf{T}\boldsymbol{x} - \boldsymbol{x}^t\mathsf{T}^t\mathsf{N}^{-1}\boldsymbol{d}\\ &= \boldsymbol{x}^t(\mathsf{S}^{-1}+\mathsf{T}^t\mathsf{N}^{-1}\mathsf{T})\boldsymbol{x} - 2\boldsymbol{x}^t\mathsf{T}^t\mathsf{N}^{-1}\boldsymbol{d},\end{aligned} \tag{A.5}$$

where, in the last transition, we neglect also the terms that do not include $\boldsymbol{x}$. We also use the identity $\boldsymbol{a}^t\mathsf{C}\boldsymbol{b} = \boldsymbol{b}^t\mathsf{C}\boldsymbol{a}$, which is valid for a symmetric matrix C, in order to gather the terms that are linear in $\boldsymbol{x}$.

This expression for $P(\boldsymbol{x}\mid\boldsymbol{d})$ can be written as a Gaussian distribution by "completing the square": We are looking for a matrix F and a vector $\boldsymbol{c}$ such that

$$\begin{aligned}P(\boldsymbol{x}\mid\boldsymbol{d}) &= \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{c})^t\mathsf{F}^{-1}(\boldsymbol{x}-\boldsymbol{c})\right]\\ &\propto \exp\left[-\frac{1}{2}\left(\boldsymbol{x}^t\mathsf{F}^{-1}\boldsymbol{x} - 2\boldsymbol{x}^t\mathsf{F}^{-1}\boldsymbol{c}\right)\right].\end{aligned} \tag{A.6}$$

Comparing terms in Eqs. (A.5) and (A.6), we find that the terms that are quadratic in $\boldsymbol{x}$ enforce

$$\mathsf{F}^{-1} = \mathsf{S}^{-1} + \mathsf{T}^t\mathsf{N}^{-1}\mathsf{T}. \tag{A.7}$$

Inserting this into the terms that are linear in $\boldsymbol{x}$, we find

$$\boldsymbol{c} = (\mathsf{S}^{-1}+\mathsf{T}^t\mathsf{N}^{-1}\mathsf{T})^{-1}\mathsf{T}^t\mathsf{N}^{-1}\boldsymbol{d}. \tag{A.8}$$

Thus, the posterior of $\boldsymbol{x}$ is a Gaussian distribution with covariance given by Eq. (A.7) and mean (and mode) given by Eq. (A.8).

In order to draw a sample, $\tilde{\boldsymbol{x}}$, from this distribution, we can in principle use the standard prescription for sampling from multivariate Gaussian distributions, as summarized in the previous section. However, inverting the covariance matrix, $\mathsf{S}^{-1}+\mathsf{T}^t\mathsf{N}^{-1}\mathsf{T}$, is once again a $O(n^3)$ operation. To circumvent this problem, we instead consider the same equation in the form

$$(\mathsf{S}^{-1}+\mathsf{T}^t\mathsf{N}^{-1}\mathsf{T})\boldsymbol{x} = \mathsf{T}^t\mathsf{N}^{-1}\boldsymbol{d}. \tag{A.9}$$

Since the matrix on the left-hand side is both symmetric and semi-positive definite, this equation can be solved iteratively by

Conjugate Gradients; for a brilliant review of this algorithm, see Shewchuk (1994). Additionally, to obtain the correct covariance structure, one can simply add one random zero-mean covariance term for each element in the covariance matrix to the right-hand side of the equation,

$$(\mathsf{S}^{-1} + \mathsf{T}^t\mathsf{N}^{-1}\mathsf{T})\boldsymbol{x} = \mathsf{T}^t\mathsf{N}^{-1}\boldsymbol{d} + \mathsf{T}^t\mathsf{N}^{-1/2}\boldsymbol{\eta}_1 + \mathsf{S}^{-1/2}\boldsymbol{\eta}_2. \tag{A.10}$$

With this definition, $\langle \boldsymbol{x} \rangle = \boldsymbol{c}$, and $\langle \boldsymbol{x}\boldsymbol{x}^t \rangle = (\mathsf{S}^{-1} + \mathsf{T}^t\mathsf{N}^{-1}\mathsf{T})^{-1} = \mathsf{F}$, as desired.

A fully analogous calculation may be done also with a non-zero prior mean, $\boldsymbol{m}$, in which case an additional term is introduced on the right-hand side of Eq. (A.9),

$$(\mathsf{S}^{-1} + \mathsf{T}^t\mathsf{N}^{-1}\mathsf{T})\boldsymbol{x} = \mathsf{T}^t\mathsf{N}^{-1}\boldsymbol{d} + \mathsf{S}^{-1}\boldsymbol{m} + \mathsf{T}^t\mathsf{N}^{-1/2}\boldsymbol{\eta}_1 + \mathsf{S}^{-1/2}\boldsymbol{\eta}_2. \tag{A.11}$$

The relative strength of the data and prior terms is thus effectively determined by the overall signal-to-noise ratio of the data as measured by S and N, and in the limit of vanishing signal-to-noise (i.e., $\mathsf{N}^{-1} \rightarrow 0$), $\langle \boldsymbol{x} \rangle = \boldsymbol{m}$, as desired. Note, also, that S quantifies the covariance of the *fluctuations around the mean*, not the co-variance of the entire field $\boldsymbol{x}$ itself. In the limit of $\mathsf{S} \rightarrow 0$ (or, equivalently, $\mathsf{S}^{-1} \rightarrow \infty$), we therefore also have $\langle \boldsymbol{x} \rangle = \boldsymbol{m}$. Thus, the magnitude of S represents a direct handle for adjusting the strength of the prior.

Appendix A.3: Inversion sampling

The samplers discussed in the two previous sections only concerns Gaussian distributions. In contrast, the so-called *inversion* sampler is a completely general sampler that works for all univariate distributions.

Let $P(x)$ be a general probability distribution for some random variable x. The inversion sampler is then defined as follows:

1. Compute $P(x)$ over a grid in x, making sure to probe the tails to sufficient accuracy.
2. Compute the cumulative probability distribution, $F(x) = \int_{-\infty}^{x} P(x')\,\mathrm{d}x'$.
3. Draw a random uniform variate, $\eta \sim U[0, 1]$.
4. Solve the nonlinear equation $\eta = F(x)$ for x.

Clearly, this is a computationally very expensive algorithm, noting that it actually requires the user to map the full distribution, $P(x)$, in the first step. This typically requires a preliminary bisection search to first identify a sufficiently wide region in x to cover all significant parts of P. Then another 50–100 evaluations are required to grid the (log-)probability distribution.

However, the facts that this sampler requires no manual tuning, and that it produces independent samples, make it an attractive component in many Gibbs samplers; typically, the overall computational cost of the entire Gibbs chain is dominated by completely different operations.

Paper V

BeyondPlanck VI. Noise characterization and modelling

H. T. Ihle, M. Bersanelli, C. Franceschet, E. Gjerløw and the BeyondPlanck collaboration

A&A, to be submitted, arXiv: 2011.06650,

V

Astronomy & Astrophysics manuscript no. noise_paper
November 16, 2020
©ESO 2020

BeyondPlanck VI. Noise characterization and modelling

H. T. Ihle[11]⋆, M. Bersanelli[4,9,10], C. Franceschet[4,10], E. Gjerløw[11], K. J. Andersen[11], R. Aurlien[11], R. Banerji[11], S. Bertocco[8], M. Brilenkov[11], M. Carbone[14], L. P. L. Colombo[4], H. K. Eriksen[11], J. R. Eskilt[11], M. K. Foss[11], U. Fuskeland[11], S. Galeotta[8], M. Galloway[11], S. Gerakakis[14], B. Hensley[2], D. Herman[11], M. Iacobellis[14], M. Ieronymaki[14], H. T. Ihle[11], J. B. Jewell[12], A. Karakci[11], E. Keihänen[3,7], R. Keskitalo[1], G. Maggio[8], D. Maino[4,9,10], M. Maris[8], A. Mennella[4,9,10], S. Paradiso[4,9], B. Partridge[6], M. Reinecke[13], M. San[11], A.-S. Suur-Uski[3,7], T. L. Svalheim[11], D. Tavagnacco[8,5], H. Thommesen[11], D. J. Watts[11], I. K. Wehus[11], and A. Zacchei[8]

[1] Computational Cosmology Center, Lawrence Berkeley National Laboratory, Berkeley, California, U.S.A.
[2] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, U.S.A.
[3] Department of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland
[4] Dipartimento di Fisica, Università degli Studi di Milano, Via Celoria, 16, Milano, Italy
[5] Dipartimento di Fisica, Università degli Studi di Trieste, via A. Valerio 2, Trieste, Italy
[6] Haverford College Astronomy Department, 370 Lancaster Avenue, Haverford, Pennsylvania, U.S.A.
[7] Helsinki Institute of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland
[8] INAF - Osservatorio Astronomico di Trieste, Via G.B. Tiepolo 11, Trieste, Italy
[9] INAF-IASF Milano, Via E. Bassini 15, Milano, Italy
[10] INFN, Sezione di Milano, Via Celoria 16, Milano, Italy
[11] Institute of Theoretical Astrophysics, University of Oslo, Blindern, Oslo, Norway
[12] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California, U.S.A.
[13] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85741 Garching, Germany
[14] Planetek Hellas, Leoforos Kifisias 44, Marousi 151 25, Greece

November 16, 2020

arXiv:2011.06650v1 [astro-ph.CO] 12 Nov 2020

ABSTRACT

We present a Bayesian method for estimating instrumental noise parameters and propagating noise uncertainties within the global BeyondPlanck Gibbs sampling framework, and apply this to *Planck* LFI time-ordered data. Following previous literature, we adopt a simple $1/f$ model for the noise power spectral density (PSD), and implement an optimal Wiener-filter (or constrained realization) gap-filling procedure to account for masked data. We then use this procedure to both estimate the gapless correlated noise in the time-domain, $\boldsymbol{n}_{\rm corr}$, and to sample the noise PSD spectral parameters, $\xi^n = \{\sigma_0, f_{\rm knee}, \alpha\}$. In contrast to previous *Planck* analyses, we only assume piecewise stationary noise within each pointing period (PID), not throughout the full mission, but we adopt the LFI DPC results as priors on α and $f_{\rm knee}$. On average, we find best-fit correlated noise parameters that are mostly consistent with previous results, with a few notable exceptions. However, a detailed inspection of the time-dependent results reveals many important findings. First and foremost, we find strong evidence for statistically significant temporal variations in all noise PSD parameters, many of which are directly correlated with satellite housekeeping data. Second, while the simple $1/f$ model appears to be an excellent fit for the LFI 70 GHz channel, there is evidence for additional correlated noise not described by a $1/f$ model in the 30 and 44 GHz channels, including within the primary science frequency range of 0.1–1 Hz. In general, most 30 and 44 GHz channels exhibit excess noise at the 2–3 σ level in each one hour pointing period. For some periods of time, we also find evidence of strong common mode noise fluctuations across the entire focal plane. Finally, we find a number of strong stripes when binning the 44 GHz correlated noise into a sky map, and we hypothesize that these may be associated with deficiencies in the gain model for this channel. Overall, we conclude that a simple $1/f$ profile is not adequate to fully characterize the *Planck* LFI noise, even when fitted hour-by-hour, and a more general model is required. These findings have important implications for large-scale CMB polarization reconstruction with the *Planck* LFI data, and understanding and mitigating these issues should be a high-priority task for future studies.

Key words. ISM: general – Cosmology: observations, polarization, cosmic microwave background, diffuse radiation – Galaxy: general

Contents

1 Introduction **2**

2 The BeyondPlanck data model and framework **3**

3 Methods **4**
3.1 Sampling correlated noise, $P(\boldsymbol{n}^{\rm corr} \mid \boldsymbol{d}, \xi^n, \boldsymbol{s}_{\rm tot}, g)$. 4
3.1.1 Ideal data 4
3.1.2 Handling masking through a conjugate gradient solver 5
3.1.3 Gap-filling by polynomial interpolation . 5
3.2 Sampling noise PSD parameters, $P(\xi^n \mid \boldsymbol{n}^{\rm corr})$ . . 6
3.2.1 Sampling the white noise level, σ_0 6
3.2.2 Sampling correlated noise parameters, $f_{\rm knee}$ and α 6
3.2.3 Priors on α and $\boldsymbol{f}_{\rm knee}$ 7

4 Mitigation of modelling errors and degeneracies **7**
4.1 Signal modelling errors and processing masks . . 7
4.2 Degeneracies with the gain 8

⋆ Corresponding author: H. T. Ihle; h.t.ihle@astro.uio.no

5 Results **10**
5.1 Posterior distributions and Gibbs chains 10
5.2 Time variability and goodness-of-fit 11

6 Systematic effects **18**
6.1 Temperature changes in the 20 K stage 18
6.2 Temperature fluctuations and $1/f$ parameters . . 19
6.3 Seasonal effects and slow drifts 20
6.4 Inter-radiometer correlations 20
6.5 Correlation with housekeeping data 22
6.6 Issues with individual radiometers 23

7 Conclusions **24**

1. Introduction

One of the main algorithmic achievements made within the field of CMB analysis during the last few decades is accurate and nearly lossless data compression. Starting from data sets that typically comprise $O(10^8 - 10^{11})$ time-ordered measurements, we are now able to routinely produce sky maps that contain $O(10^3 - 10^7)$ pixels (e.g., Tegmark 1997; Ashdown et al. 2007). From these, we may constrain the angular CMB power spectrum, which spans $O(10^3)$ multipoles (e.g., Gorski 1994; Hivon et al. 2002; Wandelt et al. 2004). Finally, from these we may derive tight constraints on a small set of cosmological parameters (e.g., Bond et al. 2000; Lewis & Bridle 2002; Planck Collaboration V 2020; Planck Collaboration VI 2020), which typically is the ultimate goal of any CMB experiment.

Two fundamental assumptions underlying this radical compression process are that the instrumental noise may be modelled to a sufficient precision, and that the corresponding induced uncertainties may be propagated faithfully to higher-order data products. The starting point for this process is typically to assume that the noise is Gaussian and random in time, and does not correlate with the true sky signal at any given time. Under the Gaussian hypothesis, the net noise contribution therefore decreases as $1/\sqrt{N_{\rm obs}}$, where $N_{\rm obs}$ is the number of observations of the given pixel, while the signal contribution is independent of $N_{\rm obs}$.

However, it is not sufficient to assume that the noise is simply Gaussian and random. One must also assume something about its statistical properties, both in terms of its correlation structure in time and its stationarity period. Regarding the correlation structure, the single most common assumption in the CMB literature is that the temporal noise power spectrum density (PSD) can be modelled as a sum of a so-called $1/f$ term and a white noise term (e.g., Bennett et al. 2013; Planck Collaboration II 2020; Planck Collaboration III 2020). The white noise term arises from intrinsic detector and amplifiers' thermal noise, and is substantially reduced by cooling the instrument to cryogenic temperatures, typically to ∼20 K for coherent receivers (as in the case of *Planck* LFI) and to 0.1–0.3 K for bolometric detectors. Traditionally, the white noise of coherent radiometers is expressed in terms of system noise temperature, $T_{\rm sys}$, per unit integration time (measured in $\mathrm{K\,Hz^{-1/2}}$), while for bolometers it is expressed as noise equivalent power, NEP ($\mathrm{W\,Hz^{-1/2}}$). The sources of the $1/f$ noise component include intrinsic instabilities in the detectors, amplifiers and readout electronics, as well as environmental effects, and, notably, atmospheric fluctuations for sub-orbital experiments. In the case of *Planck* LFI, the $1/f$ noise was dominated by gain and noise temperature fluctuations and thermal instabilities (Planck Collaboration II 2020), and was minimized by introducing the 4 K reference loads and gain modulation factor to optimize the receiver balance; see, e.g., Planck Collaboration II (2014, 2016) and BeyondPlanck Collaboration (2020) for further details.

Regarding stationarity, the two most common assumptions are either that the statistical properties remain constant throughout the entire observation period (e.g., Planck Collaboration II 2020), or that it may at least be modelled as piece-wise stationary within for instance one hour of observations (e.g., QUIET Collaboration et al. 2011). Given such basic assumptions, the effect of the instrumental noise on higher-order data products has then traditionally been assessed, and propagated, either through the use of detailed end-to-end simulations (e.g., Planck Collaboration XII 2016) or in the form of explicit noise covariance matrices (e.g., Tegmark et al. 1997; Page et al. 2007; Planck Collaboration V 2020).

The importance of accurate noise modelling is intimately tied to the overall signal-to-noise ratio of the science target in question. For applications with very high signal-to-noise ratios, detailed noise modelling is essentially irrelevant, since other sources of systematic errors dominate the total error budget. One prominent example of this is the CMB temperature power spectrum as measured by *Planck* on large angular scales (Planck Collaboration IV 2020; Planck Collaboration V 2020). Its white noise contribution can be misestimated by orders of magnitude without making any difference in terms of cosmological parameters, because the full error budget is vastly dominated by cosmic variance.

The cases for which accurate noise modelling is critically important are those with signal-to-noise ratios of order unity. For these, noise misestimation may be the difference between obtaining a tantalizing, but ultimately unsatisfying, $2\,\sigma$ result, and claiming a ground-breaking and decisive $5\,\sigma$ discovery; or, the worst-case scenario, erroneously reporting a baseless positive detection.

This regime is precisely where the CMB field is expected to find itself in only a few years from now, as the next-generation CMB experiments (e.g., CMB-S4, *LiteBIRD*, *PICO*, Simons Observatory, and many others; Abazajian et al. 2019; Suzuki et al. 2018; Sugai et al. 2020; Hanany et al. 2019; Ade et al. 2019) are currently being planned, built and commissioned in the search for primordial gravitational waves imprinted in *B*-mode polarization. The predicted magnitude of this signal is expected to be at most a few tens of nanokelvins on angular scales larger than a degree, corresponding to a relative precision of $O(10^{-8})$, and extreme precision is required for a robust detection. It will therefore become critically important to take into account all sources of systematic uncertainties, and propagate these into the final results.

The BeyondPlanck project (BeyondPlanck Collaboration 2020) is an initiative that aims to meet this challenge by implementing the first global Bayesian CMB analysis pipeline that supports faithful end-to-end error propagation, from raw time-ordered data to final cosmological parameters. One fundamental aspect of this approach is a fully parametric data model that is fitted to the raw measurements through standard posterior sampling techniques, simultaneously constraining both instrumental and astrophysical parameters. Within this framework, the instrumental noise is just one among many different sources of uncertainty, all of which are treated on the same statistical basis. The sample-based approach introduced by BeyondPlanck therefore represents a novel and third way of propagating noise uncertainties (Keihänen et al. 2020; Suur-Uski et al. 2020), comple-

mentary to the existing simulation and covariance matrix based approaches used by traditional pipelines.

As a real-world demonstration of this novel framework, the BeyondPlanck collaboration has chosen the *Planck* LFI measurements (Planck Collaboration I 2020; Planck Collaboration II 2020) as its main scientific target (BeyondPlanck Collaboration 2020). These data represent an important and realistic testbed in terms of overall data volume and complexity, and they also have fairly well-understood properties after more than a decade of detailed scrutiny by the *Planck* team (see Planck Collaboration II 2014, 2016, 2020, and references therein). However, as reported in this paper, there are still a number of subtle unresolved and unexplored issues relating even to this important and well-studied data set that potentially may have an impact on higher-level science results. Furthermore, as demonstrated by the current analysis, the detailed low-level Bayesian modelling approach is ideally suited to identify, study and, eventually, mitigate these effects.

Thus, the present paper has two main goals. The first is to describe the general algorithmic framework implemented in the BeyondPlanck pipeline for modelling instrumental noise in CMB experiments. The second goal is then to apply these methods to the *Planck* LFI observations, and characterize the performance and systematic effects of the instrument as a function of time and detector.

The rest of the paper is organized as follows. First, in Sect. 2 we briefly review the BeyondPlanck analysis framework and data model, with a particular emphasis on noise modelling aspects. In Sect. 3, we present the individual sampling steps required for noise modelling, as well as some statistics that are useful for efficient data monitoring. In Sect. 4 we discuss various important degeneracies relevant for noise modelling, and how to minimize the impact of modelling errors. Next, in Sect. 5 we give a high-level overview of the various noise posterior distributions, their correlation properties, as well as detailed specifications for each detector. In Sect. 6 we discuss anomalies found in the data, and interpret these in terms of the instrument and the thermal environment. Finally, we summarize in Sect. 7.

2. The BeyondPlanck data model and framework

The BeyondPlanck project is an attempt to build up an end-to-end data analysis pipeline for CMB experiments going all the way from raw time-ordered data to cosmological parameters in a consistent Bayesian framework. This allows us to characterize degeneracies between instrumental and astrophysical parameters in a statistically well-defined framework, from low-level instrumental quantities such as gain (Gjerløw et al. 2020), bandpasses (Svalheim et al. 2020a), far sidelobes (Galloway et al. 2020b), and correlated noise via Galactic parameters such as the synchrotron amplitude or spectral index (Andersen et al. 2020; Svalheim et al. 2020b), to the angular CMB power spectrum and cosmological parameters (Colombo et al. 2020; Paradiso et al. 2020).

The LFI dataset consists of three bands, at frequencies of roughly 30, 44, and 70 GHz. These bands have two, three, and six radiometer pairs each, respectively, which for historical reasons are numbered from 18 to 28. The two radiometers in each pair are labeled by M and S (Planck Collaboration II 2014). In BeyondPlanck, the raw uncalibrated data, $\boldsymbol{d}$, produced by each of these radiometers is modelled in time-domain as follows,

$$d_{j,t} = g_{j,t}\mathsf{P}_{tp,j}\left[\mathsf{B}^{\mathrm{symm}}_{pp',j}\sum_c \mathsf{M}_{cj}(\beta_{p'},\Delta^j_{\mathrm{bp}})a^c_{p'} + \mathsf{B}^{\mathrm{asymm}}_{pp',j}\left(s^{\mathrm{orb}}_{j,t}+s^{\mathrm{fsl}}_{j,t}\right)\right] + \\ + n^{\mathrm{corr}}_{j,t} + n^{\mathrm{w}}_{j,t}. \tag{1}$$

Here the subscript t denotes a sample index in time domain; j denotes radiometer; p denotes the pixel number; c denotes signal component; $\boldsymbol{g}$ denotes the gain; P denotes the pointing matrix; $\mathsf{B}^{\mathrm{symm}}$ and $\mathsf{B}^{\mathrm{asymm}}$ denote the symmetric and asymmetric beam matrices, respectively; $\boldsymbol{a}$ are the astrophysical signal amplitudes; $\boldsymbol{\beta}$ are the corresponding spectral parameters; Δ_{bp} are the bandpass corrections; M_{cj} is the bandpass-dependent component mixing matrix; $\boldsymbol{s}^{\mathrm{orb}}$ is the orbital dipole; $\boldsymbol{s}^{\mathrm{fsl}}$ are the far sidelobe corrections; $\boldsymbol{n}^{\mathrm{corr}}$ is the correlated noise; and $\boldsymbol{n}^{\mathrm{w}}$ is the white noise. For more details on each of these parameters see BeyondPlanck Collaboration (2020) and the other companion papers.

The goal of the Bayesian approach is now to sample from the joint posterior distribution ,

$$P(\boldsymbol{g}, \boldsymbol{n}_{\mathrm{corr}}, \xi_n, \Delta_{\mathrm{bp}}, \boldsymbol{a}, \beta, C_\ell \mid \boldsymbol{d}). \tag{2}$$

This is a large and complicated distribution, with many degeneracies. However, using Gibbs sampling we can divide the sampling process into a set of managable steps. Gibbs sampling is a simple algorithm in which samples from a joint multi-dimensional distribution are generated by iterating through all corresponding conditional distributions. Using this method, the BeyondPlanck sampling scheme may be summarized as follows (BeyondPlanck Collaboration 2020),

$$\begin{aligned}
\boldsymbol{g} &\leftarrow P(\boldsymbol{g} \mid \boldsymbol{d}, \xi_n, \Delta_{\mathrm{bp}}, \boldsymbol{a}, \beta, C_\ell) && (3)\\
\boldsymbol{n}_{\mathrm{corr}} &\leftarrow P(\boldsymbol{n}_{\mathrm{corr}} \mid \boldsymbol{d}, \boldsymbol{g}, \xi_n, \Delta_{\mathrm{bp}}, \boldsymbol{a}, \beta, C_\ell) && (4)\\
\xi_n &\leftarrow P(\xi_n \mid \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\mathrm{corr}}, \Delta_{\mathrm{bp}}, \boldsymbol{a}, \beta, C_\ell) && (5)\\
\Delta_{\mathrm{bp}} &\leftarrow P(\Delta_{\mathrm{bp}} \mid \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\mathrm{corr}}, \xi_n, \boldsymbol{a}, \beta, C_\ell) && (6)\\
\beta &\leftarrow P(\beta \mid \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\mathrm{corr}}, \xi_n, \Delta_{\mathrm{bp}}, C_\ell) && (7)\\
\boldsymbol{a} &\leftarrow P(\boldsymbol{a} \mid \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\mathrm{corr}}, \xi_n, \Delta_{\mathrm{bp}}, \beta, C_\ell) && (8)\\
C_\ell &\leftarrow P(C_\ell \mid \boldsymbol{d}, \boldsymbol{g}, \boldsymbol{n}_{\mathrm{corr}}, \xi_n, \Delta_{\mathrm{bp}}, \boldsymbol{a}, \beta). && (9)
\end{aligned}$$

Here, $\leftarrow$ indicates sampling from the distribution on the right-hand side.

Note that for some of these steps we are not following the strict Gibbs approach of conditioning on all but one variable. Most notably for us, this is the case for the gain sampling step in Eq. (3), where we do not condition on $\boldsymbol{n}_{\mathrm{corr}}$. In effect, we instead sample the gain and correlated noise jointly by exploiting the definition of a conditional distribution,

$$P(\boldsymbol{g}, \boldsymbol{n}_{\mathrm{corr}} \mid \boldsymbol{d}, \cdots) = P(\boldsymbol{g} \mid \boldsymbol{d}, \cdots)P(\boldsymbol{n}_{\mathrm{corr}} \mid \boldsymbol{d}, \boldsymbol{g}, \cdots). \tag{10}$$

This equation implies that a joint sample $\{\boldsymbol{g}, \boldsymbol{n}_{\mathrm{corr}}\}$ may be produced by first sampling the gain from the marginal distribution with respect to $\boldsymbol{n}_{\mathrm{corr}}$, and then sampling $\boldsymbol{n}_{\mathrm{corr}}$ from the usual conditional distribution with respect to $\boldsymbol{g}$. The advantage of this joint sampling procedure is a much shorter Markov correlation length as compared to standard Gibbs sampling, as discussed by Gjerløw et al. (2020).

A convenient property of Gibbs sampling is its modular nature, as the various parameters are sampled independently within each conditional distribution, but joint dependencies are still explored through the iterative scheme. In this paper, we are therefore only concerned with two of the above steps, namely Eqs. (4)

and (5). For details on the complete Gibbs chain and the other sampling steps, see BeyondPlanck Collaboration (2020) and the companion papers.

The LFI time-ordered data are divided into roughly 45 000 pointing periods, denoted PIDs (pointing ID). Most PIDs have a duration of 30–60 minutes. When sampling the correlated noise and the corresponding PSD parameters, we assume that the noise is stationary within each PID, but independent between PIDs. The gain is also assumed to be constant within each PID; however, this is not fit independently for each PID, but rather sampled smoothly on longer timescales (Gjerløw et al. 2020).

Following previous literature (Planck Collaboration II 2014; Tauber et al. 2019; Planck Collaboration II 2020), we assume that the LFI noise PSD can be described by a so-called $1/f$ model,

$$P(f) = \sigma_0^2 \left[1 + \left(\frac{f}{f_{\rm knee}}\right)^{\alpha}\right]. \tag{11}$$

Here f denotes a temporal frequency; σ_0 quantifies the white noise level of the time-ordered data[1]; α is the slope (typically negative) of the correlated noise spectrum; and the knee frequency, $f_{\rm knee}$, denotes the (temporal) frequency at which the variance of the correlated noise is equal to the white noise variance. The three PSD parameters are collectively denoted $\xi_n = \{\sigma_0, f_{\rm knee}, \alpha\}$. Note, however, that there is nothing unique or fundamental about the $1/f$ noise model, and a real-world instrument may exhibit a more complicated noise structure than this. Determining whether Eq. (11) represents a statistically acceptable model for the LFI noise is an important goal of the current paper.

3. Methods

As outlined above, noise estimation in the Bayesian BeyondPlanck framework amounts essentially to being able to sample from two conditional distributions, namely $P(\boldsymbol{n}_{\rm corr} \mid \boldsymbol{d}, \omega \setminus \boldsymbol{n}_{\rm corr})$ and $P(\xi_n \mid \boldsymbol{d}, \omega \setminus \xi_n)$. The first presentation of Bayesian noise estimation for time-ordered CMB data that was applicable to the current problem was presented by Wehus et al. (2012), and the main novel feature presented in the current paper is simply the integration of these methods into the larger end-to-end analysis framework outlined above. In addition, the current analysis also employs important numerical improvements as introduced by Keihänen et al. (2020), in which optimal mapmaking is rephrased into an efficient Bayesian language.

The starting point for both conditional distributions is the following parametric data model,

$$\boldsymbol{d} = g\boldsymbol{s}^{\rm tot} + \boldsymbol{n}^{\rm corr} + \boldsymbol{n}^{\rm wn}, \tag{12}$$

where $\boldsymbol{d}$ denotes the raw time ordered data (TOD) organized into a column vector; g is the gain; $\boldsymbol{s}^{\rm tot}$ describes the total sky signal, comprising both CMB and foregrounds, projected into time-domain; $\boldsymbol{n}^{\rm corr}$ represents the correlated noise in time domain; and $\boldsymbol{n}^{\rm wn}$ is white noise. The two noise terms are both assumed to be Gaussian distributed with covariance matrices $\mathsf{N}_{\rm corr} \equiv \langle \boldsymbol{n}_{\rm corr} \boldsymbol{n}_{\rm corr}^T \rangle$ and $\mathsf{N}_{\rm wn} \equiv \langle \boldsymbol{n}_{\rm wn} \boldsymbol{n}_{\rm wn}^T \rangle$, respectively. The complete noise PSD is then given by $P(f) = \mathsf{N}_{\rm wn} + \mathsf{N}_{\rm corr} = \sigma_0^2 + \sigma_0^2 \left(\frac{f}{f_{\rm knee}}\right)^{\alpha}$.

[1] σ_0 has different units if we are talking about the uncalibrated data, σ_0 [V], calibrated data, σ_0 [K] $\equiv \sigma_0$ [V] $/g$, or the white noise PSD, σ_0^2 [K^2 Hz^{-1}] $\equiv (\sigma_0$ [K]$)^2 \frac{2}{R_{\rm samp}}$, where $R_{\rm samp}$ is the sample rate (in Hz) of the time ordered data. Where this distinction is important, we include the units explicitly.

3.1. Sampling correlated noise, $P(\boldsymbol{n}^{\rm corr} \mid \boldsymbol{d}, \xi^n, \boldsymbol{s}_{\rm tot}, g)$

Our first goal is to derive an appropriate sampling prescription for the time-domain correlated noise conditional distribution, $P(\boldsymbol{n}^{\rm corr} \mid \boldsymbol{d}, \xi^n, \boldsymbol{s}_{\rm tot}, g)$. To this end, we start by defining the signal-subtracted data, $\boldsymbol{d}'$, directly exploiting the fact that g and $\boldsymbol{s}^{\rm tot}$ are currently conditioned upon,[2]

$$\boldsymbol{d}' \equiv \boldsymbol{d} - g\boldsymbol{s}^{\rm tot} = \boldsymbol{n}^{\rm corr} + \boldsymbol{n}^{\rm wn}. \tag{13}$$

Since both $\boldsymbol{n}^{\rm corr}$ and $\boldsymbol{n}^{\rm wn}$ are assumed Gaussian with known covariance matrices, the appropriate sampling equation for $\boldsymbol{n}^{\rm corr}$ is also that of a multivariate Gaussian distribution, which is standard textbook material; for a brief review, see Appendix A in BeyondPlanck Collaboration (2020). In particular, the maximum likelihood (ML) solution for $n_t^{\rm corr}$ is given by the so-called Wiener-filter equation,

$$\left(\mathsf{N}_{\rm corr}^{-1} + \mathsf{N}_{\rm wn}^{-1}\right) \boldsymbol{n}^{\rm corr} = \mathsf{N}_{\rm wn}^{-1} \boldsymbol{d}', \tag{14}$$

while a random sample of $\boldsymbol{n}^{\rm corr}$ may be found by solving the following equation,

$$\left(\mathsf{N}_{\rm corr}^{-1} + \mathsf{N}_{\rm wn}^{-1}\right) \boldsymbol{n}^{\rm corr} = \mathsf{N}_{\rm wn}^{-1} \boldsymbol{d}' + \mathsf{N}_{\rm wn}^{-1/2} \boldsymbol{\eta}_1 + \mathsf{N}_{\rm corr}^{-1/2} \boldsymbol{\eta}_2, \tag{15}$$

where $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are two independent vectors of random variates drawn from a standard Gaussian distribution, $\boldsymbol{\eta}_{1,2} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$.

3.1.1. Ideal data

Assuming for the moment that both $\mathsf{N}_{\rm corr}$ and $\mathsf{N}_{\rm wn}$ are diagonal in Fourier space, we note that Eq. (15) may be solved in a closed form in Fourier space,

$$n_f^{\rm corr} = \frac{d'_f + C\left(\mathsf{N}_{\rm wn}^{1/2}(f) w_1 + \mathsf{N}_{\rm wn}(f) \mathsf{N}_{\rm corr}^{-1/2}(f) w_2\right)}{1 + \mathsf{N}_{\rm wn}(f)/\mathsf{N}_{\rm corr}(f)}, \tag{16}$$

for any non-negative frequency f, where the correlated noise TOD has been decomposed as $n_f^{\rm corr} = \sum_t n_t^{\rm corr} \mathrm{e}^{-2\pi i f t}$. For completeness, C is a constant factor that depends on the Fourier convention of the numerical library of choice,[3] and $w_{1,2}$ are two independent random complex samples from a Gaussian distribution,

$$w_{1,2} \equiv \frac{\eta_R + i\eta_I}{\sqrt{2}}, \tag{17}$$

where $\eta_{R,I} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$.

Figure 1 shows three independent realizations of $\boldsymbol{n}^{\rm corr}$ that all correspond to the same signal-subtracted *Planck* 30 GHz TOD segment. Each correlated noise sample is essentially a Wiener-filtered version of the original data, and traces as such the slow variations in the data, with minor variations corresponding the two random fluctuation terms in Eq. (16), as allowed by the white noise level present in the data. We can also see that there are gaps in the data, which we will need to deal with.

[2] When a parameter appears on the right-hand side of a conditioning bar in a probability distribution, it is assumed known to infinite precision. It is therefore for the moment a constant quantity, and not associated with any stochastic degrees of freedom or uncertainties.

[3] We use the `FFTW` library, in which case $C = \sqrt{n_{\rm samples}}$, where $n_{\rm samples}$ is the number of time samples.

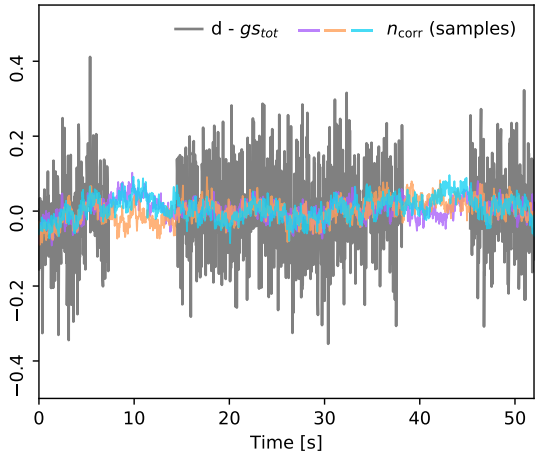

Fig. 1. Illustration of three constrained correlated noise realizations (colored curves) drawn from $P(\boldsymbol{n}^{\rm corr} \mid \boldsymbol{d}, \xi^n, \boldsymbol{s}_{\rm tot}, g)$ for the *Planck* 30 GHz data (grey curve). Regions for which parts of the data have been masked, either due to a processing mask or flagged data, are marked as white gaps.

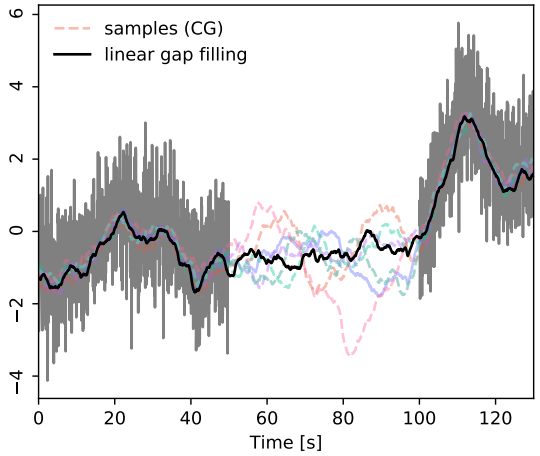

Fig. 2. Illustration of the limitation of the linear gap-filling procedure for simulated data with extreme noise properties and large gaps. In general, the linear gap-filling procedure tends to underestimate the fluctuations in $n^{\rm corr}$ on long timescales.

3.1.2. Handling masking through a conjugate gradient solver

When writing down an explicit solution of Eq. (15) in Eq. (16), we assumed that both $\mathsf{N}_{\rm corr}$ and $\mathsf{N}_{\rm wn}$ were diagonal in Fourier space. However, as illustrated in Fig. 1, real observations have gaps, either because of missing or flagged data. The most typical example of missing data is the application of a processing mask that removes all samples with too strong foreground contamination, either from Galactic diffuse sources or from extragalactic point sources.

We can represent these gaps in our statistical model by setting the white noise level for masked samples to infinity. This ensures that Eqs. (14) and (15) are still well defined, albeit somewhat harder to solve. The new difficulty lies in the fact that while $\mathsf{N}_{\rm wn}$ is still diagonal in the time domain, it is no longer diagonal in the Fourier domain. This problem may be addressed in two ways. Specifically, we can either solve Eqs. (14) and (15) directly, using an iterative method such as the conjugate gradient (CG) method (Wehus et al. 2012; Keihänen et al. 2020), or we can fill any gap in $\boldsymbol{d}'$ with a simpler interpolation scheme, for instance a polynomial plus white noise, and then use Eq. (16) directly. Clearly, the former method is mathematically superior, as it results in a statistically exact result. However, the CG method is in general not guaranteed to converge due to numerical round-off errors, and since the current algorithm is to be applied millions of times in a Monte Carlo environment, the second approach is useful as a fallback solution for the few cases for which the exact CG approach fails.

As shown by Keihänen et al. (2020), Eq. (15) may be recast into a compressed form using the Sherman-Morrison-Woodbury formula, effectively separating the masked from the unmasked degrees of freedom, and the latter may then be handled with the direct formula in Eq. (16). This approach, in addition to having a lower computational cost per CG iteration, also needs fewer iterations to converge compared to the untransformed equation. We adopt this approach without modifications for the main BeyondPlanck pipeline.

Returning to Fig. 1, we note that the correlated noise samples have significant larger variance within the gaps than in the data-dominated regime. As a result, one should expect to see a slightly higher conditional χ^2 inside the processing mask in a full analysis than outside, since $\boldsymbol{n}^{\rm corr}$ will necessarily trace the real data less accurately in that range. This is in fact seen in the main BeyondPlanck analysis, as reported by BeyondPlanck Collaboration (2020) and Suur-Uski et al. (2020). However, when marginalizing over all allowed correlated noise realizations, the final uncertainties will be statistically appropriate, due to the fluctuation terms in Eq. (15).

3.1.3. Gap-filling by polynomial interpolation

As mentioned above, the CG algorithm does not always converge, and for Monte Carlo applications that will run millions of times without human supervision, it is useful to establish a robust fallback solution. For this purpose, we adopt the basic approach of simply interpolating between the values on each side of a gap. Specifically, we compute the average of the non-masked points among the 20 points on each side of the gap, and interpolate linearly between these two values. In addition, we add a white noise component to $\boldsymbol{d}'$, based on $\mathsf{N}^{\rm wn}$, to each masked sample.

An important limitation of the linear gap-filling procedure is associated with estimation of the noise PSD parameters, ξ^n. As described in Sect. 3.2, these parameters are estimated directly from $\boldsymbol{n}^{\rm corr}$ by Gibbs sampling. A statistically suboptimal sample of $\boldsymbol{n}^{\rm corr}$ may therefore also bias ξ^n, which in turn may skew $\boldsymbol{n}^{\rm corr}$ even further. If the gaps are short, then this bias is usually negligible, but for large gaps it can be problematic. This situation is illustrated in Fig. 2, which compares the linear gap filling procedure with the exact CG approach. In general, the linear method tends to underestimate the fluctuations on large timescales within the gap.

Because of the close relative alignment of the *Planck* scanning strategy with the Galactic plane that takes place every six months (Planck Collaboration I 2011), some pointing periods happen to have larger gaps than others. For these, two long masked regions occur every minute, when the telescope points toward the Galactic plane. Any systematic bias introduced by the gap-filling procedure itself will then not be randomly distributed in the TOD, but rather systematically contribute to the same modes, with a specific period equal to the satellite spin rate. For these, the statistical precision of the CG algorithm is particularly important to avoid biased noise parameters.

Overall, the linear gap filling procedure should only be used when strictly necessary. In practice, we use it only when the CG solver fails to converge within 30 iterations, which happens in less than 0.03 % of all cases.

Another simpler and more accurate gap filling procedure is suggested by Keihänen et al. (2020): We may simply fill the gaps in $\boldsymbol{d}'$ with the previous sample of the correlated noise, and then add white noise fluctuations. This corresponds to Gibbs sampling over the white noise as a stochasic parameter, which is statistically fully valid. However, this approach requires us to store the correlated noise TOD in memory between consecutive Gibbs iterations. Since memory use is already at its limit (Galloway et al. 2020a), this method is not used for the main BeyondPlanck analysis. However, for systems with more available RAM, this method is certainly preferable over simple linear interpolation.

3.2. Sampling noise PSD parameters, $P(\xi^n \mid \boldsymbol{n}^{\rm corr})$

The second noise-related conditional distribution in the BeyondPlanck Gibbs chain is $P(\xi^n \mid \boldsymbol{n}^{\rm corr})$, which describes the noise PSD. As discussed in Sect. 2, in this paper we model this function in terms of a $1/f$ spectrum as defined by Eq. (11). We emphasize, however, that any functional form for $P(f)$ may be fitted using the methods described below. Figure 3 illustrates the PSD of the different components, and our task is now to sample each of the noise PSD parameters $\xi^n = \{\sigma_0, f_{\rm knee}, \alpha\}$, corresponding to the dashed blue line in this figure.

3.2.1. Sampling the white noise level, σ_0

We start with the white noise level, which by far is the most important noise PSD parameter in the system. We first note from Eq. (11) that if α is close to zero, the correlated and white noise terms are perfectly degenerate. Even for $\alpha \approx -1$ there is a significant degeneracy between the two components for a finite-length TOD.

Of course, for other parameters in the full Gibbs chain, only the combined $P(f)$ function is relevant, and not each component separately. At the same time, and as described by BeyondPlanck Collaboration (2020), marginalization over the two terms within other sampling steps happens using two fundamentally different methods: While white noise marginalization is performed analytically through a diagonal covariance matrix, marginalization over correlated noise is done by Monte Carlo sampling of $\boldsymbol{n}^{\rm corr}$. It is therefore algorithmically advantageous to make sure that the white noise term accounts for as much as possible of the full noise variance, as this will lead to an overall shorter Markov chain correlation length.

For this reason, we employ a commonly used trick in radio astronomy for estimating the white noise level, and define this to be

$$\sigma_0^2 \equiv \frac{\mathrm{Var}(r_{i+1} - r_i)}{2}, \tag{18}$$

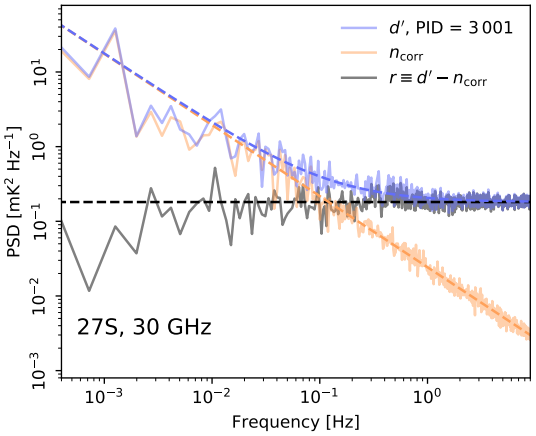

Fig. 3. Comparison of temporal PSDs for different components. The blue curve shows the PSD of the signal-subtracted data; the orange curve shows the fitted correlated noise PSD; and the gray line shows the PSD of the residual TOD. The dashed curves correspond to the best fit $1/f$-noise model, with (blue) and without (orange) white noise.

where $\boldsymbol{r} \equiv \boldsymbol{d}' - \boldsymbol{n}^{\rm corr}$. By differencing consecutive samples, any residual temporal correlations are effectively eliminated, and will therefore not bias the determination of σ_0.

This method is equivalent to fixing the white noise level to the highest frequencies in Fig. 3. Formally speaking, this means that σ_0 should not be considered a free parameter within the Gibbs chain, but rather a derived quantity fixed by the data, $\boldsymbol{d}$, the gain, g, the signal model, $\boldsymbol{s}^{\rm tot}$, and the correlated noise, $\boldsymbol{n}^{\rm corr}$. However, this distinction does not carry any particular statistical significance with respect to other parameters, and we will in the following therefore discuss σ_0 on the same footing as any of the other noise parameters.

3.2.2. Sampling correlated noise parameters, $f_{\rm knee}$ and α

With σ_0^2 fixed by Eq. (18), the other noise parameters, $f_{\rm knee}$ and α, are sampled from their exact conditional distributions. Since we assume that also the correlated noise component is Gaussian distributed, the appropriate functional form is that of a multivariate Gaussian,

$$P(f_{\rm knee}, \alpha \mid \sigma_0, \boldsymbol{n}^{\rm corr}) \propto \frac{\mathrm{e}^{-\frac{1}{2}(\boldsymbol{n}^{\rm corr})^T \mathsf{N}_{\rm corr}^{-1} \boldsymbol{n}^{\rm corr}}}{\sqrt{|\mathsf{N}_{\rm corr}|}} P(f_{\rm knee}, \alpha), \tag{19}$$

where $\mathsf{N}_{\rm corr} = \mathsf{N}_{\rm corr}(f_{\rm knee}, \alpha)$, and $P(f_{\rm knee}, \alpha)$ is an optional prior. This may be efficiently evaluated in Fourier space as

$$-\ln P = \sum_{f=f_{\rm min}}^{f_{\rm max}} \left[\frac{|n_f^{\rm corr}|^2}{\mathsf{N}_{\rm corr}(f)} + \ln \mathsf{N}_{\rm corr}(f) \right] - \ln P(f_{\rm knee}, \alpha), \tag{20}$$

where $\mathsf{N}_{\rm corr}(f) = \sigma_0^2 \left(\frac{f}{f_{\rm knee}}\right)^{\alpha}$.

To explore this joint distribution, we iteratively Gibbs sample over $f_{\rm knee}$ and α, using an inversion sampler for each of the two conditional distributions, $P(f_{\rm knee} \mid \alpha, \sigma_0, n^{\rm corr})$ and $P(\alpha \mid f_{\rm knee}, \sigma_0, n^{\rm corr})$; see Appendix A in BeyondPlanck Collaboration (2020) for details regarding the inversion sampler.

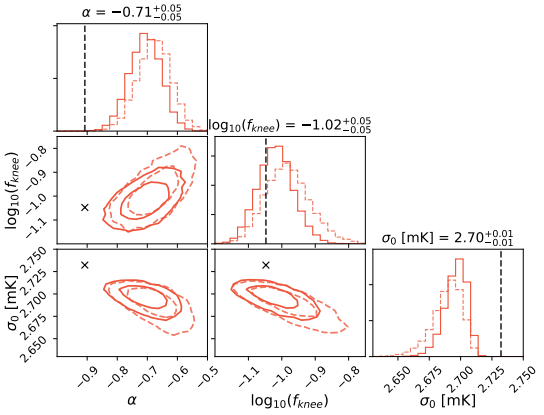

Fig. 4. Distribution of noise parameters for PID 10101 of radiometer 24S, one of the 44 GHz channels, for fixed $\boldsymbol{s}_{\rm tot}$ and g. Dashed red lines correspond to results obtained without an active prior, while the solid line corresponds to results after including the active priors on $f_{\rm knee}$ and α from Eqs. (21) and (22). Black crosses indicate the best fit values derived by the DPC pipeline for this radiometer.

If we naively apply our statistical model, all frequencies should in principle be included in the sum in Eq. (20). At the same time, we note that frequencies well above $f_{\rm knee}$ ideally should carry very little statistical weight, since the correlated noise variance then by definition is smaller than the white noise variance. This means that the sampled $\boldsymbol{n}^{\rm corr}$ is almost completely determined by the prior (i.e., the previous values of $\sigma_0, f_{\rm knee}, \alpha$), at these high frequencies. The sum in Eq. (20), on the other hand, is completely dominated by those high frequencies. The result of this is an excessively long Markov chain correlation length when including all frequencies in Eq. (20); the inferred values of α and $f_{\rm knee}$ will always be extremely close to the previous values.

One way to avoid these long correlation lengths would be not to condition on $\boldsymbol{n}^{\rm corr}$ at all, but rather use the likelihood for $\boldsymbol{d}'$ to sample α and $f_{\rm knee}$ (and sample $\boldsymbol{n}^{\rm corr}$ afterwards). This is equivalent to sampling ξ^n from the marginal distribution with respect to $\boldsymbol{n}^{\rm corr}$, and fully analogous to how the degeneracy between g and $\boldsymbol{n}^{\rm corr}$ is broken through joint sampling. However, for real world data, residual signal or systematics may leak into $\boldsymbol{d}'$, in particular at frequencies around and above the satellite scanning frequency. While some of these systematics may also leak into $\boldsymbol{n}^{\rm corr}$, in general $\boldsymbol{n}^{\rm corr}$ is cleaner, especially at frequencies below $f_{\rm knee}$, where $\boldsymbol{n}^{\rm corr}$ is dominated by the random sampling terms.

A useful solution that both makes the correlated noise parameters robust against modelling errors and results in a short Markov chain correlation length is therefore to condition on $\boldsymbol{n}^{\rm corr}$ above some pre-specified frequency. In practice, we therefore choose to only include frequencies below $f_{\rm max} = 2f_{\rm knee}^{\rm DPC}$ when evaluating Eq. (20), where $f_{\rm knee}^{\rm DPC}$ is the knee frequency determined by the *Planck* LFI Data Processing Center (DPC, Planck Collaboration II 2020). That is, we only use the part of $\boldsymbol{n}^{\rm corr}$ where we are able to measure the $1/f$ slope with an appreciable signal to noise ratio. For the lower frequency cutoff in Eq. (20), we adopt $f_{\rm min} > 0$, and only exclude the overall mean per PID.

3.2.3. Priors on α and $f_{\rm knee}$

As described by Planck Collaboration II (2020), the official *Planck* LFI DPC analyses assume the noise PSD to be stationary throughout the mission. Here we allow these parameters to vary from PID to PID, in order to accommodate possible changes in the thermal environment of the satellite. However, since the duration of a single PID is typically one hour or shorter, there is only a limited number of large-scale frequencies available to estimate the correlated noise parameters, and this may in some cases lead to significant degeneracies between α and $f_{\rm knee}$. In particular, if $f_{\rm knee}$ is low (which of course is the ideal case), α is essentially unconstrained. To avoid pathological cases, it is therefore useful to impose priors on each of these parameters, under the assumption that the system should be relatively stable as a function of time.

Specifically, we adopt a log-normal prior for $f_{\rm knee}$,

$$-\ln P(f_{\rm knee}) = \frac{1}{2}\left(\frac{\log_{10} f_{\rm knee} - \log_{10} f_{\rm knee}^{\rm DPC}}{\sigma_{f_{\rm knee}}}\right)^2 + \ln f_{\rm knee}, \quad (21)$$

where $f_{\rm knee}^{\rm DPC}$ is the DPC result for a given radiometer (Planck Collaboration II 2020) and $\sigma_{f_{\rm knee}} = 0.1$. For α, we adopt a Gaussian prior of the form

$$-\ln P(\alpha) = \frac{1}{2}\left(\frac{\alpha - \alpha^{\rm DPC}}{\sigma_\alpha}\right)^2, \quad (22)$$

where $\alpha^{\rm DPC}$ again is the DPC result for the given radiometer and $\sigma_\alpha = 0.2$. Figure 4 shows a comparison of the posterior distributions with (solid lines) and without (dashed lines) active priors for a typical example.

The prior widths have been chosen to be sufficiently loose that the overall impact of the priors is moderate for most cases. The priors are in practice only used to exclude pathological cases. Technically speaking, we also impose absolute upper and lower limits for each parameter, as this is needed for gridding the conditional distribution within the inversion sampler. However, the limits are chosen to be sufficiently wide so that they have no significant impact on final results.

4. Mitigation of modelling errors and degeneracies

When applying the methods described above to real-world data as part of a larger Gibbs chain, several other degeneracies and artifacts may emerge beyond those discussed above. In this section, we discuss some of the main challenges for the current setup, and we also describe solutions to break or mitigate these issues.

4.1. Signal modelling errors and processing masks

First, we note that the correlated noise component is by nature entirely instrument specific, and depends directly on the thermal stability of the detectors. It is therefore difficult to impose any strong spatial priors on $\boldsymbol{n}^{\rm corr}$, beyond the loose PSD priors described above, and these provide only very weak constraints in map-domain. The correlated noise is from first principles the least known parameter in the entire model, and its allowed parameter space is able to describe a wide range of different TOD combinations, without inducing a significant likelihood penalty relative to the $1/f$ model. As a result, a wide range of systematic

errors or model mismatches may be described quite accurately by modifying $\boldsymbol{n}^{\rm corr}$, rather than ending up in the residual,

$$\boldsymbol{r} \equiv \boldsymbol{d} - \boldsymbol{n}^{\rm corr} - g\boldsymbol{s}_{\rm tot}. \quad (23)$$

Colloquially speaking, the correlated noise component may in many respects be considered the "trash can" of CMB time-ordered analysis, capturing anything that does not fit elsewhere in the model. This is both a strength and a weakness. On the one hand, the flexibility of $\boldsymbol{n}^{\rm corr}$ protects against modelling errors for other (and far more important) parameters in the model, including the CMB parameters. On the other hand, in many cases it is preferable that modelling errors show up as χ^2 excesses, so that they can be identified and mitigated, rather than leaking into the correlated noise. To check for different types of modelling errors, it is therefore extremely useful to inspect both χ^2's and binned sky maps of $\boldsymbol{r}_\nu$ and $\boldsymbol{n}^{\rm corr}$ for artifacts. For an explicit example of this, see the discussion of data selection for BeyondPlanck in Suur-Uski et al. (2020), where these statistics are used as efficient tools to identify bad observations.

In general, the most problematic regions of the sky are those with bright foregrounds, either in the form of diffuse Galactic emission or strong compact sources. If residuals from such foregrounds are present in the signal-subtracted data, $\boldsymbol{d}'$, while estimating the correlated noise TOD, the correlated noise Wiener filter in Eq. (15) will attempt to fit these in $\boldsymbol{n}^{\rm corr}$, and this typically results in stripes along the scanning path with a correlation length defined by the ratio between $f_{\rm knee}$ and the scanning frequency.

To suppress such artifacts, we impose a processing mask for each frequency, as discussed in Sect. 3.1. In the current analysis, we define these masks as follows:

1. We bin the time-domain residual in Eq. (23) into an IQU pixelized sky map for each frequency (as defined by Eq. (77) in BeyondPlanck Collaboration 2020), and smooth this map to an angular resolution of 10° FWHM.
2. We take the absolute value of the smoothed map, and then smooth again with a 30′ beam to account for pixels which the raw residual map changes sign.
3. We then compute the maximum absolute value for each pixel over each of the three Stokes parameters. The resulting maps are shown in Fig. 5 for each of the three *Planck* LFI frequencies.
4. These maps are then thresholded at values well above the noise level, and these tresholded maps form the main input to the processing masks.
5. To remove particularly bright compact objects that may not be picked up by the smooth residual maps described above, we additionally remove all pixels with high free-free and/or AME levels, as estimated in an earlier analysis.

The final processing masks are shown in Fig. 6, and allow 73, 81, and 77 % of the sky to be included while fitting correlated noise at 30, 44, and 70 GHz, respectively.

4.2. Degeneracies with the gain

The brightest component of the entire BeyondPlanck signal model is the Solar CMB dipole, which has an amplitude of 3 mK. This component plays a critical role in terms of gain estimation (Gjerløw et al. 2020), and serves as the main tool to determine relative calibration differences between detectors. Both the gain and CMB dipole parameters are of course intrinsically unknown quantities, and must be fitted jointly. Any error in the determination of these will therefore necessarily result in a nonzero

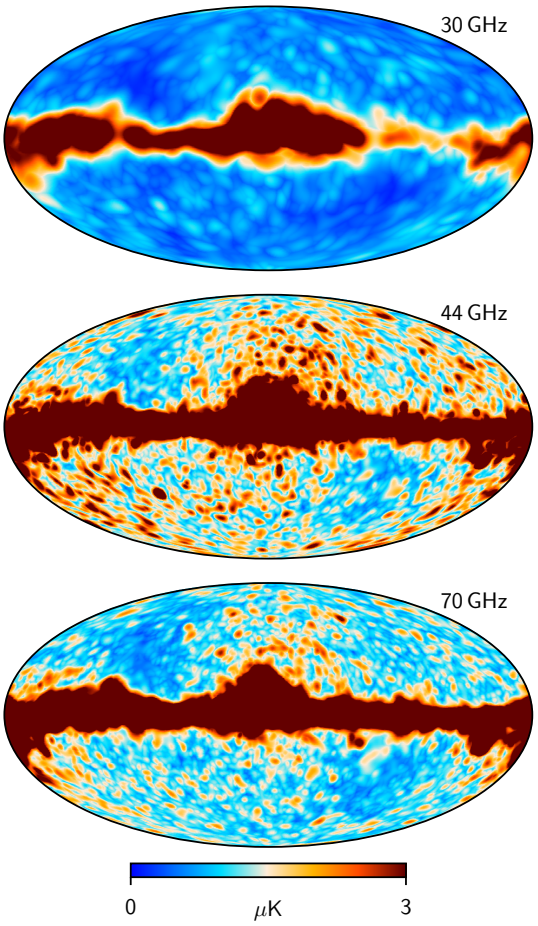

Fig. 5. Residual maps, $\boldsymbol{r}_\nu$, for each of the three *Planck* LFI frequencies, smoothed to a common angular resolution of 10° FWHM.

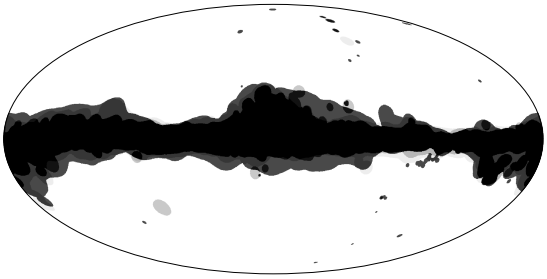

Fig. 6. Processing masks used for correlated noise sampling. Different shades of gray indicate different frequency masks. The allowed 30 GHz sky fraction (light) is $f_{\rm sky} = 0.73$; the 44 GHz sky fraction (intermediate) is $f_{\rm sky} = 0.81$; and the 70 GHz sky fraction (dark) is $f_{\rm sky} = 0.77$.

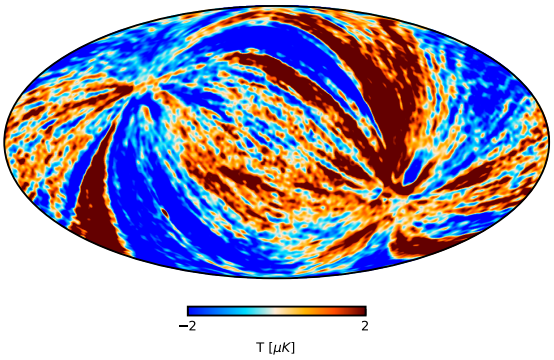

Fig. 7. Correlated noise intensity sample for the 30 GHz band when fitting a model that assumes constant gains throughout the mission. Map has been smoothed to an angular resolution of 2.5° FWHM.

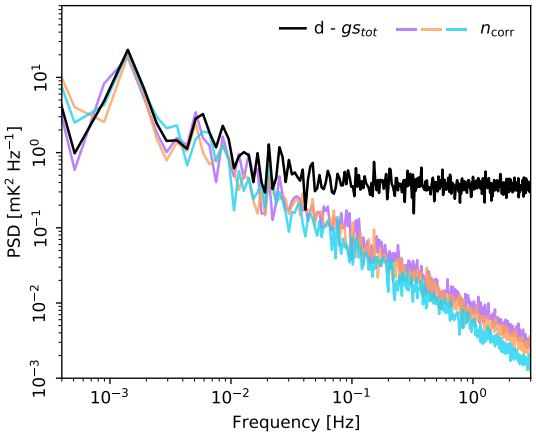

Fig. 8. Three subsequent samples (colored curves) of the correlated noise PSD for 25M, one of the 44 GHz radiometers. The black line shows the PSD of the signal-subtracted data.

residual, in the same manner as Galactic foregrounds described above, and this may therefore potentially also bias $\boldsymbol{n}^{\rm corr}$. Unlike the Galactic residuals, however, it is not possible to mask the CMB dipole, since it covers the full sky. The correlated noise component is therefore particularly susceptible to errors in either the gain or CMB dipole parameters, and residual large-scale dipole features in the binned $\boldsymbol{n}^{\rm corr}$ map is a classic indication of calibration errors. To illustrate the effect of an incorrect gain model, Fig. 7 shows a 30 GHz correlated noise sample when assuming that the gain is constant throughout the entire *Planck* mission.

The gain also has a direct connection with the white noise level, σ_0. This manifests itself in different ways, depending on the choice of units adopted for σ_0. When expressed in units of volts, the white noise level is simply given by the radiometer equation,

$$\sigma_0[\mathrm{V}] \propto g_{\rm phys} T_{\rm sys}, \tag{24}$$

Table 1. Distribution of posterior mean noise parameters for each radiometer. Error bars represent variation over time of the posterior mean values, and not the width of the posterior distribution for any given PID.

	Detector	α	$f_{\rm knee}$ [mHz]	σ_0 [mK]
30 GHz	27M	$-0.84^{+0.05}_{-0.06}$	210^{+22}_{-20}	$1.562^{+0.012}_{-0.012}$
	27S	$-0.84^{+0.06}_{-0.06}$	127^{+16}_{-12}	$1.708^{+0.010}_{-0.013}$
	28M	$-0.88^{+0.08}_{-0.11}$	137^{+22}_{-16}	$1.793^{+0.011}_{-0.013}$
	28S	$-1.00^{+0.11}_{-0.17}$	43^{+6}_{-5}	$1.638^{+0.008}_{-0.012}$
44 GHz	24M	$-1.03^{+0.10}_{-0.11}$	27^{+3}_{-3}	$3.155^{+0.012}_{-0.013}$
	24S	$-0.78^{+0.07}_{-0.07}$	92^{+29}_{-12}	$2.704^{+0.013}_{-0.018}$
	25M	$-1.02^{+0.11}_{-0.12}$	$19.5^{+2.7}_{-2.4}$	$2.832^{+0.010}_{-0.015}$
	25S	$-1.03^{+0.08}_{-0.08}$	45^{+5}_{-5}	$2.687^{+0.019}_{-0.015}$
	26M	$-0.97^{+0.07}_{-0.08}$	63^{+8}_{-5}	$3.264^{+0.018}_{-0.012}$
	26S	$-0.86^{+0.09}_{-0.09}$	50^{+8}_{-5}	$2.848^{+0.013}_{-0.012}$
70 GHz	18M	$-1.08^{+0.15}_{-0.19}$	$14.7^{+3.2}_{-2.3}$	$4.57^{+0.03}_{-0.03}$
	18S	$-1.18^{+0.16}_{-0.14}$	$18.2^{+3.3}_{-2.7}$	$4.18^{+0.02}_{-0.03}$
	19M	$-1.20^{+0.15}_{-0.16}$	$11.6^{+2.4}_{-1.8}$	$5.193^{+0.019}_{-0.036}$
	19S	$-1.10^{+0.13}_{-0.14}$	$13.7^{+2.5}_{-2.0}$	$4.962^{+0.018}_{-0.036}$
	20M	$-1.10^{+0.16}_{-0.23}$	$7.6^{+2.3}_{-1.4}$	$5.258^{+0.024}_{-0.028}$
	20S	$-1.25^{+0.13}_{-0.20}$	$5.5^{+1.8}_{-1.0}$	$5.571^{+0.014}_{-0.051}$
	21M	$-1.35^{+0.15}_{-0.12}$	38^{+8}_{-6}	$4.033^{+0.015}_{-0.016}$
	21S	$-1.15^{+0.15}_{-0.18}$	$13.0^{+2.7}_{-1.9}$	$5.018^{+0.027}_{-0.025}$
	22M	$-1.33^{+0.14}_{-0.23}$	$9.5^{+3.4}_{-2.1}$	$4.381^{+0.018}_{-0.019}$
	22S	$-1.25^{+0.21}_{-0.27}$	14^{+8}_{-5}	$4.746^{+0.024}_{-0.024}$
	23M	$-1.04^{+0.10}_{-0.12}$	30^{+4}_{-3}	$4.493^{+0.024}_{-0.021}$
	23S	$-1.19^{+0.08}_{-0.08}$	60^{+6}_{-6}	$4.815^{+0.019}_{-0.024}$

where $g_{\rm phys}$ is the actual physical gain of the radiometer, and $T_{\rm sys}$ is the system temperature (BeyondPlanck Collaboration 2020). In calibrated units of $\mathrm{K_{CMB}}$, however, the white noise level is

$$\sigma_0[\mathrm{K}] \propto \frac{g_{\rm phys}}{g_{\rm model}} T_{\rm sys}, \tag{25}$$

where $g_{\rm model}$ is the gain estimate in our model. When considering the evolution of the noise parameters as a function of time, we then note that σ_0[V] will correlate with the physical gain, which depends strongly on the thermal environment at any given time. On the other hand, if our gain model is correct, i.e., $g_{\rm model} \approx g_{\rm phys}$, these fluctuations will cancel in temperature units, and σ_0[K] should instead correlate with the system temperature, $T_{\rm sys}$. The system temperature also depends on the physical temperature, $T_{\rm phys}$, as the amplifiers' noise and waveguide losses increase with temperature. These were measured in pre-flight tests to be at a level $dT_{\rm sys}/dT_{\rm phys} \approx 0.2$–$0.5$ K/K, depending on the radiometer (Terenzi et al. 2009). In conclusion, if we observe a sudden change in σ_0[K] that is not present in σ_0[V], this might indicate a problem in the gain model. We also expect that changes in σ_0[K] reflect genuine variations of the white noise level, mainly driven by changes in the 20 K stage. In the following, we will plot σ_0 as a function of time in both units of volts and kelvins, and use these to disentangle gain and system temperature variations.

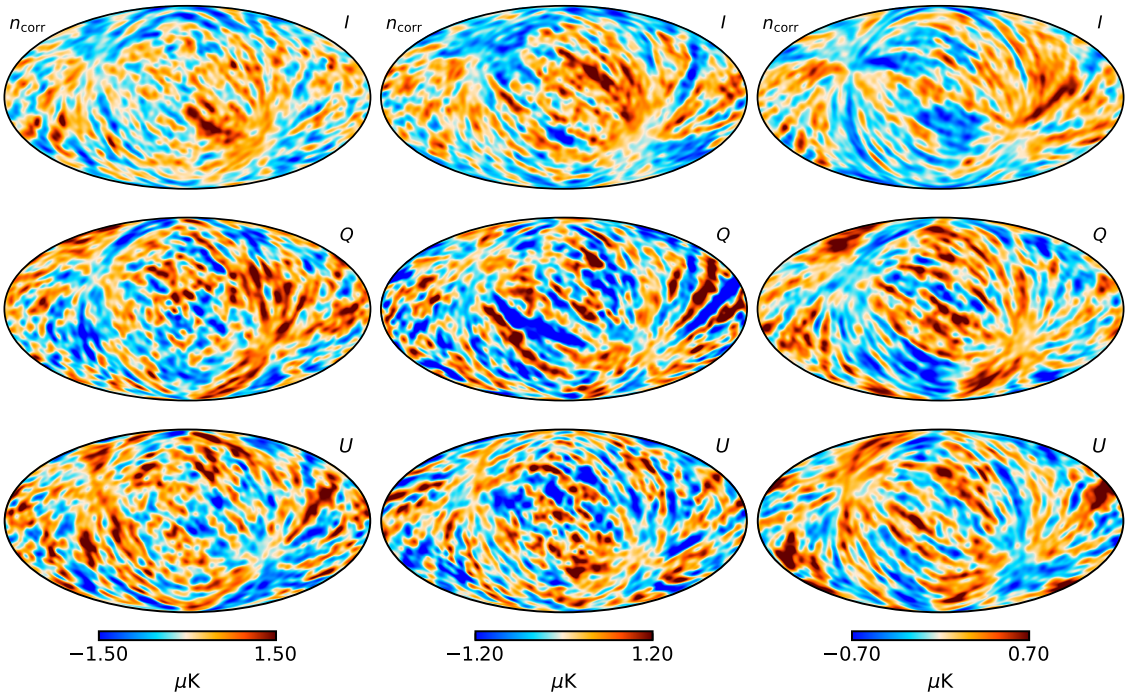

Fig. 9. Maps of a single Gibbs sample of the correlated noise added over all radiometers in the 30 GHz (left), 44 GHz (middle) and 70 GHz (right) bands. From top to bottom, rows show Stokes I, Q and U, respectively. Maps have been smoothed to a common angular resolution of 5° FWHM.

5. Results

We are now ready to present the main results obtained by applying the methods described above to the *Planck* LFI data within the BeyondPlanck Gibbs sampling framework (BeyondPlanck Collaboration 2020), as summarized in terms of the posterior distributions for each of the noise parameters. In total, six independent Gibbs chains were produced in the main BeyondPlanck analysis, each chain including 200 samples, for a total computational cost of about 230 000 CPU-hours or three weeks continuous wall-time (BeyondPlanck Collaboration 2020; Galloway et al. 2020a).

5.1. Posterior distributions and Gibbs chains

First, we recall that at every step in the Gibbs chain, we sample the correlated noise parameters for each pointing period and each radiometer, both the time-domain realization $\boldsymbol{n}^{\rm corr}$ and the PSD parameters ξ^n. To visually illustrate the resulting variations from sample to sample in terms of PSDs, Fig. 8 shows three subsequent spectrum samples for a single pointing period for the 25M radiometer. We see that the correlated noise follows the data closely at low frequencies, while at high frequencies the PSD is effectively extrapolated based on the current model. The scatter between the three colored curves shows the typical level of variations allowed by the combination of white noise and degeneracies with other parameters in the model.

Figure 9 shows the pixel-space correlated noise corresponding to a single Gibbs sample, obtained after binning $\boldsymbol{n}^{\rm corr}$ for all radiometers and all PIDs into an IQU map. Columns show different frequency maps (30, 44, and 70 GHz), and rows show different Stokes parameters (I, Q, and U). Overall, we see that the morphology of each map is dominated by stripes along the *Planck* scanning strategy, as expected for correlated $1/f$ noise, and we do not see any obvious signatures of either residual foregrounds in the Galactic plane, nor CMB dipole leakage at high latitudes. This suggests that the combination of the data model and processing masks described above performs reasonably well. We also note that the peak-to-peak values of the total correlated noise maps are $O(1\,\mu\mathrm{K})$, which is of the same order of magnitude as the predicted signal from cosmic reionization (Planck Collaboration IV 2020). Thus, correlated noise estimation plays a critical role for large-scale polarization reconstruction, while it is negligible for CMB temperature analysis.

While the 30 and 70 GHz maps visually appear to be isotropic and random, we do see signatures of excess striping in the 44 GHz channel, in particular in Stokes Q. Several $>1\,\mu$K stripes extend from the Eastern parts of the map, near the Galactic plane, through the Southern hemisphere. We have not yet been able to identify the origin of these stripes, but note that possible explanations include gain model errors for selected PID ranges, for instance in the form of one or more unmodelled amplifier gain jumps possibly triggered by thermal or electrical instabilities, or a sub-optimal processing mask. As reported by BeyondPlanck Collaboration (2020); Colombo et al. (2020); Paradiso et al. (2020), these artifacts lead to significant map-based χ^2 excesses in the final CMB polarization analysis, and thereby strongly limit our ability to constrain the optical depth of reionization with the current model. Our findings are consistent with

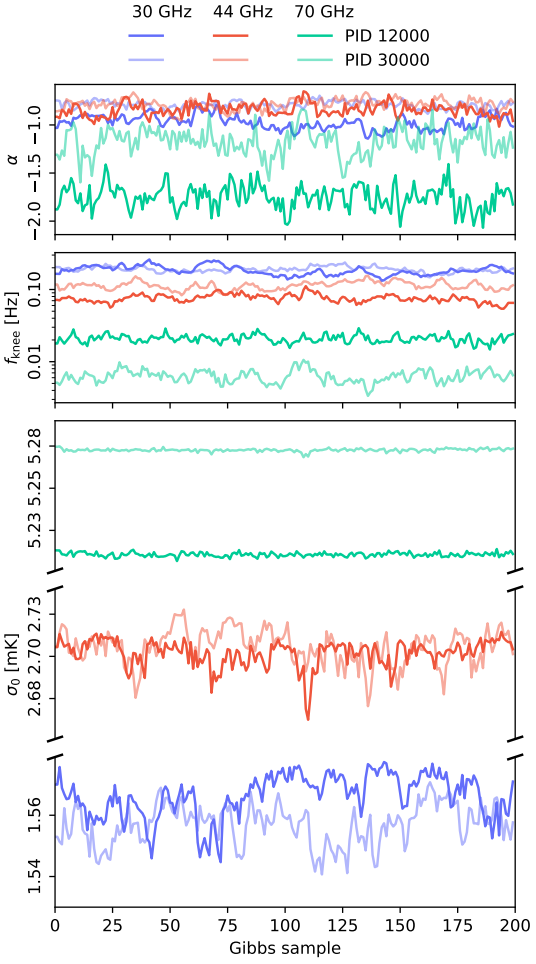

Fig. 10. Gibbs samples of noise parameters for two different PIDs for detectors 27M, 24S and 20M of the 30, 44 and 70 GHz bands respectively.

the difficulty experienced in treating the 44 GHz data for polarization analysis in all the official LFI DPC releases. However, the BeyondPlanck processing allows us to visually identify the structure of the excess correlated noise in the maps. Understanding the origin of these stripes and mitigating their impact represents a top priority for the next version of the BeyondPlanck processing.

For both $\boldsymbol{n}^{\rm corr}$ and ξ^n, the main result of the BeyondPlanck pipeline are the full ensembles of Gibbs samples. These are too large to visualize in their entirety here, and are instead provided digitally, including animations of the correlated noise maps as a function of iteration.[4] In the following, we will therefore focus on ξ^n, and as an example Fig. 10 displays one of the full Gibbs chains for two different PIDs for one radiometer from each LFI frequency band. We see that the Gibbs chains appear both stable and well-behaved. Some chains have longer Markov chain autocorrelation lengths than others, as expected from their different levels of degeneracies both within the noise model itself, and between the noise and the signal or gain. Overall, the Markov chain properties appear sound in all cases, with relatively short burn-in and correlation lengths. We conservatively remove the first 50 samples from each chain to account for burn-in.

The main results are shown in Figs. 11–16, which summarize the noise PSD parameters for each LFI radiometer in terms of distributions of posterior means (top section; histograms made from the posterior means for all PIDs) and as average quantities as a function of PID (bottom section). The former are useful to obtain a quick overview of the mean behavior of a given radiometer, while the latter is useful to study its evolution in time. Blue, red, and green correspond to 30, 44, and 70 GHz radiometers, respectively. Mean ξ^n values are tabulated in Table 1, while the average noise properties of all radiometers in each band are plotted as a function of time in Fig. 17.

Regarding mean values, we see that the 30 GHz radiometers generally have fairly high knee frequencies, $f_{\rm knee} \sim 100$ mHz, and shallow power law slopes, $\alpha \sim -0.85$. The 70 GHz channels, on the other hand, have lower knee frequencies, $f_{\rm knee} \sim 20$ mHz, and steeper slopes, $\alpha \sim -1.2$. The 44 GHz channels generally fall between these two extremes.

The dashed lines in Figs. 11–16 show the *Planck* LFI DPC values for each parameter (Planck Collaboration II 2020), which are assumed to be constant throughout the mission. Generally speaking, these agree well with the results presented here. The main exception is the 30 GHz white noise level, σ_0, for which we on average find 2 % lower values. It is difficult to precisely pinpoint the origin of these differences, but we do note that Galactic foregrounds are particularly bright at 30 GHz. One possible hypothesis is therefore that these are fitted slightly better in the joint and iterative BeyondPlanck approach, as compared to the linear pipeline DPC approach.

5.2. Time variability and goodness-of-fit

Perhaps the single most important and visually immediate conclusion to be drawn from these plots is the fact that the noise properties of the LFI instrument vary significantly in time. This is evident in all three frequency channels and all radiometers. Furthermore, by comparing the time evolution between different radiometers, we observe many common features, both between frequencies and, in particular, among radiometers within the same frequency. Many of these may be associated with specific and known changes in the thermal environment of the satellite, and can be traced using thermometer housekeeping data; this will be a main topic for the next section.

The bottom panels in Figs. 11–16 show a χ^2 per PID of the following form,

$$\chi^2 \equiv \frac{\sum_{i=1}^{n_{\rm samp}} \left(\frac{r_i}{\sigma_0}\right)^2 - n_{\rm samp}}{\sqrt{2n_{\rm samp}}} \tag{26}$$

where $n_{\rm samp}$ is the number of samples, and r_i is the residual for sample i as defined by Eq. (23). Thus, this quantity measures the normalized mean-subtracted χ^2 for each PID, which should, for ideal data and $n_{\rm samp} \gg 1$, be distributed according to a standard Gaussian distribution.

Starting with the 70 GHz channel, which generally is the most well-behaved, we see that the χ^2 fluctuates around zero for most channels, with a standard deviation of roughly unity.

[4] http://beyondplanck.science

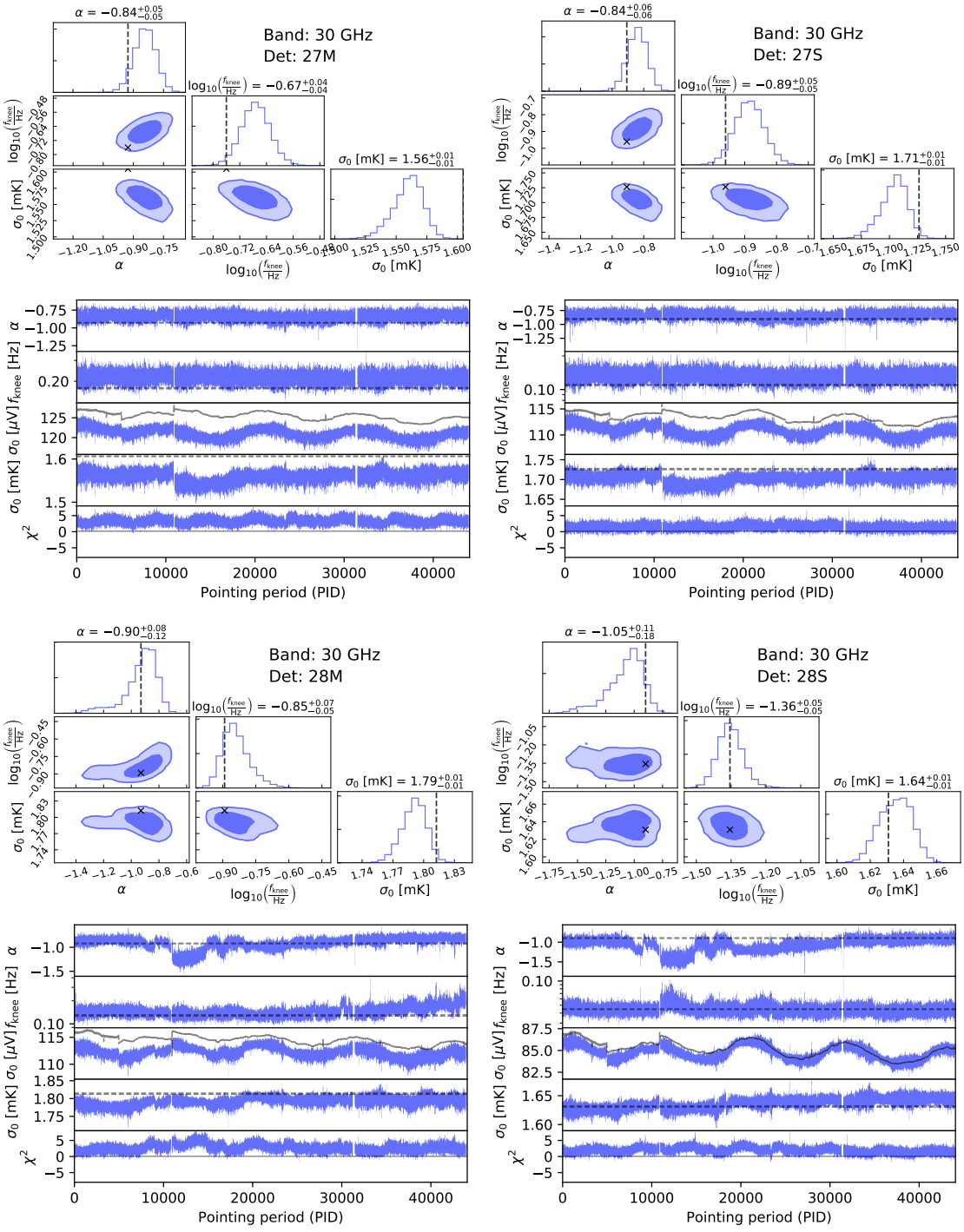

Fig. 11. Noise characterization of the *Planck* LFI 30 GHz radiometers; 27M (*top left*), 27S (*top right*); 28M (*bottom left*), and 28S (*bottom right*). For each radiometer, the top figure shows distributions of noise parameters PSD, $\xi^n = \{\sigma_0, f_{\rm knee}, \alpha\}$, averaged over all Gibbs samples for the full mission. The bottom figure shows the time evolution of the posterior mean of the noise parameters, and the bottom panel shows the evolution in reduced normalized χ^2 in units of σ. Black dashed curves and crosses show corresponding values as derived by, and used in, the official *Planck* LFI DPC pipeline.

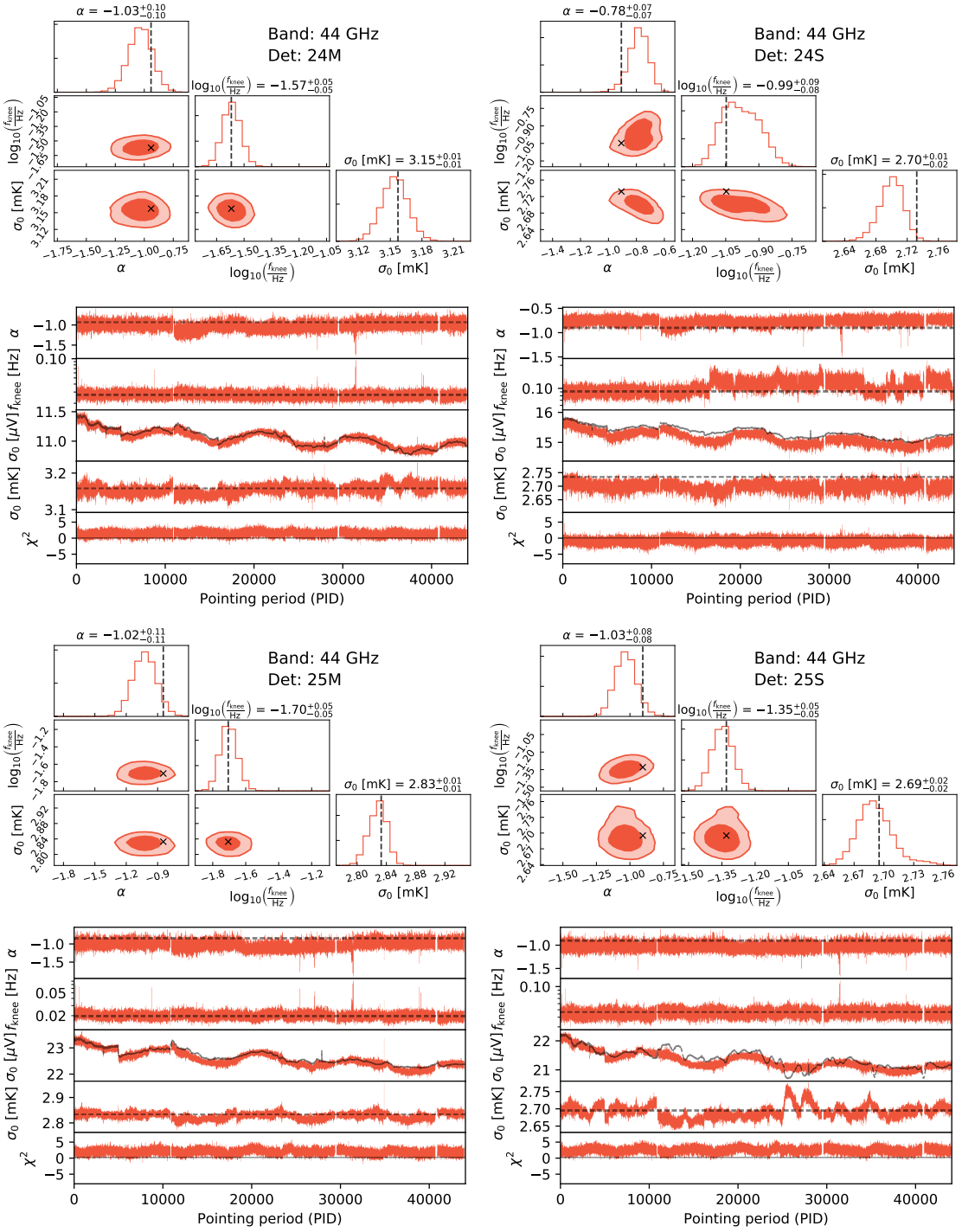

Fig. 12. Noise characterization of the *Planck* LFI 44 GHz radiometers; 24M (*top left*), 24S (*top right*); 25M (*bottom left*), and 25S (*bottom right*). For each radiometer, the top figure shows distributions of noise parameters PSD, $\xi^n = \{\sigma_0, f_{\rm knee}, \alpha\}$, averaged over all Gibbs samples for the full mission. The bottom figure shows the time evolution of the posterior mean of the noise parameters, and the bottom panel shows the evolution in reduced normalized χ^2 in units of σ. Black dashed curves and crosses show corresponding values as derived by, and used in, the official *Planck* LFI DPC pipeline.

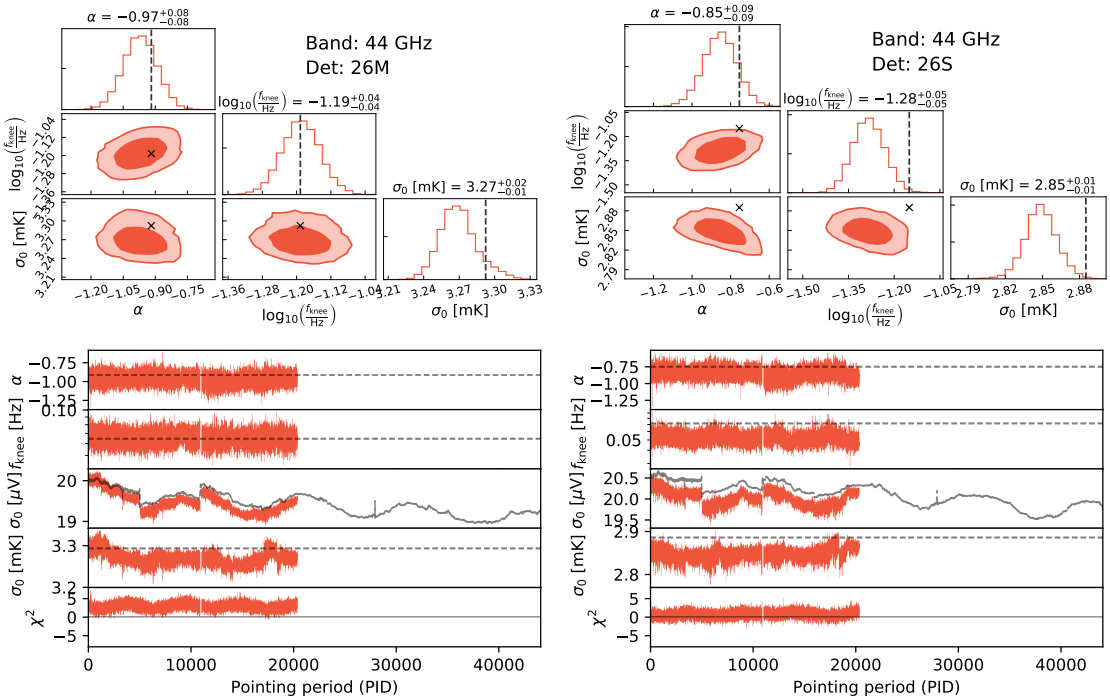

Fig. 13. Noise characterization of the *Planck* LFI 44 GHz radiometers; 26M (*left*), 26S (*right*). For each radiometer, the top figure shows distributions of noise parameters PSD, $\xi^n = \{\sigma_0, f_{\rm knee}, \alpha\}$, averaged over all Gibbs samples for the full mission. The bottom figure shows the time evolution of the posterior mean of the noise parameters, and the bottom panel shows the evolution in reduced normalized χ^2 in units of σ. Black dashed curves and crosses show corresponding values as derived by, and used in, the official *Planck* LFI DPC pipeline.

The most notable feature is a short gap in the 18M and 18S radiometer distributions. As discussed by Suur-Uski et al. (2020), the χ^2 for these two radiometers shows an excess in this range, and these data are therefore omitted from the main analysis.

This is also the case for the 26M and 26S 44 GHz radiometers, to a higher extent. This particular case is considered explicitly in the next section, where it is shown that the 26S radiometer appears unstable for the second half of the mission, with significantly higher χ^2 and $f_{\rm knee}$ values. These data are therefore also removed from the main BeyondPlanck analysis (Suur-Uski et al. 2020).

In general, the 30 and 44 GHz channels appear less stable than the 70 GHz channels in terms of overall χ^2. These are neither centered on zero, nor stationary in time, but rather tend to show statistically significant excesses of 2–3 σ per PID, with internal temporal variations at the 1 σ level. These excesses will be discussed in Sect. 6, but we note for now that they must be expected to significantly impact higher-level analyses, in particular large-scale CMB polarization analysis, as they strongly suggest that a simple $1/f$ noise model is incomplete. As such, the predicted noise bias will necessarily be underestimated. Consequently, establishing a more complete noise model for the 30 and 44 GHz channels is a top priority for a next-generation BeyondPlanck analysis.

As a typical illustration of such χ^2 failures, Fig. 18 shows the PSD for a range of 18 PIDs for the 28M 30 GHz radiometer. Here the $1/f$ model is not able to fit the real correlated noise to sufficient statistical accuracy at intermediate temporal frequencies, between 0.1 and 10 Hz, but rather shows a generally flatter trend. Similar behavior is seen in most 30 and 44 GHz radiometers, while the 70 GHz radiometers are better behaved, probably simply because of their lower $f_{\rm knee}$ values.

Turning our attention to the ξ^n parameters, we see even larger variability than in the χ^2. First, we note a period of significant instability in most channels between PIDs 8–20 000, but most strikingly in the 70 GHz α estimates. This feature will be discussed in more detail in Sect. 6, where it is explicitly shown to be correlated with thermal variations. We note, however, that the $1/f$ noise model seems flexible enough to adjust to these particular changes, as no associated excess χ^2 is observed in the same range.

Next, when considering the white noise level, σ_0, given in units of volts or kelvins, we see the pattern anticipated in the previous section. The uncalibrated white noise in units of volts follows the slow drifts of the gain, which typically manifests itself in slow annual gain oscillations. In contrast, the calibrated noise in units of $\mathrm{K_{CMB}}$ is far more stable. An important exception to this is 25S, which exhibits large variations in $\mathrm{K_{CMB}}$, in particular around PIDs 25–30 000. This might suggest a problem with the gain model for this particular radiometer, and this could possibly also be associated with the strong stripes in $\boldsymbol{n}^{\rm corr}$ noted above.

Other significant features include sharp jumps in $f_{\rm knee}$, for instance as seen in 24S. These typically coincide with external events, for instance during cooler maintainence (see Sect. 6 and Gjerløw et al. 2020).

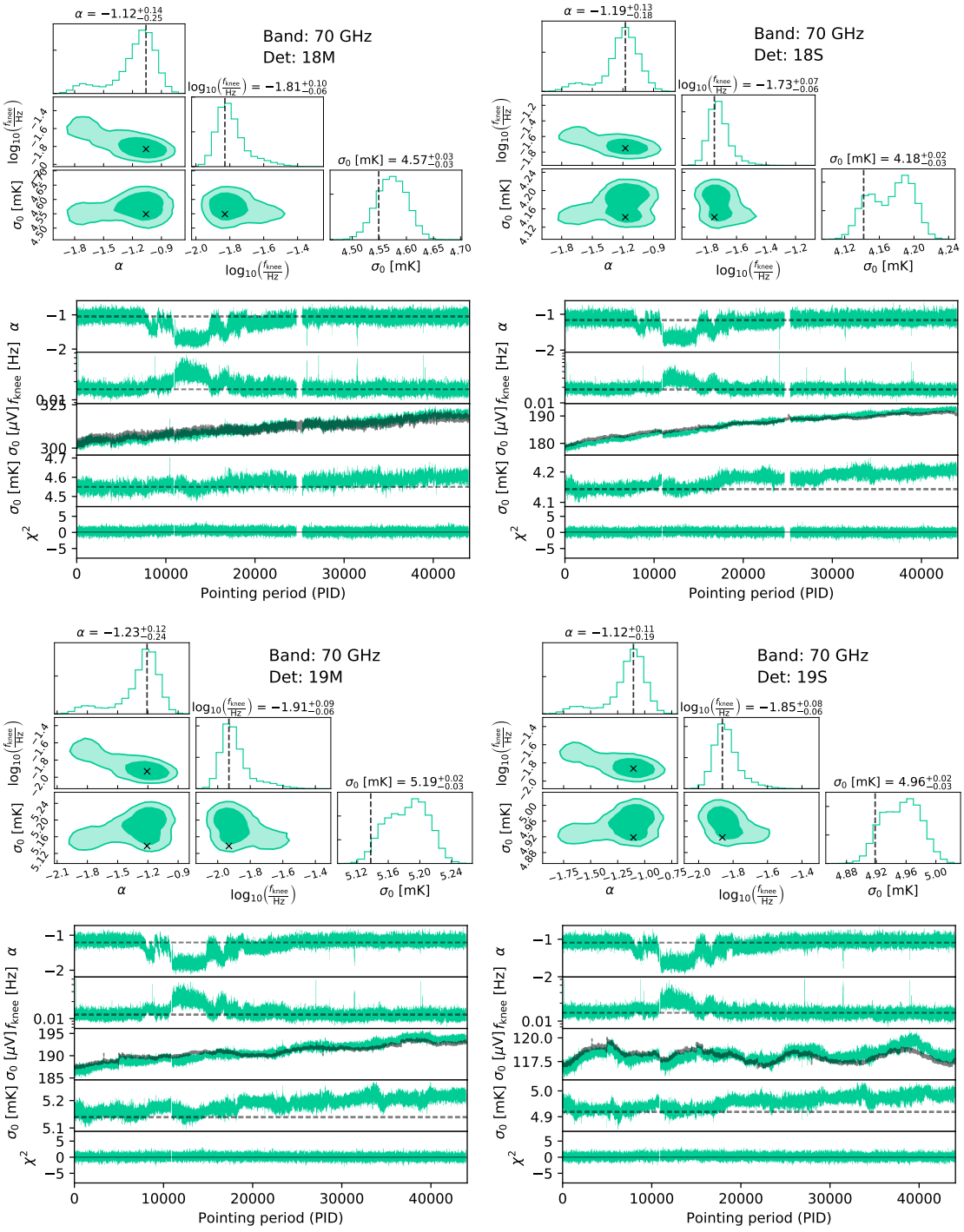

Fig. 14. Noise characterization of the *Planck* LFI 70 GHz radiometers; 18M (*top left*), 18S (*top right*); 19M (*bottom left*), and 19S (*bottom right*). For each radiometer, the top figure shows distributions of noise parameters PSD, $\xi^n = \{\sigma_0, f_{\rm knee}, \alpha\}$, averaged over all Gibbs samples for the full mission. The bottom figure shows the time evolution of the posterior mean of the noise parameters, and the bottom panel shows the evolution in reduced normalized χ^2 in units of σ. Black dashed curves and crosses show corresponding values as derived by, and used in, the official *Planck* LFI DPC pipeline.

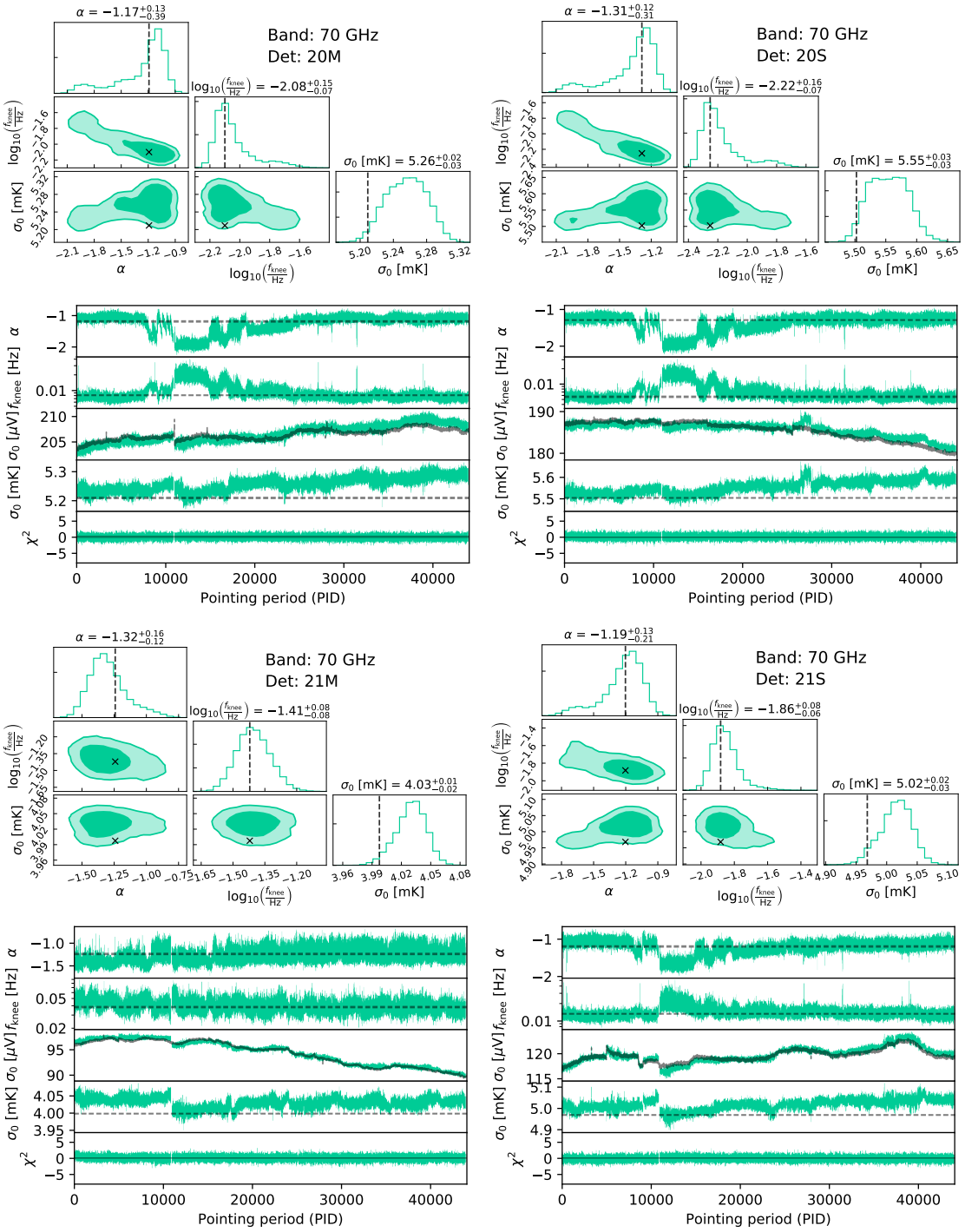

Fig. 15. Noise characterization of the *Planck* LFI 70 GHz radiometers; 20M (*top left*), 20S (*top right*); 21M (*bottom left*), and 21S (*bottom right*). For each radiometer, the top figure shows distributions of noise parameters PSD, $\xi^n = \{\sigma_0, f_{\rm knee}, \alpha\}$, averaged over all Gibbs samples for the full mission. The bottom figure shows the time evolution of the posterior mean of the noise parameters, and the bottom panel shows the evolution in reduced normalized χ^2 in units of σ. Black dashed curves and crosses show corresponding values as derived by, and used in, the official *Planck* LFI DPC pipeline.

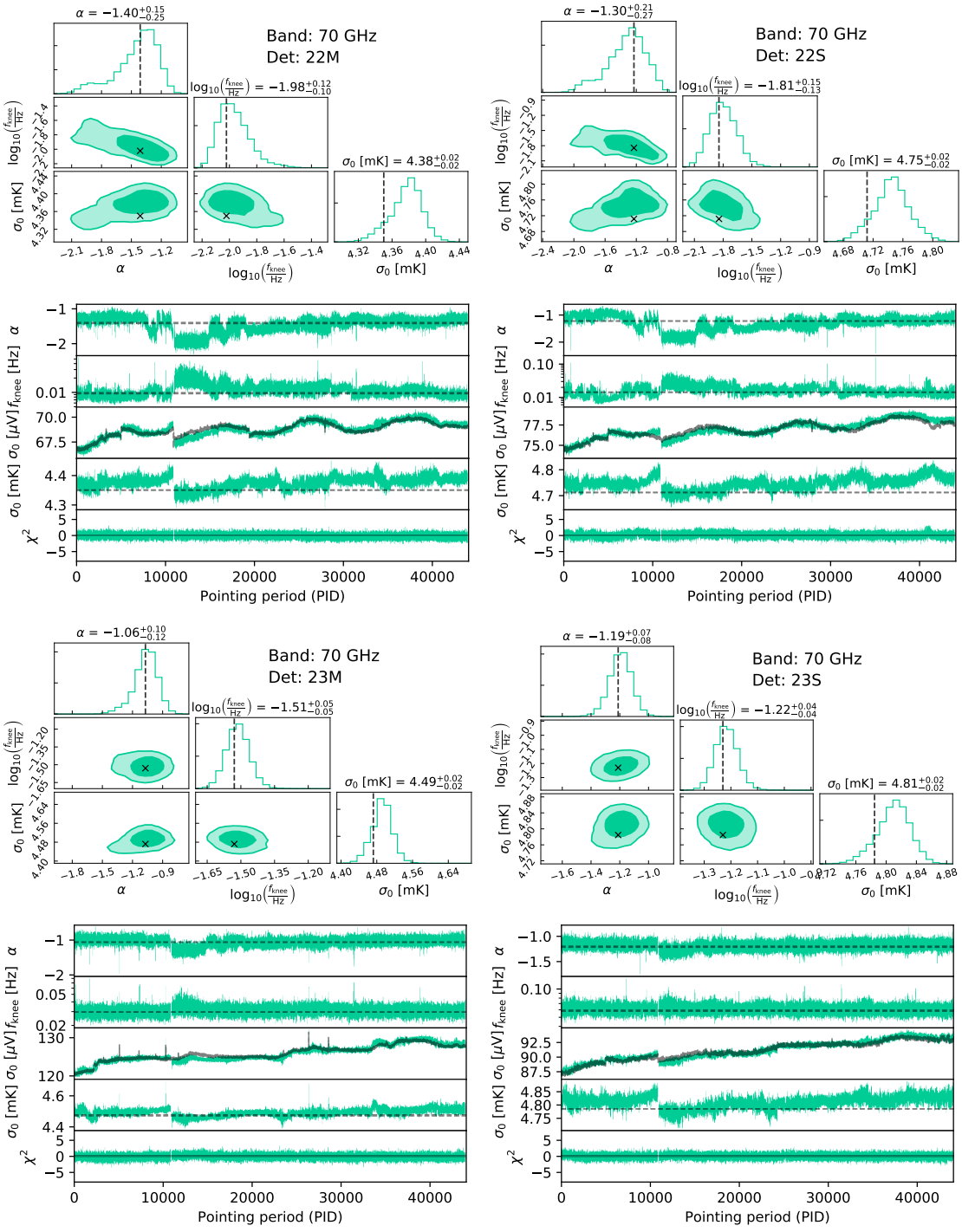

Fig. 16. Noise characterization of the *Planck* LFI 70 GHz radiometers; 22M (*top left*), 22S (*top right*); 23M (*bottom left*), and 23S (*bottom right*). For each radiometer, the top figure shows distributions of noise parameters PSD, $\xi^n = \{\sigma_0, f_{\rm knee}, \alpha\}$, averaged over all Gibbs samples for the full mission. The bottom figure shows the time evolution of the posterior mean of the noise parameters, and the bottom panel shows the evolution in reduced normalized χ^2 in units of σ. Black dashed curves and crosses show corresponding values as derived by, and used in, the official *Planck* LFI DPC pipeline.

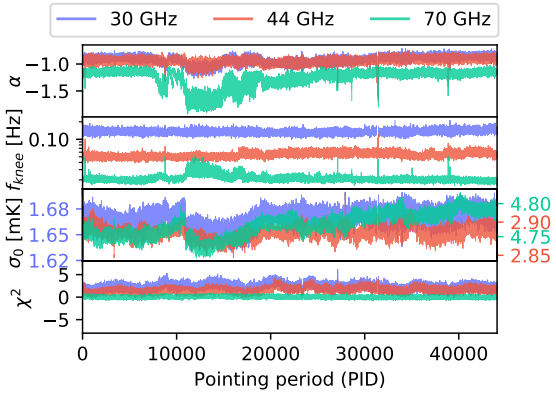

Fig. 17. Posterior mean noise parameters averaged over all radiometers in each band, for the full mission.

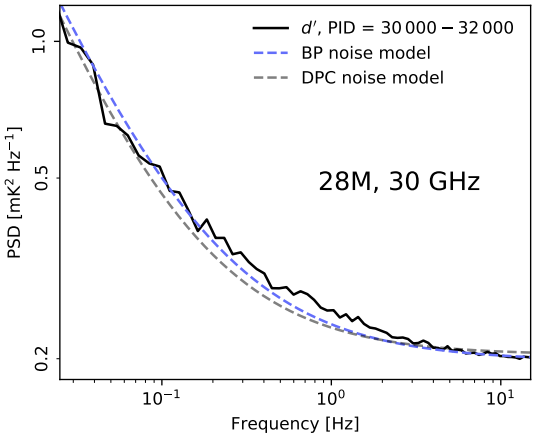

Fig. 18. PSD of signal-subtracted data from radiometer 28M, averaged over 18 PIDs (at intervals of 100 PIDs) in the range 30 000–32 000 (black). The dashed lines show the mean BeyondPlanck (dashed blue) and LFI DPC (dashed gray) noise models for the same data. An excess is seen at intermediate frequencies between 0.1 and 10 Hz.

6. Systematic effects

Previous LFI analyses have assumed a stationary noise model with three fixed parameters (σ_0 [K], $f_{\rm knee}$, and α) for each of the 22 radiometers. In contrast, each of these parameters is in BeyondPlanck estimated for every PID, increasing the total number of PSD noise parameters from 66 to about 3 million. This increase of information allows us to capture the effects of evolution in the radiometer responses and local thermal environment, as well as subtle interactions between them. In this section, we will use this new information to characterize potential residual systematic effects in the data, and, as far as possible, associate these with independent housekeeping data or known satellite events. An overview of the measurements from eight temperature sensors that are particularly important for LFI is provided in Fig. 19.

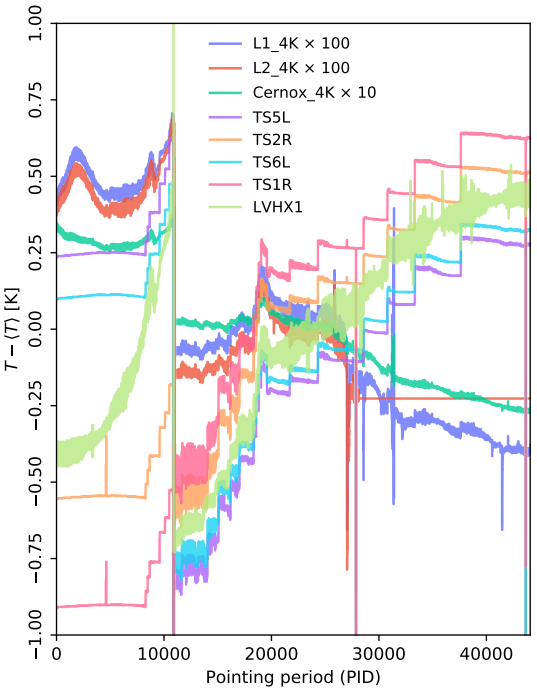

Fig. 19. High-level comparison of measurements from eight temperature sensors that are particularly relevant for LFI. Sensors TS5L, TS6L, TS1R, TS2R are installed in the 20 K focal plane frame, while LVHX1 is the liquid-vapour heat exchanger providing 18K to HFI; sensors L1_4K, L2_4K and Cernox_4K are on the HFI 4 K stage supporting the LFI 4 K reference loads. The step-like increases in the 20 K stage are visible both before and after the sorption cooler switchover event (near PID 11 000). For details on the locations of the various temperature sensors, see Fig. 21 of Bersanelli et al. (2010) and Fig. 18 of Lamarre et al. (2010). For visualization purposes, the mean value has been subtracted from each data set, and some have been scaled by one or two orders of magnitude, as indicated in the legend.

For details on the locations of the various temperature sensors, see Fig. 21 of Bersanelli et al. (2010) and Fig. 18 of Lamarre et al. (2010).

6.1. Temperature changes in the 20 K stage

A key element for the LFI thermal environment was the *Planck* sorption cooler system (SCS), which provided the 20 K stage to the LFI front-end and the 18K pre-cooling stage to HFI. The SCS included a nominal and a redundant unit (Planck Collaboration II 2011). In August 2010 (around PID 11 000), a heat switch of the nominal cooler unit reached its end-of-life, and the SCS was therefore switched over to the redundant cooler.[5] This "switchover" event implied a major redistribution of the temperatures in the LFI focal plane, with variations at ∼1 K level, for two main reasons. First, the efficiency of the newly active redundant cooler led to an overall decrease of the absolute tempera-

[5] This operation took place at PID 10911, corresponding to Operation Day (OD) 454.

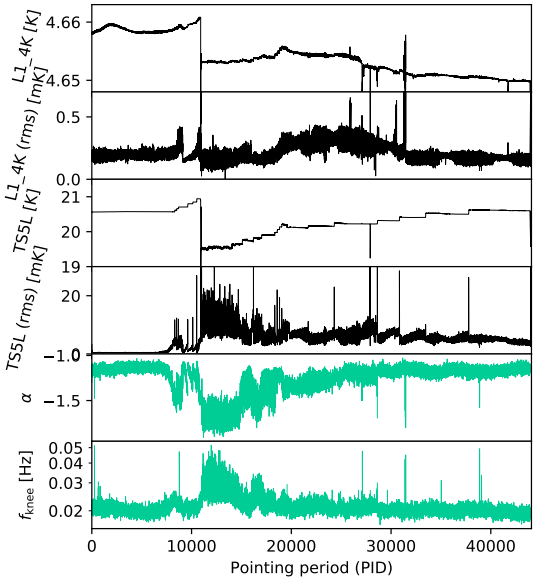

Fig. 20. Average correlated noise properties of the 70 GHz radiometers (bottom two panels) compared with 4 K and 20 K temperature sensor read-outs (top four panels) for the full mission.

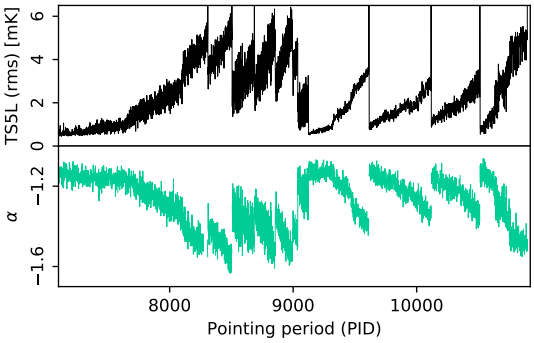

Fig. 21. Same as Fig. 20, but zoomed in on PIDs 7000–11 000.

ture. Second, because of the different location of the interface between the focal plane structure and the cold-end for the redundant cooler, a change of temperature gradients appeared across the focal plane.

Since the SCS dissipated significant power, changes in its configuration produced measurable thermal effects in the entire *Planck* spacecraft, and most directly in the 20 K stage. In the period preceding the switchover, starting around PID 8000, a series of power input adjustments were commanded to reduce thermal fluctuations in the 20 K stage while optimizing the sorption cooler lifetime, which generated a number of step-like increases in the LFI focal plane temperature. These are measured by all the LFI temperature sensors located in the 20 K focal plane unit, as shown in Fig. 19.

Following the switchover, in the period with PIDs 11–15 000, a significant increase of 20 K temperature fluctuations was observed. These excess fluctuations were understood as due to residual liquid hydrogen sloshing in the inactive cooler and affecting the cold-end temperature. The issue was resolved by heating the unit and letting the residual hydrogen evaporate. Afterwards, to optimize the performance and lifetime of the operating cooler, several periodic, step-like adjustments were again introduced in the operational parameters of the cooler. This resulted in a semi-gradual, monotonic increase of the LFI focal plane temperature from switchover to end of mission of ∼1.3 K.

In Fig. 19 the sudden discontinuity at switchover (PID 11 000) is visible for all temperature sensors, and the stepwise up-ward trend driven by SCS operational adjustments can be seen in all 20 K sensors. These temperature variations directly affected the LFI noise performance for most radiometers, as observed in the lower panels of Figs. 11–16. To see this, it may be useful to concentrate on a well-behaved case (e.g., radiometers 22 or 23, Fig. 16) and then recognize the same features in other radiometers.

The effect of the SCS switchover shows up as a sharp discontinuity also in the white noise levels near PID 11 000. The sudden decrease of the focal plane temperature of about 1 K implies a change in radiometer gain, as well as a genuine reduction in radiometer noise. This leads to a decrease not only of σ_0 [V] but also of σ_0 [K]. Furthermore, due to the change in cold-end interface, the temperature drop at switchover was larger on the top-right-hand side of the focal plane (as defined by the view in Fig. 6 of BeyondPlanck Collaboration 2020) than in other regions. In particuclar, we see in Figs. 11–16 that the drop in σ_0 [K] is particularly pronounced for radiometers 21, 22, 23, 27 (both M and S), which are all located in that portion of the focal plane.

Using again Fig. 16 as a guide, we can also recognize the effect of the incremental increase of focal plane temperature due to sorption cooler adjustments, both before and after switchover. The increasing physical temperature of the focal plane drives a corresponding increase of σ_0 [K], which is visible for most of the radiometers in Figs. 11–16. However, we cannot exclude that part of the observed slow increase of white noise is due to aging effects degrading the intrinsic noise performance of the front-end amplifiers. To disentangle these two components would require a more detailed thermal and radiometric model.

6.2. Temperature fluctuations and $1/f$ parameters

In Fig. 20 (top four panels) we report the value and rms of representative temperature sensor of the 4 K and 20 K stages (L1_4K and TS5L). During the thermal instability period that followed the switchover, the noise properties of essentially all the 70 GHz radiometers markedly changed their $1/f$ noise behavior (with the only notable exception of 21M). This is highlighted in the lower two panels of Fig. 20, which show the averaged values of α and $f_{\rm knee}$ for all the 70 GHz radiometers. The correlation between $1/f$ noise parameters and temperature fluctuations is excellent, with higher fluctuations producing an increase in $f_{\rm knee}$ and a steepening (i.e., more negative) slope α. The latter is a typical behavior of thermally driven instabilities, which tend to transfer more power to low frequencies, and thus steepen the $1/f$ tail. This behavior shows up also in the individual 70 GHz radiometers (Figs. 14–16).

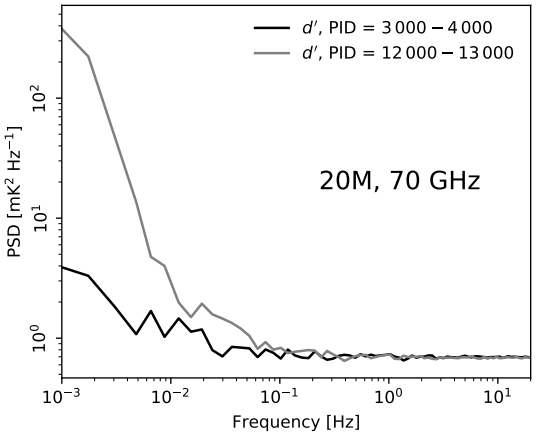

Fig. 22. PSD of signal subtracted data from radiometer 20M, averaged over 10 PIDs in the ranges 3000–4000 (black) and 12 000–13 000 (grey).

Figure 21 is a zoom into the pre-switchover period (PID 7000–11 000) of the upper plot. Here we see the effect of some of the step-wise adjustments in the sorption cooler operation, whose main effect is to temporarily reduce the temperature fluctuations. The observed tight correlation with the steepening of the slope is striking.

For a specific example of how noise property variations modify the noise PSD, Fig. 22 shows the average PSD for 10 PIDs between 3000–4000 (black) compared to 10 PIDs between 12 000–13 000 (grey) for the 70 GHz 20M radiometer. We see large increases in power at low frequencies, and a shift in the knee frequency.

These correlations appear more weakly in the 30 and 44 GHz radiometers (see Fig. 17). In particular, there is no correlation with the knee frequency. This behavior could be partly explained by the fact that, by mechanical design, the front-end modules (FEMs) of the 30 and 44 GHz are less thermally coupled to the frame and cooler front-end; or it could be indicative of an additional source of non-thermal correlated noise that dominates the slope and knee frequency of these channels. This could be the case also for the 70 GHz radiometer 21M, for which the lack of correlation cannot be explained in terms of poor thermal coupling.

These hypotheses are supported by Fig. 18, which compared the PSD of the 30 GHz 28M signal-subtracted data, averaged over 18 PIDs in a typical stable period, with both the Beyond-Planck and LFI DPC noise models for the same period. None of the models are able to properly describe the observed data. The deviation indicates that there is an excess power in the frequency range between 0.1 and 5 Hz.

6.3. Seasonal effects and slow drifts

The changing Sun-satellite distance during the yearly *Planck* orbit around the Sun produced a seasonal modulation of the solar power absorbed by the spacecraft. The corresponding effect on the LFI thermal environment was negligible for the actively-controlled front end, as demonstrated by the lack of yearly modulation in the 20 K temperature sensors (see Fig. 19 and upper panel of Fig. 20). However, the 300 K environment and the passive cooling elements (V-groove radiators) were affected by a ∼1 % seasonal modulation (see Fig. 6 of Planck Collaboration I 2014).

Since the radiometer back-end modules (BEMs) provided a major contribution to the radiometer gain g, and these are located in the 300 K service module (SVM), the thermal susceptibility of the BEMs coupled with local thermal changes is expected to induce radiometer gain variations. On the other hand, since the BEMs are downstream relative to the >30 dB amplification from the FEMs, their contribution to the noise temperature, $T_{\rm sys}$, is negligible. Therefore we may expect the LFI uncalibrated signal (and the uncalibrated noise σ_0 [V]) to show a seasonal modulation due to thermally-driven BEM gain variations, with essentially no degeneracy with $T_{\rm sys}$.

Figures 11–16 show that several LFI radiometers exhibit such modulation in the uncalibrated white noise, σ_0 [V], throughout the four year survey. For all of these, the modulation disappears in σ_0 [K], indicating that our gain model properly captures this effect. We also observe that the sign of the modulation is opposite for the 70 GHz and the 30–44 GHz radiometers. Furthermore, all radiometers that exhibit seasonal modulation also show a systematic slow drift of σ_0 [V] throughout the mission with the same sign as the initial modulation (which corresponds to increasing physical temperatures in the SVM). Since the spacecraft housekeeping recorded a slow overall increase in temperature throughout the mission ($\Delta T \approx 5$ K), the observed drift of σ_0 [V] is qualitatively consistent with the hypothesis of BEM susceptibility as the origin of the effect.

For each radiometer, the amplitude of the modulation depends on the details of the thermal susceptibility of the LFI elements down-stream relative to the third V-groove, including waveguide losses, BEM components, particularly low-noise amplifiers (LNAs), detector diodes, data acquisition electronics (gain and offset), etc. The dominant element is the BEM, whose thermal susceptibility was measured in the LFI pre-launch test campaign for the 30 and 44 GHz radiometers (Villa et al. 2010). The change in BEM output voltage, $\Delta V_{\rm out}$, as a function of the variation in BEM physical temperature, $\Delta T_{\rm BEM}$, can be written as

$$\Delta V_{\rm out} \propto \phi_{\rm BEM} \Delta T_{\rm BEM} \left(T_{\rm sys} + T_{\rm in} \right), \tag{27}$$

where $T_{\rm in}$ is the input signal temperature (either sky or reference load) and $\phi_{\rm BEM}$ is a transfer function quantifying the BEM thermal susceptibility. The measured values of $\phi_{\rm BEM}$ (Villa et al. 2010) were slightly negative for all the 30 and 44 GHz radiometers, ranging from −0.01 to −0.02, and this is consistent with both the observed overall drift and the seasonal effect. No such ground tests could be done for the 70 GHz instrument. However, in-flight tests during commissioning (Cuttaia & Terenzi 2011) revealed that the sign of $\phi_{\rm BEM}$ for the 70 GHz radiometers was opposite to those of 30 and 44 GHz, which is consistent with our interpretation.

6.4. Inter-radiometer correlations

So far, we have mostly considered noise properties as measured separately for each radiometer. However, given the significant sensitivity to external environment parameters discussed above, it is also interesting to quantify correlations between detectors. As a first measure of this, we plot in Fig. 23 the cross-correlation of $\boldsymbol{n}_{\rm corr}$ averaged over all pairs of radiometers within each frequency band as a function of PID. As expected from the previous

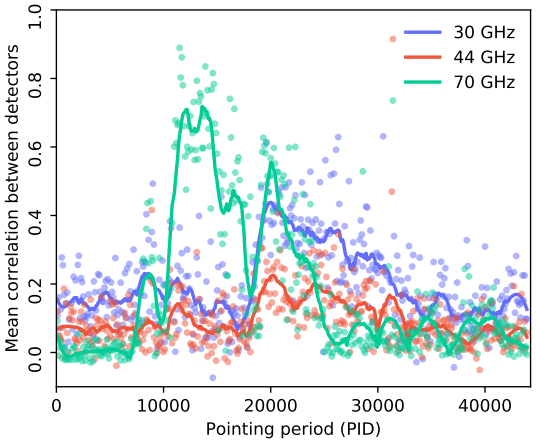

Fig. 23. Average correlation between the correlated noise in the different radiometers in each frequency band for the full mission.

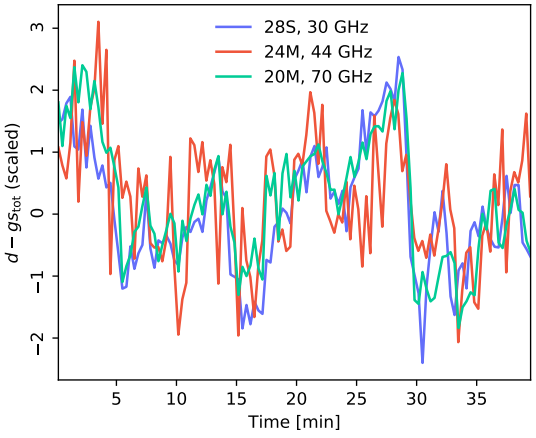

Fig. 24. Signal subtracted data from radiometers from all three bands for PID 12 301. The data is averaged over a 20 second timescale and scaled to fit in the same plot.

discussion, we find a large common correlation for the 70 GHz channel that peaks in the post-switchover period. Similar coherent patterns are seen in the 30 and 44 GHz channels, but at somewhat lower levels.

As a specific example of such common mode noise, Fig. 24 shows the signal subtracted timestreams for one radiometer from each band for PID 12 301, which is representative for the period of maximum correlation. Here we see that the same large scale fluctuations are present in all three bands. In Fig. 25 we show the average cross correlation between time streams of all 70 GHz radiometers for the same PID. We compare the average correlation between the correlated noise components, $\boldsymbol{n}_{\rm corr}$, with the correlation between the residuals, $\boldsymbol{d}' - \boldsymbol{n}_{\rm corr}$. We see that even though the correlations between the $\boldsymbol{n}_{\rm corr}$ components are large, the residuals are highly uncorrelated. This is an indication

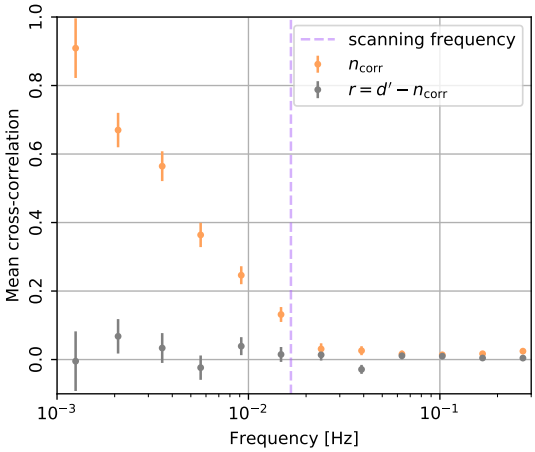

Fig. 25. Average cross correlation between timestreams of all 70 GHz radiometers for PID 12301. Orange points show average correlation between the correlated noise components, while the grey points shows average correlation between the residuals. We see that even though the correlation between the correlated noise components is large, the residuals are completely uncorrelated, indicating that this correlated signal does not leak into the rest of the pipeline, but is all incorporated into the correlated noise.

that the common mode signal is efficiently described by $\boldsymbol{n}_{\rm corr}$, and it therefore does not leak into the rest of the BeyondPlanck pipeline.

Figure 26 shows a global correlation matrix of all the noise parameters for all the LFI radiometers throughout the mission. A number of interesting features can be recognized in this diagram:

1. We note that all 70 GHz radiometers exhibit an internally coherent trend, where $f_{\rm knee}$ and α behave essentially as a common mode for the entire 70 GHz array, with the only exception being 21M. This coherent behavior reflects the common thermal origin of the $1/f$ noise of the 70 GHz radiometers, as discussed in Sect. 6.2. We also see that σ_0 [K] shows a similar common mode behavior for the 70 GHz radiometers and, to a lesser extent, it correlates also with the σ_0 [K] of the 44 and 30 GHz radiometers. This is indicative of the fact that changes in the LFI radiometers' sensitivity are driven by the global LFI thermal environment, most importantly by the slow increase in temperature at the 20 K temperature stage.
2. For 30 and 44 GHz we do not observe the same common mode behavior for $f_{\rm knee}$ and α as for the 70 GHz. Rather, we see positive correlation (red pixels in Fig. 26) between $f_{\rm knee}$ and α within each single radiometer. This suggests that (a) the dominant source of $1/f$ noise is independent for each 30 and 44 GHz radiometers, and (b) for a given radiometer, as $f_{\rm knee}$ increases, the slope becomes flatter (i.e., α becomes less negative). This behavior further supports the hypothesis that the dominant source of correlated noise in the 30 and 44 GHz is not of thermal origin.
3. Finally, we observe an anti-correlation between $f_{\rm knee}$ and σ_0 (as a common mode at 70 GHz and individually for 30 and 44 GHz). Slightly larger values of $f_{\rm knee}$ for lower σ_0 can be understood in terms of the correlated fluctuations becoming subdominant near $f_{\rm knee}$ when the white noise increases during the mission time.

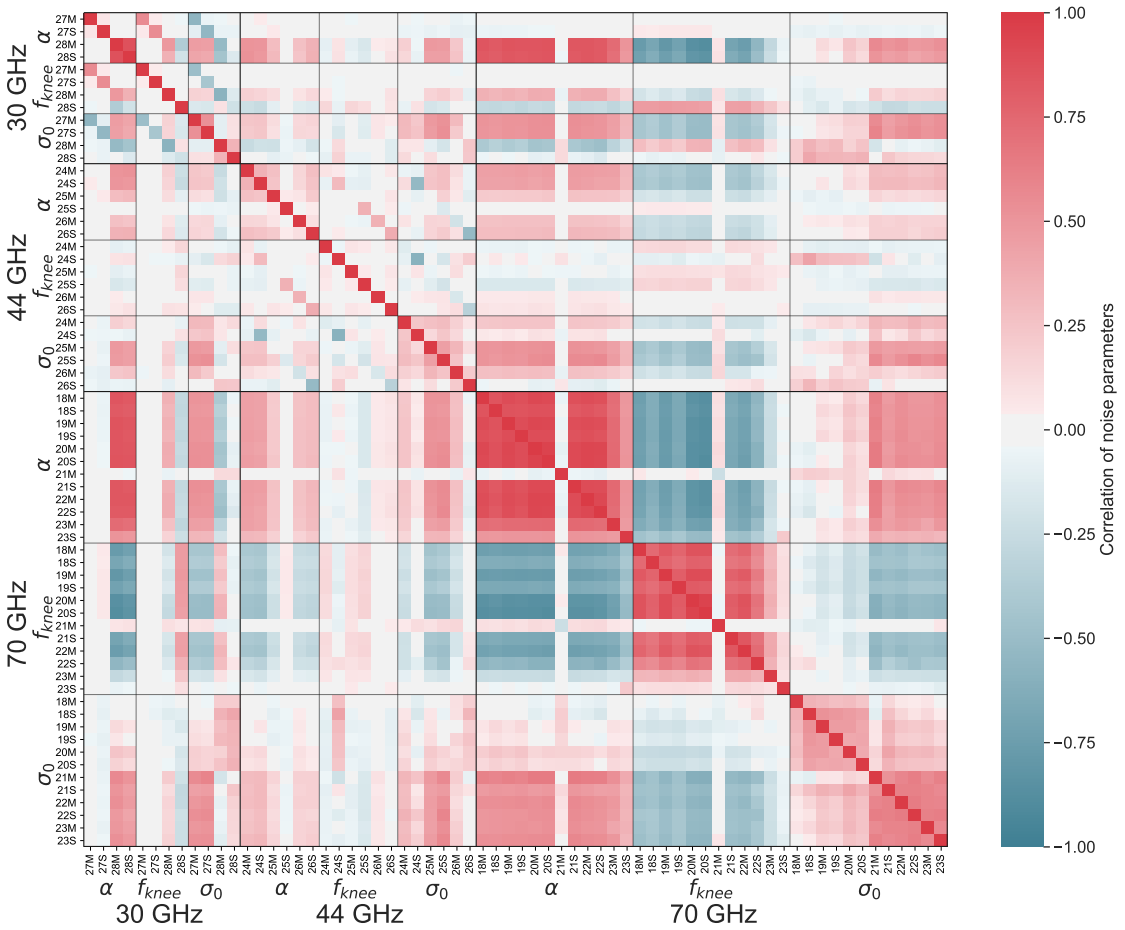

Fig. 26. Average over all Gibbs samples of the noise parameters, $\xi^n = \{\sigma_0, f_{\rm knee}, \alpha\}$, for each PID. We then find the correlation in time between these averages for the different bands and detectors. The results here are for the calibrated white noise level, σ_0 [K].

6.5. Correlation with housekeeping data

Next, we correlate the LFI noise parameters with housekeeping data, and in particular with temperature sensor that are relevant for LFI. This is summarized in Fig. 27, showing the correlation coefficients with respect to several sensors that monitor the 20 K stage (TS5L, TS2R, TS6L, TS1R, LVHX1) and the 4 K stage (L1_4K, L2_4K, Cernox_4K). Some significant patterns appear that can be interpreted in terms of the general instrument behavior:

1. For the 70 GHz radiometers, both the rms and the peak-to-peak of the 20 K temperature sensor fluctuations correlate with $f_{\rm knee}$ and anti-correlate with α (i.e., they prefer a steeper power-law slope). This is particularly evident for the TS5L and TS6R sensors, which are located nearby the 70 GHz array. This indicates that the $1/f$ noise of the 70 GHz radiometers is dominated by residual thermal fluctuations in the 20 K stage. A similar trend can be seen also at 30 GHz in the two horn-coupled receivers 28M and 28S. However, the 44 GHz channels show no sign of this behavior. Combined with the lack of correlation with the 4 K sensors, this is consistent with the hypothesis that the $1/f$ noise of the 44 GHz (and partly the 30 GHz) radiometers is dominated by non-thermal fluctuations.
2. Weaker correlations are seen between the various noise parameters and the 4 K temperature sensors. The lack of significant correlation of the rms and peak-to-peak of 4 K sensors with any of the $1/f$ parameters, $f_{\rm knee}$ and α, is an indication that the 4 K reference loads do not contribute significantly to the radiometers correlated noise.
3. A strong anti-correlation (correlation) of the gain g with the absolute value of the 20 K sensors for the 70 GHz (30–44 GHz) radiometers is observed. Based on the discussion in Sect. 6.3, this pattern can be understood by noting that the 20 K stage temperature systematically increased throughout the mission, driven by sorption cooler adjustments. The same monotonic trend was also on-going in the 300 K stage, which controls the BEM amplifiers. This is thus a spurious

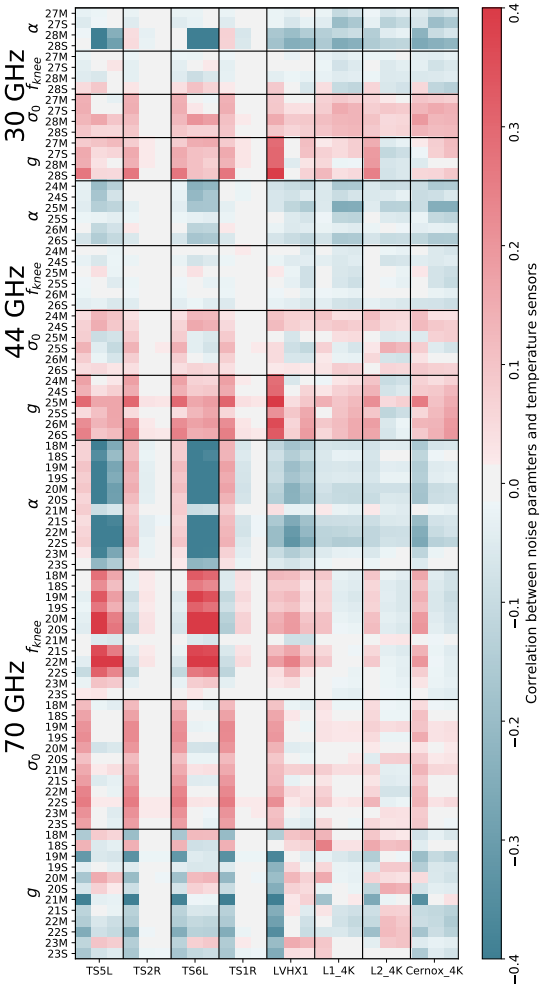

Fig. 27. Correlation in time, for the complete mission, between noise parameters and the temperature sensors. For each sensor we show the results (from left to right) from the mean temperature, the temperature rms and the peak-to-peak temperature of each sensor within each pointing period. The results here are for the calibrated white noise level, σ_0 [K]. We have imposed a mild highpass-filter in time of the different datasets in order to avoid random correlations on the very longest timescales.

correlation, for which the increasing back-end temperature actually leads to lower (higher) values of g for the 70 GHz (30–44 GHz) radiometers.

6.6. Issues with individual radiometers

In addition to the overall behavior and correlations that are common to many or most radiometers, there are issues that only seem to affect individual radiometers. Here we point out two special cases, namely 26S and 21M.

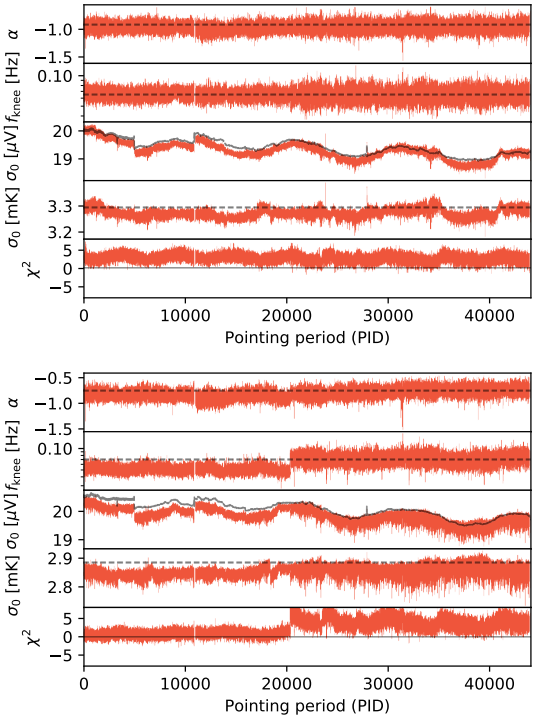

Fig. 28. Evolution of posterior mean noise parameters for radiometers 26M (top) and 26S (bottom). Black curves show corresponding values as derived by, and used in, the official *Planck* LFI DPC pipeline. Note here that the PIDs after about PID 20 000 were not included in the main BeyondPlanck run, so the results there are from a much shorter earlier test run, and are hence more noisy.

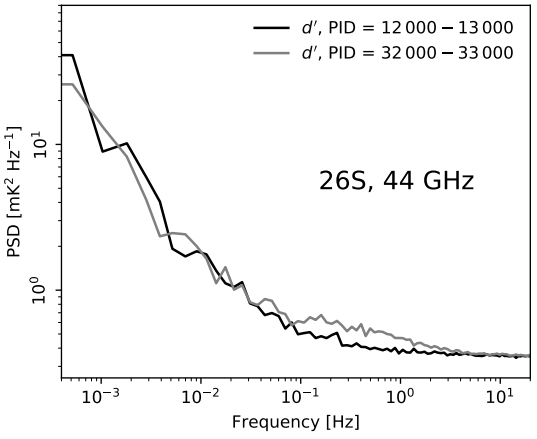

Fig. 29. PSD of signal subtracted data from radiometer 26S, averaged over 10 PIDs (at intervals of 100 PIDs) in the ranges 12 000–13 000 (black) and 32 000–33 000 (grey). We see that there is significantly more power in the frequency range 10^{-1}–10^{1} Hz in the later period.

First, as discussed above, we often find excess noise power in the 30 and 44 GHz channels in the signal-subtracted data at intermediate frequencies, ~0.1–5 Hz. This excess cannot be described with a $1/f$ noise model, and leads to high χ^2 values. The most extreme example of this is the 44 GHz 26S radiometer, as shown in Fig. 28. This figure is identical to the bottom panel Fig. 13, except that masked PIDs are now not omitted. Here we see a jump in $f_{\rm knee}$ around PID 20 800, after which the χ^2 is consistently high with a mean of about $5\,\sigma$. This is elucidated in Fig. 29, which compares the noise PSD averaged over 10 PIDs in the 12 000–13 000 range with a corresponding average evaluated in the 32 000–33 000 range. We see an excess in power at intermediate frequencies that is not possible to fit with the $1/f$ noise model. Considering that the *Planck* spin period is 60 s, temporal frequencies of 0.1–1 Hz correspond to angular scales of 6–60° on the sky. This unmodelled noise therefore represents a significant contaminant with respect to large-scale CMB polarization reconstruction, which is one of the main scientific targets for the current BeyondPlanck analysis. We therefore choose to exclude all data from 26S after after PID 20 800 (Suur-Uski et al. 2020). To avoid excessive temperature-to-polarization leakage, we also exclude 26M data for the same region. In total, this represents 17 % of the data full 44 GHz data volume, and the BeyondPlanck 44 GHz frequency map therefore has a higher white noise level than the corresponding official *Planck* products; but with a more complete noise description.

The sudden degradation of 26S at around PID 21 000 has no simultaneous counterpart in any other LFI radiometer, including the coupled 26M which exhibits a normal behaviour (see Fig. 28). This suggests a singular event within the 26S itself, or in the bias circuits serving its RF components. Since we do not observe significant changes in the radiometer output signal level and no anomalies are seen in the LNAs currents, it is unlikely that the problem resides in the HEMT amplifiers. A more plausible cause would be a degradation of the phase switch performance, possibly due to ageing, instability of the input currents, or loss of internal tuning balance (Mennella et al. 2010; Cuttaia et al. 2009). Indeed sub-optimal operation of the phase switches would not significantly change the signal output level, but is known to introduce excess $1/f$ noise, as verified during the ground testing and in-flight commissioning phase.

The second anomalous case is the 70 GHz 21M radiometer. While the noise properties of the other 70 GHz channels are internally significantly correlated, this particular channel does not show similar correlations. The reason for the different behavior of 21M is still not fully understood. However, as shown for PID 2201 in Fig. 30, this particular radiometer exhibits a typical "popcorn" or "random telegraph" noise, i.e., a white noise jumping between two different offset states. During ground testing this behavior was noted in the undifferenced data of LFI21 and LFI23 and ascribed to bimodal instability in the detector diodes. The effect was then recognized in-flight and this prevented proper correction of ADC nonlinearity effect (Planck Collaboration III 2014). However, because the timescale of diode jumps between states (typically a few minutes) is longer than the differencing between sky and reference load (0.25 ms, corresponding to the phase switch frequency of 4 kHz), the effect is efficiently removed in the differenced data. In the current analysis, we actually observe popcorn behavior in the differenced data, suggesting either an increased instability of the affected diode in 21M (possibly due to aging), or a different origin of the effect. Popcorn noise has been also found in some HFI channels (Planck HFI Core Team 2011). We have not seen any sign of popcorn noise in any of the other LFI channels besides 21M,

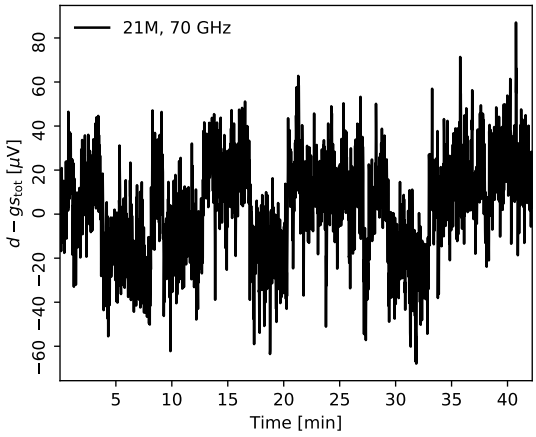

Fig. 30. Signal subtracted data from radiometer 21M for PID 2201. The data are averaged over a one second timescale.

but we have also not performed a deep dedicated search for it. However, the fact that the χ^2 distribution for channel 21M appears acceptable suggests that this effect, even if surviving in the differenced data stream, happens at a sufficiently long timescale that $\boldsymbol{n}_{\rm corr}$ is able to absorb it, preventing it from leaking into other astrophysical components.

7. Conclusions

This paper has two main goals. First, it aims to describe Bayesian noise estimation within a global CMB analysis framework (BeyondPlanck Collaboration 2020). As such, this work represents the first real-world application and demonstration of methods originally introduced by Wehus et al. (2012), while at the same time taking advantage of important numerical improvements introduced by Keihänen et al. (2020). The second main goal is to apply this method to the *Planck* LFI measurements to characterize their noise properties at a more fine-grained level than done previously (Planck Collaboration II 2020).

An important question regarding the original work of Wehus et al. (2012) was whether the method would be practical for real-world observations, or whether it was too computationally expensive to be useful in a real pipeline. We are now in a position to conclusively answer this question: As summarized by (BeyondPlanck Collaboration 2020; Galloway et al. 2020a), the noise estimation step in the BeyondPlanck pipeline accounts for 38 % of the total runtime, or 48 CPU-hours per sample for the 70 GHz channel, most of which is spent Fourier transforming the raw time-ordered data. As such, exact Bayesian noise estimation certainly is an expensive pipeline component—but it is by no means not prohibitive. Additionally, it is important to note that Bayesian correlated noise sampling replaces both traditional mapmaking and noise covariance matrix evaluations (Keihänen et al. 2020; Suur-Uski et al. 2020), which are two of the most expensive procedures in a traditional CMB analysis pipeline (Planck Collaboration I 2020), and, in fact, this leads to lower computational requirements overall. As an example, we note that a full BeyondPlanck Gibbs sample (which includes both low- and high-level processing with all LFI channels) costs 163 CPU-hours, while producing a single component of the full

Planck Full Focal Plane (FFP) simulation of the 70 GHz channel costs 9360 CPU-hours. Likewise, we also note that the current BeyondPlanck analysis was run on an in-house cluster with 416 cores and 9 TB of RAM, while the *Planck* simulations were produced on a large national computing center with $O(10^5)$ cores (Planck Collaboration I 2020; Planck Collaboration Int. LVII 2020). We believe that the computational speed of this method alone should make it an attractive option for other CMB experiments, not to mention the possibilities of performing joint exact Bayesian analysis.

As far as LFI-specific results are concerned, the current results point toward generally complex noise behaviour with subtle contributions from origins that have not yet been fully accounted for. Most notably, the noise properties of each LFI radiometer vary significantly in time, and depend sensitively on the thermal environment of the instrument. For the 70 GHz channel, for which the correlated noise amplitude (and knee frequency) is generally low, most of these variations may be described in terms of a simple $1/f$ model with time-dependent parameters. With very few exceptions, the time-domain χ^2 of this channel is statistically acceptable throughout the mission.

However, for the 30 and 44 GHz channels a more complex picture has emerged. For these, the χ^2 is generally high by 2–3 σ per PID, and with significant variations as a function of time. Multiple observations suggest a yet undetected source of non-thermal correlated noise in the 44 GHz (and at a lower level in the 30 GHz radiometers) that is responsible for mild, and possibly time-varying, deviations from the simple $1/f$ model. Inspection of individual PIDs indicates the presence of excess power between 0.1 and 5 Hz, well above the *Planck* scanning frequency of 0.017 Hz, thereby affecting the angular scales that are relevant for large-scale CMB polarization science. Our analysis suggests that these effects are not due to temperature fluctuations, but rather associated with other effects, such as electrical instabilities or other environmental issues. We have carried out a preliminary investigation by correlating the LFI radiometers whose LNA bias were supplied by common electronics groups,[6] but we have found no compelling evidence of correlations or anomalies. Many other electrical effects must be studied by exploiting all the available housekeeping information. Most of the spikes in the rms of the temperature sensors (see, e.g., Fig. 20) are readily understood as due to commanded cooler adjustments, but a few of them deserve further investigation. The influence of transient perturbations should also be systematically investigated, including the possible effect of cosmic rays and solar flares.

A complete and quantitative analysis will require a detailed thermal model of the full instrument that includes the back-end unit and interfaces with the V-grooves, coupled with thermal susceptibility parameters of the relevant components (LNAs, OMTs, waveguides, BEMs, detector diodes, data acquisition electronics). Such a detailed study is beyond the scope of this work, but this is now made possible through the present study, and it will become a primary target for future BeyondPlanck LFI analysis.

A separate issue appears to be associated with the gain model, which for some radiometers exhibits larger temporal variations than might be expected from housekeeping data. This might be connected with a number of strong stripes in the 44 GHz correlated noise map, which affects large fractions of the sky at the 1 μK level, which are highly relevant for large-scale CMB polarization reconstruction. Of course, problems with the 44 GHz channel have been reported by the *Planck* team ever since the first data release (Planck Collaboration VI 2014; Tauber et al. 2019; Planck Collaboration II 2020), and it was therefore omitted from the main *Planck* CMB polarization analysis (Planck Collaboration IV 2020). The current method and results do not yet resolve these problems, but hopefully they shine new light on the issue that might help resolve it through further modelling. In particular, understanding and mitigating the χ^2 excesses at 30 and 44 GHz (possibly by generalizing the $1/f$ noise model) and the stripes at 44 GHz (possibly through more robust gain modelling) should be a top priority for future BeyondPlanck analyses.

In general, the detailed BeyondPlanck modelling approach allows us to highlight a number of subtle systematic patterns in the LFI radiometers that were already noted and reported in previous analyses, but only now, for the first time, have been possible to elucidate and understand in greater detail. Examples are a detailed characterization of the nature of seasonal modulations and long term drifts, and correlations between instrument noise parameters with temperature sensor read-out information. These methods are likely to play a central role in the analysis of future high-sensitivity CMB *B*-mode experiments, for instance *LiteBIRD* (Sugai et al. 2020).

Acknowledgements. We thank Prof. Pedro Ferreira for useful suggestions, comments and discussions, and Dr. Diana Mjaschkova-Pascual for administrative support. We also thank the entire *Planck* and *WMAP* teams for invaluable support and discussions, and for their dedicated efforts through several decades without which this work would not be possible. The current work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement numbers 776282 (COMPET-4; BeyondPlanck), 772253 (ERC; bits2cosmology), and 819478 (ERC; Cosmoglobe). In addition, the collaboration acknowledges support from ESA; ASI and INAF (Italy); NASA and DoE (USA); Tekes, Academy of Finland (grant no. 295113), CSC, and Magnus Ehrnrooth foundation (Finland); RCN (Norway; grant nos. 263011, 274990); and PRACE (EU).

References

Abazajian, K., Addison, G., Adshead, P., et al. 2019, arXiv e-prints, arXiv:1907.04473
Ade, P., Aguirre, J., Ahmed, Z., et al. 2019, J. Cosmology Astropart. Phys., 2019, 056
Andersen et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Ashdown, M. A. J., Baccigalupi, C., Balbi, A., et al. 2007, A&A, 471, 361
Bennett, C. L., Larson, D., Weiland, J. L., et al. 2013, ApJS, 208, 20
Bersanelli, M., Mandolesi, N., Butler, R. C., et al. 2010, A&A, 520, A4
BeyondPlanck Collaboration. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Bond, J. R., Jaffe, A. H., & Knox, L. 2000, ApJ, 533, 19
Colombo et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Cuttaia, F., Mennella, A., Stringhetti, L., et al. 2009, Journal of Instrumentation, 4, 2013
Cuttaia, F. & Terenzi, L. 2011, Planck-LFI Technical Note, 098
Galloway et al. 2020a, A&A, in preparation [arXiv:201x.xxxxx]
Galloway et al. 2020b, A&A, in preparation [arXiv:201x.xxxxx]
Gjerløw et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Gorski, K. M. 1994, ApJ, 430, L85
Hanany, S., Alvarez, M., Artis, E., et al. 2019, arXiv e-prints, arXiv:1902.10541
Hivon, E., Górski, K. M., Netterfield, C. B., et al. 2002, ApJ, 567, 2
Keihänen et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Lamarre, J., Puget, J., Ade, P. A. R., et al. 2010, A&A, 520, A9
Lewis, A. & Bridle, S. 2002, Phys. Rev. D, 66, 103511
Mennella, A., Bersanelli, M., Butler, R. C., et al. 2010, A&A, 520, A5
Page, L., Hinshaw, G., Komatsu, E., et al. 2007, ApJS, 170, 335
Paradiso et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Planck HFI Core Team. 2011, A&A, 536, A4
Planck Collaboration I. 2011, A&A, 536, A1
Planck Collaboration II. 2011, A&A, 536, A2
Planck Collaboration I. 2014, A&A, 571, A1
Planck Collaboration II. 2014, A&A, 571, A2
Planck Collaboration III. 2014, A&A, 571, A3

[6] There were four such power groups in LFI, feeding the radiometers associated with the following horn sets: (19-20-28), (18-26), (21-22-24-27), (23-25).

Planck Collaboration VI. 2014, A&A, 571, A6
Planck Collaboration II. 2016, A&A, 594, A2
Planck Collaboration XII. 2016, A&A, 594, A12
Planck Collaboration I. 2020, A&A, 641, A1
Planck Collaboration II. 2020, A&A, 641, A2
Planck Collaboration III. 2020, A&A, 641, A3
Planck Collaboration IV. 2020, A&A, 641, A4
Planck Collaboration V. 2020, A&A, 641, A5
Planck Collaboration VI. 2020, A&A, 641, A6
Planck Collaboration Int. LVII. 2020, A&A, in press [arXiv:2007.04997]
QUIET Collaboration, Bischoff, C., Brizius, A., et al. 2011, ApJ, 741, 111
Sugai, H., Ade, P. A. R., Akiba, Y., et al. 2020, Journal of Low Temperature Physics, 199, 1107
Suur-Uski et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Suzuki, A., Ade, P. A. R., Akiba, Y., et al. 2018, Journal of Low Temperature Physics, 193, 1048
Svalheim et al. 2020a, A&A, in preparation [arXiv:201x.xxxxx]
Svalheim et al. 2020b, A&A, in preparation [arXiv:201x.xxxxx]
Tauber, J. A., Nielsen, P. H., Martín-Polegre, A., et al. 2019, A&A, 622, A55
Tegmark, M. 1997, ApJ, 480, L87
Tegmark, M., Taylor, A. N., & Heavens, A. F. 1997, ApJ, 480, 22
Terenzi, L., Salmon, M. J., Colin, A., et al. 2009, Journal of Instrumentation, 4, 2012
Villa, F., Terenzi, L., Sandri, M., et al. 2010, A&A, 520, A6
Wandelt, B. D., Larson, D. L., & Lakshminarayanan, A. 2004, Phys. Rev. D, 70, 083511
Wehus, I. K., Næss, S. K., & Eriksen, H. K. 2012, ApJS, 199, 15

Paper VI

BeyondPlanck VII. Bayesian estimation of gain and absolute calibration for CMB experiments

E. Gjerløw, H. T. Ihle, S. Galeotta and the BeyondPlanck collaboration

A&A, to be submitted, arXiv: 2011.08082,

Astronomy & Astrophysics manuscript no. BP_gains
November 17, 2020
©ESO 2020

arXiv:2011.08082v1 [astro-ph.CO] 16 Nov 2020

BeyondPlanck VII. Bayesian estimation of gain and absolute calibration for CMB experiments

E. Gjerløw[11]⋆, H. T. Ihle[11], S. Galeotta[8], K. J. Andersen[11], R. Aurlien[11], R. Banerji[11], M. Bersanelli[4,9,10], S. Bertocco[8], M. Brilenkov[11], M. Carbone[14], L. P. L. Colombo[4], H. K. Eriksen[11], M. K. Foss[11], C. Franceschet[4,10], U. Fuskeland[11], M. Galloway[11], S. Gerakakis[14], B. Hensley[2], D. Herman[11], M. Iacobellis[14], M. Ieronymaki[14], J. B. Jewell[12], A. Karakci[11], E. Keihänen[3,7], R. Keskitalo[1], G. Maggio[8], D. Maino[4,9,10], M. Maris[8], S. Paradiso[4,10], B. Partridge[6], M. Reinecke[13], A.-S. Suur-Uski[3,7], T. L. Svalheim[11], D. Tavagnacco[8,5], H. Thommesen[11], D. J. Watts[11], I. K. Wehus[11], and A. Zacchei[8]

[1] Computational Cosmology Center, Lawrence Berkeley National Laboratory, Berkeley, California, U.S.A.
[2] Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08544, U.S.A.
[3] Department of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland
[4] Dipartimento di Fisica, Università degli Studi di Milano, Via Celoria, 16, Milano, Italy
[5] Dipartimento di Fisica, Università degli Studi di Trieste, via A. Valerio 2, Trieste, Italy
[6] Haverford College Astronomy Department, 370 Lancaster Avenue, Haverford, Pennsylvania, U.S.A.
[7] Helsinki Institute of Physics, Gustaf Hällströmin katu 2, University of Helsinki, Helsinki, Finland
[8] INAF - Osservatorio Astronomico di Trieste, Via G.B. Tiepolo 11, Trieste, Italy
[9] INAF/IASF Milano, Via E. Bassini 15, Milano, Italy
[10] INFN, Sezione di Milano, Via Celoria 16, Milano, Italy
[11] Institute of Theoretical Astrophysics, University of Oslo, Blindern, Oslo, Norway
[12] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Drive, Pasadena, California, U.S.A.
[13] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Str. 1, 85741 Garching, Germany
[14] Planetek Hellas, Leoforos Kifisias 44, Marousi 151 25, Greece

November 17, 2020

ABSTRACT

We present a Bayesian calibration algorithm for CMB observations as implemented within the global end-to-end BeyondPlanck framework, and apply this to the *Planck* Low Frequency Instrument (LFI) data. Following the most recent *Planck* analysis, we decompose the full time-dependent gain into a sum of three nearly orthogonal components: One absolute calibration term, common to all detectors; one time-independent term that can vary between detectors; and one time-dependent component that is allowed to vary between one-hour pointing periods. Each term is then sampled conditionally on all other parameters in the global signal model through Gibbs sampling. The absolute calibration is sampled using only the orbital dipole as a reference source, while the two relative gain components are sampled using the full sky signal, including the orbital and Solar CMB dipoles, CMB fluctuations, and foreground contributions. We discuss various aspects of the data that influence gain estimation, including the dipole/polarization quadrupole degeneracy, processing masks, and anomalous jumps in the instrumental gain. Comparing our solution to previous pipelines, we find good agreement in general, with relative deviations of $-0.84\,\%$ ($-0.67\,\%$) for 30 GHz, $-0.14\,\%$ ($0.02\,\%$) for 44 GHz and $-0.69\,\%$ ($-0.08\,\%$) for 70 GHz, compared to *Planck* 2018 (NPIPE). These deviations are within expected error bounds, and we attribute them to differences in data usage and general approach between the pipelines. In particular, we note that the BeyondPlanck calibration is performed globally, which results in better inter-frequency consistency than previous estimates. Additionally, *WMAP* observations are used actively in the BeyondPlanck analysis, and this both breaks internal degeneracies in the *Planck* data set and results in better overall agreement with *WMAP*. Although our presentation and algorithm are currently oriented toward LFI processing, the general procedure is fully generalizable to other experiments, as long as the CMB dipole signal can be used for calibration.

Key words. ISM: general – Cosmology: observations, polarization, cosmic microwave background, diffuse radiation – Galaxy: general

1. Introduction

The cosmic microwave background (CMB) anisotropies are among the most important observables available to cosmologists, and accurate determination of their statistical properties has been a main goal for a multitude of collaborations and experiments during the last three decades (e.g., Smoot et al. 1992; de Bernardis et al. 2000; Kovac et al. 2002; Bennett et al. 2013; Planck Collaboration I 2020, and references therein). The BeyondPlanck project (BeyondPlanck Collaboration 2020) is an initiative aiming to establish a common multi-experiment analysis platform that supports global end-to-end Bayesian analysis of raw time-ordered data (TOD) produced by such experiments, as well as seamless propagation of low-level uncertainties into all high-level products, including frequency and component maps, the CMB angular power spectra, and cosmological parameters. As a first demonstration, we apply this framework to the *Planck* LFI data, as presented in this and a suite of companion papers.

A fundamentally important step in any CMB analysis pipeline is photometric calibration—the process of mapping the instrument readout to the incoming physical signal. In general, this procedure involves comparing some specific feature in the

⋆ Corresponding author: E. Gjerløw; eirik.gjerlow@astro.uio.no

measured data with a known calibration model, for instance the CMB dipole or astrophysical foreground signal (or both), or by comparing the total measured power with a reference load with a known physical temperature (e.g., Planck Collaboration V 2016).

The multiplicative factor that converts between sky signal and detector readout is called the gain. This factor typically depends on the local environment of the detectors, such as the ambient temperature, and is therefore in principle different for each sample. However, as long as the detectors are thermally stable on reasonably long time scales, it is usually a good approximation to assume that the gain is constant over some short period of time, or at least that it is smoothly varying in time. For instance, the *WMAP* team adopted and fitted a six-parameter model for the gain, using housekeeping data such as focal plane temperature measurements to interpolate in time (Jarosik et al. 2011). For LFI, we will assume that the gain factor is constant throughout each *Planck* pointing period (PID) – the timescale during which the satellite scans a given "circle" on the sky; these last roughly an hour each. We will also assume that the gain is varying smoothly between neighboring PIDs, except during a small set of events during which the instrument was actively modified by the mission control center, for instance during cooler maintenance.

The *Planck* LFI Data Processing Centre (DPC) (Planck Collaboration II 2014; Planck Collaboration V 2016; Planck Collaboration II 2020) used an onboard 4 K reference load to support the 30 GHz calibration for the early results, while for the other channels, and for all channels in later releases, they relied primarily on the CMB dipole signal. Gain fluctuations and correlations were modelled and suppressed by boxcar averaging over a signal-to-noise dependent window size. The *Planck* HFI DPC (Planck Collaboration VIII 2014, 2016; Planck Collaboration III 2020) also used the CMB dipole signal for gain estimation, but in this case they assumed a constant gain factor throughout the whole mission, relying on the excellent thermal stability of the *Planck* instrument. Apparent gain variations were instead assumed to arise from non-linearities in the analog-to-digital conversion module, which then allowed for a deterministic correction. A similar approach has also been adopted by the recent SROLL2 re-analysis initiative (Delouis et al. 2019).

In the most recent official analysis (NPIPE; Planck Collaboration Int. LVII 2020), the *Planck* team adopted the LFI gain model for all channels up to and including the 857 GHz channel. A novelty introduced in that analysis, however, was a decomposition of the gain factor into two nearly orthogonal components: an absolute (or baseline) gain factor, which was assumed to be constant for the entire mission, and a detector-specific gain mismatch factor that could vary both in time and between detectors. This approach allowed estimation of each component separately, using calibrators that are better suited to each component. For example, the low signal-to-noise (but well-understood) orbital dipole was used to calibrate the absolute gain factor due to the long integration time involved in estimating this particular component. Solving for all relevant factors was then performed jointly with other relevant quantities.

In this paper, we adopt the NPIPE approach, and decompose the full gain into the above-mentioned components, and we estimate these jointly with all other parameters in the full model. Thus, the main novel feature presented in this paper is the integration of the gain estimation procedure within a larger Gibbs framework, as summarized by BeyondPlanck Collaboration (2020), which performs joint estimation of all relevant parameters in a statistically consistent manner, including the CMB and astrophysical foreground sky signal.

The rest of the paper is structured as follows: In Sect. 2, we aim to build intuition regarding gain estimation, presenting the general data model that we use and highlighting various important features of this model, as applied to real-world LFI observations. Next, in Sect. 3 we describe our main gain estimation procedure, before showing results in Sect. 4, and comparing these with those derived by other pipelines. Finally, we summarize in Sect. 5, with an eye toward future experiments and applications.

2. Data and modelling considerations

We start our presentation with a general discussion of the gain-related data model, and how to account for various complications that arise when fitting this to real-world data.

2.1. Data model

As described by BeyondPlanck Collaboration (2020), the main goal of the BeyondPlanck analysis framework is to develop an end-to-end Bayesian analysis platform for CMB data, starting from raw time-ordered data. As for most Bayesian problems, the key step in our approach is therefore to write down an explicit parametric model for the observed data from a given detector, $d_{t,i}$, where t is the index denoting the sample,[1] and i is the index denoting the detector in question. In the current analysis, we adopt the following high-level model,

$$d_{t,i} = g_{t,i} s^{\mathrm{tot}}_{t,i} + n^{\mathrm{corr}}_{t,i} + n^{\mathrm{wn}}_{t,i}, \tag{1}$$

where $n^{\mathrm{corr}}_{t,i}$ and $n^{\mathrm{wn}}_{t,i}$ are correlated and white noise, respectively, $g_{t,i}$ is the gain factor, and $s^{\mathrm{tot}}_{t,i}$ denotes the total signal. Here, $s^{\mathrm{tot}}_{t,i}$ is given in kelvin, while $d_{t,i}$ is the instrument readout, which is measured in volt, meaning that the unit for $g_{t,i}$ becomes [V/K]. The total signal can further be decomposed into

$$\begin{aligned} s^{\mathrm{tot}}_{t,i} &= s^{\mathrm{sky}}_{t,i} + s^{\mathrm{orb}}_{t,i} + s^{\mathrm{fsl}}_{t,i} \\ &= P_{tp,i} \mathsf{B}^{\mathrm{symm}}_{pp',i} s^{\mathrm{sky}}_{p'} + P_{tp,i} \mathsf{B}^{\mathrm{asymm}}_{pp',i} s^{\mathrm{orb}}_{p'} + P_{tp,i} \mathsf{B}^{\mathrm{asymm}}_{pp',i} s^{\mathrm{fsl}}_{p'} \\ &= P_{tp,i} \left[\mathsf{B}^{\mathrm{symm}}_{pp',i} s^{\mathrm{sky}}_{p'} + \mathsf{B}^{\mathrm{asymm}}_{pp',i} \left(s^{\mathrm{orb}}_{p'} + s^{\mathrm{fsl}}_{p'} \right) \right] \end{aligned} \tag{2}$$

In this expression, P is the pointing matrix, which contains the mapping between the pointing direction the instrument, p and the sample index t; $\mathsf{B}^{\mathrm{symm}}$ and $\mathsf{B}^{\mathrm{asymm}}$ denote convolution with the symmetric and asymmetric beams, which quantify the fraction of the total signal coming from direction p' when the instrument is pointing towards p; s^{sky} is the sky signal (including the Solar dipole); s^{orb} is the orbital dipole (to be discussed below); and s^{fsl} represents signal leakage through the far-sidelobes. For further details regarding any of these objects, we refer the interested reader to BeyondPlanck Collaboration (2020) and references therein.

The main topic of the current paper is estimating $g_{t,i}$. In this respect, it is important to note that all other free parameters in the data model, including $s^{\mathrm{sky}}_{t,i}$ and $n^{\mathrm{corr}}_{t,i}$, are also unknown, and must be estimated jointly with $g_{t,i}$. Casting this statement into

[1] Here, a "sample" means the detector readout at every $1/f$ seconds, where f is the sampling frequency of the instrument. The whole set of these samples is called the time-ordered data (TOD). The sampling frequency for the three LFI instruments are 32.5 Hz, 46.5 Hz, and 78.8 Hz for the 30 GHz, 44 GHz, and 70 GHz instrument, respectively.

Bayesian terms, we wish to sample from the *posterior distribution*,[2]

$$P(\boldsymbol{g}, \boldsymbol{s}^{\rm tot}, \boldsymbol{s}^{\rm orb}, \boldsymbol{n}^{\rm corr}, \dots \mid \boldsymbol{d}). \tag{3}$$

That is, we aim to model the *global* state of the instrument and data, and map out the probability of various points in parameter space by sampling from this distribution. This may at first glance seem like a intractable problem. However, a central component of the BeyondPlanck framework is parameter estimation through *Gibbs sampling*. According to the theory of Gibbs sampling, samples from a joint posterior distribution may be drawn by iteratively sampling from all *conditional* distributions. In other words, when sampling the gain, we may assume that the sky signal and correlated noise parameters are perfectly known. And likewise, when sampling the sky signal or correlated noise parameters, we may assume that the gain is perfectly known. The correlations between these various parameters are then probed by performing hundreds or thousands of iterations of this type.

Thus, for the purposes of calibration alone, we do not need to be concerned with many aspects that indirectly affect the gain, such as CMB dipole or correlated noise estimation (Planck Collaboration II 2014; Planck Collaboration V 2016; Planck Collaboration II 2020). Instead, we are here concerned only with defining an adequate model for $\boldsymbol{g}$, and expressing this in a way that minimizes degeneracies with parameters in the Gibbs chain.

As discussed above, the gain is generally not constant in time. A very conservative (and somewhat naïve) model would therefore be to assume that the gain is in fact different for every sample t. However, this model clearly does not take into account our full knowledge about the instrument (Planck Collaboration XXVIII 2014). In particular, we do know that the gain is expected to correlate with the detector temperature, and this temperature does not change significantly on timescales of just one sample. Rather, based on available housekeeping data, a good assumption is that the gain is constant within a given pointing period (PID, or *scan*) – which is defined as a collection of samples measured over a period of about an hour, during which the instrument spins about its axis once per minute while keeping the spin axis vector stationary. Between each scan, the instrument performs a slight adjustment of the satellite spin axis, ensuring that new sky areas are covered in consecutive pointing periods.

To reflect the assumption of constant gain within each scan, we rewrite our data model as follows,

$$d_{t,i} = g_{q,i} s^{\rm tot}_{t,i} + n^{\rm corr}_{t,i} + n^{\rm wn}_{t,i}, \tag{4}$$

where q now denotes PID. Thus, t is used to indicate a specific sample, while q represents a collection of samples.

From Eq. (4), we immediately note the presence of two important degeneracies, involving the sky signal and noise, respectively. If we attempt to fit for g, $s^{\rm tot}$, and $n^{\rm corr}$ simultaneously, without knowing anything about any of them, we see that a given solution, say, $\{g_0, s^{\rm tot}_0, n^{\rm corr}_0\}$, will result in an identical goodness-of-fit as another solution $\{g_1, s^{\rm tot}_1, n^{\rm corr}_1\}$, as long as either

$$g_1 = g_0 \frac{s^{\rm tot}_0}{s^{\rm tot}_1}, \tag{5}$$

or

$$n^{\rm corr}_1 = n^{\rm corr}_0 + g_0 s^{\rm tot}_0 - g_1 s^{\rm tot}_1. \tag{6}$$

In other words, the gain is multiplicatively degenerate with the signal, and additively degenerate with the correlated noise. Such degeneracies are mainly a computational problem, since with two degenerate parameters in a Gibbs chain, exploring the resulting distributions takes a much larger number of samples than for uncorrelated parameters. A main topic of this paper is how to break these degeneracies in a statistically self-consistent and computationally efficient manner.

2.2. Absolute versus relative gain calibration

So far, we have been talking about the calibration of a given detector in isolation, which relates to what we call *absolute* calibration. Absolute calibration refers to correctly determining the "true" value of the gain, and is important for accurately determining the emitted intensity of astrophysical components, such as the CMB.

Another closely related concept is *relative* calibration,[3] which quantifies calibration differences between radiometers. Because of *Planck*'s scanning strategy, which only provides weak cross-linking[4] (Planck Collaboration I 2011), it is impossible to estimate the three relevant Stokes parameters (the intensity, I, and two linear polarization parameters, Q and U) independently for each detector. Rather, the polarization signal is effectively determined by considering pairwise differences between detector measurements, while properly accounting for their relative polarization angle differences at any given time. Therefore, any instrument characterization error that induces spurious detector differences will be partially interpreted by the analysis pipeline as a polarization signal. If our relative gain calibration is wrong, such differences will be introduced.

Given the high sensitivity of current and future CMB experiments, the gain must be estimated to a fractional precision better than $O(10^{-3})$ for robust CMB temperature analysis, and better than $O(10^{-4})$ for robust polarization analysis. Accurate relative calibration is thus even more important than accurate absolute calibration, and this will, as discussed in the next section, inform the choices we make on how to estimate these two types of calibration.

2.3. The Solar and orbital CMB dipoles

One of the most powerful ways to break the signal/gain degeneracy mentioned in the previous section is to observe a source of known brightness. If that source happens to be significantly stronger than other sources in the same area of the sky, we could fix $s^{\rm tot}_{t,i}$ in Eq. 4 and the gain would essentially be determined by the ratio of the data to the known source brightness.

Unfortunately, the number of available astrophysical calibration sources that may be useful for CMB calibration purposes is very limited, given the stringent requirements discussed in

[2] Here, and elsewhere, boldface quantities generally mean vectors. Which vector space they belong to will to a large degree be evident from the subscripts – in this case, there are no subscripts, meaning that the vectors contain all samples from all detectors.

[3] Note that our definition differs slightly from the LFI DPC definition of relative calibration. In their nomenclature, relative calibration refers to temporal fluctuations of the gain around the mean within a given detector.

[4] A given point on the sky, once observed by a detector, is not observed by the same detector at a different angle before a long time has passed, during which several environmental parameters of the detector may have changed.

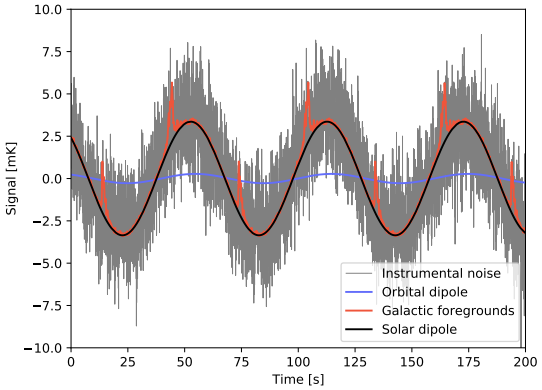

Fig. 1: Comparison of different contributions to the time-ordered data seen by *Planck* at 30 GHz, for a PID whose orientation is close to perpendicular to the dipole axis. The blue and black curves show the orbital and Solar CMB dipoles, respectively, while the red line shows contributions from small-scale CMB fluctuations and Galactic foregrounds. The gray line shows instrumental noise.

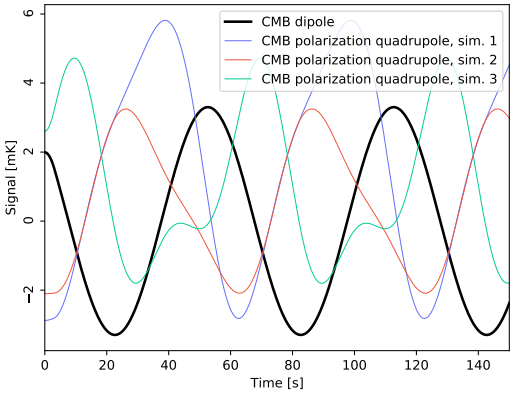

Fig. 2: Comparison of the CMB temperature dipole (thick black line) observed through the *Planck* scanning strategy with three random polarization quadrupole simulations (thin colored lines); the latter have been scaled by a factor of 10^4 for visualization purposes.

Sect. 2.2. For instance, the brightness temperature of individual planets within the Solar system is only known to about 5 % (Planck Collaboration VIII 2016), while few other local sources are known with a precision better than 1 %.

The key exception is the CMB dipole. The peculiar velocity of the *Planck* satellite relative to the CMB rest frame induces a strong apparent dipole on the sky due to the Doppler effect. Specifically, photons having an anti-parallel velocity relative to the satellite motion are effectively blue-shifted, while photons with a parallel velocity are redshifted.

It is useful to decompose the peculiar spacecraft velocity into two components; the motion of the Solar system relative to the CMB rest frame, $\boldsymbol{v}_{\rm solar}$, and the orbital motion of the *Planck* satellite relative to the Solar system barycenter, $\boldsymbol{v}_{\rm orbital}$. Thus, the total velocity of the satellite relative to the CMB rest frame is $\boldsymbol{v}_{\rm tot} = \boldsymbol{v}_{\rm solar} + \boldsymbol{v}_{\rm orbital}$. Taking into account the full expression for the relativistic Doppler effect, the induced dipole reads

$$s^{\rm dip}(\boldsymbol{x}, t) = T_{\rm CMB}\left(\frac{1}{\gamma(t)(1-\boldsymbol{\beta}(t)\cdot\boldsymbol{x})} - 1\right), \quad (7)$$

where $\boldsymbol{\beta} = \boldsymbol{v}_{\rm tot}/c$, and $\gamma = (1-|\boldsymbol{\beta}|^2)^{-1/2}$. The total dipole is time dependent because of the motion of the satellite over the course of the mission. We can similarly define a Solar dipole, $s^{\rm solar}(\boldsymbol{x})$ and an orbital dipole, $s^{\rm orb}(\boldsymbol{x}, t)$, which are induced by only the Solar and orbital velocities alone, respectively. Both dipoles play crucial roles in CMB calibration; the orbital dipole for absolute calibration and the Solar dipole for relative calibration.

Starting with the orbital dipole, we note that this depends only on the satellite's velocity with respect to the Sun. This is exceedingly well measured through radar observations, and known with a precision better than 1 cm s^{-1} (Godard et al. 2009). For an orbital speed of 30 km s^{-1}, this results in an overall relative precision better than $O(10^{-6})$. However, Eq. (7) also depends on the CMB monopole, which is measured by *COBE*-FIRAS to 2.72548 K ± 0.57 mK (Fixsen 2009), corresponding to a relative uncertainty of 0.02 % or $O(10^{-4})$. Thus, the absolute calibration of any current and future CMB experiment cannot be determined with a higher absolute precision than $O(10^{-4})$ until a next-generation CMB spectral distortion experiment, for instance PIXIE (Kogut et al. 2011), is launched. Still, this precision is more than sufficient for *Planck* calibration purposes.

The second CMB dipole component corresponds to the Sun's motion with respect to the CMB rest frame. While this velocity is intrinsically unknown, one may estimate this from the relative amplitude of the Solar and orbital dipoles. This is illustrated in Fig. 1, which compares the orbital and Solar dipole signals (blue and black curves) with contributions from Galactic foreground emission and instrumental noise at 30 GHz for about three minutes of time-ordered observations. The Solar dipole is effectively determined by the relative amplitude ratio between the black and blue curves in this figure.

Based on this approach, the most recent *Planck* analyses have determined that the Solar CMB dipole amplitude is about 3.36 mK, corresponding to Solar velocity of about 370 km s^{-1} (Planck Collaboration I 2020; Planck Collaboration Int. LVII 2020). For comparison, large-scale CMB polarization fluctuations typically exhibit variations smaller than $O(1\,\mu{\rm K})$ (Planck Collaboration IV 2020), and consequently the relative calibration of different detectors must be better than $O(10^{-4})$ to avoid the Solar CMB dipole to significantly contaminate large-scale polarization. Achieving this level of precision in the presence of correlated noise, Galactic foregrounds, far sidelobe contamination and other sources of systematic uncertainties is the single most difficult challenge associated with large-scale CMB polarization science.

2.4. The degeneracy between the CMB temperature dipole and polarization quadrupole

In Sect. 2.1 we noted that the gain is multiplicatively degenerate with the signal, and additively degenerate with the noise. Within this broad categorization, there are also some particularly important gain/signal degeneracies that are worth highlighting, and perhaps the most prominent example is that with respect to the CMB polarization quadrupole. To illustrate this, consider the case in which two detectors report different CMB dipole am-

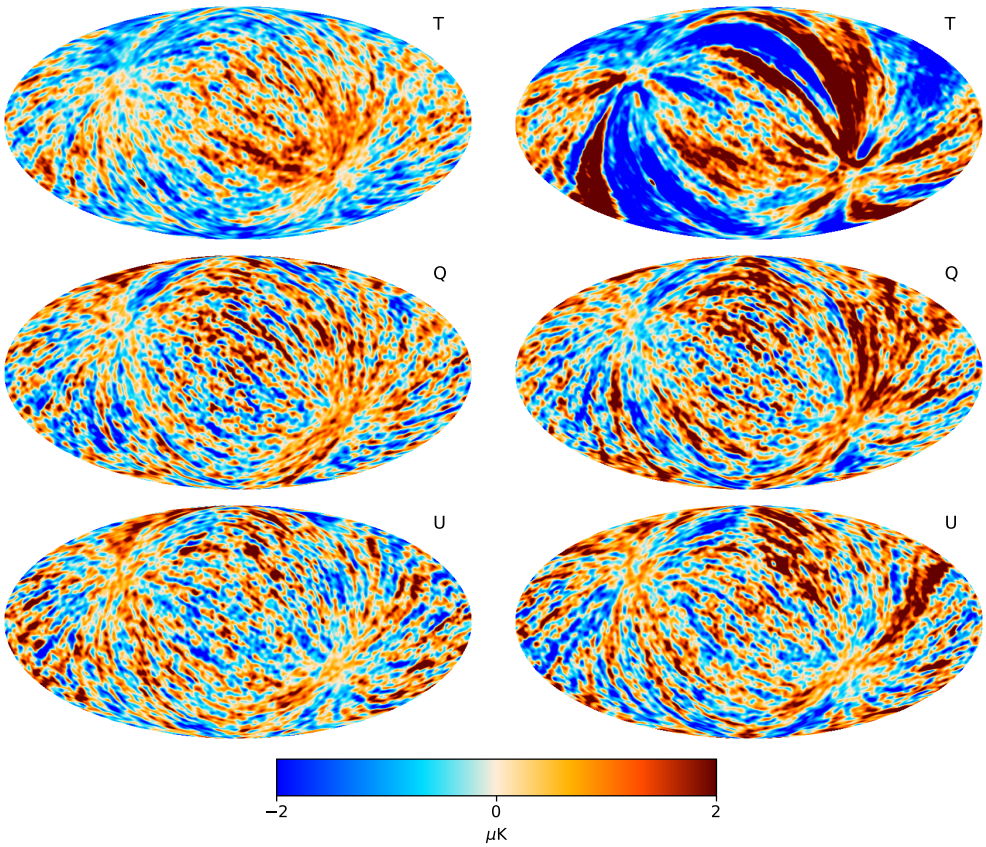

Fig. 3: Correlated noise maps for the 30 GHz channel in a Gibbs chain that includes (*left panel*) or neglects (*right panel*) gain time-dependencies. All maps are smoothed to a common resolution of 2.5° FWHM.

plitude signals; how could such a difference be explained? One possible explanation is a calibration mismatch, i.e., that the absolute calibration of one or both detectors is mis-estimated.

Another possible explanation could be a large polarization CMB *quadrupole* signal. Due to the scanning strategy adopted by *Planck*, in which the same ring is observed repeatedly for one hour, a polarization quadrupole can easily appear with a dipolar signature, depending on the particular phase orientation of the mode in question. This is illustrated in Fig. 2. The black thick line shows the CMB temperature dipole as a function of time, while the colored thin lines show three random polarization quadrupole simulations, all observed with the *Planck* scanning strategy. Out of the three random quadrupole simulations, two have a time-domain behaviour that very closely mimics the CMB temperature dipole, and in the presence of noise and instrumental effects, it would be exceedingly difficult to distinguish between the two models.

This is a perfect recipe for a degenerate system, and one that carries the potential of contaminating any large-scale polarization reconstruction. It is, however, important to note that this particular degeneracy appears with a very specific morphology, and affects only a handful of spatial polarization modes, as defined by projecting the CMB Solar dipole onto the *Planck* scanning strategy. Recognizing the importance of this degeneracy, previous *Planck* analyses have adopted different strategies of resolving the issue. For instance, both the *Planck* LFI DPC and `NPIPE` analyses have opted to disregard the CMB polarization component completely during the calibration phase (Planck Collaboration II 2016, 2020; Planck Collaboration Int. LVII 2020). This may be at least partially justified for LFI on theoretical grounds by noting that the CMB polarization variance on large angular scales predicted by current best-fit ΛCDM models is $\lesssim 0.1\,\mu\mathrm{K}^2$, which is comparable to, or below, the overall noise. For the significantly more sensitive HFI instrument, this assumption is not adequate, and the recent `NPIPE` analysis therefore explicitly estimates a transfer function to account for this effect (Planck Collaboration Int. LVII 2020).

In the following, we adopt a similar strategy as *Planck* LFI, but with an important difference: Rather than excluding the entire CMB polarization component from the calibration procedure, we only exclude the single CMB polarization quadrupole mode. This ensures that all higher-order CMB moments are formally estimated without any effective transfer function, while the quadrupole formally still does have a non-zero transfer function. However, given the high LFI noise level and the low ΛCDM prediction this is in practice negligible, as already demonstrated

by previous *Planck* DPC analyses. It is also worth noting that the real CMB temperature quadrupole amplitude is significantly lower than the ΛCDM prediction (de Oliveira-Costa et al. 2004), and one may therefore expect also the polarization quadrupole to be low, as these are correlated. Still, for cosmological parameters derived from large-scale *Planck* polarization data, one is well advised to compare estimates with and without the CMB polarization quadrupole included, and ensure that these are statistically consistent. Of course, this particular mode is also associated with a high cosmic variance, and a small bias regarding this mode is in practice unlikely to have any measurable effect.

2.5. Processing masks and PID selection

The Gibbs sampling framework used by BeyondPlanck requires an explicit parametric model that describes CMB, foregrounds, and the instrument. If this model turn out to be an insufficient representation of the actual data, the Gibbs sampling framework will attempt to fit eventual modelling errors using the parameters that are at its disposal. Ideally, such unexplained contributions should end up as an excess residual in $\boldsymbol{r} = \boldsymbol{d} - \boldsymbol{g}\boldsymbol{s}^{\mathrm{tot}} - \boldsymbol{n}^{\mathrm{tot}}$, but in practice they often also contaminate the other model parameters, such as the CMB. The correlated noise, $\boldsymbol{n}^{\mathrm{corr}}$, is one such parameter that, because of its relatively unconstrained and global structure, ends up absorbing a wide range of modelling errors, as discussed by Ihle et al. (2020). Furthermore, as already noted in Sect. 2.1, there is a tight degeneracy between the correlated noise and the gain, and $\boldsymbol{n}^{\mathrm{corr}}$ is therefore a sensitive monitor for gain errors. Figure 3 illustrates this in terms of one arbitrarily selected 30 GHz $\boldsymbol{n}^{\mathrm{corr}}$ sample from the main BeyondPlanck analysis (BeyondPlanck Collaboration 2020). The left column shows such a sample in the default model, in which the gain is allowed to vary from PID to PID, while the right column shows the same when enforcing a time-independent gain. While the maps in the left column are visually dominated by scan-aligned random stripes, as expected for $\boldsymbol{n}^{\mathrm{corr}}$, the maps in the right column (in particular the top row) shows large excesses with a dipolar pattern along each *Planck* scanning ring. This is the archetypical signature of gain modelling errors, and this clearly demonstrates the need for a time-variable gain model. At the same time, there is also a clear quadrupolar pattern in the default configuration, with a positive excess along the Galactic plane and a negative excess near the Galactic poles. This structure is visually consistent with a near sidelobe modelling residual, in the sense that the Galactic foreground signal is slightly over-smoothed compared to the prediction of the nominal beam model, and the resulting residual is picked up by the correlated noise component. This is not surprising, considering that about 1 % of the full LFI 30 GHz beam solid angle is unaccounted for in the GRASP beam model (Planck Collaboration II 2020), and some of this missing power may be in the near sidelobes. Fortunately, we also see from the same plot that the impact of this effect is modest, and accounts for only about 1 μK at 30 GHz.

More generally, because we have an incomplete understanding of both the instrument and the microwave sky, modelling errors will at some level always be a concern when estimating both gain and correlated noise. Furthermore, these modelling errors will typically be stronger near the Galactic plane or bright compact sources, where foreground uncertainties are large. For this reason, it is customary to apply a processing mask while estimating these quantities, omitting the parts of the sky that are least understood from the analysis. In BeyondPlanck, we define processing masks based on data-minus-signal residual maps for each frequency (Ihle et al. 2020), and these are shown in Fig. 4.

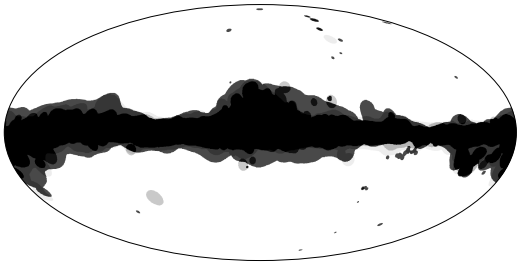

Fig. 4: Processing masks used in gain and correlated noise estimation. Different shades of grey indicate different frequency masks. The allowed 30 GHz sky fraction (light) is 0.73, the 44 GHz sky fraction (intermediate) is 0.81, and the 70 GHz sky fraction (dark) is 0.77. Reproduced from Ihle et al. (2020).

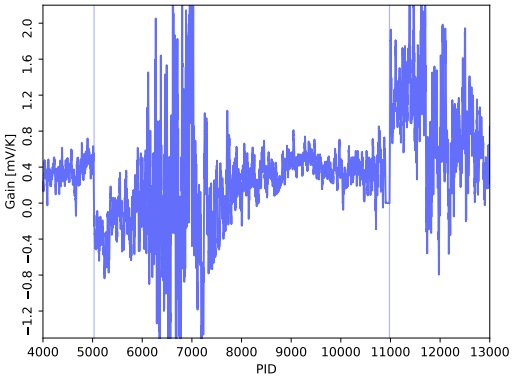

Fig. 5: Examples of jumps seen in the gain factors. The line shows a lightly smoothed gain solution for the 27S detector, which highlights two sharp jumps.

In addition, as discussed by Suur-Uski et al. (2020), we also exclude a number of PIDs from the main analysis, for similar reasons as for applying processing masks. Most of these PIDs, however, do not correspond to particularly problematic areas of the sky, but rather to unmodelled instrumental changes or systematic errors. For instance, all data collected during either cooler maintenance or major satellite maneuvers are rejected. Another important example is the 44 GHz LFI 26S radiometer, for which the correlated noise levels increased dramatically during the second half of the mission, possibly because of some subtle electrical issue, and these data are therefore omitted from the current analysis (Ihle et al. 2020). Excluded PIDs will show up as gaps in all PID plots in this paper.

2.6. Gain jumps

As noted from the beginning of the *Planck* experiment (see, e.g., Planck Collaboration XXVIII 2014), the physical gain of the instrument exhibits several sharp jumps. These jumps are related to changes in the thermal environment of the instrument – an example of such an event is the turning off of the 20 K sorption cooler. Not all of the events are well understood, and can mainly be traced after an initial gain estimation. We show examples of

Table 1: PIDs of gain jumps considered in the BeyondPlanck analysis. We indicate whether a jump comes from the NPIPE analysis or whether it comes from finding sharp jumps in the 20 K onboard temperature sensor readout.

PID	Jump classification
3352	NPIPE
5030	NPIPE
5484	NPIPE
8309	20 K
8503	20 K
8606	20 K
9613	20 K
10117	20 K
10512	20 K
10911	NPIPE
14061	20 K
15957	NPIPE
16204	20 K
16455	NPIPE
17024	20 K
18338	20 K
21484	NPIPE
25654	NPIPE
27110	NPIPE
27343	NPIPE
30387	NPIPE
32763	NPIPE
38591	NPIPE
43929	NPIPE

such jumps in Fig. 5 and list the jumps that we have considered in our data analysis in Table 1. These jumps mainly enter the analysis when we consider the smoothing of time-dependent gains in Sect. 3.5.

The jumps we consider are a combination of the jumps noted in the NPIPE analysis (Planck Collaboration Int. LVII 2020), as well as those found in an inspection of the readouts of the 20 K temperature sensors onboard the instrument, where similar jumps are found. The gain is expected to correlate with this temperature, so we take such jumps as being indicative of possible gain jumps in our model as well.

All of the jumps are considered to be present in all detectors, so we have not entered a more in-depth analysis of whether a given jump is actually present in every detector. Whereas such an analysis could be interesting in its own right, for the purposes of gain estimation the statistical cost of introducing a discontinuity into the model is very minimal, and we therefore make every jump a global one.

2.7. Breaking degeneracies through multi-experiment analysis

As described in BeyondPlanck Collaboration (2020), BeyondPlanck includes as part of its data selection several external data sets that are necessary to break fundamental degeneracies within the model. One particularly important example in this respect is the inclusion of low-resolution *WMAP* polarization data. In the same way that the *WMAP* experiment was unable to measure a few specific polarization modes on the sky due to peculiarities in its scanning strategy (Jarosik et al. 2007), *Planck* is also unable constrain some modes as defined by its scanning strategy (Planck Collaboration II 2020). However, because the *WMAP* and *Planck* scanning strategies are intrinsically very different, their degenerate modes are not the same, and, therefore all sky modes may be measured to high precision when analyzing both data sets jointly.

This will be explicitly demonstrated in Sect. 4.3, where we compare the BeyondPlanck sky maps to those derived individually from each experiment. The morphology of these frequency difference maps correspond very closely to the correction templates produced respectively by the *WMAP* and *Planck* teams (Jarosik et al. 2007; Planck Collaboration II 2020), and BeyondPlanck is statistically consistent with both data sets. Agreement is a direct and natural consequence of performing a joint fit, and there is no need for additional explicit template corrections for BeyondPlanck.

At the same time, it is also important to note that only the *Planck* LFI data are currently modelled in terms of time-ordered data, whereas the *WMAP* sky maps and noise covariance matrices are analyzed as provided by the *WMAP* team. Therefore, if there should be unknown systematics present in *WMAP*, those errors will necessarily also propagate into the various BeyondPlanck products. An important near-future goal is therefore to integrate also the *WMAP* time-ordered data into this framework. This work is already on-going, as discussed by Watts et al. (2020), but a full *WMAP* TOD-based analysis lies beyond the scope of the current work.

3. Methodology

As discussed in Sect. 2.1, our main goal in this paper is to draw samples from $P(\boldsymbol{g} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{n}^{\rm corr}, \boldsymbol{d}, \ldots)$, the conditional distribution of $\boldsymbol{g}$ given all other parameters. In this section, we describe each of the various steps involved in this process.

3.1. Correlated noise degeneracies and computational speed-up

Before we present our main sampling algorithms for $\boldsymbol{g}$, we recall from Sect. 2.1 that $\boldsymbol{g}$ is additively degenerate with $\boldsymbol{n}^{\rm corr}$. In a Gibbs sampling context, strong degeneracies lead to very long Markov correlation lengths as the Gibbs sampler attempts to explore the degenerate space between the two parameters. In order to save computing power and time, it is therefore better to sample $\boldsymbol{g}$ and $\boldsymbol{n}^{\rm corr}$ *jointly*, such that for a given iteration of the main Gibbs chain, we instead sample directly from $P(\boldsymbol{g}, \boldsymbol{n}^{\rm corr} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$.[5]

A joint sample may be produced by means of univariate distributions through the definition of a conditional distribution,

$$P(x_1 \mid x_2) \equiv \frac{P(x_1, x_2)}{P(x_1)} \Rightarrow P(x_1, x_2) = P(x_1)P(x_2 \mid x_1). \tag{8}$$

Thus, sampling from the joint distribution $P(\boldsymbol{g}, \boldsymbol{n}^{\rm corr} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$ is equivalent to first sampling $\boldsymbol{g}$ from its *marginal* distribution with respect to $\boldsymbol{n}^{\rm corr}$, $P(\boldsymbol{g} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$, and then subsequently sampling $\boldsymbol{n}^{\rm corr}$ from its *conditional* distribution with respect to $\boldsymbol{g}$, $P(\boldsymbol{n}^{\rm corr} \mid \boldsymbol{g}, \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$. These two steps must be performed immediately after one another, or else we would introduce an inconsistency in the Gibbs chain with respect to the other parameters.

Note that $P(\boldsymbol{n}^{\rm corr} \mid \boldsymbol{g}, \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$ is unchanged compared to the original Gibbs prescription, and no changes are required to

[5] Although this might seem somewhat counter-intuitive in the context of Gibbs sampling, joint sampling formally corresponds to reparametrizing $\{\boldsymbol{g}, \boldsymbol{n}^{\rm corr}\}$ into *one* parameter in the Gibbs chain.

sample from that particular distribution (see Ihle et al. 2020, for more details on this sampling process). When it comes to $P(\boldsymbol{g} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$, we refer to Appendix A.2 of BeyondPlanck Collaboration (2020), whose sampling equations we will use throughout this paper. We note that the data model used in that appendix is the same general form as Eq. (4), and that sampling from $P(\boldsymbol{g} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$ is exactly analogous to what is shown in that appendix, as long as we make the identification $\boldsymbol{n} \rightarrow \boldsymbol{n}^{\rm corr} + \boldsymbol{n}^{\rm wn}$. As the covariance matrix of a sum of independent Gaussian variables ($\boldsymbol{n}^{\rm corr}$ and $\boldsymbol{n}^{\rm wn}$) is also Gaussian, with a covariance matrix given by the sum of the individual covariance matrices, we can in what follows use the results of the above-mentioned appendix to sample from $P(\boldsymbol{g} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$ as long as we let $\mathsf{N} \rightarrow \mathsf{N}^{\rm corr} + \mathsf{N}^{\rm wn}$.

Computationally speaking, sampling from $P(\boldsymbol{g} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$ instead of $P(\boldsymbol{g} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{n}^{\rm corr}, \boldsymbol{d}, \ldots)$ is numerically equivalent to a more expensive noise covariance matrix evaluation.[6] To mitigate this additional cost, we note that the gain is assumed to be slowly varying in time, and, in particular, constant within each PID. All time-domain operations may therefore be carried out using co-added low-resolution data with negligible loss of precision. In practice, all TOD operations are in the following carried out at a sample rate of 1 Hz, leading to a computational speed-up of about two orders of magnitude.

3.2. Absolute gain calibration with the orbital dipole

Next, we also recall from Sect. 2.1 that the gain is multiplicatively degenerate with the total sky signal. At the same time, we note that the orbital CMB dipole is known to exquisite precision, and this particular component is therefore the ideal calibrator for CMB satellite experiments. However, its relatively low amplitude as compared with instrumental noise renders it incapable of tracking short-term gain variations, and, when fitted jointly with astrophysical foregrounds, it is also not sufficiently strong to determine relative calibration differences between detectors at the precision required for large-scale polarization reconstruction. Therefore, to minimize sensitivity to potential residual systematic and modelling errors, it is advantageous to estimate the absolute calibration using the orbital dipole alone, but use the full signal model (including the bright Solar CMB dipole) when estimating relative and time-dependent gain variations.

This motivates splitting the gain into two components,

$$g_{q,i} = g_0 + \gamma_{q,i}, \tag{9}$$

where g_0 is now independent of both time and detectors, following Planck Collaboration Int. LVII (2020). Our goal is then to use only the orbital CMB dipole to estimate g_0, and later use the full sky signal to estimate $\gamma_{q,i}$. Thus, with this reparametrization, we go from sampling from $P(\boldsymbol{g} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$ to sampling from $P(g_0, \boldsymbol{\gamma} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$. As usual, drawing samples from this joint distribution can be done by Gibbs sampling, so that we first sample g_0 from $P(g_0 \mid \boldsymbol{\gamma}, \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$ and then $\boldsymbol{\gamma}$ from $P(\boldsymbol{\gamma} \mid g_0, \boldsymbol{s}^{\rm tot}, \boldsymbol{d}, \ldots)$.

We should note that estimating g_0 using only the orbital dipole formally represents a violation of the Gibbs formalism, as we no longer draw this particular parameter from its exact conditional distribution. This is one of many examples for which we "buy" stability with respect to systematic errors at the price of increased statistical uncertainties. This is similar to the application of a processing mask when estimating the zero-levels of a CMB sky map (e.g., Planck Collaboration Int. XLVI 2016), or fitting correlated noise parameters using only a sub-range of all available temporal frequencies (e.g., Ihle et al. 2020). In all such cases, parts of the data are disregarded in order to prevent potential systematic errors from contaminating the parameter in question.

For the split in Eq. (9) to be valid, we must ensure that $\sum_{q,i} \gamma_{q,i} = 0$, such that $\gamma_{q,i}$ represents only deviations from the absolute calibration. For technical reasons, it turns out that this will be easier to do if we also reparametrize $\gamma_{q,i}$,

$$\gamma_{q,i} = \Delta g_i + \delta g_{q,i}, \tag{10}$$

where Δg_i represents the detector-specific constant gain, and $\delta g_{q,i}$ denotes deviations from Δg_i per scan. We can then separately enforce $\sum_i \Delta g_i = 0$ and $\sum_q \delta g_{q,i} = 0$ for each detector i, which is computationally cheaper than enforcing both constraints simultaneously.

Thus, we split the gain into three nearly independent variables, and explore their joint distribution by Gibbs sampling. The overarching goal for this section, then, is to derive sampling algorithms for each of the three associated conditional distributions,

$$P(g_0 \mid \Delta g_i, \delta g_{q,i}, \boldsymbol{d}_i, \boldsymbol{s}_i^{\rm tot}, \mathsf{N}_i, \ldots) \tag{11}$$

$$P(\Delta g_i \mid g_0, \delta g_{q,i}, \boldsymbol{d}_i, \boldsymbol{s}_i^{\rm tot}, \mathsf{N}_i, \ldots) \tag{12}$$

$$P(\delta g_{q,i} \mid g_0, \Delta g_i, \boldsymbol{d}_i, \boldsymbol{s}_i^{\rm tot}, \mathsf{N}_i, \ldots). \tag{13}$$

We now consider each of these in turn.

3.3. Sampling the absolute calibration, g_0

To sample the absolute calibration using the orbital dipole alone, we need to define a data model that depends only on g_0 and $s^{\rm orb}$. We do this by first subtracting the full signal model as defined by the current joint parameter state, and then add back only the orbital dipole term,

$$\begin{aligned} r_{t,i} &\equiv d_{t,i} - (g_0^{\rm curr} + \Delta g_i + \delta g_{q,i}) s_{t,i}^{\rm tot} + g_0^{\rm curr} s_{t,i}^{\rm orb} \\ &= g_0 s_{t,i}^{\rm orb} + n_{t,i}^{\rm tot}. \end{aligned} \tag{14}$$

Here $g_0^{\rm curr}$ denotes the absolute gain at the *current* iteration in the Gibbs chain, i.e., before drawing a new value for g_0.

As noted earlier, working with this residual and using the previous sample of g_0 to estimate the current sample does represent a breaking of the Gibbs formalism, since the statistically exact residual for g_0 would be

$$d_{t,i} - (\Delta g_i + \delta g_{q,i}) s_{t,i}^{\rm tot} = g_0 s_{t,i}^{\rm tot} + n_{t,i}^{\rm tot}. \tag{15}$$

However, in this case we would also be calibrating g_0 on the total sky signal instead of just the orbital dipole. Thus, we trade mathematical rigour and statistical uncertainties for stronger robustness with respect to systematic effects.

As discussed in Sect. 3.1, the noise term in Eq. (14) includes both correlated and white noise, and the appropriate covariance matrix is therefore $\mathsf{N} = \mathsf{N}^{\rm corr} + \mathsf{N}^{\rm wn}$. Given this fact, Eq. (14) corresponds to a simple univariate Gaussian model as a function of g_0, and the appropriate sampling algorithm is discussed in Appendix A.2 of BeyondPlanck Collaboration (2020). Applying

[6] Although not shown here, sampling from $P(\boldsymbol{g} \mid \boldsymbol{s}^{\rm tot}, \boldsymbol{n}^{\rm corr}, \boldsymbol{d}, \ldots)$ would follow the exact same procedure, but with a noise covariance matrix given by $\mathsf{N}^{\rm wn}$ instead of $\mathsf{N}^{\rm wn} + \mathsf{N}^{\rm corr}$. $\mathsf{N}^{\rm wn}$ is a diagonal matrix, while $\mathsf{N}^{\rm corr}$ is not, and since most operations are less heavy, computationally speaking, when diagonal matrices are involved, the resulting sampling process would also be lighter in that case.

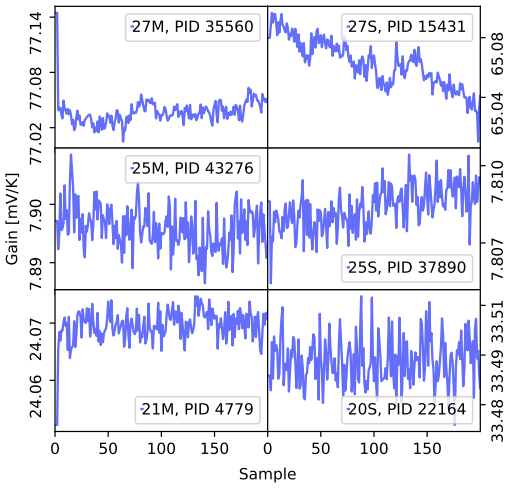

Fig. 6: Gibbs chains of the total gain for selected detectors and PIDs.

Table 2: Smoothing window size limits for the LFI detectors. All detectors within the same frequency band share the same window size limits.

Detector	$w^{\min}$	$w^{\max}$
30 GHz	400	1200
44 GHz	300	1500
70 GHz	400	1800

that general procedure to our special case, we may write down the following sampling equation for $\hat{g}_0$,[7]

$$\hat{g}_0 = \frac{\sum_i (\boldsymbol{s}_i^{\rm orb})^T \mathsf{N}_i^{-1} \boldsymbol{r}_i}{\sum_i (\boldsymbol{s}_i^{\rm orb})^T \mathsf{N}_i^{-1} \boldsymbol{s}_i^{\rm orb}} + \frac{\eta}{\sqrt{\sum_i (\boldsymbol{s}_i^{\rm orb})^T \mathsf{N}_i^{-1} \boldsymbol{s}_i^{\rm orb}}}, \tag{16}$$

where $\eta \sim N(0,1)$ is a random number drawn from a standard normal distribution. Here, and elsewhere, a T superscript indicates the matrix transpose operator. The first term in this equation is the (Wiener filter) mean of the distribution $P(g_0 \mid \boldsymbol{r}_i, \mathsf{N}_i)$, while the second term ensures that g_0 has the correct variance.

3.4. Sampling detector-dependent calibration

For Δg_i, we proceed similarly as with g_0, with two exceptions. First, Δg_i now represents the relative calibration between detectors, and, as discussed in Sect. 2.1, we need to use a stronger calibration signal than the orbital dipole to avoid significant polarization leakage. Secondly, we have to impose the constraint $\sum_i \Delta g_i = 0$.

[7] Note that we do not apply any priors on g_0 in this paper, which corresponds to $\mathsf{S}^{-1} = 0$, adopting the notation of BeyondPlanck Collaboration (2020), where S is the prior covariance of g_0. The remaining notational differences between Eq. (16) and Eq. (A.10) in that paper arises from our organizing all vectors and matrices in terms of independent detectors, using the fact that $\boldsymbol{n}^{\rm tot}$ is assumed to be independent between detectors; this may not be strictly true in practice, as discussed by Ihle et al. (2020), and future analyses may prefer to account for the full joint matrix.

We start by defining the following residual,

$$r_{t,i} \equiv d_{t,i} - (g_0 + \delta g_{q,i}) s_t^{\rm tot} = \Delta g_i s_t^{\rm tot} + n_t^{\rm tot} \tag{17}$$

for each detector. This equation is structurally similar to Eq. (14), with the main difference being that the total sky signal, which is dominated by the Solar dipole, is retained on the right-hand side. Otherwise, Eq. (17) still represents a Gaussian model, and we should be able to proceed similarly as for g_0 when drawing from the conditional distribution. We do, however, need to ensure that $\sum_i \Delta g_i = 0$, and this will significantly impact the form of the target distribution. The numerically most convenient method for imposing such a constraint is through the method of Lagrange multipliers.

In general, the method of Lagrange multipliers allows the user to minimize a function $f(\boldsymbol{x})$ under some set of constraints that may be formulated as $g(\boldsymbol{x}) = 0$. Without these constraints, one would of course determine x by solving $df/dx = 0$. With additional external constraints, however, it is convenient to instead define the so-called Lagrangian,

$$\mathcal{L}(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) + \lambda g(\boldsymbol{x}), \tag{18}$$

and set the corresponding partial derivatives with respect to $\boldsymbol{x}$ and λ equal to zero. It is readily seen that $\partial\mathcal{L}/\partial\lambda = 0$ corresponds directly to $g(\boldsymbol{x}) = 0$, which is precisely the desired constraint.

Our primary target distribution is

$$\begin{aligned} P(\Delta g \mid \boldsymbol{r}, \boldsymbol{s}^{\rm tot}, \mathsf{N}) &\propto P(\boldsymbol{r} \mid \Delta g, \boldsymbol{s}^{\rm tot}, \mathsf{N}) P(\Delta g) \\ &\propto \exp\left(\sum_i \left(\boldsymbol{r}_i - \Delta g_i \boldsymbol{s}^{\rm tot}\right)^T \mathsf{N}_i^{-1} \left(\boldsymbol{r}_i - \Delta g_i \boldsymbol{s}^{\rm tot}\right) \right) \end{aligned} \tag{19}$$

where the first line follows from Bayes' theorem, and the second follows from the fact that we assume vanishing covariance between detectors, and that $\boldsymbol{r}_i$ is Gaussian distributed with a mean of $\Delta g_i \boldsymbol{s}_i^{\rm tot}$ and covariance N_i. We are of course free to minimize the logarithm of this function instead of the function itself, which makes things easier as it takes the exponential away. We may therefore define the following Lagrangian,

$$\mathcal{L}(\Delta g_i, \lambda) = \sum_i (\boldsymbol{r}_i - \Delta g_i \boldsymbol{s}^{\rm tot})^T \mathsf{N}_i^{-1} (\boldsymbol{r}_i - \Delta g_i \boldsymbol{s}^{\rm tot}) + \lambda \sum_i \Delta g_i, \tag{20}$$

where λ is the Lagrange multiplier.

To optimize this function, we take the derivative with respect to Δg_i and λ to obtain two coupled equations. The first equation takes the form

$$\begin{aligned} &\frac{\partial \mathcal{L}}{\partial \Delta g_i} = 0 \\ \Rightarrow\ & -2(\boldsymbol{r}_i - \Delta g_i \boldsymbol{s}_i^{\rm tot})^T \mathsf{N}_i^{-1} \boldsymbol{s}_i^{\rm tot} + \lambda = 0 \\ \Rightarrow\ & \Delta g_i (\boldsymbol{s}_i^{\rm tot})^T \mathsf{N}_i^{-1} \boldsymbol{s}_i^{\rm tot} + \frac{1}{2}\lambda = (\boldsymbol{r}_i)^T \mathsf{N}_i^{-1} s_i^{\rm tot}, \end{aligned} \tag{21}$$

while the second simply reads

$$\begin{aligned} &\frac{\partial \mathcal{L}}{\partial \lambda} = 0 \\ \Rightarrow\ & \sum_i \Delta g_i = 0. \end{aligned} \tag{22}$$

Jointly solving these linear equations for Δg_i provides estimates with the correct mean. What we require, however, is a sample from the appropriate distribution, and not mean estimates. We must therefore add a fluctuation term, as in Eq. (16).

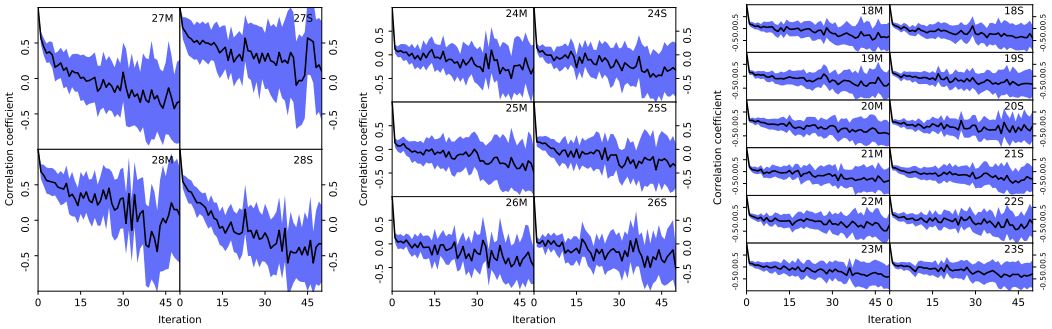

Fig. 7: Correlation coefficients as a function of distance between Gibbs samples for 30 (*left panel*), 44 (*middle panel*), and 70 GHz (*right panel*) detectors. The black thick line shows the mean value for all PIDs, while the blue band shows the 1σ error bars. The colored thin lines show the 1 % of the PIDs with the highest square uncertainty weighted deviation from the mean.

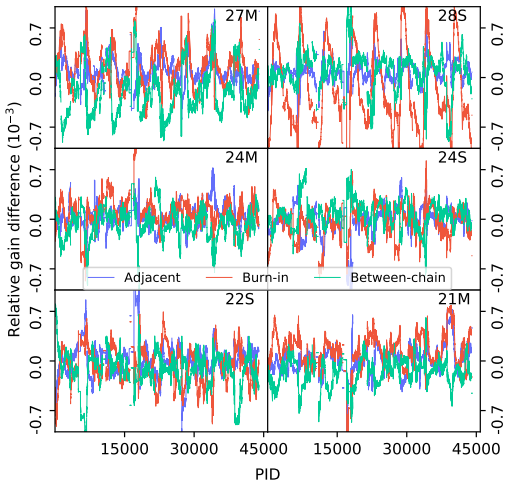

Fig. 8: Relative differences between the last sample in one of the BeyondPlanck chains and the sample adjacent to the last sample (blue line), the initial sample (the burn-in difference; red line) and the last sample in another chain (green line), for selected detectors.

To do so, we note that if it were not for λ, Eq. (21) would have the exact same form as Eq. (16), with $\boldsymbol{s}^{\rm tot}$ substituted for $\boldsymbol{s}^{\rm orb}$. Comparing Eq. (21) with Eq. (16), we then see that the final equation for the desired sample must be

$$\Delta\hat{g}_i(\boldsymbol{s}_i^{\rm tot})^T\mathsf{N}_i^{-1}\boldsymbol{s}_i^{\rm tot} + \frac{1}{2}\lambda = (\boldsymbol{r}_i)^T\mathsf{N}_i^{-1}\boldsymbol{s}_i^{\rm tot} + \eta\sqrt{(\boldsymbol{s}_i^{\rm tot})^T\mathsf{N}_i^{-1}\boldsymbol{s}_i^{\rm tot}}, \quad (23)$$

where, as usual, $\eta \sim N(0,1)$.

Casting this in terms of a linear system with $n_{\rm detector} + 1$ unknowns, this may be solved straightforwardly with standard numerical libraries. For a two-detector example, the resulting system of equations takes the form

$$\begin{bmatrix} (\boldsymbol{s}^{\rm tot})^T\mathsf{N}_1^{-1}\boldsymbol{s}^{\rm tot} & 0 & \frac{1}{2} \\ 0 & (\boldsymbol{s}^{\rm tot})^T\mathsf{N}_2^{-1}\boldsymbol{s}^{\rm tot} & \frac{1}{2} \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \Delta\hat{g}_1 \\ \Delta\hat{g}_2 \\ \lambda \end{bmatrix} = \quad (24)$$

$$\begin{bmatrix} (\boldsymbol{r}_1)^{\rm tot}\mathsf{N}_1^{-1}s_1^{\rm tot} + \eta_1\sqrt{(\boldsymbol{s}_1^{\rm tot})^T\mathsf{N}_1^{-1}\mathbf{s}_1^{\rm tot}} \\ (\boldsymbol{r}_2)^{\rm tot}\mathsf{N}_2^{-1}s_2^{\rm tot} + \eta_2\sqrt{(\boldsymbol{s}_2^{\rm tot})^T\mathsf{N}_2^{-1}\mathbf{s}_2^{\rm tot}} \\ 0 \end{bmatrix}. \quad (25)$$

3.5. Sampling temporal gain variations

Finally, we consider the temporal gain variations, $\delta g_{q,i}$. As before, we write down the following residual,

$$r_{t,i} \equiv d_{t,i} - (g_0 + \Delta g_i)s_{t,i}^{\rm tot} = \delta g_{q,i}s_{t,i}^{\rm tot} + n_{t,i}^{\rm tot}, \quad (26)$$

where we again employ the total signal as a calibrator. The only difference with respect to Eq. (17) is that $\delta g_{q,i}$ now contains multiple elements per detector, and is now a vector in PID space. We can make this point more explicit by writing

$$\boldsymbol{r}_i \equiv \boldsymbol{d}_i - (g_0 + \Delta g_i)\boldsymbol{s}_i^{\rm tot} = \mathsf{T}_i\delta\boldsymbol{g}_i + \boldsymbol{n}_i^{\rm tot}, \quad (27)$$

where T is an $n_{\rm samp} \times n_{\rm scan}$-matrix that contains $s_{t,i}^{\rm tot}$ in element (t,q) for all values of t in scan q. All other elements are zero. Thus, T projects $\delta\boldsymbol{g}$ into the $n_{\rm samp}$-dimensional space of $\boldsymbol{r}_i$ and $\boldsymbol{n}_i^{\rm tot}$.

Once again following the procedure in Equation A.3 in BeyondPlanck Collaboration (2020), we may write down the following sampling equation,

$$\mathsf{T}_i^T\mathsf{N}_i^{-1}\mathsf{T}_i\,\delta\hat{\boldsymbol{g}}_i = \mathsf{T}_i^T\mathsf{N}_i^{-1}\boldsymbol{r}_i + (\mathsf{T}_i^T\mathsf{N}_i^{-1}\mathsf{T}_i)^{\frac{1}{2}}\eta, \quad (28)$$

where $\eta \sim N(0,1)$ is a random Gaussian vector of length $n_{\rm PID}$.

In its current form, Eq. (28) assumes that $\delta g_{q,i}$ is uncorrelated between scans. As discussed by Planck Collaboration VIII (2014), Planck Collaboration VIII (2016), Planck Collaboration III (2020), and Planck Collaboration Int. LVII (2020), this is not an efficient model for the *Planck* LFI data, because the gain is primarily determined by the thermal environment of the instrument, and this is quite stable in time. It is therefore advantageous, and in practice necessary, to enforce some form of smoothing between $\delta g_{q,i}$ to obtain robust results.

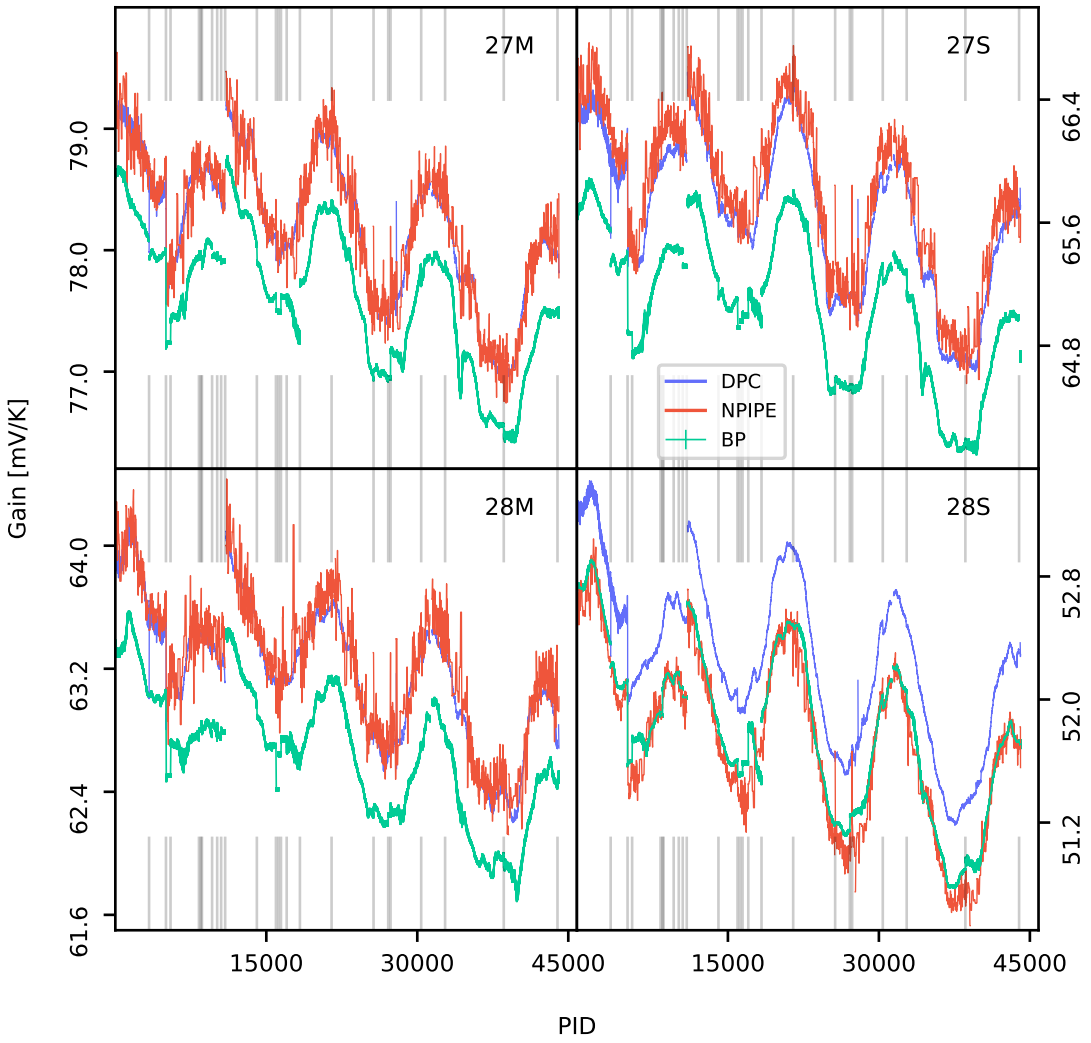

Fig. 9: Comparison of gain estimates for the 30 GHz detectors for *Planck* 2018, NPIPE, and BeyondPlanck. The light grey lines are the locations of hard-coded gain jumps at which we set hard stops for the smoothing algorithm, detailed in Table 1. The width of the BeyondPlanck line is given by the Monte Carlo uncertainty of the chains.

The proper Bayesian way of handling this is to assign a covariance matrix to $\delta g_{q,i}$, and then sample jointly from the likelihood and prior for $\delta g_{q,i}$, as is done for both the instrumental noise and CMB fluctuations (BeyondPlanck Collaboration 2020; Ihle et al. 2020; Colombo et al. 2020). However, due to time constraints, this is left for future work. In this paper, we instead adopt a similar approach as the LFI DPC and NPIPE (Planck Collaboration II 2020; Planck Collaboration Int. LVII 2020), and smooth $\delta g_{q,i}$ using a variable-width smoothing window,

$$\overline{\delta g}_{q,i} = \left(\sum_{q-n_{q,i}}^{q+n_{q,i}} w_{q,i}\right)^{-1} \sum_{q-n_{q,i}}^{q+n_{q,i}} w_{q,i}\delta g_{q,i} \tag{29}$$

where $w_{q,i} \equiv (\mathsf{T}_i^T \mathsf{N}_i^{-1} \mathsf{T}_i)^{-1/2}$ is the inverse noise variance weight of $\delta g_{q,i}$, and $n_{q,i}$ is the width of the smoothing window. Also following the LFI DPC, we let $n_{q,i}$ vary in time, depending on relative strength of the sky signal calibrator. Specifically, we define

$$m_{q,i} = \mathrm{Var}(s^{\mathrm{sky}}_{t,i}, t \in q) \tag{30}$$

to be the empirical variance of $s^{\mathrm{sky}}_{t,i}$ as seen by detector i in scan q, and we define a minimum and maximum width, n^{min} and n^{min}, for each frequency channel, as listed in Table 2. We then interpolate linearly between these limits based on the local signal

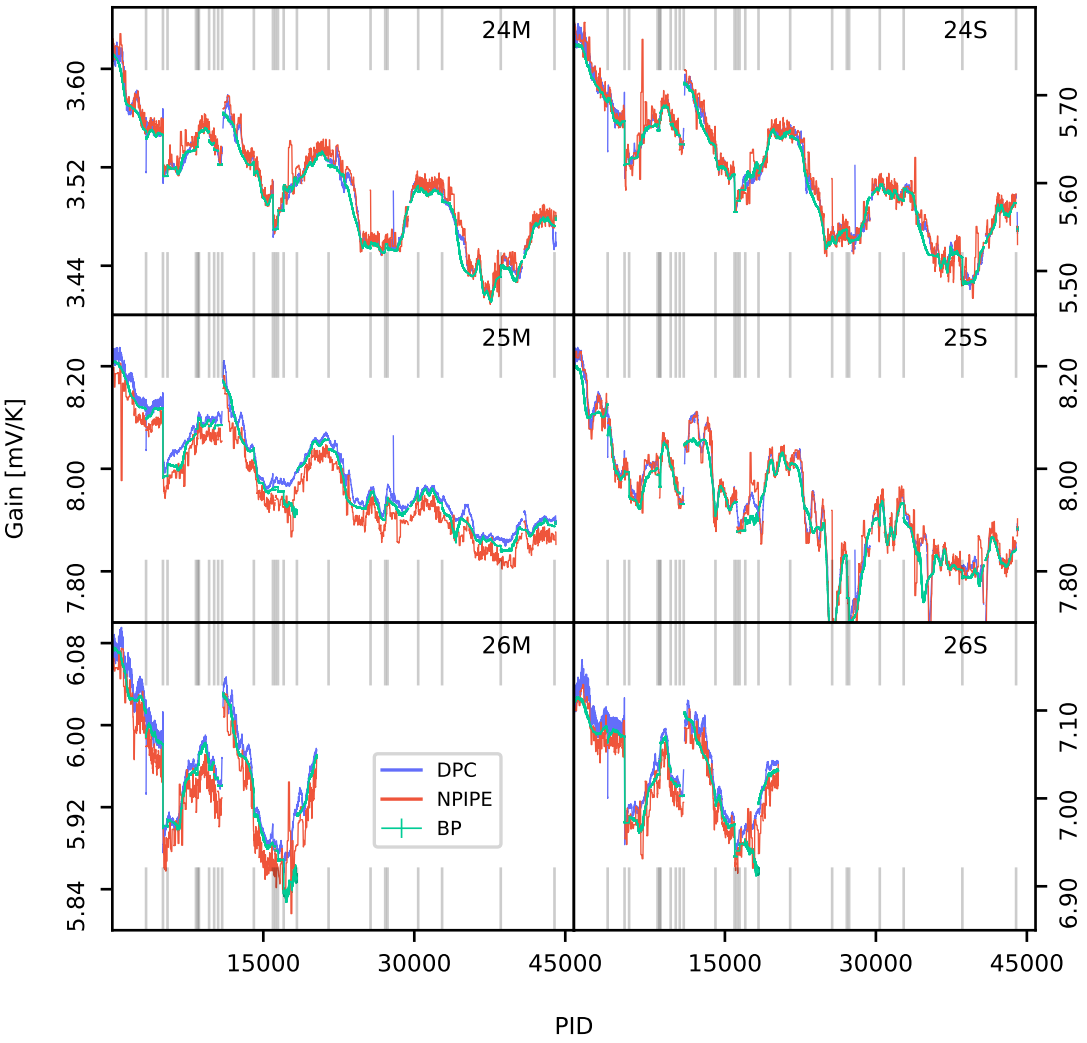

Fig. 10: Comparison of gain estimates for the 44 GHz detectors. The last half of PIDs for detectors 26M and 26S were not included in the calibration dataset. See Fig. 9 for details.

strength,

$$n_{q,i} = n_i^{\min} + \frac{m_{q,i} - m^{\mathrm{low}}}{m^{\mathrm{high}} - m^{\mathrm{low}}}(n_i^{\max} - n_i^{\min}), \tag{31}$$

where $m^{\mathrm{low}} = 1\,\mu$K and $m^{\mathrm{high}} = 5\,\mu$K; these values are adopted directly from the LFI DPC.

Finally, we recall that we require $\sum_q \delta g_{q,i} = 0$, which is imposed by subtracting the mean after smoothing $\delta g_{q,i}$ as described above. This does result in a statistically slightly sub-optimal estimate, as it gives equal weight to all scans, rather than assigning inverse noise variance weights as in Eq. (25). Solving the time-dependent system with Lagrange multipliers would be computationally prohibitively expensive, as it includes more than 44 000 individual PIDs per detector. We note that the current solution is statistically unbiased, even if slightly sub-optimal, and we leave implementation of the exact Wiener-filter approach to future work.

4. Results

We are now finally ready to present gain estimates for each *Planck* LFI radiometer, as estimated within the end-to-end Bayesian BeyondPlanck analysis framework.

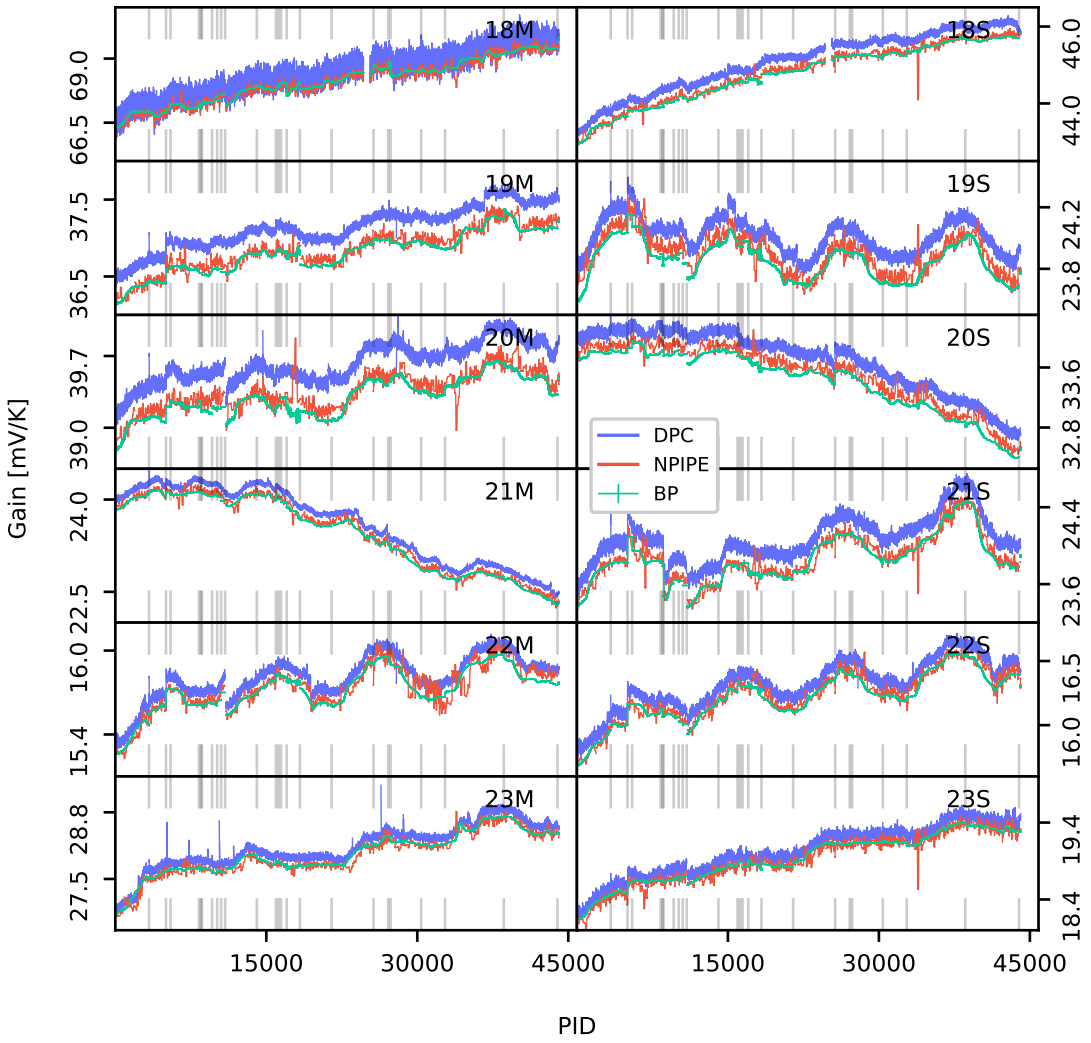

Fig. 11: Comparison of gain estimates for the 70 GHz detectors. See Fig. 9 for details.

4.1. Gain posterior distributions

We start by considering the sampling efficiency of the Monte Carlo chains produced with the above algorithm in terms of mixing and Markov chain correlation lengths. Figure 6 shows total gain as a function of Gibbs iteration for some representative radiometers and PIDs. Overall, most chains are stable and mix well, but there are some cases for which the gain solution is slowly drifting, and do not properly converge to a stationary value within the length of the Gibbs chain. A particularly bad example is the 30 GHz LFI 27M radiometer, PID 15 431. These situations are in general more common for the 30 GHz channel than the higher frequencies, because of the higher foreground signal in this channel. They are also more common for PIDs for which the *Planck* scanning strategy is closely aligned with the Galactic plane, since a relatively larger fraction of the data are rejected by the processing mask.

To globally characterize these issues, Fig. 7 shows the correlation lengths across all PIDs. The black line shows the estimated mean correlation length, whereas the blue bands show the estimated standard deviation. Thus, for 44 and 70 GHz essentially all cases have a correlation length shorter than 10–15 samples, and therefore mix well. However, the correlation lengths for the 30 GHz channel are slightly worse, especially for the 27S and 28M radiometers, where samples separated by 20–30 steps have a mean correlation of roughly 10 %. This behaviour is likely to be improved in future processing, when better foreground modelling will allow a smaller processing mask.

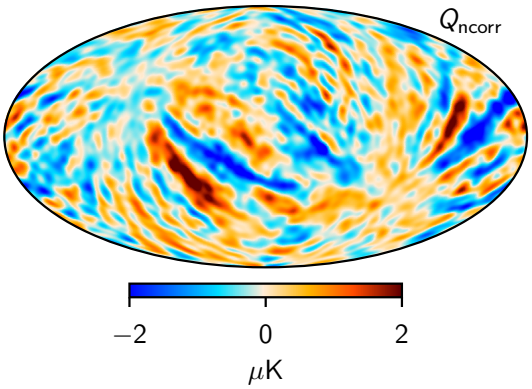

Fig. 12: Map of the correlated noise, $\boldsymbol{n}^{\rm corr}$, of the Q Stokes parameter for the 44 GHz frequency channel, smoothed to an effective angular resolution of 5° FWHM. Reproduced from Suur-Uski et al. (2020).

In Fig. 8, we show relative differences between the last sample of the chain and the first drawn sample (red line), and we compare that to the similar relative difference with the second-to-last sample in the chain (blue line), as well as an in-between chain comparison (green line). We see that even the first sample of the chain is as close to the final solution as the next-to-last sample, and the conditional burn-in period with respect to the gain does not significantly affect our results. Long-term burn-in is caused indirectly through correlations with other parameters in the system. Because of these external correlations, we follow BeyondPlanck Collaboration (2020), and omit the first 50 samples when presenting the final gain estimates.

In Figs. 9–11, we compare the the gain factors derived by BeyondPlanck, NPIPE, and *Planck* 2018 for each detector. For BeyondPlanck, the widths of each curve represent 1 σ posterior confidence regions as evaluated directly from the Gibbs chains (after omitting 50 samples for burn-in), while for the other two solutions we only show final best-fit estimates.

Overall, the largest differences between BeyondPlanck and the other two pipelines are observed in the 30 GHz channel. In particular, we find that the BeyondPlanck gain model is systematically lower than the 2018 model by about 0.84 % and than NPIPE by about 0.67 % for this channel, which translates into frequency maps that are about 0.84 % (or 0.67 %) *brighter*. We also see that the NPIPE and 2018 models agree very well for three of the radiometers, while 28S is an outlier, for which NPIPE is close to BeyondPlanck.

The 30 GHz channel is the most difficult to calibrate among all three LFI channels, because of its brighter foreground signal, and the different ways in which the three pipelines handle this fact makes the abovementioned gain solution differences less surprising: The NPIPE process treats this channel separately, in that this channel is analyzed without priors on polarized foregrounds. The resulting map is then subsequently used as a spatial polarization prior for the 44 and 70 GHz channels (Planck Collaboration Int. LVII 2020). In comparison, the 2018 approach also assumes vanishing CMB polarization during calibration, but this approach make no distinction between the orbital and Solar dipole with respect to absolute gain calibration (as both NPIPE and BeyondPlanck do), but rather assumes that the fitted foreground model is sufficiently accurate. In contrast, the BeyondPlanck pipeline does not treat the 30 GHz channel differently in any way, and also does not assume that the CMB polarization signal vanishes (except for the single quadrupole mode, as discussed in Sect. 2.4). Instead, it uses *WMAP* information to support the foreground modelling, and to constrain the poorly measured modes in LFI. Overall, these algorithmic differences lead to the observed deviations between the various solutions.

The relative differences are smaller for 44 and 70 GHz: −0.14 % (44 GHz) and −0.69 % (70 GHz) between BeyondPlanck and the *Planck* 2018 model, and 0.02 % (44 GHz) and -0.08 % (70 GHz) between BeyondPlanck and NPIPE. For 44 and 70 GHz the agreement between BeyondPlanck and NPIPE is generally excellent, while the 2018 model generally has slightly higher absolute calibration than other two in the 70 GHz channel. The main visual difference between BeyondPlanck and NPIPE lies in the fact that BeyondPlanck smoothes the gain solution after estimation, while NPIPE does not.

4.2. 44 GHz correlated noise stripes

As discussed by Suur-Uski et al. (2020), the BeyondPlanck 44 GHz correlated noise maps exhibit a number of strong "stripes" in the 44 GHz channel. This effect is shown in Fig. 12 for the Stokes' Q parameter, as reproduced from Suur-Uski et al. (2020). Significant effort has been spent on identifying the origin of these stripes, but no firm conclusions have been reached yet. One example of the tests that we have done is to reduce the smoothing window widths listed in Table 2 by a factor of two. In this case, we do note that the stripes disappear from $\boldsymbol{n}_{\rm corr}$, but at the same time, we also see large spikes appearing in the gain solution for PID ranges for which the scanning direction is aligned with the dipole minimum, and thus for which the signal-to-noise is low for purposes of gain estimation.

Another line of investigation has been to modify the processing masks shown in Fig. 4, noting that strong point sources or signal gradients can induce stripes like this. However, we have not yet been able to construct any mask that solves this problem, or, indeed, modifying the general behaviour of the stripes notably.

A third hypothesis that has been considered is the potential presence of undetected large gain jumps, similar to those listed in Table 1. In the current implementation, this list is adopted without modification from previous analysis pipelines. Noting that the 44 GHz radiometers generally appear more susceptible to non-thermal artefacts than the other two frequency channels, it is reasonable to speculate that some subtle hardware failures could affect one or more of the 44 GHz radiometers, similar to what is seen in the 44 GHz LFI 26S radiometer around PID 20 800. We have, however, not yet been able to identify any obvious gain jumps of this type, and this is therefore still only speculation.

Understanding and mitigating these stripes is clearly one of the top priorities for future BeyondPlanck analysis, as they currently appear to limit our ability to extract reliable CMB polarization information from the Southern Galactic hemisphere (Colombo et al. 2020; Paradiso et al. 2020).

4.3. Comparison with external data

To understand the combined impact of the various gain model differences discussed above, it is useful to compare the final BeyondPlanck frequency maps with externally processed observations, both from *WMAP* and *Planck*. In this respect, we note that

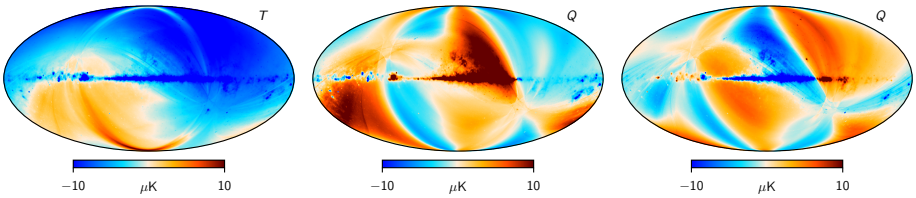

Fig. 13: Gain residual template for the LFI 30 GHz channel, produced by the *Planck* DPC through manual iteration between calibration, mapmaking and component separation (Planck Collaboration II 2020).

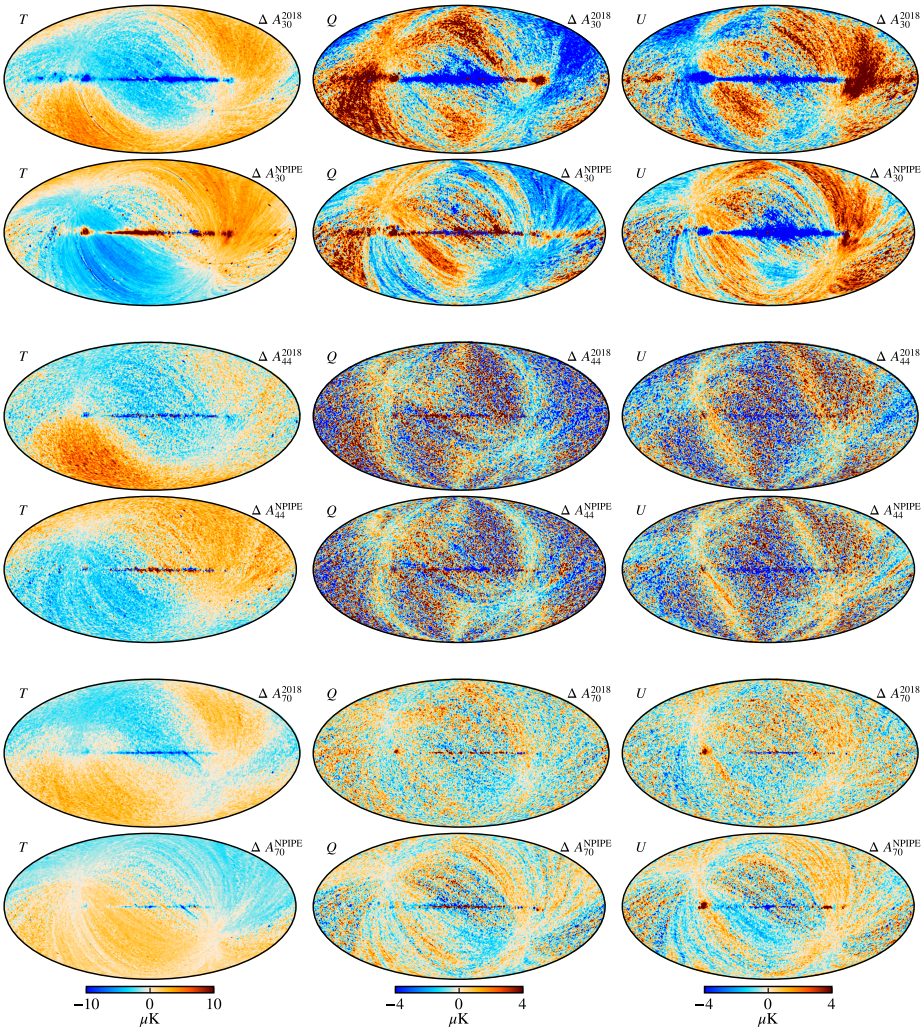

Fig. 14: Differences between BeyondPlanck and 2018 or NPIPE frequency maps, smoothed to a common angular resolution of 2° FWHM. Columns show Stokes T, Q and U parameters, respectively, while rows show pair-wise differences with respect to the pipeline indicated in the panel labels. A constant offset has been removed from the temperature maps, while all other modes are retained. The 2018 maps have been scaled by their respective beam normalization prior to subtraction. Reproduced from BeyondPlanck Collaboration (2020).

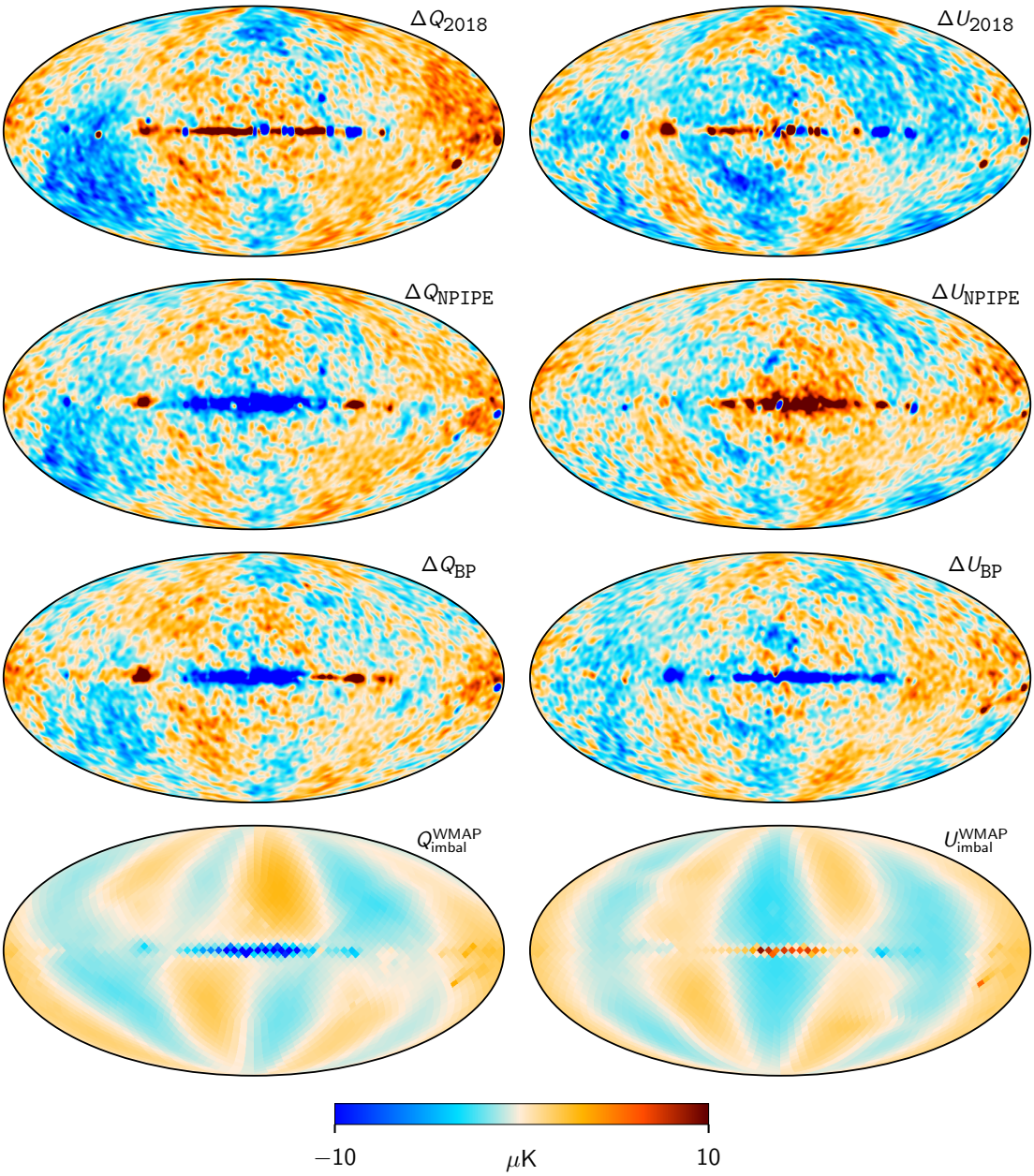

Fig. 15: Difference maps between the *Planck* 30 GHz and *WMAP K*-band maps for *Planck* 2018 (*first row*), NPIPE (*second row*), and BeyondPlanck (*third row*). All maps have been smoothed to a common angular resolution of 3° FWHM before evaluating the differences. The *WMAP K*-band map has been scaled by a factor of 0.495 to account for different center frequencies, assuming a synchrotron spectral index of $\beta_s = -3.1$. The bottom row shows one of the *WMAP K*-band transmission imbalance templates discussed by Jarosik et al. (2007), which accounts for known poorly measured modes in the *WMAP* data.

both the *Planck* 2018 and *WMAP* data sets are associated with sets of correction templates that track known systematic effects (or poorly measured modes) in the respective sky maps. For the *Planck* 2018 30 GHz channel, this template is shown in Fig. 13. As discussed by Planck Collaboration II (2020), this template was produced by iterating between calibration and component separation, and therefore trace uncertainties in the gain model due to foreground uncertainties. Furthermore, due to limited time, only four full iterations of this type was completed for the *Planck* 2018 analysis, and one must therefore expect that there are still residuals of this type present in the final sky maps at some level.

With this in mind, we show in Fig. 14 BeyondPlanck–NPIPE and BeyondPlanck–*Planck* 2018 difference maps for Stokes *I*, *Q*, and *U* (columns), for all three LFI frequencies (rows). The first, third and fifth rows show differences with respect to *Planck* 2018, while the second, fourth and sixth rows show differences with respect to NPIPE. Several features in these difference maps are interesting from the calibration perspective. Starting with the NPIPE temperature difference maps, we see that all three channels are dominated by a clean dipole-like residual aligned with the Solar CMB dipole. This shows that the BeyondPlanck and NPIPE temperature maps are morphologically very similar, but have different absolute calibration. We also see that the temperature map difference between BeyondPlanck and NPIPE exhibits a flip in the dipole direction going from 30 to 44 to 70 GHz. This sign change is consistent with the differences in the calibration factors between 44 and 70 GHz reported in Table 10 in Planck Collaboration Int. LVII (2020), finding a difference of 0.31 % between the absolute calibration of the 44 and 70 GHz channels. Since the CMB Solar dipole has an amplitude of about 3360 μK, this relative difference translates into an absolute temperature difference of roughly 10 μK in the observed sky signal, which is fully consistent with the dipole differences we see in Fig. 14. In comparison, the *Planck* 2018 temperature difference maps show a more prominent quadrupolar structure with a morphology that might resemble the effect of bandpass mismatch leakage (Planck Collaboration X 2016; Planck Collaboration Int. LVII 2020).

For polarization, the most striking differences are seen in the 30 GHz channel, for which variations at the 4 μK level are seen over large fractions of the sky. Furthermore, these residuals correlated very closely with the LFI DPC gain template shown in Fig. 13, suggesting that they are indeed caused by foreground-induced gain residuals. The same patterns are also seen in the NPIPE difference maps, but with a notably lower level.

For the 44 GHz maps, the visually most obvious feature is a high white noise level, which is due to BeyondPlanck rejecting data from horn 26 for more than half the mission because of noise instabilities (Ihle et al. 2020; Suur-Uski et al. 2020). In addition, we note that the Stokes *Q* difference maps show correlated noise stripes similar to those highlighted in Fig. 12. However, we also note that these structures have different amplitudes in the NPIPE and *Planck* 2018, and these stripes are therefore present in at least one of the other pipelines as well, and possibly both.

Figure 15 shows a similar comparison between the various *Planck* 30 GHz maps and the *WMAP* *K*-band channel (Bennett et al. 2013). In this case, all maps have been smoothed to a common angular resolution of 3° FWHM, and the *K*-band map has been scaled by a factor of 0.495 to account for the different center frequencies of the two maps while adopting a synchrotron-like spectral index of $\beta_s = -3.1$. From top to bottom, the first three rows show difference maps with respect to *Planck* 2018, NPIPE, and BeyondPlanck.

Overall, we see a clear progression in agreement with respect to *WMAP* *K*-band, in the sense that BeyondPlanck shows smaller residuals than NPIPE, which in turn shows smaller residuals than *Planck* 2018. Furthermore, we note that the strong residuals traced by the LFI gain template in Fig. 13 are most pronouced in the *Planck* 2018 map.

At the same time, we also observe significant coherent large-scale features in the difference map between BeyondPlanck and *K*-band. To at least partially understand these, we show the *WMAP* transmission imbalance templates derived by Jarosik et al. (2007) in the bottom row of Fig. 14. These templates trace poorly measured modes due to the differential nature of the *WMAP* instrument. Although corrections for this effect are applied to the final *K*-band sky maps, the uncertainty on the template amplitudes is estimated to 20 %. Considering the tight correlation between the BeyondPlanck–*WMAP* difference map and the transmission imbalance template, it seems clear that at least a significant fraction of the remaining residual may be explained in terms of this effect. Of course, this also suggests that a future joint analysis between *Planck* and *WMAP* in time-domain will be able to constrain the *WMAP* transmission imbalance parameters to much higher precision, and *Planck* data can thereby be used to break an important internal degeneracy in *WMAP*. As reported by Watts et al. (2020), this work has already started, but a full exploration of time-ordered *WMAP* data is outside the scope of the current analysis. We also emphasize that the current BeyondPlanck analysis only use low-resolution *WMAP* polarization data for which a full covariance matrix is available, and these modes are appropriately downweighted in those matrices.

5. Conclusions

We have presented the BeyondPlanck approach to gain calibration within the larger Commander Gibbs sampling framework. This framework relies directly on the Solar and orbital dipoles for relative and absolute calibration, respectively, and accounts for astrophysical foreground and instrumental confusion through global modelling.

One critically important difference with respect to previous *Planck* LFI analysis efforts is the fact that we actively use external data to break internal *Planck* degeneracies, and in particular *WMAP*observations. This significantly alleviates the need for imposing strong algorithmic priors during the calibration process. Most notably, while both the *Planck* 2018 and NPIPE pipelines assumed CMB polarization to be negligible on all angular scales during the calibration phase, we only assume that the CMB quadrupole is negligible. The reason we still make this assumption is that the *Planck* scanning strategy renders the CMB quadrupole very nearly perfectly degenerate with the CMB Solar dipole coupled to subtle gain fluctuations; a hypothetical future and well-designed satellite mission should not require this prior, as long as its scanning strategy modulates the CMB dipole on sufficiently short time-scales and with good cross-linking.

Overall, we find good agreement between the BeyondPlanck and previous gain models. The biggest differences are observed in the LFI 30 GHz channel, with gain variations of 0.84 % between *Planck* 2018 and BeyondPlanck. These differences result in subtle but significant temperature and polarization residuals. When comparing these with external *WMAP* *K*-band observations, it seems clear that the BeyondPlanck LFI maps are generally cleaner than previous renditions with respect to gain residuals. At the same time, we emphasize that these differences are also consistent with previously published error estimates, as presented by the *Planck* 2018 and NPIPE teams themselves. For in-

stance, the morphology of the *Planck* 2018 polarization residuals matches previously published LFI DPC gain residual templates (Planck Collaboration II 2020), and the NPIPE absolute calibration differences are fully consistent with internal NPIPE estimates (Planck Collaboration Int. LVII 2020). These results are thus neither novel nor surprising, but they simply highlight the inherent advantages of global analysis, using complementary data sets to break internal degeneracies.

Finally, we note that even though the procedures outlined in this paper have been aimed at modelling the LFI detectors, there is nothing about the data model or methodology that is unique for LFI. The method should be directly applicable for other data sets and experiments as well, and, indeed, a preliminary *WMAP* analysis is already underway (Watts et al. 2020).

Acknowledgements. We thank Prof. Pedro Ferreira for useful suggestions, comments and discussions, and Dr. Diana Mjaschkova-Pascual for administrative support. We also thank the entire *Planck* and *WMAP* teams for invaluable support and discussions, and for their dedicated efforts through several decades without which this work would not be possible. The current work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement numbers 776282 (COMPET-4; BeyondPlanck), 772253 (ERC; bits2cosmology), and 819478 (ERC; Cosmoglobe). In addition, the collaboration acknowledges support from ESA; ASI and INAF (Italy); NASA and DoE (USA); Tekes, Academy of Finland (grant no. 295113), CSC, and Magnus Ehrnrooth foundation (Finland); RCN (Norway; grant nos. 263011, 274990); and PRACE (EU).

References

Bennett, C. L., Larson, D., Weiland, J. L., et al. 2013, ApJS, 208, 20
BeyondPlanck Collaboration. 2020, A&A, in preparation [arXiv:2011.05609]
Colombo et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
de Bernardis, P., Ade, P. A. R., Bock, J. J., et al. 2000, Nature, 404, 955
de Oliveira-Costa, A., Tegmark, M., Zaldarriaga, M., & Hamilton, A. 2004, Phys. Rev. D, 69, 063516
Delouis, J. M., Pagano, L., Mottet, S., Puget, J. L., & Vibert, L. 2019, A&A, 629, A38
Fixsen, D. J. 2009, ApJ, 707, 916
Godard, B., Croon, M., Budnik, F., & Morley, T. 2009, in Proceedings of the 21st International Symposium on Space Flight Dynamics (ISSFD), Toulouse
Ihle et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Jarosik, N., Barnes, C., Greason, M. R., et al. 2007, ApJS, 170, 263
Jarosik, N., Bennett, C. L., Dunkley, J., et al. 2011, ApJS, 192, 14
Kogut, A., Fixsen, D. J., Chuss, D. T., et al. 2011, J. Cosmology Astropart. Phys., 2011, 025
Kovac, J. M., Leitch, E. M., Pryke, C., et al. 2002, Nature, 420, 772
Paradiso et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Planck Collaboration I. 2011, A&A, 536, A1
Planck Collaboration II. 2014, A&A, 571, A2
Planck Collaboration VIII. 2014, A&A, 571, A8
Planck Collaboration XXVIII. 2014, A&A, 571, A28
Planck Collaboration II. 2016, A&A, 594, A2
Planck Collaboration V. 2016, A&A, 594, A5
Planck Collaboration VIII. 2016, A&A, 594, A8
Planck Collaboration X. 2016, A&A, 594, A10
Planck Collaboration I. 2020, A&A, 641, A1
Planck Collaboration II. 2020, A&A, 641, A2
Planck Collaboration III. 2020, A&A, 641, A3
Planck Collaboration IV. 2020, A&A, 641, A4
Planck Collaboration Int. XLVI. 2016, A&A, 596, A107
Planck Collaboration Int. LVII. 2020, A&A, in press [arXiv:2007.04997]
Smoot, G. F., Bennett, C. L., Kogut, A., et al. 1992, ApJ, 396, L1
Suur-Uski et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]
Watts et al. 2020, A&A, in preparation [arXiv:201x.xxxxx]