# Hamiltonian Monte Carlo modelling of satellite magnetic data

Master's Thesis

Michael Madelaire
August 2019

Supervisor:
Chris Finlay

DTU Space
National Space Institute

**DTU Space**
**National Space Institute**
**Technical University of Denmark**

Elektrovej, building 327
DK - 2800 Kgs. Lyngby
Phone +45 4525 9500
www.space.dtu.dk

# Abstract

The Earth's magnetic field is generated by multiple sources. It is used by humans for navigation and orientation purposes, and its sources provide information on the physical properties and dynamics of Earth's interior and its environment. Classically, sources of the geomagnetic field are divided into internal and external components. When observing the field from Earth's surface the internal sources are a superposition of both the core field and the lithospheric field. Due to the nature of the potential field representation it is not easy to separate the internal sources from each other. For this reason, the core and lithospheric fields has up to now only been crudely separated at a certain spherical harmonics degree.

This thesis began with the aim of inferring the core field probabilistically. In order to accomplish this it was proved necessary to co-estimate both the core and lithospheric fields. The probabilistic inference method applied in this thesis is a Hamiltonian Monte Carlo scheme for sampling model space. By introducing an auxiliary momentum, this technique is able to utilize gradient information to quickly converge toward the mode of the posterior probability distribution even in high dimensional spaces. The sampling algorithm is implemented using the Stan open-source software package which compiles the problem into a highly optimized C++ program. Additionally the No-U-Turn algorithm is implemented; this automatically tunes the step-size and the amount of steps which can be amongst the largest drawbacks when using Hamiltonian Monte Carlo. The prior information regarding the core field is a time series of the spherical harmonic coefficients, truncated at degree 30, generated by a core dynamo simulation. This allows the creation of a multivariate prior distribution for the spherical harmonic coefficients representing the core field. The lithospheric prior probability distribution assumes the spherical harmonic coefficients independent with amplitude set based on a single forward realization of the lithospheric field based on geological information. This is assumed to be very weak prior information. The information on the core and lithospheric fields in the prior probability distributions are updated using observations of the magnetic field made by the Swarm satellite trio between August and November 2018. Applying this method it was found to be possible to co-estimate the core and lithospheric fields up to a spherical harmonic truncation degree of 22. When compared to a reference core field (CHAOS-6), between spherical harmonic degree 1-13, the resulting predictions are found to be very similar. Additionally, it was possible to reconstruct major lithospheric anomalies, despite using only 2000 observed data-points, and the lithospheric prior contained no information concerning spatial structures. Comparisons of the inferred lithospheric model with a reference lithospheric model (LCS-1), between spherical harmonic degree 15 and 20, showed agreement over major lithospheric anomalies except for Australia where the flux patches had opposite polarity. The inferred lithospheric model has a weak zonal pattern that may be a consequence of the weak diagonal lithospheric prior employed.

Overall, the performance of the Hamiltonian Monte Carlo algorithm is found to be very encouraging and a successful separation of the core and lithospheric fields was achieved. In order to utilize its full potential more prior information concerning the lithospheric field and a larger collection of core field prior realizations is needed. In the longer term it will also be important to extend the model to include a time-dependent core field and to make use of longer time series of satellite data.

# Preface

This thesis was prepared at the National Space Institute at the Technical University of Denmark in fulfillment of the requirements for acquiring a master's degree in Earth and Space Physics and Engineering.

DTU, August 16, 2019

Michael Madelaire (s144117)

# Acknowledgements

I would like to thank my supervisor, Chris Finlay, for the comments and suggestions he has provided. I would also like to thank Magnus D. Hammer for providing a data-set of Swarm observations and error estimates. Further, my thanks go to Dr. Julien Aubert for providing the time series of a core dynamo simulation that was used as part of the prior information. Additionally, I would like to thank Andreas Nilsson for giving me an introduction on how to define models in Stan. Finally, I would like to thank the Stan community for answering all my questions, ranging from theory to methodology.

# Contents

# CHAPTER 1

# Introduction

The Earth is surrounded by a magnetic field that extends far beyond its atmosphere. Interaction between Earth's magnetic field and the solar wind creates an abrupt transition between the interplanetary magnetic field and Earth's magnetic field, known as the magnetopause. The domain enclosed by the magnetopause contains all sources that contribute to the Earth's magnetic field. Classically, from the perspective of observations at Earth's surface the magnetospheric and ionospheric currents are seen as external sources. While the internal sources are the core and lithospheric field, illustrated in figure 1.1. This necessarily changes with respect to the point of reference.
Contributions from the magnetosphere and ionosphere differ quite distinctly from the core and lithospheric field. They are powered by the solar wind and therefore change on a timescale of minutes and hours (Olsen and Stolle 2012). This allows induction of currents in the mantle, which creates its own secondary magnetic field and is an internal contribution.
The core field changes on a timescale of years and contribute ∼95 % of the field measured at Earth's surface while the lithospheric field only contribute a few percent, but it varies geographically (Olsen and Stolle 2012).

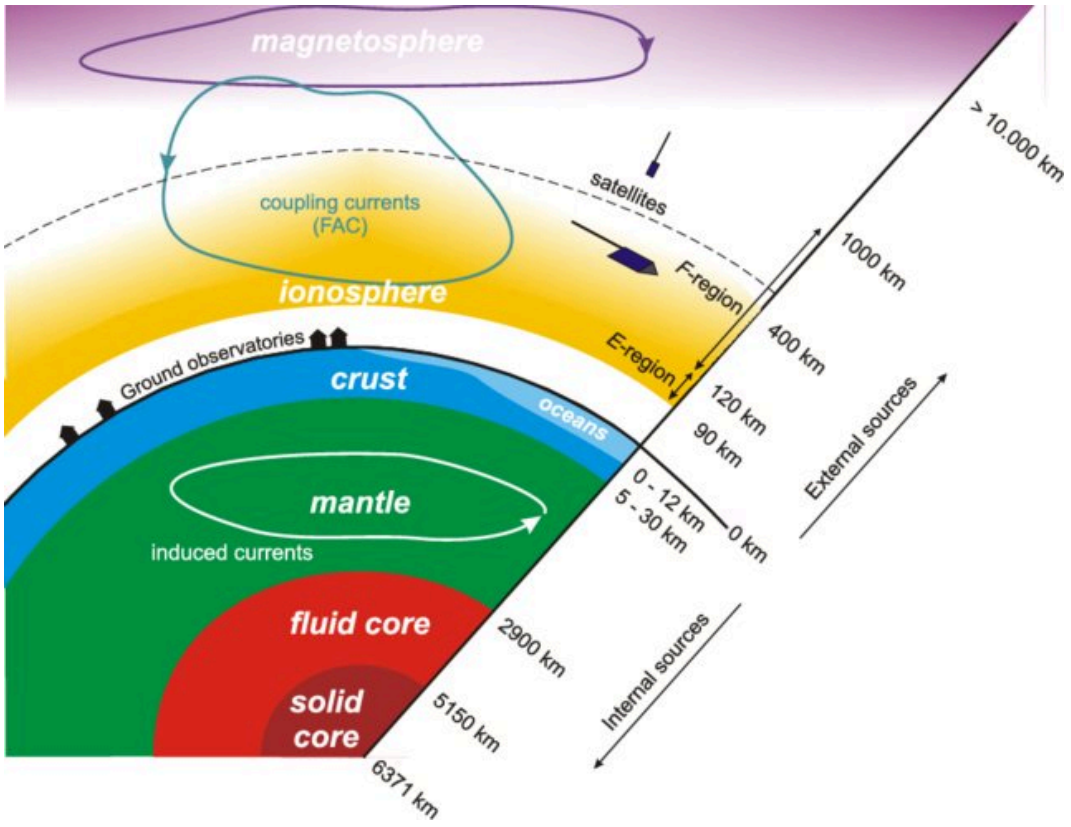



**Figure 1.1:** Illustration of the locations of the sources that contribute to Earth's magnetic field. Image taken from Olsen et al. 2010.

## 1.1 Core and lithospheric field

When interested in isolating the core and/or lithospheric field, one typically considers a period of time with low solar activity. In this way external sources and related induced currents are minimized. Data selected in this manner is called geomagnetically quiet data. There are several geomagnetic indices used to define whether or not conditions are geomagnetically quiet. Quiet data is typically selected on Earth's night side when certain thresholds of the geomagnetic indices are not exceeded (Olsen and Stolle 2012).
Assuming this type of data selection is enough to remove the largest effects from the magnetosphere and ionosphere it still leaves a superposition of internal sources (Baratchart and Gerhards 2017). In figure 1.2 the power spectrum of Earth's internal field is shown as a function of spherical harmonic degree and wavelength, see sections 2.1 and 3.3.6 for definitions on spherical harmonics and power spectrum. The core field is generated deep within the core of the planet and therefore dominates the lower harmonics. As the wavelength decreases so does the power. At approximately degree 15 the lithospheric field takes over as the dominating source. The superpositioned field is often crudely separated at spherical harmonic degree 14.

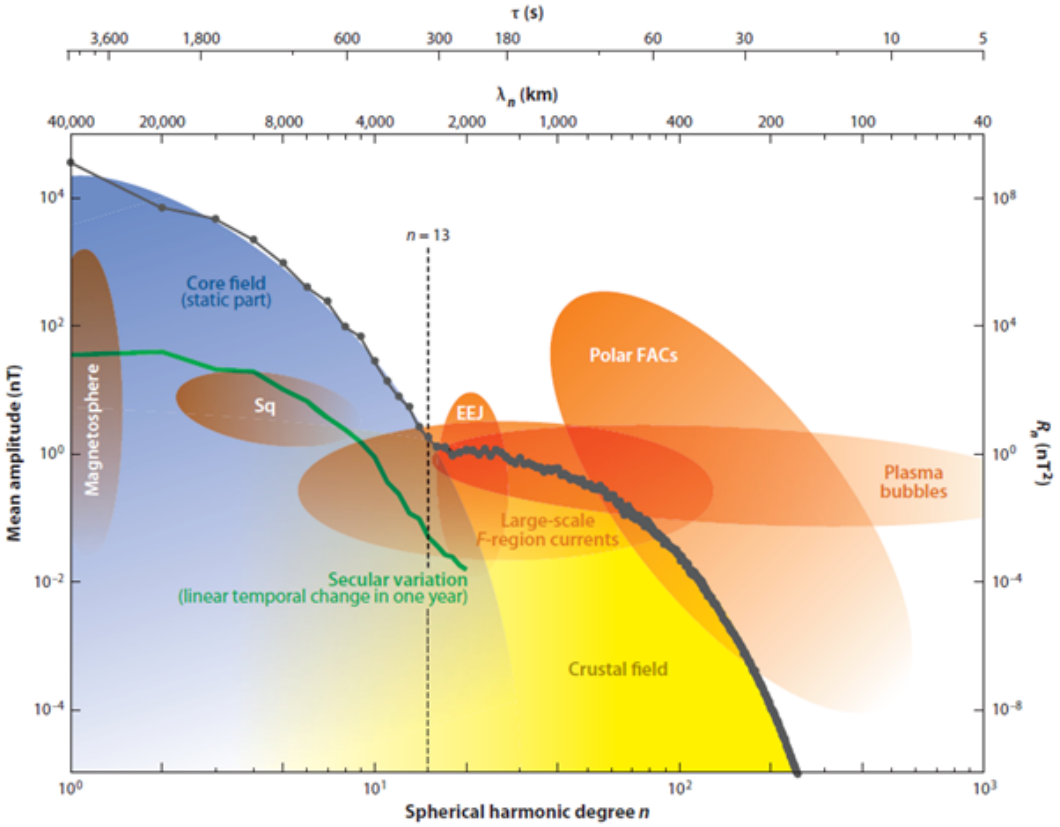



**Figure 1.2:** Power spectrum of the Earth's geomagnetic field as generated by the CHAOS-4 model at 350 km altitude (Olsen and Stolle 2012).

## 1.2 The problem: Source separation

Separating the core field by truncating the spherical harmonic expansion, equation 2.6, at degree 14 as suggested above will not only contaminate it with contributions from the lithospheric field, but also lose shorter wavelength information of the core field.
This method was suggested several decades ago (Lowes 1974) when the core field was said to be dominant to at least degree 11. Later the core was said to dominate to degree 13 while the

lithospheric field was dominant from degree 15 and above (Langel and Estes 1982, Cain et al. 1989). A newer study puts these boundaries at spherical harmonics degree 12 and 15 while the degrees between are some intermediate zone (Voorhies et al. 2002).
Studies of Mars' lithospheric field show that there is a contribution from lithospheric sources at lower spherical harmonic degrees (Voorhies et al. 2002). It is therefore reasonable to think that this behaviour is also present on Earth.

The question at hand is thus how to effectively separate the two sources. This thesis seeks to contribute with a new method of modelling the core field. By introducing prior information of the correlation between spherical harmonic coefficients, considering separately the core and lithospheric sources it might be possible to constrain how much they fit the lithospheric field. The ability to distinguish between sources of the geomagnetic field will ultimately allow the study of the core's small scale structure and the lithospheric field's large scale structure. This is interesting for further understanding on the Earth's interior.

## 1.3 Approach of the thesis

Classically, the geomagnetic field is modelled using a regularized least squares approach (Bloxham et al. 1989, Olsen et al. 2006) minimizing prediction errors while finding a suitable trade-off between model complexity and prediction errors.
The approach taken in this thesis will be Bayesian allowing for the consistent introduction of prior information. Solving the problem probabilistically will also provide uncertainties on the resulting field models which is crucial when evaluating any scientific result.
Although the probabilistic approach has been around for some time its uses has not yet become common in the community. For example, its use has been suggested to analyse fluid motion in the Earth's core (Mosegaard and Rygaard-Hjalsted 1999). A probabilistic approach for modelling the historical field between 1840-2010 was presented in Gillet et al. 2013. Newer techniques include simultaneous modelling and separation of the geomagnetic field contributions (Holschneider et al. 2016). Only observatory data has been used so far and thus unable to separate core and lithospheric sources. The paleomagnetic community has also begun to move in the direction of probabilistic modelling due to the sparse nature of their observations and uncertainties related to lock-in and age-depth in sediment records (Nilsson et al. 2018).
In the following chapters it will be revealed that the inverse problem solved is linear, the prior and likelihood is Gaussian and the resulting posterior distribution is Gaussian as well. At this point the reader might question the necessity of using a Bayesian approach. Here it should be noted that this thesis tests a new platform that if successful can be extended to include non-linear problems and non-Gaussian distributions in future studies.

## 1.4 Outline of the thesis

Here is a brief outline of how the thesis will proceed along with summaries of what can be found in each chapter.

Chapter 2 gives the theoretical background of the geomagnetic field and how it is expressed mathematically through spherical harmonics, including the necessary assumptions. Additionally, the chapter gives a brief introduction to the Bayesian approach to inverse problems. Followed by a more in-depth presentation of Hamiltonian Monte Carlo and the associated No-U-

Turn algorithm.

Chapter 3 presents the software implementation of Hamiltonian Monte Carlo that is chosen for this thesis, while discussing alternative options. The chapter continues with an example of how Bayesian models are implemented, followed by a presentation of the models applied in this thesis along with several hyperparameters. Finally, the chapter presents the diagnostic tools that will be used.

Chapter 4 presents and illustrates the data that will be used. First the prior information about the core and lithospheric fields is presented. Followed by an explanation of how synthetic data used for benchmark tests is made. Finally, the real satellite data is presented along with the data selection criteria that has been applied and the method used for data error estimation.

Chapter 5 contains the results. First of benchmark tests, with synthetic data, that are used for justifying the selected hyperparameters are presented. This is followed by attempts at estimating the core field using real satellite data. Finally, the chapter presents benchmark tests of co-estimation with synthetic data followed by attempts using real satellite data.

Chapter 6 discusses and compares the final results of chapter 5 to well known models. Additionally, limitations to the performance discovered throughout the thesis is presented. This is followed by suggestions for future work and improvements. Chapter 7 concludes on the results presented and the method applied and gives an outlook for future studies.

# CHAPTER 2

# Theory

In this section the immediate theory necessary to understand the methodology is presented. It will contain a short section on how the Earth's geomagnetic field can be expressed in terms of spherical harmonics, along with a brief introduction to the Bayesian approach to inverse problems. Finally, there will be a more in depth explanation of Hamiltonian Monte Carlo and the self adjusting algorithms in NUTS.

## 2.1 The geomagnetic field and spherical harmonics

A magnetic field can be expressed by Maxwell's equations

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J} + \varepsilon_0 \mu_0 \frac{\partial \boldsymbol{E}}{\partial t} \tag{2.1}$$

Here $\boldsymbol{B}$ is the magnetic field, $\mu_0$ is the magnetic vacuum permability, $\boldsymbol{J}$ is the electric current density, $\varepsilon_0$ is the vacuum permability and $\boldsymbol{E}$ is the electric field.
Equation 2.1 can be reduced by applying the quasi-static approximation. Meaning that the displacement current can be neglected if the changes are assumed sufficiently slow, resulting in equation 2.2.

$$\boldsymbol{\nabla} \times \boldsymbol{B} = \mu_0 \boldsymbol{J} \tag{2.2}$$

Sufficiently slow implies $T >> \frac{L}{c}$. Where $T$ and $L$ are typical time and length scales, respectively, $c$ is the speed of light. In the context of Earth's geomagnetic field $L \approx 40.000$ km, thus the approximation holds if $T >> 1/8$ s. Note that the typical timescale of secular variations is much larger.
It can further be assumed that there are no free currents, $\boldsymbol{J} = 0$, if data selection is done carefully. As a consequence the curl of $\boldsymbol{B}$ is zero which implies that it can be written as function of a potential field, $V$, equation 2.3.

$$\boldsymbol{B} = -\boldsymbol{\nabla} V \tag{2.3}$$

By utilizing that the divergence of a magnetic field is zero and the newly derived equation 2.3 Laplace's equation can be derived as

$$\begin{aligned} & \boldsymbol{\nabla} \cdot \boldsymbol{B} = 0 \\ => \quad & \boldsymbol{\nabla} \cdot (-\boldsymbol{\nabla} V) = 0 \\ => \quad & -\boldsymbol{\nabla}^2 V = 0 \end{aligned} \tag{2.4}$$

Within the domain of the above assumptions the magnetic field $\boldsymbol{B}$ is fully described by the potential field $V$. The solution to equation 2.4 can be found through separation of variables and is the spherical harmonic (**SH**) expansion, equation 2.5.

$$\begin{aligned} V(r,\theta,\lambda) = a \sum_{n=1}^{\infty} \sum_{m=0}^{n} \Bigg( & \left[ g_n^m \cos(m\lambda) + h_n^m \sin(m\lambda) \right] \left( \frac{a}{r} \right)^{n+1} + \\ & \left[ q_n^m \cos(m\lambda) + s_n^m \sin(m\lambda) \right] \left( \frac{r}{a} \right)^{n} \Bigg) P_n^m \left( \cos(\theta) \right) \end{aligned} \tag{2.5}$$

Here $a$ is the Earth's radius 6371 km, $r$ is radius, $\theta$ is latitude, $\lambda$ is longitude, $n$ is the SH-degree, $m$ is the SH-order, $(g_n^m, h_n^m, q_n^m, s_n^m)$ are the Gauss coefficients and $P_n^m(\cos(\theta))$ is the Legendre polynomial.
Equation 2.5 describes both the internal $(g_n^m, h_n^m)$ and external $(q_n^m, s_n^m)$ sources. The external sources can be neglected when only the internal is of interest. Thus the solution reduces to

$$V(r,\theta,\lambda) = a\sum_{n=1}^{\infty}\sum_{m=0}^{n}\left([g_n^m\cos(m\lambda) + h_n^m\sin(m\lambda)]\left(\frac{a}{r}\right)^{n+1}\right)P_n^m(\cos(\theta)) \tag{2.6}$$

The magnetic components, equations 2.7, can be derived by differentiating the potential, equation 2.6, with respect to $r$, $\theta$ and $\phi$.

$$\begin{aligned}
B_r &= \sum_{n=1}^{\infty}\sum_{m=0}^{n}\left[-(n+1)\left(g_n^m\cos(m\phi) + h_n^m\sin(m\phi)\right)\left(\frac{a}{r}\right)^{n+2}P_n^m(\cos(\theta))\right]\\
B_\theta &= a\sum_{n=1}^{\infty}\sum_{m=0}^{n}\left[\left(g_n^m\cos(m\phi) + h_n^m\sin(m\phi)\right)\left(\frac{a}{r}\right)^{n+1}\frac{dP_n^m(\cos(\theta))}{d\theta}\right]\\
B_\phi &= a\sum_{n=1}^{\infty}\sum_{m=0}^{n}\left[\left(-g_n^m\sin(m\phi) + h_n^m\cos(m\phi)\right)\left(\frac{a}{r}\right)^{n+1}P_n^m(\cos(\theta))\right]
\end{aligned} \tag{2.7}$$

The forward problem is given by rewritting, equations 2.7, into matrix form, such that all three components of the magnetic field are given as

$$\boldsymbol{B}_i = \boldsymbol{G}_i\boldsymbol{m} \tag{2.8}$$

where $i$ denotes the spherical coordinate components $r$, $\theta$ and $\phi$. Thus the observations $\boldsymbol{B}_i$ have size $1\times N$, the Gauss coefficients $\boldsymbol{m}$ size $1\times M$ and the data kernel $\boldsymbol{G}_i$ size $N\times M$.
For simplicity the three components are combined such that

$$\begin{gathered}
\boldsymbol{d} = \boldsymbol{G}\boldsymbol{m}\\
\boldsymbol{d} = \begin{bmatrix}\boldsymbol{B}_r\\ \boldsymbol{B}_\theta\\ \boldsymbol{B}_\phi\end{bmatrix}\quad,\quad \boldsymbol{G} = \begin{bmatrix}\boldsymbol{G}_r\\ \boldsymbol{G}_\theta\\ \boldsymbol{G}_\phi\end{bmatrix}
\end{gathered} \tag{2.9}$$

making $\boldsymbol{d}$ and $\boldsymbol{G}$ size $1\times 3N$ and $3N\times M$, respectively.
Finally, the inverse problem being solved in this thesis, equation 2.10, is found by isolating $\boldsymbol{m}$.

$$\boldsymbol{m} = \boldsymbol{G}^{-1}\boldsymbol{d} \tag{2.10}$$

## 2.2 The Bayesian approach to inverse problems

With the inverse problem established, equation 2.10 it is time to define the method with which it will be solved. In this section the very basics of the Bayesian approach to solving inverse problems will be presented.
Consider two parameters $\boldsymbol{d}, \boldsymbol{m}$. They have a joint probability that can be expanded using the product rule, equation 2.11.

$$p(\boldsymbol{d},\boldsymbol{m}) = p(\boldsymbol{d}|\boldsymbol{m})p(\boldsymbol{m}) = p(\boldsymbol{m}|\boldsymbol{d})p(\boldsymbol{d}) \tag{2.11}$$

In the scenario where $\boldsymbol{d}$ and $\boldsymbol{m}$ are data and model coefficients, respectively, it is the probability of the model given the data, $p(\boldsymbol{m}|\boldsymbol{d})$, that is of interest. Isolate it and Bayes' theorem is derived, equation 2.12.

$$p(\boldsymbol{m}|\boldsymbol{d}) = \frac{p(\boldsymbol{d}|\boldsymbol{m})p(\boldsymbol{m})}{p(\boldsymbol{d})} \tag{2.12}$$

$p(\boldsymbol{m})$ is the prior probability defined by the prior information about the parameters in $\boldsymbol{m}$.
$p(\boldsymbol{d}|\boldsymbol{m})$ defines the probability of the data given a model. This is typically determined by assigning some distribution to the observational errors. The product of $p(\boldsymbol{d}|\boldsymbol{m})$ and $p(\boldsymbol{m})$ defines the shape of the posterior probability, $p(\boldsymbol{m}|\boldsymbol{d})$.
$p(\boldsymbol{d})$, commonly referred to as the evidence, serves to normalize the posterior distribution. Whether or not $p(\boldsymbol{d})$ is included relies heavily of the purpose. When it comes to parameter estimation it can be neglected. When comparing two different sets of prior information understanding which results in the largest probability mass can be a significant help.
Applying the Bayesian approach does not result in a single answer, but a distribution. The "best" solution is the one with the largest posterior probability and if the posterior is Gaussian, or otherwise symmetric, it is the mode (Aster et al. 2013).

If the problem at hand only has a couple of dimensions, and $p(\boldsymbol{d}|\boldsymbol{m})$ and $p(\boldsymbol{m})$ are rather simple, then ensuring the entire posterior distribution is fully explored is trivial since all possible values of the model parameters could be tested. This is also called an exhaustive search. When a model is more complex it is simply not an option to explore all possibilities due to an unrealistic computational time. Instead a smarter approach is needed to explore areas of interest such as the mode and the area around it.

The process of evaluating a subset of probability space in order to represent the full posterior distribution is called sampling. One of the first techniques was a so called random walk, which simply "walks" around and chooses points to be evaluated pseudo randomly. One could then stop sampling when the posterior distribution no longer changed, indicating that it might have converged. For high dimensional cases or multimodal distributions this is very ineffective, but has led to the development of some very powerful sampling tools.

## 2.3 Hamiltonian Monte Carlo

The inverse problem has now been established along with the fundamentals of the Bayesian approach. The question left unanswered in the previous section is what sampling algorithm is to be used. The SH inverse problem presented quickly increases in dimensionality with increasing SH-degree. The classical MCMC approaches can therefore not be used. Instead the Hamiltonian Monte Carlo (**HMC**) algorithm will be applied. The following section is a walkthrough of the theory behind HMC and how it conceptually can be understood. This section will closely follow Betancourt 2018.

HMC is a Bayesian inference tool and is therefore used to sample a posterior distribution. The purpose of any sampling algorithm is to converge towards the mode while exploring the area around it. The mode itself and the area closest to it has a high probability density, but a small volume resulting in little information also referred to as probability mass. Oppositely, the tail of the distribution has a large volume, but low probability density again resulting in a small probability mass. Thus the area of interest is between the tail and mode, referred to as the typical set, and is an area of large probability mass. The HMC algorithm combines two techniques with the purpose of quickly converging toward the mode and exploring the typical set. The relative size of the typical set, with respect to the surrounding area, decreases with an increasing dimensionality. This effect is often referred to as the curse of dimensionality. In low dimensional cases the typical set is fairly large making classical MCMC approaches likely to work well, but in high dimensional cases an algorithm such as Metropolis-Hastings would be inefficient. The sampler randomly suggests the next step which after evaluation is accepted

or rejected. At any instance there will be more points leading away from the mode than towards it. The sampler will eventually converge toward the mode, but with an extremely low acceptance rate and not within the finite amount of time available in practical applications. HMC circumvents this problem by exploiting the differential geometry of the posterior distribution. Imagine the posterior distribution as a topographic map where contour lines define areas of equal probability. With that information it is trivial to converge toward the mode reducing time spent in the tail, but the mode is not of much interest. The question is thus how to converge toward the mode while staying on and following the contour lines of the typical set.

A good analogy is orbital motion. Imagine a physical system as illustrated in figure 2.1. A satellite placed at a distance from the Earth with no initial velocity will be pulled towards the Earth and eventually crash, similar to the sampler getting stuck at the mode if only utilizing gradient information. Applying a momentum will cause it not to fall straight toward the Earth. If the momentum is insufficient or too large, figures B and C, the satellite will either crash or leave orbit. With just the right amount of momentum the satellite will orbit the Earth perfectly, figure D. If the Earth is replaced with the mode, satellite with sampler, gravity with gradient and orbital path with the typical set then we are back to the original problem. Thus adding an auxiliary momentum to the sampler will create a vector field which will lead the sampler around the typical set, as shown in figure D.

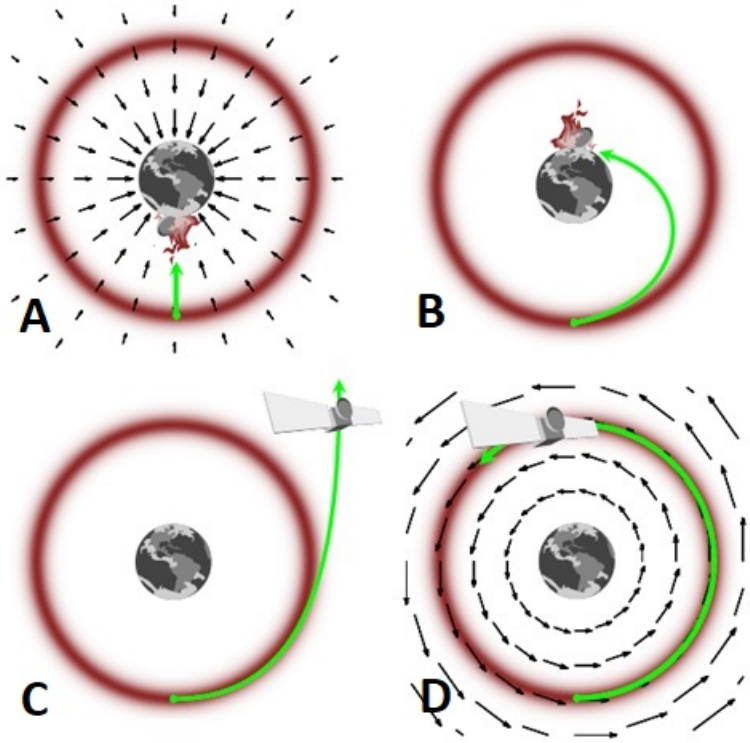



**Figure 2.1:** Conceptual illustration of HMC. A: The sampler only follows gradient information and propagates toward the mode. B: Momentum is added, but not enough, and the sampler propagates toward the mode. C: Momentum is added, too much, and the sampler leaves the typical set and propagates out into the tail. D: Momentum is added, just the right amount, and the sampler stays in the typical set. Image taken from Betancourt 2018.

Consider a model $\boldsymbol{m}$ in model space $\mathbb{M}$ and data $\boldsymbol{d}$ in data space $\mathbb{D}$. Prior to introducing an auxiliary momentum every point in model space can be described by a vector $\boldsymbol{m}$ of dimension $N$. After adding momentum, $\boldsymbol{p} \in \mathbb{P}$, the space has doubled in size. Thus all points are now described by a $2N$-dimensional vector. The $2N$-dimensional space is referred to as phase space,

equation 2.13 (Fichtner et al. 2019).

$$\mathbb{X} = \mathbb{M} \times \mathbb{P} \tag{2.13}$$

In phase space the posterior distribution is a canonical distribution defining the probability of the system given a certain energy. The joint probability of $\boldsymbol{m}$ and $\boldsymbol{p}$ can be expanded by the product rule, equation 2.14.

$$p(\boldsymbol{m}, \boldsymbol{p}) = p(\boldsymbol{p}|\boldsymbol{m})p(\boldsymbol{m}) \tag{2.14}$$

By defining $\boldsymbol{p}$ to be conditional on $\boldsymbol{m}$ the posterior distribution can be retrieved by marginalizing over $\boldsymbol{p}$. The joint probability can also be expressed by an invariant Hamiltonian function, equation 2.15.

$$p(\boldsymbol{m}, \boldsymbol{p}) = e^{-H(\boldsymbol{m},\boldsymbol{p})} \tag{2.15}$$

$H(\boldsymbol{m}, \boldsymbol{p})$ describes the probabilistic structure of phase space and thereby the typical set. Often $H$ is referred to as the energy. By inserting equation 2.14 into 2.15 and isolating for $H$ it can be rewritten using $K$ and $U$, equation 2.16.

$$\begin{aligned} H(\boldsymbol{m}, \boldsymbol{p}) &= -\log\left[p(\boldsymbol{m}, \boldsymbol{p})\right] \\ H(\boldsymbol{m}, \boldsymbol{p}) &= -\log\left[p(\boldsymbol{p}|\boldsymbol{m})\right] - \log\left[p(\boldsymbol{m})\right] \\ H(\boldsymbol{m}, \boldsymbol{p}) &= K(\boldsymbol{m}, \boldsymbol{p}) + U(\boldsymbol{m}) \end{aligned} \tag{2.16}$$

$K(\boldsymbol{m}, \boldsymbol{p})$ and $U(\boldsymbol{m})$ are often referred to as kinetic and potential energies, respectively. As mentioned, $H(\boldsymbol{m}, \boldsymbol{p})$, captures the geometry of the typical set and can therefore be used to determine a vector field, as shown in figure 2.1. This vector field can be generated from Hamilton's equations

$$\begin{aligned} \frac{d\boldsymbol{m}}{dt} &= \frac{\partial H}{\partial \boldsymbol{p}} = \frac{\partial K}{\partial \boldsymbol{p}} \\ \frac{d\boldsymbol{p}}{dt} &= -\frac{\partial H}{\partial \boldsymbol{m}} = -\frac{\partial K}{\partial \boldsymbol{m}} - \frac{\partial U}{\partial \boldsymbol{m}} \end{aligned} \tag{2.17}$$

Following the vector field for a time, $t$, generates a path $\Phi_t(\boldsymbol{m}, \boldsymbol{p})$. The path projected back from phase space to model space will describe a path within the typical set, given that the initial $\boldsymbol{m}$ also was in the typical set. This concept gives rise to a simple sampling algorithm; A random $\boldsymbol{m}$, inside the typical set, is chosen as the starting point. Then a momentum is determined by sampling $\boldsymbol{p} \sim p(\boldsymbol{p}|\boldsymbol{m})$. Integrate over Hamilton's equation for a time $t$, leaving a new joint probability $p(\boldsymbol{m}, \boldsymbol{p})$. Project from phase space into model space by marginalizing over $\boldsymbol{p}$, leaving a new sample of $\boldsymbol{m}$. Repeat by using the new sample as starting point.
This gives rise to some questions such as what $K$ and $U$ are, and how long the integration time $t$ should be. $U$ is the negative logarithm of the posterior distribution, which is the product of the prior and likelihood.
$K$, on the other hand is more complicated, but by rethinking phase space it is made easier. Previously the posterior distribution was suggested visualized as a contour map of probability densities. A similar thought is now employed to the phase space, but here the concentric lines represent an energy level $E$. Thus the inverse function of $H$ will point to a $2N-1$-dimensional plane of equal energy, equation 2.18.

$$H^{-1}(E) = \{\boldsymbol{m}, \boldsymbol{p}|H(\boldsymbol{m}, \boldsymbol{p}) = E\} \tag{2.18}$$

In this way all positions in phase space can be described by an energy, $E$, and a position, $\theta_E$, on that energy level. The canonical distribution can be reworked as equation 2.19.

$$p(\boldsymbol{m}, \boldsymbol{p}) = p(\theta_E|E)p(E) \tag{2.19}$$

Consequently, exploring phase space divides into two parts. A deterministic exploration of an energy level along the Hamiltonian path. And a stochastic transition between energy levels. Seen in this way $p(E|\boldsymbol{m})$ is a distribution of energies given a position. Instead of projecting to and from model space a new energy level can be sampled from $p(E|\boldsymbol{m})$.

Fast and effective exploration of an energy level depends a lot on how uniform and regular the levels are. This strongly depends on the choice of the kinetic energy function, which along with the posterior will shape the geometry of phase space and the energy levels within.
There are an infinite amount of possible kinetic energy functions that could be used, but typically Euclidean-Gaussian Kinetic Energies (**EGKE**) are chosen. The Euclidean metric, $\boldsymbol{g}$, of a system can be used to determine quantities such as distances between parameters, equation 2.20.

$$\Delta\left(\boldsymbol{m}-\boldsymbol{m}'\right)=\left(\boldsymbol{m}-\boldsymbol{m}'\right)^{T}\boldsymbol{g}\left(\boldsymbol{m}-\boldsymbol{m}'\right) \tag{2.20}$$

But it can also be used to rotate and scale the system, equation 2.21.

$$\boldsymbol{M}=\boldsymbol{RSgS}^{T}\boldsymbol{R}^{T} \tag{2.21}$$

Here $\boldsymbol{S}$ and $\boldsymbol{R}$ scale and rotate, respectively. And $\boldsymbol{M}$, commonly called the mass matrix, can transform the entire model space. If the mass matrix is a good approximation of the true posterior covariance matrix it can decorrelate the system and transform it into a Gaussian space. When transforming model space it is important to remember that the Hamiltonian is invariant and therefore the momentum rotates and scales equally and opposite to it, $\boldsymbol{p}'=\sqrt{\boldsymbol{M}^{-1}}\boldsymbol{p}$. The EGKE is given as in equation 2.22. Selecting a correct mass matrix will help create uniform energy levels in phase space reducing integration time and increasing the general sample speed.

$$K(\boldsymbol{m},\boldsymbol{p})=\frac{1}{2}\boldsymbol{p}^{T}\boldsymbol{M}^{-1}\boldsymbol{p}+\log\left(|\boldsymbol{M}|\right)+const \tag{2.22}$$

Beside EGKE there are other kinetic energy families. Riemannian-Gaussian kinetic energies depends on position allowing for local corrections. This is especially effective if the posterior distribution is not Gaussian.

The integration time, $t$, determines how much of the Hamiltonian trajectory is traversed and therefore how efficient it is. If $t$ is too large then the sampler will complete the path and return to an already explored area. Setting $t$ too low will result in not exploring enough. A single, well tuned, integration time does not exists since it is dependent on the energy level set. Consequently the optimal integration time will increase as the sampler moves from the typical set towards the tail. To optimize exploration it is therefore necessary to determine a unique $t$ for each level.

## 2.4 NUTS and automatic tuning

In the previous section HMC was explained and it was mentioned that the ideal integration time depends on the individual energy level set. This is one of the biggest challenges when using HMC. Fortunately an automatic system for determining it has been developed. It is called the No-U-Turn Sampler (**NUTS**) (Hoffman and Gelman 2014).
The NUTS adaptation to HMC involves two simple boundary conditions that both needs to be satisfied before the exploration of a trajectory is terminated.

$$\begin{aligned}\boldsymbol{p}_{+}(t)^{T}\left(\boldsymbol{m}_{+}(t)-\boldsymbol{m}_{-}(t)\right)<0\\ \boldsymbol{p}_{-}(t)^{T}\left(\boldsymbol{m}_{-}(t)-\boldsymbol{m}_{+}(t)\right)<0\end{aligned} \tag{2.23}$$

Here $\boldsymbol{m}_-(t)$ can be viewed as the initial position when going from model to phase space at $t = 0$. $\boldsymbol{m}_+(t)$ is the position as it evolves with time. The criterion can be illustrated as a line between $\boldsymbol{m}_-(t)$ and $\boldsymbol{m}_+(t)$ on the energy level set, figure 2.2a. When the momentum of each point is anti-aligned and perpendicular to the line connecting them, then the termination criterion is on the boundary of being satisfied. Note that $\boldsymbol{p}_-(t)$ is shown as negative, meaning that they have to be aligned in the illustration.

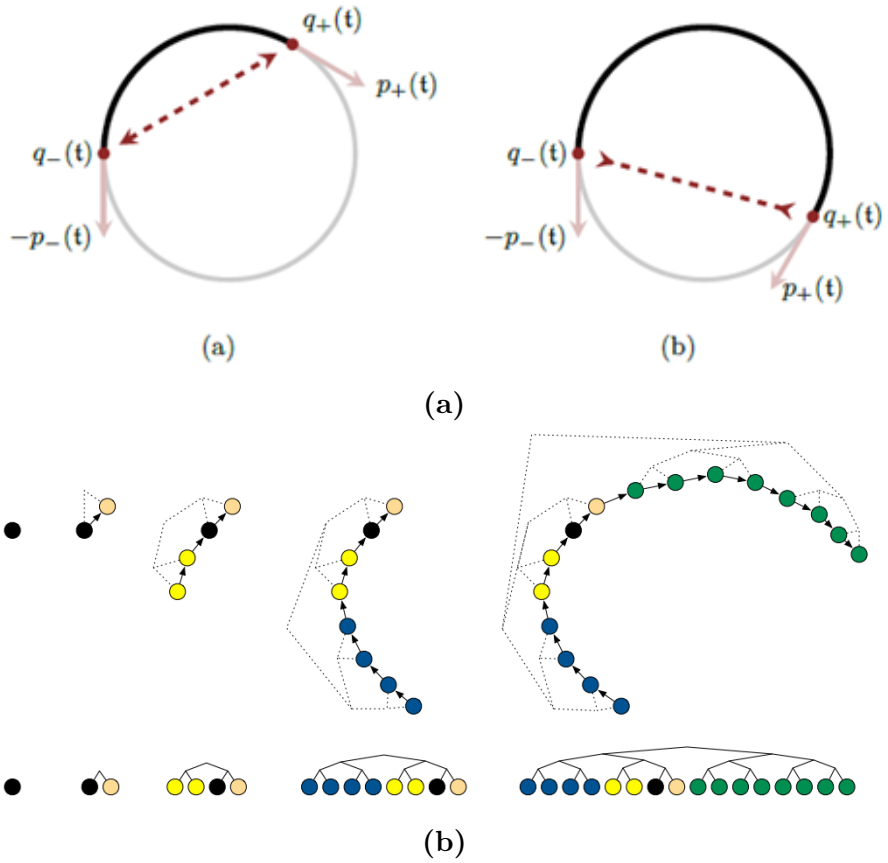



**Figure 2.2:** 2.2a: Illustration of the NUTS termination criterion for the dynamic integration time $t$. Note that $\boldsymbol{q} = \boldsymbol{m}$. Image taken from Betancourt 2018. 2.2b: Illustration of the integration scheme employed by the NUTS algorithm to preserve time reversibility. From the initial point (black) there is integrated randomly forwards or backwards starting with one leapfrog step which then doubles every time. Image taken from Hoffman and Gelman 2014.

The integration time, $t$, defines a continuous regime, while in practice it is discrete. It is thus necessary to break the integration time up into a step-size, $\varepsilon$, and an amount of steps, $L$. Any numerical integrator wont necessarily preserve the Hamiltonian, but can stray from the energy level. Implementing a symplectic integrator can solve this problem since it will preserve the Hamiltonian, although it can diverge if it experiences an area of high curvature. In NUTS the leapfrog integrator is implemented (Hoffman and Gelman 2014).
The step-size chosen should be large enough to avoid unnecessary computations while small enough such that the sampler does not stray from the trajectory. An optimal $\varepsilon$ is chosen by scaling it to fit a certain acceptance rate. The target acceptance rate should be between 0.6 to 0.9 (Monnahan et al. 2017).
The introduction of an acceptance rate is necessary because the chosen integrator cannot pre-

serve the Hamiltonian perfectly. A point is thus accepted with probability

$$\min\left[1, \frac{e^{-H(\boldsymbol{m}_t,\boldsymbol{p}_t)}}{e^{-H(\boldsymbol{m}_0,\boldsymbol{p}_0)}}\right] = \min\left[1, e^{-H(\boldsymbol{m}_t,\boldsymbol{p}_t)+H(\boldsymbol{m}_0,\boldsymbol{p}_0)}\right] \tag{2.24}$$

where subscript 0 and $t$ denotes the starting point and point of evaluation (Fichtner et al. 2019).

With the current method the point at which the trajectory is terminated will become the next sample of $\boldsymbol{m}$, regardless of it being optimal. This is caused by the integration only happening in one direction. By introducing an integration scheme that integrates both forwards and backwards in time the next sample can be sampled from all points within the explored trajectory.

NUTS uses a doubling scheme where the sign of the momentum changes randomly, illustrated in figure 2.2b. From the starting point (black) one leapfrog step is taken forwards or backwards. Then two steps are taken either forwards or backwards. The doubling continues until the termination criterion, equation 2.23, is satisfied. The amount of doublings are referred to as the tree-depth due to the method sometimes being called a binary tree.
With time reversibility preserved a sample from the trajectory can be sampled with probability

$$\frac{e^{-H(\boldsymbol{m},\boldsymbol{p})}}{\sum_{i=1}^{L'} e^{-H(\boldsymbol{m}_i,\boldsymbol{p}_i)}} \tag{2.25}$$

where $L'$ is the amount of accepted points.

# CHAPTER 3

# Methods

This chapter presents what implementation of HMC has been chosen along with possible alternatives. Additionally, an example of how models are implemented in the software is given followed by a presentation of the different prior distributions utilized. Finally, the chapter lists the diagnostics that will be used to evaluate the performance of a run.

## 3.1 Probabilistic inversion software: STAN

In section 2.3 and 2.4 the rather elegant idea behind HMC was presented. But implementing ideas in practice are not always simple. Illustrated nicely by the problem of a dynamic integration time in section 2.4. The following will explain what software implementation of the theory was chosen and what alternatives there exists.

### 3.1.1 Choosing the implementation

The chosen software used in this thesis is Stan (Carpenter et al. 2017), a free open-source C++ program for Bayesian inference named after Stanislaw Ulam.
Stan depends on the Stan math library (Carpenter et al. 2015) that contains everything from matrix operations to automatic differentiation schemes and probability distributions. The library is optimized in several ways such as vectorization to minimize computations making it highly efficient.

There exists several interfaces for Stan. *CmdStan* allows the user to call the program through command line shell. Following the name pattern *PyStan* is for Python and *RStan* for R. There also exists wrappers of *CmdStan* allowing it to be used in Matlab, Julia, Stata and Mathematica.

The chosen interface is *CmdStan* v2.19.1. *CmdStan*, unlike its counterparts, has the possibility to specify and retrieve the mass matrix. The ability to specify a mass matrix can speed up convergence. Although it might not be very relevant for doing a single run it is highly cost effective when doing multiple runs testing different parameters. Additionally, being able to examine the estimated mass matrices can prove very important in evaluating the success of a run.

Other Bayesian inference tools do exist, a natural question is therefore why choose Stan. Well-known Bayesian inference packages such as BUGS (Lunn et al. 2009) or JAGS (Plummer 2009) do not have HMC. There does exist other programs with HMC implementations such as Edward (Tran et al. 2017), LaplacesDemon (Statisticat 2016) and PyMC (Salvatier et al. 2016).
Edward is a Python library with a wide range of sampling algorithms including HMC, but it does not include the NUTS adaptation.
LaplacesDemon is an R package and as the geomagnetic community is moving toward Python

it would be preferred not to use another software.
PyMC does have HMC with the NUTS adaptation. PyMC uses Theano (Al-Rfou et al. 2016), a Python library for mathematical optimization, as its backend making it very similar to Stan. Both require the user to define a way to evaluate the log posterior. They use almost the same statistical syntax and are compiled into C/C++ programs for optimization. PyMC prides itself on being the only probabilistic programming software that allows the user to specify models directly in Python (Salvatier et al. 2016), although this is also possible in *PyStan*. That said, PyMC is slightly more advanced than *PyStan* in that through a minor circumvention it is possible to define initial mass matrices which is not yet possible in *PyStan*. Seeing that all Stan interfaces are build on Stan, it is only a question of making the feature available since it is supported.

All in all, Stan and PyMC are equal when it comes to functionality. In the end Stan was chosen because it has far larger and more detailed documentation and examples along with a large and active community where developers frequently help.

### 3.1.2 STAN model example

In Stan, defining models has to be done in the Stan modelling language. It is quite straight forward and consists of well defined code blocks. For illustrational purposes an example is given in figure 3.1, for a linear regression problem $y = ax + b$, or written in matrix form as in equation 3.1. It is assumed both model parameters and data errors are Gaussian distributed.

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{Gm} + \boldsymbol{res} \\ a &\sim \mathcal{N}(\mu_a, \sigma_a) \\ b &\sim \mathcal{N}(\mu_b, \sigma_b) \\ res &\sim \mathcal{N}(0, \sigma_{data}) \end{aligned} \tag{3.1}$$

When writing a model in Stan, there are seven possible code blocks, three of which are mandatory, marked with green in the example.
The first mandatory block is **data** where all data necessary for the program to run has to be specified. In this example that constitutes of observations `y`, data kernel `G`, length of data `n`, mean and standard deviation of the prior distribution `mu_prior` and `sigma_prior` and finally the measurement error `sigma_data`. The program will read these from a file supplied by the user. It is therefore important that names in the data file and the model are identical.
The second block is **parameters** where model parameters are initialized. They have to be initialized in this block and no calculations can take place. Here the `m` vector contains $a$ and $b$.
Finally, the **model** block is where the log posterior is calculated. Here it is possible to define variables such as the vector `res` that contain residuals of a single model realization. Next the log prior of $a$ and $b$ and likelihood are calculated. The calculations are close to identical, but in the example they are done in three different ways to illustrate the possibilities. Stan has many built-in distributions, such as Gaussian, making it very intuitive since it uses a statistical syntax. The prior probability of $a$, marked with a red 1, is written in the statistical syntax and is the most compact format. The built-in functions can also take vector input making it possible to assign a Gaussian distribution to all elements in a vector without using a loop. The prior probability of $b$, marked by a red 2, is written how the compact format is interpreted. It is simply a function that returns the log probability of the prior and adds it to the target distribution. With that in mind it is very easy to create user-defined distributions. In this example the likelihood, marked with a red 3, is given by a user-defined Gaussian distribution.

The user can define functions in the **functions** block, not necessarily probability distributions. If creating a probability distribution it simply has to determine the log of the custom distribution keeping in mind that it should be robust so to avoid underflow. Simply calculating the prior and then taking the logarithm will in many cases yield fatal errors when the sampler at some point evaluates $\log(\sim 0)$. Similarly, inversions of matrices has to be done robustly; the covariance matrix when defining a multivariate Gaussian distribution.
Throughout a run the sampler evaluates the posterior probability many thousands of times. It is therefore smart to define constants when possible to not waste computational resources. In figure 3.1 three such constants arise and these should be defined in the **transformed data** block and not in the **model** block. The same goes for all other derivatives of data from the **data** block. With knowledge of the **functions** and **transformed data** blocks the function `likelihood_lpdf` can be specified and applied in the **model** block to return the log likelihood.

Two optional code blocks are not represented in figure 3.1. In the **transformed parameters** block new parameters can be defined from data, existing parameters and/or other variables. Such a block can be useful when sampling in a transformed model space, i.e. eigenspace, and having to project back into the original model space, when evaluating likelihood.
Finally, there is the **generated quantities** block allowing the calculation of variables that depend on data and parameters. This will only be done at the end of an iteration not at every leapfrog step. It could simply be the forward problem or statistics of interest.
A more in-depth explanation of all blocks and possibilities can be found in the Stan users guide (Stan Development Team 2019c).

### 3.1.3 Hyperparameters

With the Stan model defined it is simply a question of compiling and calling it. It is called by passing a data-file and specifying several optional hyperparameters. Since there is no convergence criteria when using NUTS it is necessary to specify how many iterations it has to run. Specifically how many iterations it is going to use on warm-up and how many are to be used for actual sampling, post warm-up. The default is 1000 in either case.

The warm-up period itself is split into three phases, two of which are fast adaptations and one is a slow adaptation. Fast and slow refer to whether local or global information is used. In the fast adaptation phases only the step-size is changed using the dual averaging algorithm presented in Hoffman and Gelman 2014. This algorithm has several tuneable parameters and these will be kept at their default values.
In the slow adaptation phase both the step-size and the mass matrix are adapted. The mass matrix is approximated using Welford's algorithm (Stan Development Team 2019a), which allows for the (co)variance to be estimated throughout several iterations without holding them all in memory.

The first warm-up phase, figure 3.2, is a fast adaptation and is per default 75 iterations. Here the step-size is altered while the sampler is supposed to converge toward the typical set. The second warm-up phase is a slow adaptation and is separated into smaller sections of increasing size. By default they start at 25 iterations and double for every new section. At the end of every section the mass matrix is updated and the step-size is allowed to converge throughout the following section. The mass matrix adaptation is memoryless so that only samples from the latest slow adaptation section are used.
The third warm-up phase is again a fast adaptation. The newly adapted mass matrix is kept

```
functions {
  real likelihood_lpdf(vector res, real n_log_two_pi, real log_det_data,
                                                    real inv_sigma_sqr) {
    real log_likelihood;
    log_likelihood = n*log_two_pi;
    log_likelihood += -0.5*log_det_data;
    log_likelihood += -0.5*pow(res'*res,2)*inv_sigma_sqr;
  }
}

data {
  int<lower=0> n;                          // Length of data vectors
  vector[n] y;                             // Observed data
  real sigma_data;                         // Uncertainty of data
  matrix[n, 2] G;                          // Data kernel
  vector[2] mu_prior;                      // Mean a and b
  vector[2] sigma_prior;                   // Standard deviation of a and b
}

transformed data {
  real n_log_two_pi = -n/2*log(2*pi());    // Log of initial Gaussian term
  real log_det_data = n*log(sigma_data);   // Log-determinant of the
                                           // diagonal data covariance matrix
  real inv_sigma_sqr = 1/pow(sigma_data,2); // Inverse of the data variance
}

parameters {
  vector[2] m;                             // Initialize model parameters a and b
}

transformed parameters {

}

model {
  vector[n] res = G*m-y;                   // Calculate residuals

  // Prior
  m[1]~normal(mu_prior[1], sigma_prior[1]); 1
  target += normal_lpdf(m[2] | mu_prior[2], sigma_prior[2]); 2

  // Likelihood
  target += log_likelihood(res, log_two_pi, log_det_data, inv_sigma_sqr); 3
}

generated quantities {

}
```

**Figure 3.1:** Example of a Stan model for sampling a linear regression problem. There are three mandatory blocks colored with green, and four optional blocks colored with either blue or grey. The grey color is used for blocks that are not in use, but simply illustrates where they naturally would occur in a model. Orange text are comments and the red numbers, 1, 2 and 3, mark the prior of $a$ and $b$ along with the likelihood. These specific lines are marked to show three different methods of defining a Gaussian distribution in Stan.

constant while the step-size is allowed to converge to its final value. Per default this phase lasts 50 iterations. For reasons explained in section 5.1 this will be increased to 400.

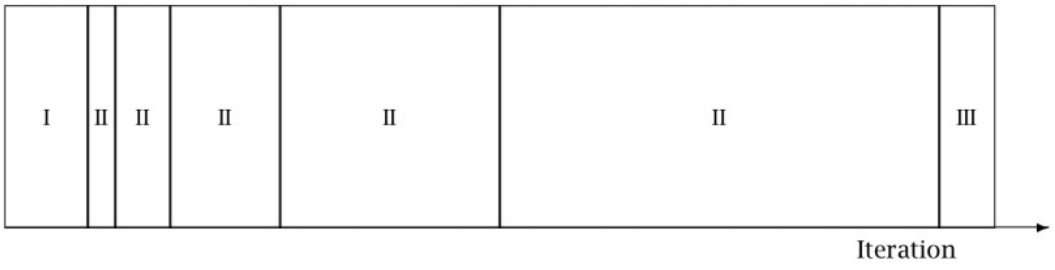



**Figure 3.2:** Illustration of the three phases in the warm-up scheme used in Stan's implementation of HMC. The first and third phase are fast adaptations while the second is slow. The second phase is split up into smaller sections that double in size. Image taken from Stan Development Team 2019a.

The mass matrix is by default assumed diagonal, but can be changed to a dense matrix. Using the dense mass matrix requires more warm-up because all off-diagonal elements of the posterior covariance have to be estimated.
The kinetic energy function is assumed to be EGKE, see section 2.3 for definition. Riemannian-Gaussian Kinetic Energies are not implemented in any of the mentioned software, but is in development (Stan Development Team 2019a).

The purpose of the NUTS adaptation is to automatically determine the integration time which is given by both step-size and the amount of steps. The amount of steps is by default truncated at tree-depth 10, equivalent to $2^{10} = 1024$ leapfrog steps. This is done to prevent stuck samplers to continue infinitely. The step-size can be initialized, but the default is one.
Several parameters control the NUTS adaptation such as the target acceptance rate which influence the step-size. It is by default set to 80 % and will not be tested or changed.

Custom initial states can also be passed such as a starting point and a mass matrix. By default a starting value for each parameter will be sampled uniformly between $[-2, 2]$ while the mass matrix will be initialized as the identity matrix.
In this thesis custom initializations will be made and the performance difference between these and the default will be discussed in section 5.1. The starting point can be initialized as the least squares solution to the inverse problem, equation 2.10, which is defined as

$$\boldsymbol{m} = \left(\boldsymbol{G}^T\boldsymbol{G}\right)^{-1}\boldsymbol{G}^T\boldsymbol{d} \tag{3.2}$$

The mass matrix can be initialized as the prior covariance matrix as defined in section 4.1 and 4.2.

If the initial state is suitable the warm-up iterations can be reduced to zero resulting in immediate post warm-up sampling. If the starting point is off but the step-size and the mass matrix are correct then adaptation can be disengaged and the sampler can through a short warm-up period propagate towards the typical set. Additionally, it is possible to only adapt the step-size by eliminating phase two of the warm-up, but it is not possible to only adapt the mass matrix.

Finally, multiple chains of the HMC sampler can be initialized at once. By default four chains are used and it will not be altered. There are two benefits of using multiple chains; The sampling process can be sped up by distributing the desired amount of post warm-up samples

onto multiple chains. Unfortunately a Markov chain can not be parallelized and therefore each chain can only run on one processor and each chain have to go through its own warm-up. Additionally, using multiple chains can allow the user to determine if the posterior distribution is multimodal or not. If a single chain is used it is likely to only sample one of the modes and the diagnostics would report it as a success. If multiple chains are used and they sample different modes the $\widehat{R}$ diagnostic, discussed in section 3.3.1, would detect that the chains did not mix properly.

## 3.2 Prior probability distributions and likelihood

This section will present in detail how the inverse problem an associated prior and likelihood distributions are implemented as a Stan model. First there will be an explanation of the models basis-form, which is everything expect the prior distribution. Afterwards the three priors applied in this thesis will be presented systematically.

### 3.2.1 The model basis-form

The Stan model applied in this thesis is very similar to the example shown in figure 3.1. The structure of the basis-form, figure 3.3, will be explained in the following.

The size of the data and coefficients are passed to the model along with the observations and the data kernel. Definitions of these can be found in section 2.1 along with the forward problem, equation 2.9.
Similar to the example model the likelihood will be assumed Gaussian distributed, which in the multivariate form is

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^k \left|\boldsymbol{\Sigma}\right|}} e^{\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]} \tag{3.3}$$

Here $\boldsymbol{\mu}$ is the mean, $\boldsymbol{\Sigma}$ is the covariance matrix, equation 3.4, and $k$ is the length of $\boldsymbol{x}$. In the case of the likelihood $\boldsymbol{x}$ is the residuals between observations and predictions and therefore $\boldsymbol{\mu} = 0$. Additionally, the data errors are assumed independent and therefore $\boldsymbol{\Sigma}$ is diagonal.

$$\operatorname{cov}(m_i, m_j) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[\frac{1}{2}(m_i - m_j)^2\right] \tag{3.4}$$

When determining the log of equation 3.3 it can be split up into three convenient terms

$$\log\left(p(\boldsymbol{x})\right) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{3.5}$$

Note that when $\boldsymbol{x}$ is assumed independent Gaussian it is most optimal to use the multivariate expression with a diagonal covariance matrix instead of using a loop. The first two terms in equation 3.5 are constants and are defined in the **transformed data** block, figure 3.3. While the log likelihood evaluation occurs in the **model** section, similar to the example in figure 3.1.

```
data {
  // Observations and forward problem
  int N;                                    // Length of data vectors
  int M;                                    // Amount of model coefficients
  vector[3*N] B;                            // Observations
  matrix[3*N, M] G;                         // Data kernel

  // Likelihood
  vector[3*N] varLike;                      // Error estimate on observations

  // Prior
      // Information about mean and spread of any of the three priors
}

transformed data {

  // Likelihood constants
  real kL1 = -(3*N)/2*log(2*pi());          // Log of initial Gaussian term
  real kL2 = -0.5*sum(log(varLike));        // Log-determinant

  // Prior constants
      // Constants regarding the log prior
}

parameters {
  vector[M] mCore;                          // Initialize model coefficients
}

model {
  vector[3*N] res = G*mCore-y;                    // Calculate residuals

  // Prior
      // The prior distriobution

  // Likelihood
  target += kL1;
  target += kL2;
  target += -0.5*(res'*res./varLike);
}
```

**Figure 3.3:** The basis-form of all models applied in this thesis. The color scheme is identical to the example in figure 3.1. In this basis-format no information about prior distributions is given. Only data import, the forward problem and log likelihood calculation are defined.

### 3.2.2 Independent Gaussian

The simplest prior belief, that is being tested, is viewing the SH-coefficients as independent and Gaussian distributed. The Gaussian distribution is generally to be a good fit as seen in section 4.1, but the coefficients are definitely not independent, the affect of which is shown in section 5.

Implementing this prior is identical to the likelihood in the basis-form, equation 3.3. Vectors of mean and variances are passed to the model though the **data** block, appendix 8.2. With the prior information defined two constants are determined in the **transformed data** block and the log probability of the prior is calculated in the **model** block.
The prior information used to determine the mean and variances required to implement such a distribution is presented in section 4.1 and 4.2.

### 3.2.3 Multivariate Gaussian

When implementing the multivariate Gaussian prior, equation 3.3, a few more steps have to be taken with regards to the covariance matrix. In the case of the likelihood and independent prior the covariance has been diagonal making determination of the inverse trivial. But with a non-diagonal covariance matrix calculating the inverse can numerically be difficult when elements are close to zero. Therefore an LDLT-decomposition is used to provide a robust and reliable inversion.

A requirement of the LDLT-decomposition is that the matrix, being decomposed, is positive definite. This is assured when examining the core dynamo simulation used as prior information, section 4.1.

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T \tag{3.6}$$

$\boldsymbol{L}$ is a lower triangular matrix and $\boldsymbol{D}$ is a diagonal matrix. The inverse of $\boldsymbol{\Sigma}$ can now be defined as

$$\boldsymbol{\Sigma}^{-1} = \left(\boldsymbol{L}^T\right)^{-1} \boldsymbol{D}^{-1}\boldsymbol{L}^{-1} \tag{3.7}$$

Also the log determinant can be determined from the decomposition as

$$\log\left(|\boldsymbol{\Sigma}|\right) = \sum_{i=1}^{M} \log\left(D_{i,i}\right) \tag{3.8}$$

Both of these constants are determined outside the Stan model and passed to it, for which reason they show up in the **data** block, appendix 8.3. In practice the decomposition is done using the *ldl()* function in the Python library *SciPy* (Jones et al. 2001).

### 3.2.4 Co-estimation

The forward problem is as stated in equation 2.9. When attempting to co-estimate the core and lithospheric field the SH coefficients are divided in two, into contributions belonging to either source, equation 3.9.

$$\boldsymbol{m} = \boldsymbol{m}_{core} + \boldsymbol{m}_{litho} \tag{3.9}$$

The forward problem can be expanded to

$$\boldsymbol{d} = \boldsymbol{G}\left(\boldsymbol{m}_{core} + \boldsymbol{m}_{litho}\right) \tag{3.10}$$

Here $\boldsymbol{m}_{core}$ and $\boldsymbol{m}_{litho}$ refer to the core and lithospheric contributions, respectively. Thus the model space when co-estimating will be doubled in size.

Another consequence is the need for two prior probabilities. The one associated with the core field is a multivariate Gaussian distribution just as presented in section 3.2.3. The lithospheric field prior on the other hand will assume the coefficients independent and Gaussian distributed, similar to the approach in section 3.2.2. Reasoning for the lithospheric prior being independent is explained in section 4.2. The implementation of the co-estimation model can be found in appendix 8.4.

### 3.2.5 Independent GMM

The last prior is one based on a two component Gaussian Mixture Model (**GMM**). The distributions of the SH coefficients belonging to the lower SH-degree, as generated from the core dynamo simulation presented in section 4.1, are found not to be represented well by a Gaussian distribution. In order to justify the assumption of them being Gaussian a comparison will be made with the more complex GMM prior.

A two component GMM is given as

$$p(\boldsymbol{x}) = \frac{\boldsymbol{A}_1}{\sqrt{2\pi\boldsymbol{\sigma}_1^2}} e^{\left[-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^2}{2\boldsymbol{\sigma}_1^2}\right]} + \frac{\boldsymbol{A}_2}{\sqrt{2\pi\boldsymbol{\sigma}_2^2}} e^{\left[-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_2)^2}{2\boldsymbol{\sigma}_2^2}\right]} \tag{3.11}$$

Where $\boldsymbol{\mu}_i$ is the mean, $\boldsymbol{\sigma}_i$ is the standard and $\boldsymbol{A}_i$ is weight scaling the Gaussian distribution. The subscript 1 and 2 refers to the two Gaussian distributions that are combined. It is required that $\boldsymbol{A}_1 + \boldsymbol{A}_2 = 1$. The GMM fits that are shown in section 4.1 are made by applying the Levenberg-Marquardt algorithm so to solve the optimization problem of fitting the given distribution. In practice this is done using the *curve_fit()* function in the Python library *SciPy* (Jones et al. 2001).

Implementing a GMM prior in Stan still requires the user to define the logarithm of it. Doing this directly as

$$\log(p(\boldsymbol{x})) = \log\left(\boldsymbol{c}_1 e^{[\boldsymbol{k}_1]} + \boldsymbol{c}_2 e^{[\boldsymbol{k}_2]}\right) \tag{3.12}$$

can result in over- or underflow. Here $\boldsymbol{c}_i = \frac{\boldsymbol{A}_i}{\sqrt{2\pi\boldsymbol{\sigma}_i^2}}$ and $\boldsymbol{k}_i = -\frac{(\boldsymbol{x}-\boldsymbol{\mu}_i)^2}{2\boldsymbol{\sigma}_i^2}$. To ensure robustness when determining the logarithm equation 3.12 can be rewritten as

$$\log(p(\boldsymbol{x})) = \boldsymbol{k}_3 + \log\left(\boldsymbol{c}_1 e^{[\boldsymbol{k}_1 - \boldsymbol{k}_3]} + \boldsymbol{c}_2 e^{[\boldsymbol{k}_2 - \boldsymbol{k}_3]}\right) \tag{3.13}$$

where $\boldsymbol{k}_3 = \max(\boldsymbol{k}_1, \boldsymbol{k}_2)$. Opposite to the previous models this will not be implemented by hand, but by using the *log_sum_exp()* function provided by Stan, appendix 8.5.

## 3.3 Diagnostics

After a Stan program is complete it is possible to retrieve diagnostics native to Stan. These include E-BFMI, $\hat{R}$, MCSE and $n_{eff}$ which will be explained in the following, along with some additional diagnostics related to the geomagnetic field modelling problem at hand.

### 3.3.1 $\widehat{R}$

The $\hat{R}$ parameter, also called the potential scale reduction parameter (Gelman et al. 2013), indicates whether or not the HMC chains have converged. Meaning whether or not the posterior

distribution would change if more samples were taken. It is determined as the ratio between the estimated posterior variance and the average variance within each chain, equation 3.14, where $i \in 1, .., M$ and $M$ is the number of parameters in the model. Ideally $\hat{R}$ should be one, but empirically it has been shown that values above 1.1 are problematic (Betancourt 2017).

$$\hat{R} = \sqrt{\frac{\widehat{\text{var}}(m_i|\boldsymbol{d})}{W}} \tag{3.14}$$

The estimated posterior variance, $\widehat{\text{var}}(m_i|\boldsymbol{d})$, is determined from both the variance within each chain, $W$, and the variance between chains, $B$. The influence of $B$ decreases with an increasing amount of samples per chain, $n_s$, equation 3.15.

$$\widehat{\text{var}}(m_i|\boldsymbol{d}) = \frac{n_s - 1}{n_s} W + \frac{1}{n_s} B \tag{3.15}$$

The within-chain variance is the mean of the sample variance, $s_k^2$, between all chains, equation 3.16, here $n_c$ is the amount of chains and subscript $k$ runs over all chains. The between-chain variance is the sample variance of the average in each chain, $\overline{m}_{i,.,k}$, compared with the combined averaged of all samples, $\overline{m}_{i,.,.}$, equation 3.16. Subscript $j$ is the sample number and . indicates a subscripts entire range.

$$\begin{gathered} W = \frac{1}{n_c} \sum_{k=1}^{n_c} s_k^2 \quad , \quad B = \frac{n_s}{n_c - 1} \sum_{k=1}^{n_c} (\overline{m}_{i,.,k} - \overline{m}_{i,.,.})^2 \\ s_k^2 = \frac{1}{n_s - 1} \sum_{j=1}^{n_s} (m_{i,j,k} - \overline{m}_{i,.,k})^2 \quad , \quad \overline{m}_{i,.,k} = \frac{1}{n_s} \sum_{j=1}^{n_s} m_{i,j,k} \quad , \quad \overline{m}_{i,.,.} = \frac{1}{n_c} \sum_{k=1}^{n_c} \overline{m}_{i,.,k} \end{gathered} \tag{3.16}$$

In situations where $\hat{R}$ is not well-behaved, below 1.1, results should not be used. Note that the cause of bad performance is not explained by this diagnostic, but simply states that the posterior has not converged.

### 3.3.2 E-BFMI

Other diagnostics such as the Energy Bayesian Fraction of Missing Information (**E-BFMI**) (Betancourt 2018) can provide a deeper understanding of the performance of the HMC sampler. If the posterior distribution has a heavy tail then the typical set will be stretched. As a result the integration time will be significantly increased when exploring the typical set close to the tail. Given a good kinetic energy function integration time could be decreased. If instead the kinetic energy is poorly estimated resulting in a small step-size the sampler can continue on that trajectory forever. Such behaviour can be described with E-BFMI, equation 3.17.

$$\widehat{\text{E-BFMI}} \equiv \frac{\sum_{i=1}^{n_s} (E_i - E_{i-1})^2}{\sum_{i=0}^{n_t} (E_i - \overline{E})^2} \tag{3.17}$$

Here $E$ is the energy, also known as the Hamiltonian, as introduced in section 2.3. Note how $i$ begins from zero in the denominator, zero indicates the starting point and is not normally counted as a sample.

Sample runs with a low E-BFMI values are problematic. Because it is a fairly new diagnostic only empirical thresholds have been found. The threshold in Stan is set to 0.2 (Betancourt 2017) while newer documentations state values below 0.3 are problematic (Betancourt 2018). Independent of the threshold the best suggestion for improving a low E-BFMI is reparameterization of the problem, discussed in Stan Development Team 2019c.

### 3.3.3 Tree-depth

Unsuccessful runs do not always have complex problems behind them. If the problem being solved is high dimensional or the kinetic energy function is poorly estimated then the maximum tree-depth can be exceeded. This, in of itself, is not a fatal error. If the trajectory is terminated prematurely during the warm-up period the sampler can still recover, but convergence through sub-optimal choices can take longer. If the trajectory is terminated prematurely during the post warm-up sampling it is more problematic. In that case the run should be remade with a larger maximum tree-depth or a longer warm-up period.

### 3.3.4 Sample independence

When a run is completed and diagnostics and results are to be extracted it is important that the samples are independent. In contrast to classical MCMC algorithms HMC has the potential to provide an independent sample every iteration. Although possible, it is not always the case and therefore the amount of independent samples has to be estimated, from now on referred to as Effective Sample Size (**ESS**) following the Stan literature (Stan Development Team 2019b). The ESS of a single chain is given as in equation 3.18. Here $\rho_l$ is the autocorrelation of lag $l$.

$$\text{ESS} = \frac{n_s}{1 + 2\sum_{l=1}^{\infty} \rho_l} \tag{3.18}$$

The real autocorrelation can not be calculated, but has to be estimated (Betancourt 2017). For a single time series the autocorrelation if defined as

$$\widehat{\rho}_l = \frac{\sum_{i=l+1}^{n_s}(m_l - \overline{m})(m_{i-l} - \overline{m})}{\sum_{i=1}^{n_s}(m_i - \overline{m})^2} \tag{3.19}$$

When working with multiple chains a combined autocorrelation can be computed as in equation 3.20. Here $\hat{\rho}_{l,k}$ is the estimated autocorrelation at lag $l$ in chain $k$.

$$\widehat{\rho}_l = 1 - \frac{W - \frac{1}{n_c}\sum_{k=1}^{n_c} \hat{\rho}_{l,k}}{\widehat{\text{var}}(m_i|\boldsymbol{d})} \tag{3.20}$$

With an estimate of the autocorrelation the ESS for multiple chains can be defined, equation 3.21.

$$\text{ESS} = \frac{n_t}{\widehat{\tau}} \tag{3.21}$$

$\widehat{\tau}$ is defined similarly to the denominator in 3.18, but truncated at a certain lag, $l_t$, because noise in the estimate, equation 3.20, increases with $l$ (Betancourt 2017). Defining $\widehat{\tau}$ as in equation 3.22 ensures a positive autocorrelation by summing pairs starting from $l = 0$ since negative values only are encountered at odd lags. Here $\widehat{P}_{l'} = \hat{\rho}_{2l'} + \hat{\rho}_{2l'+1}$. A suitable truncation, $l_t$, is defined as the largest lag such that $\widehat{P}_{l'} > 0$ meaning to sequential values of $\widehat{\rho}_l$ cannot be negative.

$$\widehat{\tau} = 1 + 2\sum_{l=1}^{2l_t+1} \widehat{\rho}_l = -1 + 2\sum_{l'=0}^{l_t} \widehat{P}_{l'} \tag{3.22}$$

Under this definition ESS can be larger than the actual amount of samples, $n_t$, which typically happens if the marginal posterior is close to Gaussian and the specific model parameter does not correlate much with other parameters (Stan Development Team 2019b).

### 3.3.5 Mean model and error estimates

Until now the diagnostics that have been defined concern the sampler itself. It can occur that these diagnostics point toward a neat and converged solution, but it is important to look at the samples themselves.

One such diagnostic is the mean model. In this thesis the model parameters will tend toward a Gaussian distribution given enough independent samples due to the prior and likelihood that will be implemented. With that reasoning the mode of the distributions equals the mean.
In Stan the mean model is given as the mean of all samples across all chains, equation 3.23

$$\mu_{post,i} = \frac{1}{n_t} \sum_{k=1}^{n_c} \sum_{j=1}^{n_s} m_{i,j,k} \tag{3.23}$$

Likewise the associated standard deviation is given over all samples, equation 3.24.

$$\sigma_{post,i} = \frac{1}{n_t - 1} \sum_{k=1}^{n_c} \sum_{j=1}^{n_s} (m_{i,j,k} - \mu_{post,i})^2 \tag{3.24}$$

In the event of a low amount of independent samples or correlated samples the calculated mean is errorneous which is taking into account by the Monte Carlo Standard Error (**MCSE**) (Brooks et al. 2011). MCSE is the error of the estimation, as defined in equation 3.25, and it is a function of the posterior standard deviation and inversely proportional to ESS. Remember that ESS depends on the within- and between-chain variance along with the correlation of the samples essentially making it very large if the samples of the posterior are correlated. Thus an MCSE value tending toward zero will tell you the samples are independent and the estimated mean model lies on the mode of the posterior distribution.

$$\text{MCSE} = \frac{\sigma_{post,i}}{\sqrt{\text{ESS}}} \tag{3.25}$$

### 3.3.6 Power spectrum

Although the necessary information is gathered within the mean model and the associated MCSE and $\boldsymbol{\sigma}_{post}$ they can be difficult to interpret and compare to other runs.
As a means of visualization the power spectrum, introduced in section 1.1, will be used. The magnetic field of interest is the core which means it is easy to see if the power diverges from its expected horizontal path, when evaluating it at the CMB. Its mathematical definition is given in equation 3.26. Here $n$ and $m$ are the SH-degree and order, $a$ is the reference radius (6371 km), $r$ is the radius of evaluation (3480 km) and $g_n^m$ and $h_n^m$ are the SH-coefficients. By examining the expression it is evident that the constant in front of the sum rapidly increases with SH-degree, if $r < a$, making the power spectrum increase drastically by small deviations in the higher harmonics. It is therefore a very good indication of how physically likely the sampled models are.

$$W_n(r) = (n+1) \left(\frac{a}{r}\right)^{2n+4} \sum_{m=0}^{n} \left[(g_n^m)^2 + (h_n^m)^2\right] \tag{3.26}$$

In the following chapter the power spectra related to models used for synthetic data, prior information and comparisons will be shown. When visualizing distributions with the power spectrum it will be done so with a shaded area representing the entire distribution. Atop of that 30 uniformly sampled realizations will be shown so to get a feeling of the spread. Finally, the mean model, including error bars showing the influence of three MCSE's will be displayed.

### 3.3.7 Visualization on maps

With the current diagnostics it is not possible to see the physical characteristics of the magnetic field associated with the sampled models. Another very important diagnostic is therefore projecting predictions, by solving the forward problem, onto a map of the Earth. The maps are made using the Python library *Basemap*[1]. In order to keep the grid equal area a Hammer projection is used.

The predictions are made on a uniform grid, which can be problematic due to it stretching when representing a spherical surface in 2D. By using a step-size of 0.1 degrees, resulting in 6,500,000 predictions, it is assumed insignificant.

These projections are the basis for several diagnostics. Such as a map of the posterior mean or random realizations. The mean can look quite good, while random realizations deviate therefrom. If the realizations deviate too much while the sample statistics are well-behaved it can suggest too little data constraint. If the realizations are similar to the mean the model parameters can be considered well determined, which is easily summarized in a Root Mean Square (**RMS**) map. The RMS maps refers to stacking the predictions from 300 random realizations atop each other and calculating the RMS at each point. In cases of a large spread the RMS-map will return either no pattern or what looks like a smoothed image of the mean. If the spread is small the RMS map will emphasize well constrained features that can be found in a majority of the random realizations.

## 3.4 Equidistant grid for synthetic data

One of the most important things when modelling is availability of sufficient data. Data constrains the model and it is therefore important to have good quality data and a sufficient amount. Without enough the prior has to be very strong otherwise it can lead to unphysical behaviour. Enough data does not only entail a large enough amount of data, but also requires it to be evenly distributed.

Several ways of going about this are described in Saff and Kuijlaars 1997. The most computationally simple projects a spiral down unto the spheres surface and selects points on that path. An implementation, in Python, of the spiral points algorithm and plot command can be found in appendix 8.1. An illustration of the result from this implementation is visualized in figure 3.4. From a simple visual inspection it appears that the distribution is even in both latitude and longitude. Thus the data will not constrain any region more than another.
The method described is quite simple and more precise methods do exist, but this method of producing uniformly sampled data is considered sufficient. The real data will be generated as described in (Hammer and Finlay 2019), which will be elaborated on in section 4.4.

[1] https://matplotlib.org/basemap/

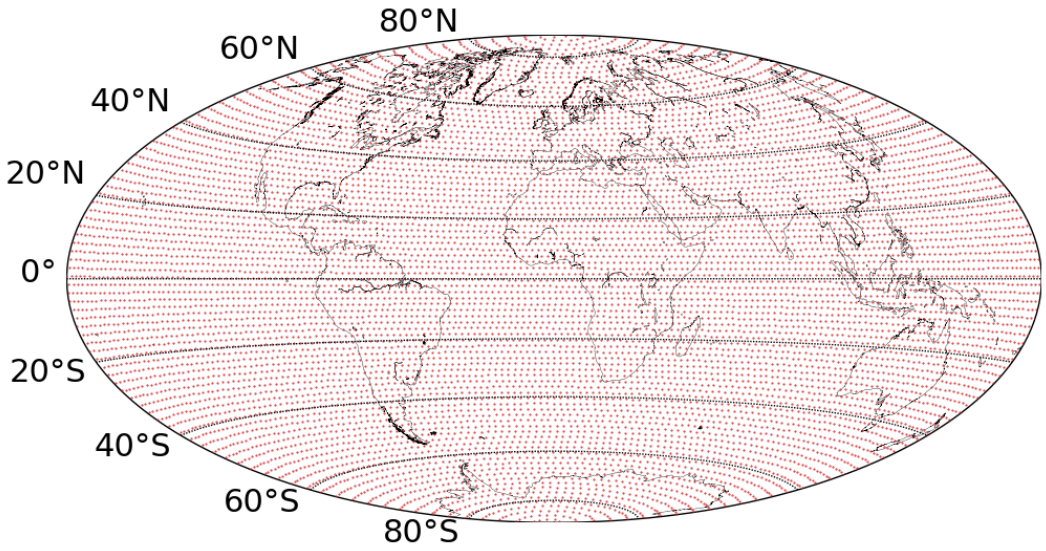



**Figure 3.4:** 10,000 points selected using a spiral point method (Saff and Kuijlaars 1997) visualized on a spherical surface.

# CHAPTER 4

# Data

This chapter is divided into four sections. The first two will present data from physics-based simulations used when creating the core and lithospheric priors. The last two will present the synthetic data used for testing and the real satellite data used to produce the primary results, along with how it has been processed.

## 4.1 Prior information from core dynamo simulations

In section 3.2 ways of practically implementing different prior distributions were explained. In this section the prior information about the core field will be presented. The data used to construct the core field priors comes from a core dynamo simulation designed to accurately represent the asymptotic conditions physical expected in Earth's core and a force balance between the Coriolis, pressure, buoyancy and Lorentz forces (Aubert et al. 2017). The simulation, here considered only regarding the magnetic field at its outer boundary out to SH-degree 30, was simulated for thousands of years and a realization extracted every 20-40 years, resulting in a total of 687 realizations. These provide a time series of each SH-coefficient from which information about their distributions can be extracted.

In the case of the independent and multivariate Gaussian priors the mean and covariance, equation 3.4, has to be defined. Ideally these statistics need to be generated from a set of independent samples, but due to the nature of how they are generated this is not the case. Especially the large wavelength harmonics tend to have a slowly decreasing autocorrelation, equation 3.19. Defining independence as when the autocorrelation function oscillates randomly between $\pm 5$ % correlation then the axial dipole, figure 4.1a, only has a single independent sample. The rate of decay in autocorrelation tends to increase with the SH-degree. The first coefficient that, approximately, remains within the 5 % threshold for each sample is $g_7^6$. That said, coefficients belonging to a higher SH-degree, but with a low SH-order, tends to have a more slow decreasing autocorrelation, as is the case for $g_1^0$.

When determining a covariance matrix the time series have to be of equal length meaning that the coefficient with the fewest independent samples is the determining factor. It was therefore decided to assume that all samples were sufficiently uncorrelated to provide useful information and to simply use all samples, in order to actually obtain a covariance matrix.

The result of having very dependent samples, as is the case in the lowest SH-degrees, can be seen on their distributions. The axial dipole, $g_1^0$, has a multimodal distribution, figure 4.1c. Here a Gaussian distribution is superimposed revealing a quite bad fit. But as the SH-degree increases the distributions tend toward Gaussian. Around $g_6^5$, figure 4.1d, and thereafter the distributions are all rather Gaussian. This is fortunate because the large wavelength harmonics are typically well constrained by data and thus a poor quality prior information at low SH-

degrees is not problematic. It is first when reaching the high degree harmonics that the prior information is going to play a significant role.

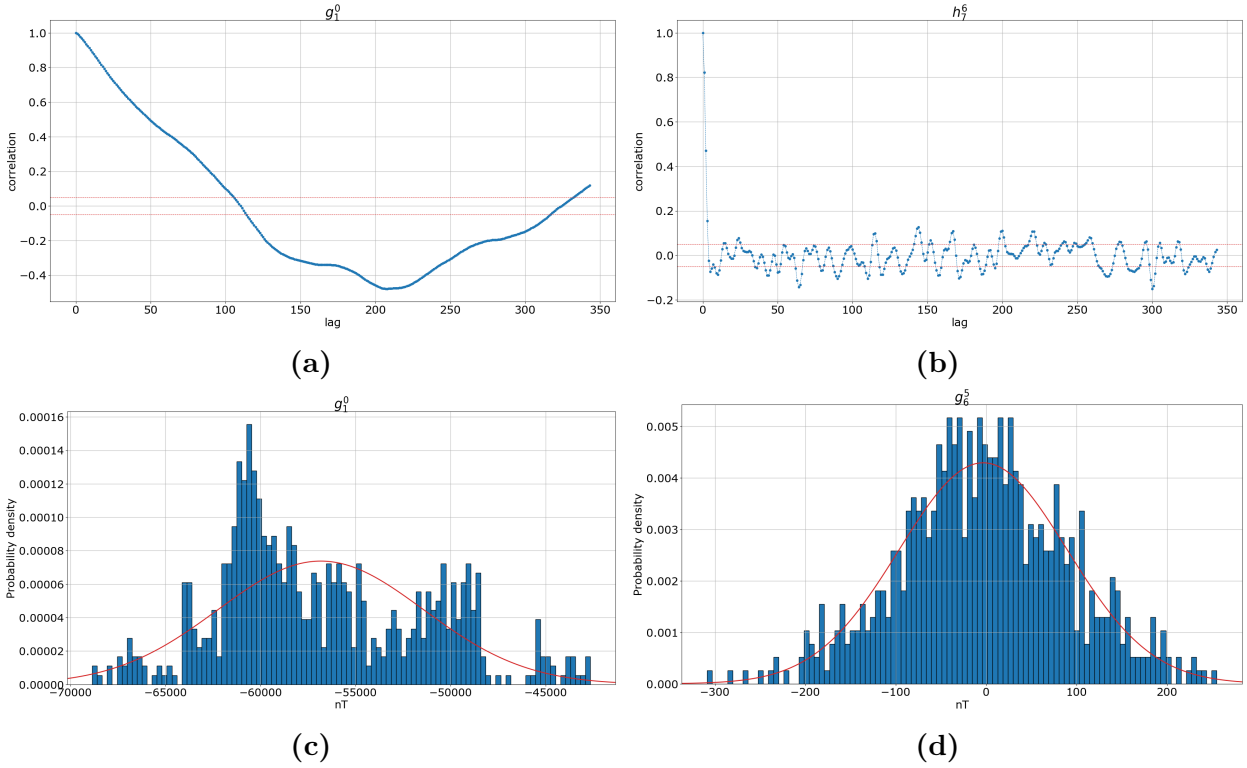



**Figure 4.1:** Statistics on SH coefficient time series generated from a core dynamo simulation to SH-degree 30. Left and right shows $g_1^0$ and $g_6^5$, respectively. Top row is autocorrelation with two red dotted lines indicating $\pm 5$ % correlation. Bottom row is their distribution with a Gaussian distribution superimposed. Around $g_6^5$, and thereafter, the distributions tend to be Gaussian.

The resulting covariance matrix, defined as in equation 3.4, is difficult to interpret due to the difference in the SH coefficients' magnitude. This manifests itself as a very high variance in the first harmonics, due to them simply being larger. In order to illustrate this the $\log_{10}$ variance is shown in figure 4.2a. Here it is clear how the variance decreases with SH-degree and quite significantly.

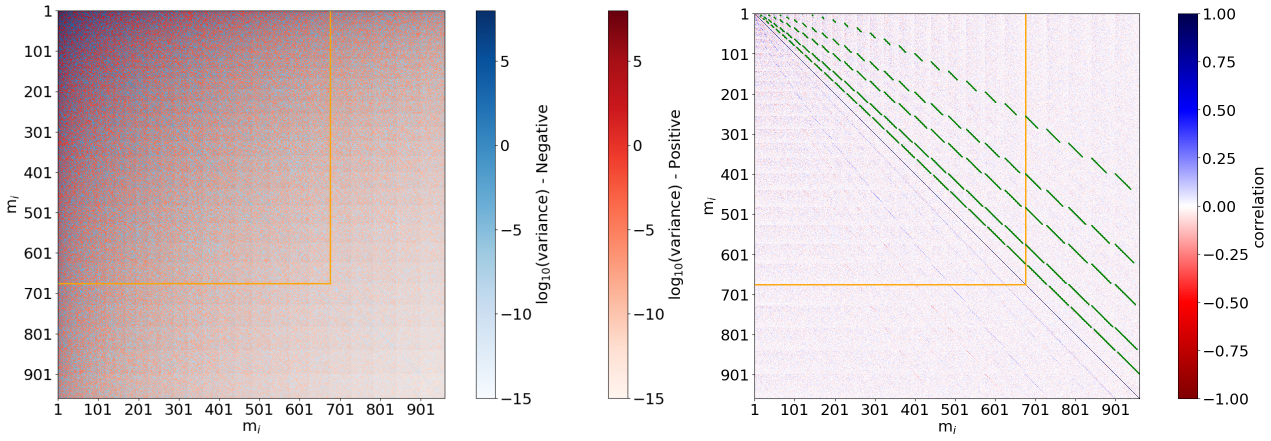
By instead calculating the correlation, defined as in equation 4.1, the structure can be examined independent of the variance size, figure 4.2b.

$$\mathrm{corr}(m_i, m_j) = \frac{\mathrm{cov}(m_i, m_j)}{\sqrt{\mathrm{cov}(m_i, m_i)\mathrm{cov}(m_j, m_j)}} \tag{4.1}$$

Two patterns emerge. There is a pattern of horizontal and vertical lines with virtually no correlation meeting on the diagonal of the correlation matrix. This pattern propagates along the diagonal with increasing distance between the lines. The pattern will be referred to as square waves. They occur due to the newly added SH-order coefficients in every SH-degree correlate very little with any coefficient prior to itself. This also means that they indicate the transition from one SH-degree to another.

Additionally, there is a cone-like pattern running parallel to the diagonal and is made up of five sequences of points. In the upper triangular region, figure 4.2b, they have been emphasized with green dots that can be compared to the untouched lower triangular region. These five

sequences will be referred to as lines, with the first line being closest to the diagonal and the fifth furthest away.



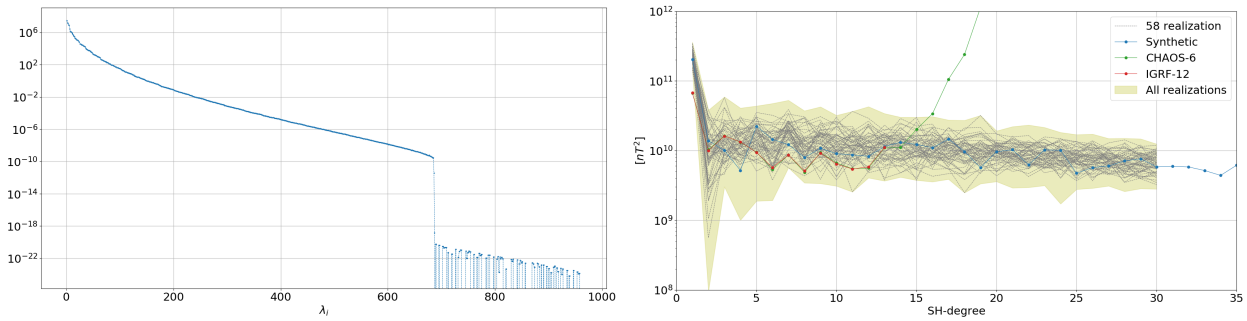

**(a)** **(b)**

**Figure 4.2:** Illustration of covariance and correlation matrix, respectively. They are based on 687 realizations of a core dynamo simulation to SH-degree 30. Note that the covariance matrix is given in $\log_{10}$ values. Since both positive and negative covariance can occur they have been colored blue and red, respectively.

The first line reveals a strong positive correlation between the last coefficients in a SH-degree and the last coefficients in the degree prior to it; $g_4^4$ and $h_4^4$ correlates strongly with $g_3^3$ and $h_3^3$, respectively. This is an exception to the low correlation in the square waves mentioned above. It is also an exception to the rest of the lines. These describe correlation between the beginning of a SH-degree and another. More specifically the second line points two degrees back, an example could be $g_4^1$ and $g_3^1$. The third, fourth and fifth points four, six and ten degrees back. The length of the sequences of green dots grows with two for each degree which is similar to the increase in coefficients in each SH-degree. It should be mentioned that the first and third line are positive, while the fourth and fifth are negative. Finally, the second changes from positive to negative, along a single sequence. There does appear to be an additional positive correlation line between the fourth and fifth line. It is very vague and therefore not highlighted, but would fit very well into the pattern.
The hope is that the correlation structure within the covariance will constrain the parameter estimation, especially at higher SH-degrees where the data will have a hard time due to noise.

In figure 4.2 both matrices have an orange line indicating the boundary if truncating at SH-degree 25. This is relevant because SH-degree 25 is the highest possible degree that has less coefficients than the amount of samples available from the dynamo simulation. As mentioned in section 3.2 the covariance matrix has to be inverted when evaluating the prior and the method of doing so has to be numerically accurate which is why LDLT-decomposition is used. A requirement of this method is that the matrix, being decomposed, is positive definite. Calculating the eigenvalues of the covariance matrix, figure 4.3a, clearly show a discontinuous drop to zero after coefficient 686 illustrating the need for truncation. As a consequence the prior information can only be used to SH-degree 25, unless more samples from the core dynamo simulation are acquired.

It is also interesting to see how the realizations look from a geophysical perspective. A power spectrum of the realizations at the CMB, figure 4.3b, shows how much the power changes throughout the simulation. The spread might be large, but what is important is that it includes the true model, be it synthetic or real satellite data, and that the intrinsic correlation structure

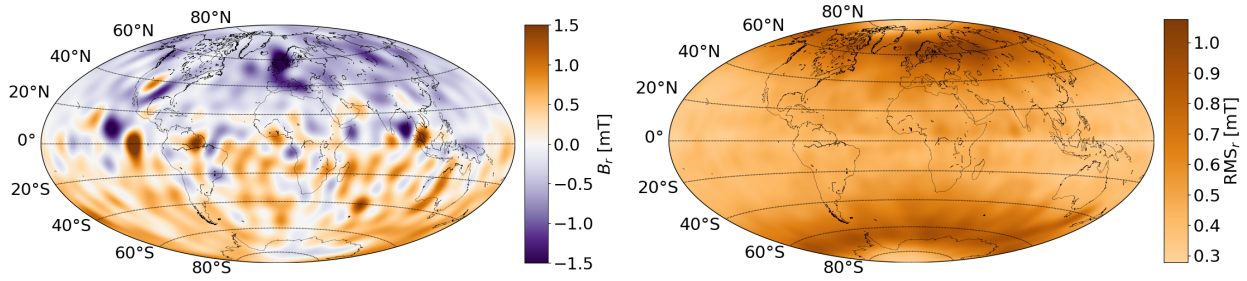



**(a)** **(b)**

**Figure 4.3:** 4.3a: Eigenvalues of the covariance matrix, figure 4.2a. The discontinuous drop occurs between 686 and 687 making it necessary to truncate at SH-degree 25. 4.3b: Mauersberger-Lowes power spectrum of CHAOS-6 (green), IGRF-12 (red) and the core dynamo simulations used to generate the covariance matrix. The faded area represents all realizations while the grey lines are 58 realizations to illustrate the distribution. The blue line is the model used to create synthetic data, section 4.3.

is a good representation of the truth. Additionally, the figure shows how the CHAOS-6 model, diverges at the CMB after SH-degree 14. This shows how sensitive the models are when downward continuing to the CMB and how important clean data is, so not to introduce too many contaminants.

The fact that we need to truncate the dynamo simulations at SH-degree 25 puts an upper limit on the detail that is possible to get from a sampled model. A map of a single realization, figure 4.4a, at the CMB shows precisely this. From the map a strong dipolar pattern can be seen with a significant amount of reversed flux patches in the southern hemisphere below Africa. From the RMS map, figure 4.4b, it is clear that none of these patterns are present in all the simulation realizations which makes sense given that the realizations represents the evolution of the geomagnetic core field and thus have no reason to be similar in the small scale structure.




**(a)** **(b)**

**Figure 4.4:** 4.4a: The radial component of a random realization from the core dynamo simulation used as prior information, truncated at SH-degree 25. 4.4b: An RMS map of the radial component from all 687 realizations. There is no well constrained small scale structure in all the realizations.

The means and covariance matrix needed to define the independent and multivariate Gaussian priors can be determined from the core dynamo simulation presented above. Creating a GMM prior, section 3.2, to better fit the coefficients distributions requires slightly more. A GMM with two components is fitted to all distributions resulting in a better fit, figure 4.5.

Note that a single Gaussian fit works quite well from $g_6^5$ and upwards, as mentioned previously. Thus the effect of the GMM should only be noticeable on the coefficients prior to $g_6^5$.

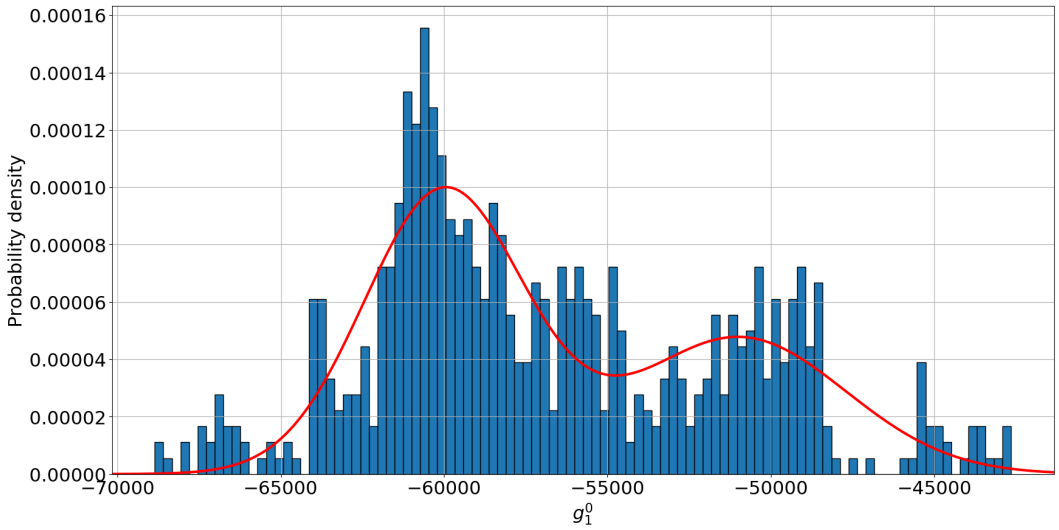



**Figure 4.5:** Distribution of $g_1^0$ from the core dynamo simulations used as prior information. Here illustrated with a two component GMM superimposed. The GMM fits will be used to define a prior that can be compared to that which assumes all distributions Gaussian.

## 4.2 Prior information on the lithospheric field

Although the primary objective is to model the geomagnetic core field attempts will also be made at co-estimation of the core and lithospheric field. In doing so a prior has to be constructed to describe the lithospheric field. Unfortunately, there are no sequences of realizations available from lithospheric simulations similar to what was used for the core field. Instead the prior distribution will be constructed from the lithospheric model presented by Masterton et al. 2013, from here on simply referred to as the Masterton model. The Masterton model is made from assigning magnetic susceptibilities and crustal thickness to different locations in the lithosphere based on laboratory magnetization and seismic measurements. By varying these it might be possible to create a series of realizations that provide information about the correlation of the model coefficients describing the lithospheric field. Unfortunately only one realization is available and there is not enough time to create a new lithospheric simulation in this project.

The Masterton model goes to SH-degree 256 allowing for extremely small scale structure as clearly seen in the figure 4.6a in which the radial component of the model is plotted. When later testing co-estimation on synthetic data, the full Masterton model will be used to create the lithospheric contribution. Note that the highest possible SH truncation degree that can be applied when sampling is 25 due to the cores prior. This limits the detail that can be reconstructed, exemplified by truncating the Masterton model at SH-degree 25 in figure 4.6b. The image seems smoothed resulting in a lower magnitude clear from the well known Bangui anomaly beneath central Africa.

In order to create the lithospheric prior the coefficients are assumed Gaussian distributed. Because the Masterton model is the only information it is assumed to be the mean while the standard deviation is simply defined to be 40 % of the mean value. From this a diagonal covariance matrix can be created. Determining the prior in such a way is likely to encapsulated

the true model due to the large spread, but unfortunately there is no information about the correlation between coefficients which is likely to be important in the source separation problem.

Illustrating the core and lithospheric prior's power spectrum alongside each other, at the CMB, reveals a steady and horizontal core field intersected by the crustal field between degree 16-17, figure 4.6c. This is as expected although the intersect occurs later than expected, as is backed up by the CHAOS-6 power being systematically larger, after degree 15. This is no mystery, it is actually well-known that the Masterton model predictions are somewhat weak despite the assigned susceptibility in the model being quite high (Masterton et al. 2013). This is no problem in the synthetic case, but when moving on to real data it can prove problematic. To circumvent this the standard deviation will be increased to 120 % of the mean value when working with real data. The large standard deviation will ensure that observation-based models such as CHAOS-6 will be encompassed by the prior, figure 4.6d.

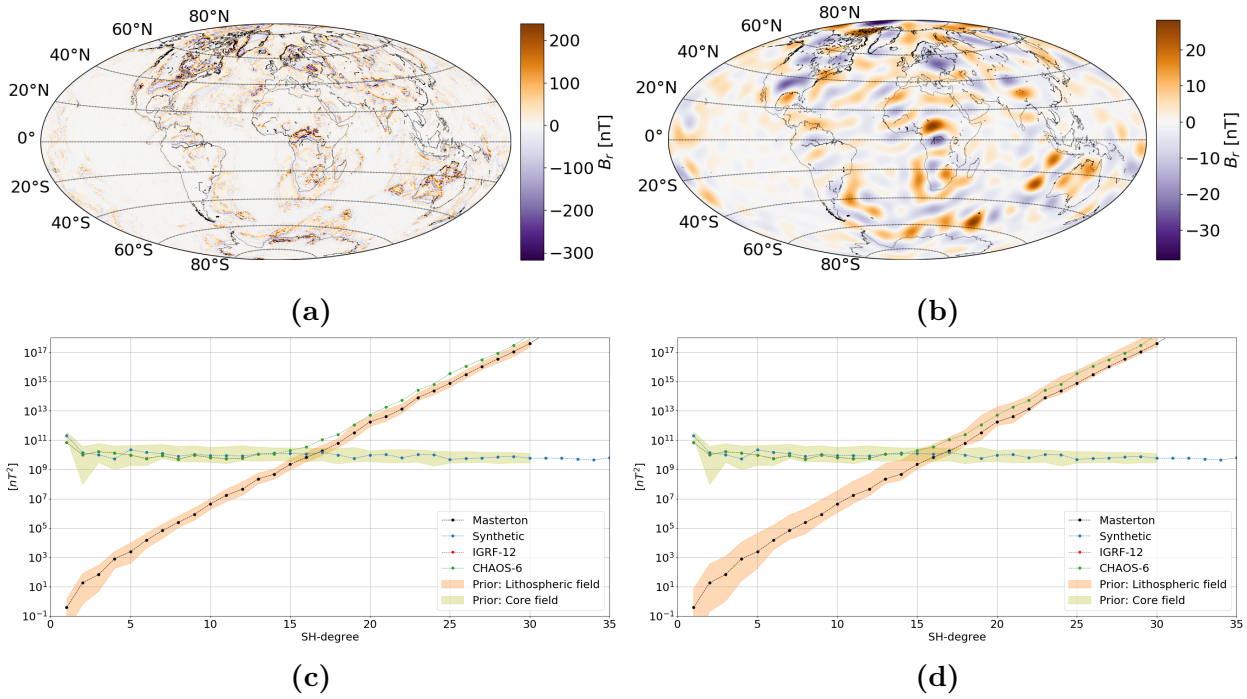



**Figure 4.6:** Maps and power spectrum of the Masterton model used as lithospheric prior information. 4.6a: Map of the radial component from the full Masterton model at Earth's surface. 4.6b: Map of the radial component from the Masterton model, truncated at SH-degree 25, at Earth's surface. 4.6c: Power spectrum of both core and lithospheric prior used for testing co-estimation on synthetic data, at the CMB. 4.6d: Power spectrum of both core and lithospheric prior used when co-estimating with real data. The lithospheric prior is wider here than in 4.6c due to the lack of power in the Masterton model.

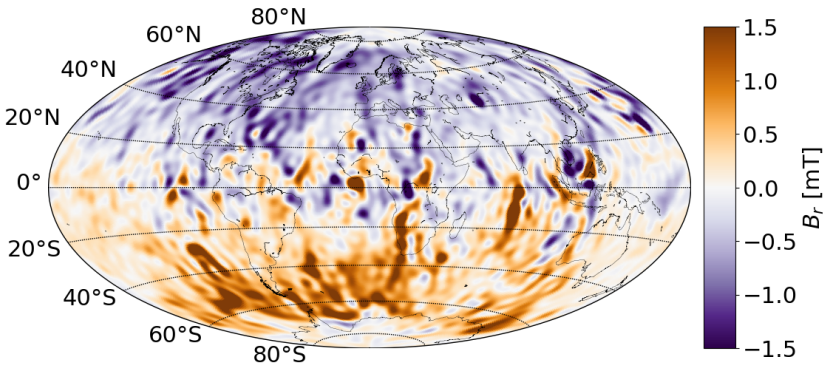
## 4.3 Synthetic data used for benchmark tests

The synthetic data used in section 5.1 to determine the hyperparameters and in section 5.2 for benchmark tests is generated from a core dynamo simulation to SH-degree 60 similar to the one described in section 4.1. The similarity is also apparent from the power spectra in figure 4.3b, where it is labelled Synthetic and will in the future be referred to as the synthetic model.

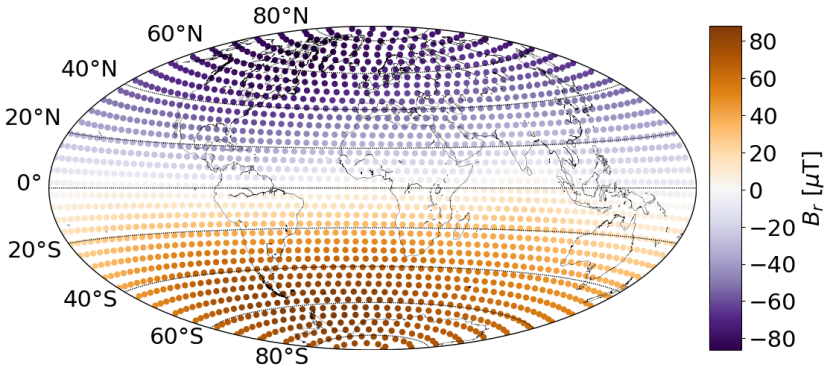
The synthetic data is created at an altitude of 400 km above the Earth's surface to mimic real satellite data. It is created on a grid as mentioned in section 3.4 and typically 2000 data-points are used for testing. The $B_r$, $B_\theta$ and $B_\phi$ components are generated at each location.

Solving the forward problem yields a field as shown in figure 4.7a, note that it is illustrated at the CMB giving rise to more detail. The radial component clearly shows a dipolar pattern, but the strength of the field varies a lot also evident from the magnitude. High intensity flux patches are prone to follow a sectoral pattern. A decrease in field strength is also present over the poles, corresponding to the placement of the tangent cylinder. The tangent cylinder is a cylinder that surrounds the inner core and can be seen as a boundary between different flow patterns (Hollerbach and Gubbins 2007). In general the field appears more intense in the southern Atlantic hemisphere. This is explained by how the core dynamo simulation was made; To account for the observed westward drift of magnetic flux patches across the Atlantic hemisphere differential growth of the inner core was implemented (Aubert et al. 2013).

Note that the actual synthetic data will be infused with Gaussian noise, $\mathcal{N}(0, 10)$ (units in nT). The noise will be applied to the magnitude which then will propagate out to the three components dependent on their size. The radial component of a synthetic data set with 2000 data-points is shown in figure 4.7b.




**(a)**




**(b)**

**Figure 4.7:** 4.7a: The radial component of the core dynamo simulation, truncated at SH-degree 60, used to generate the synthetic data illustrated at the CMB. 4.7b: The radial component of a synthetic data set with 2000 data-points generated at an altitude of 400 km above the Earth's surface and including noise.

## 4.4 Real satellite data

The real data used in this thesis is based on observations from the Swarm satellite trio. The swarm satellites were launched in 2012. Two of them fly side by side and started at an altitude of 450 km while the third started at an altitude of 530 km (Olsen and Stolle 2012). The orbital height of the satellites will slowly decrease as time passes. The satellites and their orbital path are illustrated in figure 4.8.

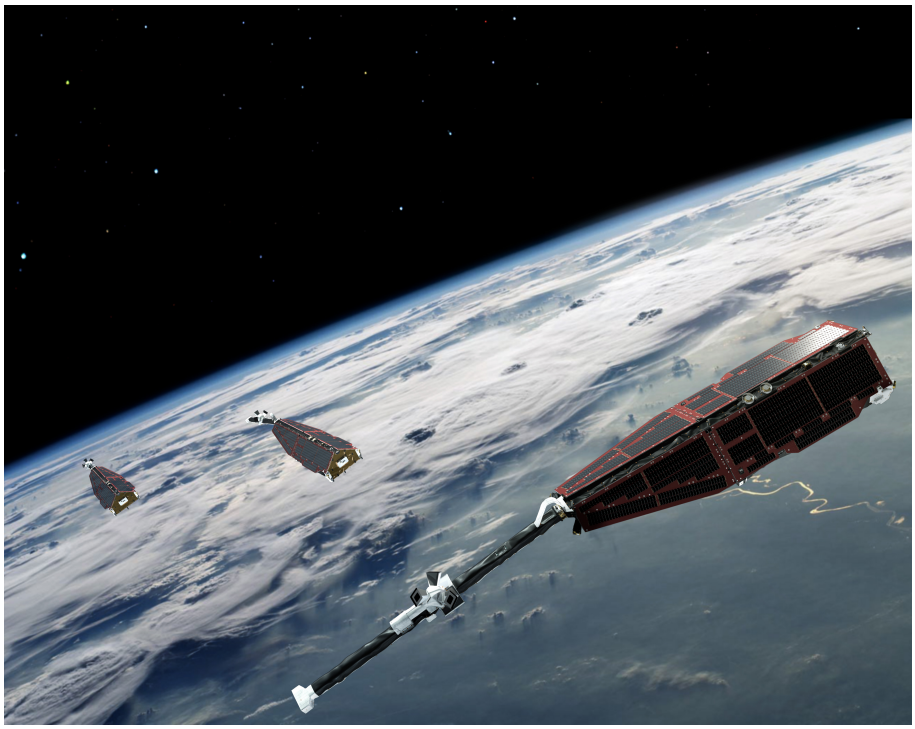
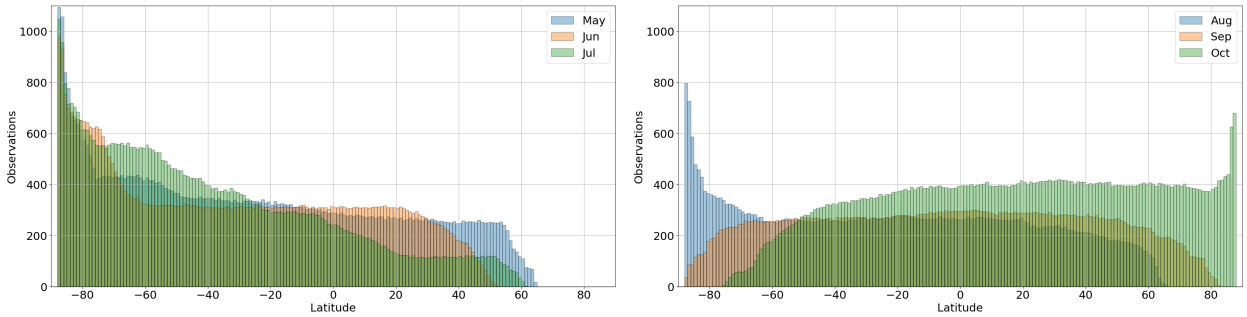


**Figure 4.8:** Illustration of the Swarm satellite trio over Earth. Image source.

A pre-processed data-set was very generously provided by Magnus D. Hammer. The data-set contains vector magnetometer observations in spherical coordinates along with error estimates of each component. The processing process is similar to that presented in Hammer and Finlay 2019, but will be summarized below.

The observations are given in a 15 seconds temporal resolution whereafter quiet time criteria are applied. Quiet time refers to periods when solar activity is low and therefore contributions by external sources are expected to be minimal.

In addition, only data from dark regions are used, meaning when the Sun is more than 10 degrees below the horizon. This criteria removes a large portion of all data, but it is necessary when examining internal sources in order to avoid solar driven sources on the day side. Regarding the quiet time criteria, it is required that $K_p$ $<2^o$ and $|dRC/dt| < 2$ nT hr$^{-1}$. $K_p$ is an index of planetary geomagnetic activity while $RC$ more directly targets the magnetospheric ring current. The disturbance due to solar activity is potentially largest under certain configurations of the Interplanetary Magnetic Field (**IMF**) that can lead to magnetic re-connection. All observations used here satisfy the criteria that $B_{IMF,z} < 0$, $|B_{IMF,y}| < 6$ nT and that the merging electric field at the magnetopause $E_m \leq 0.8$ mVm$^{-1}$ (Olsen et al. 2015). $E_m$ is given by the coupling function $0.33 v^{4/3} B_t^{2/3} sin(|\Phi|/2)$. Where $v$ is solar wind speed, $B_t = \sqrt{B_{IMF,y}^2 + B_{IMF,z}^2}$ and $\Phi = arctan\,(B_{IMF,y}/B_{IMF,z})$.




**(a)** **(b)**

**Figure 4.9:** Visualization of the distribution of observation with respect to latitude after applying the quiet time criteria. 4.9a: Latitudinal distribution between May and July 2018. 4.9b: Latitudinal distribution between August and October 2018.

Selecting quiet time data can not completely remove contributions from external sources. Magnetospheric field contributions are removed with the CHAOS-6 external model (Finlay et al. 2016). Ionospheric field contributions are removed based on the CIY4 model (Sabaka et al. 2018). Lithospheric contributions are removed using the LCS-1 model (Olsen et al. 2017). When co-estimating the lithospheric field is not removed.

When working with satellite data the density of observations with respect to latitude will be biased. It is therefore ill advised to use all data since regions of high data density will be emphasized. This problem is overcome by using an equal area grid following Hammer and Finlay 2019. The spherical surface is separated into $N$ regions of approximately equal area. Each region can contain several observations one of which will be randomly selected to populate that area. Iterating over all regions the grid will be filled. The observations retain their original coordinates making the grid only approximately equal area. It is possible that some regions will not be populated, there are two reasons for this both having to do with grid density. First, the satellites orbit inclinations of 87-88 degrees (Olsen and Stolle 2012) results in a void around the poles. If the requested amount of data, $N$, is too large some regions will lie inside the void where there are no observations. Secondly, the grid can simply be too fine compared to the density observations.

Although there are several years of Swarm data only three months are used in this thesis, to avoid secular variations and as a simple test of the new inversion methods explored here. The three months period has to be selected carefully because Earth's tilt will bias the quiet time data toward one hemisphere depending on the time of year. It is therefore important to chose a period that has global coverage. A bad example is May through July 2018, figure 4.9a. There are almost no observations above ∼60 degrees latitude which will significantly affect performance. Instead selecting August through October results in a suitable distribution over all latitudes, figure 4.9b.

The longitudinal distribution is close to uniform, but it is important to check that all longitudes are present at all latitudes otherwise there will not be global coverage. Luckily this is not a problem in the selected period.

The radial component of the real data is visualized in figure 4.10 using a grid with 10,000 points. The observations, including lithospheric contributions, have no small scale features directly visible at these altitudes only the large dipolar pattern and a weakening of the field over the southern Atlantic, figure 4.10a. It is as expected when observing the geomagnetic field

at satellite altitude.

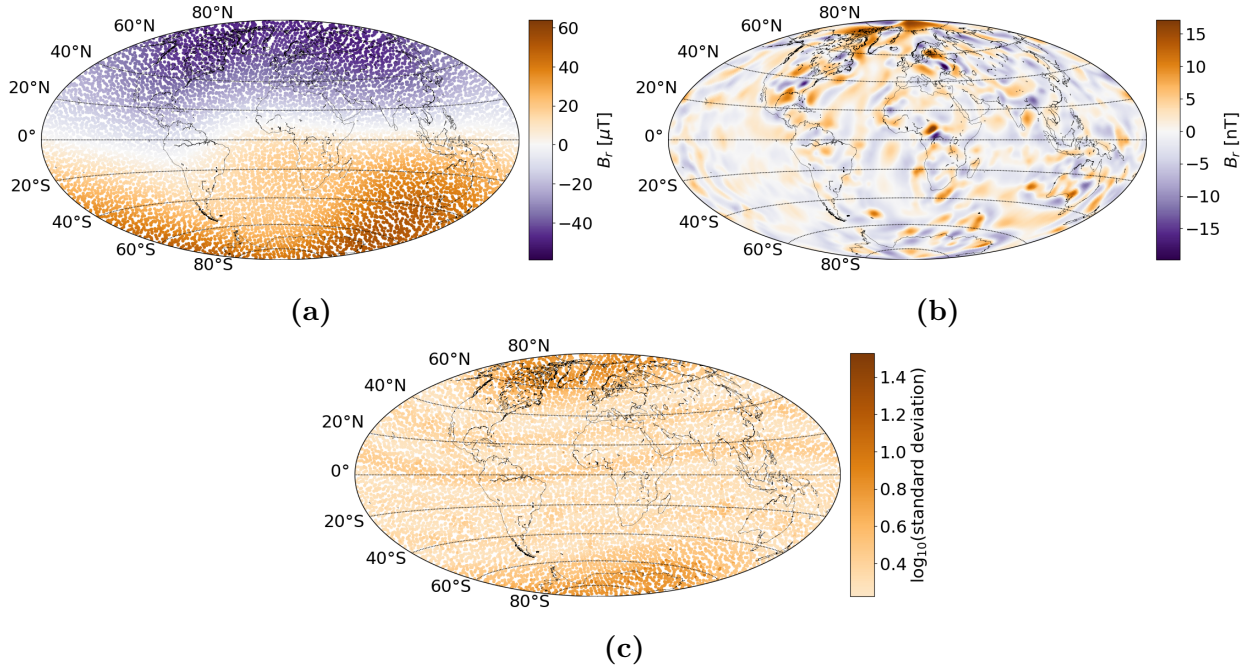



**Figure 4.10:** Visualization of real data on maps: 4.10a, the radial component of the observations from the Swarm satellite trio after quiet time selection criteria are applied. 4.10b, radial component of the LCS-1 model, at satellite altitude, which is removed from 4.10a when trying to isolate the core field in some tests. 4.10c, data variance estimates used given on a logarithmic scale to emphasize patterns.

The lithospheric field model, LCS-1, subtracted to approximate the core field, figure 4.10b, is extremely weak when compared to the magnitude of the observations, figure 4.10a. It can seem unnecessary to remove such a small contribution when the observed field is in the order of $10^5$ nT. When downward continuing to the CMB the signal is assumed to belong to the core. The small scale structures will interfere with the short wavelength harmonics leading to a significant and nonphysical power increase.

Beside observations of the three components the data-set also includes an error estimate. The method of estimating these is identical to the one presented in Hammer and Finlay 2019. Standard deviations of residuals between quiet time Swarm data, from 2013 to 2018, and CHAOS-6 are calculated for all Quasi-Dipole (**QD**) latitudes with a step-size of 2 degrees. In this way observations are assigned with error estimates solely based on their QD-latitude. Additionally, the errors are scaled with Huber weights. This ensures that observations that deviate strongly from the CHAOS-6 model will get assigned a larger error estimate.

The standard deviations are largest over the polar regions. Figure 4.10c visualizes precisely this, note that the values are logarithmic to emphasize the pattern, such as the two lines of slightly higher variance at low and mid latitudes in the northern hemisphere. The lowest of which lies on the equator in the QD reference frame.
The standard deviations assigned to the radial component are quite low, generally below 5 nT, while the ones assigned the horizontal components are a few times larger.
The distribution of the error estimates are summarized in table 4.1. The top of the table is for all error estimates while the bottom only include latitudes larger than ±60 degrees. It is clear that errors on the horizontal components are very large at high latitudes with some extreme values reaching 70 nT.

| | Mean | Min | Max | 25 % | 50 % | 75 % | 95 % |
|---|---|---|---|---|---|---|---|
| $r$ | 2.584 | 1.671 | 18.714 | 1.939 | 2.172 | 2.647 | 5.654 |
| $\theta$ | 4.474 | 2.329 | 66.576 | 2.498 | 2.714 | 3.180 | 18.128 |
| $\phi$ | 4.661 | 1.972 | 71.376 | 2.153 | 2.231 | 2.649 | 19.276 |
| | | | | | | | |
| $r$ | 5.349 | 3.135 | 18.714 | 4.293 | 5.158 | 6.311 | 7.053 |
| $\theta$ | 16.722 | 7.817 | 66.576 | 12.874 | 17.463 | 20.594 | 24.166 |
| $\phi$ | 21.482 | 6.704 | 71.376 | 16.618 | 18.964 | 24.875 | 35.277 |

**Table 4.1:** Summary of error estimates, in units of nT, assigned to the observed core field, when using a grid with 10,000 points. The upper part refers to all data while the lower are statistics for observations above ±60 degrees latitude.

Horizontal components will naturally have more noise, despite quiet time selection criteria, due to how magnetic fields are generated in the ionosphere. This is especially true in the polar regions where the ionosphere is directly connected to the magnetosphere and solar wind through field aligned currents.

If an observation has a large residual in contrast to the assigned standard deviation it will be assigned a Huber weight below one such that its error estimate is increased. The weights are truncated at one such that a weight cannot result in a reduction of an error estimate. Contrary to what one might think the radial component has the most weights smaller than one. Figures 4.11a to 4.11c show the weights, of each component, as a function of latitude along with a color scheme indicating the size of the associated residual. All weights equal to one have been discarded when making this plot and the percentage in the title indicates how many weights are below one. In all distributions the majority of weights are found at low and mid latitudes. These weights are generally smaller than those found at high latitudes. Following the orange colored points it is seen that they make a half circle over the entire latitude range. The residuals, associated with orange (3-10 nT), over low and mid latitudes are considered large and therefore small Huber weights are assigned. At high latitude orange is not enough to trigger a small Huber weight, because the error estimate already assigned is significantly larger than at lower latitudes.

The weights belonging to the horizontal components are quite evenly spread over latitude and longitude. Oppositely, the weights on the radial component are clustered in smaller groups, figure 4.11d. When comparing the positions with lithosphere models LCS-1 and Masterton, figures 4.10b and 4.6a, some appear to overlap with larger anomalies, such as west of Australia and the east coast of the USA.

When attempting co-estimation of the lithospheric field with real data LCS-1 will not be removed, but the data points that are randomly selected to populate the grid are identical to the ones used to estimate the core field. Similarly the error estimates will be reused such that the only difference is whether or not the lithospheric field has been removed.

The pattern found in figure 4.11d was later discovered to be a mistake in the calculation of the residuals. In reality the residuals were calculated as the difference between the observations with the lithospheric field removed and the CHAOS-6 without truncating it. This was discovered too late to be changed and the consequence of this is increased error estimates over areas with large lithospheric anomalies. This is not optimal, but increased error estimates should only weaken the data constraint.

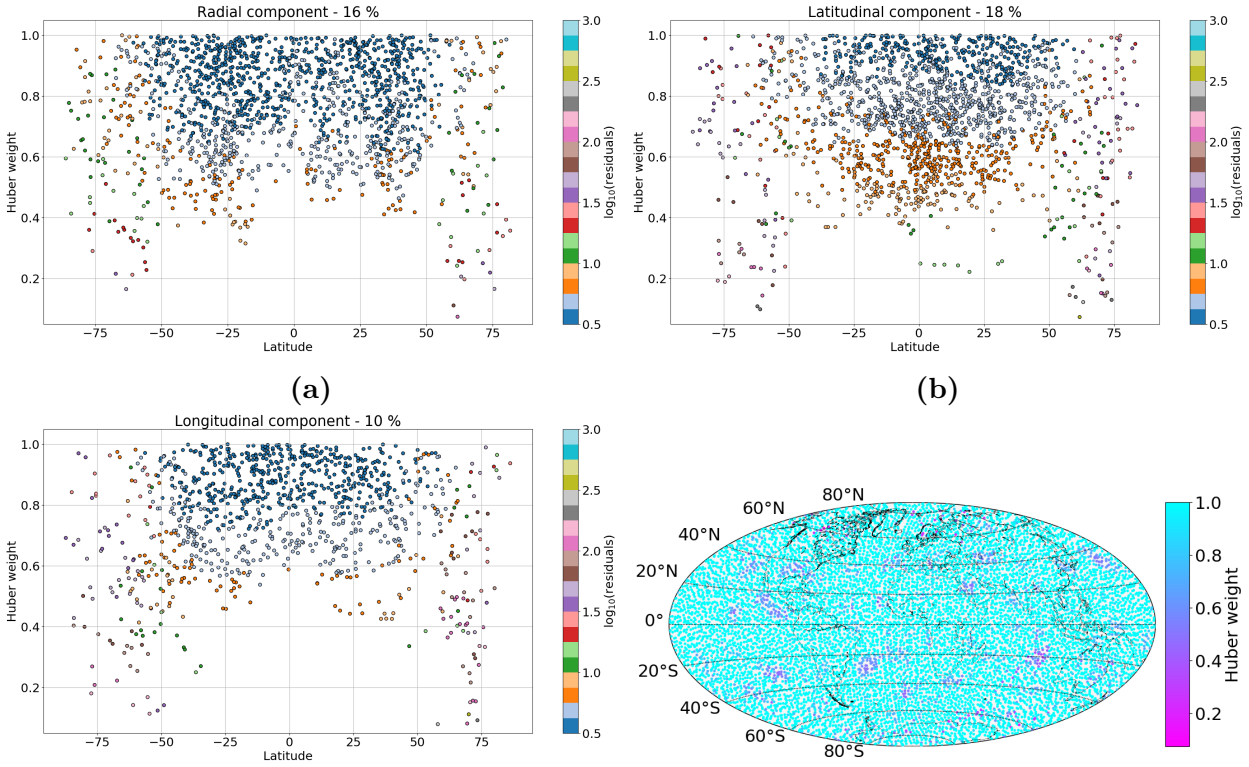



**Figure 4.11:** Visualisation of Huber weights assigned to observations that deviate from CHAOS-6 predictions. 4.11a to 4.11c: Huber weights as function of latitude with color scheme to view the residuals size. Only weights below one are illustrated and the percentage in the title summarize how many that is. 4.11d: Huber weights of the radial component shown on a map. They appear to be in clusters, some of which overlap with lithospheric anomalies. Huber weights associated with horizontal components are more uniformly distributed.

# CHAPTER 5

# Results

In this chapter the main results of the thesis will be presented. Initially there will be a section on the HMC setup and tests demonstrating how the choice of different hyperparameters affect the outcome. Then follows a section reporting on tests using synthetic data where the truth is known followed by a section with real satellite data. Finally, there is a section representing the results from attempts to co-estimate the core and lithospheric fields, showing tests with both synthetic and real satellite data.

## 5.1 Tests of the HMC setup

Although the role of all hyperparameters have been established, see section 2 and 3, the specific values chosen in this thesis have not yet been presented and justified. These choices include the choice of probability distribution used to represent the prior, how many warm-up and post warm-up iterations are needed, what the maximum tree-depth should be etc.

All tests in the following subsections will be performed on synthetic data as presented in section 4.3. One of the parameters the tests will be compared on is computational time. Table 5.1 summarizes the time spent on most tests in section 5.1.

| | GMM | | | Ind Gauss | | | Dense | | | Diag | | | No LSQ | | | No $\boldsymbol{M}$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | W | P | T | W | P | T | W | P | T | W | P | T | W | P | T | W | P | T |
| Chain 1 | 214 | 9 | 223 | 218 | 9 | 226 | 207 | 9 | 217 | 65 | 21 | 85 | 646 | 12 | 658 | 172 | 9 | 181 |
| Chain 2 | 191 | 9 | 200 | 193 | 9 | 202 | 211 | 9 | 220 | 75 | 19 | 94 | 846 | 10 | 856 | 185 | 9 | 194 |
| Chain 3 | 196 | 10 | 206 | 193 | 10 | 204 | 206 | 9 | 216 | 68 | 19 | 87 | 563 | 12 | 575 | 182 | 9 | 191 |
| Chain 4 | 217 | 9 | 226 | 212 | 9 | 221 | 194 | 9 | 204 | 74 | 18 | 92 | 523 | 11 | 534 | 178 | 9 | 187 |
| Average | 205 | 9 | 214 | 204 | 9 | 213 | 205 | 9 | 214 | 70 | 19 | 89 | 644 | 11 | 655 | 179 | 9 | 188 |

**Table 5.1:** Summary of computational time spent on models in section 5.1. The time is given in minutes and rounded to the nearest integer. W, P and T refer to warm-up, post warm-up and total time, respectively. If nothing else is stated the runs are made with a SH truncation degree of 20, 4000 warm-up iterations, 2000 post warm-up iterations, a maximum tree-depth of 10, a dense mass matrix, starting point set to the least squares solution and initialized with the mass matrix equal the covariance matrix prior probability. **GMM**: Run using the independent GMM prior. **Ind Gauss**: Run using an independent Gaussian prior. **Dense**: Run using a multivariate Gaussian prior. **Diag**: Run using a multivariate Gaussian prior with a diagonal mass matrix. **No LSQ**: Run using a multivariate Gaussian prior, without the LSQ starting point. **No $\boldsymbol{M}$**: Run using a multivariate Gaussian prior, without initializing the mass matrix.

### 5.1.1 Choice of prior distribution

The prior information is very important in HMC for achieving good results and reducing computational time by guiding the sampler towards what is believed to be the correct path. The model coefficients are assumed Gaussian distributed which is not entirely true when considering

the core dynamo realizations, figure 4.1c. To justify the assumption its performance will be compared to a prior based on a two component GMM, figure 4.5, from here on referred to as the GMM prior.

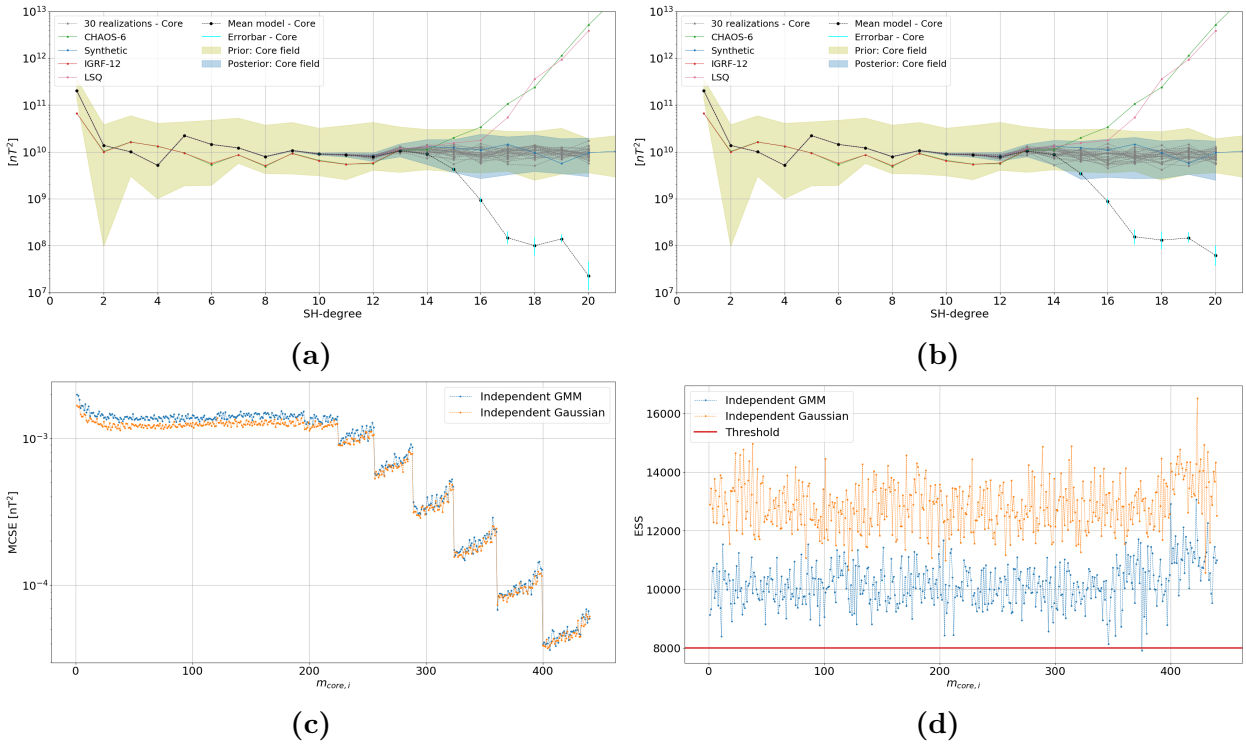



**Figure 5.1:** Comparison between assuming the SH coefficients independent and Gaussian distributed and distributed according to a two component GMM. 5.1a and 5.1b: Power spectrum, at the CMB, of the posterior; Left: Independent Gaussian. Right: Independent GMM. Note how the mean model is very low in power. 5.1c: MCSE of both runs. The structure follows the diagonal of the posterior covariance. Results using a GMM prior have a slightly higher MCSE. 5.1d: ESS of both runs. The GMM prior results in a lower ESS which is why its MCSE is higher.

In the following two tests the SH coefficients will be assumed independent. The tests were truncated at SH-degree 20, with 2000 synthetic data-points of the core field, 4000 warm-up and 2000 post warm-up iterations, a maximum tree-depth of 10 and initialized with the least squares (**LSQ**) solution, defined in section 3.1.3, along with a mass matrix equal to the covariance matrix of the prior probability distribution, figure 4.2a. They were run with four chains resulting in a total of 8000 post warm-up samples.

Both runs passed the diagnostics native to STAN, which means $\widehat{R}$ and E-BFMI, see section 3.3.1 and 3.3.2 for definitions, were within the thresholds and no post warm-up iterations were terminated prematurely.

The resulting power spectra, figure 5.1, are very similar to each other. Both of the posterior distributions are well constrained by the data until SH-degree 13 where after they widen and follow the boundary of the prior. The power of the mean model decreases significantly after SH-degree 14 suggesting that some marginal posterior distributions have mean zero, given that the random realizations behave well. Note also that the MCSE appears larger at higher harmonics, but that is only due to the logarithmic scale it is illustrated on. In fact the opposite is true as seen from figure 5.1c illustrating the MCSE for both tests. As a result of the ESS, figure 5.1d, being fairly constant the MCSE follows the posterior covariance structure closely, recall

equation 3.25. The MCSE belonging to the GMM prior is slightly higher which corresponds nicely with it having a lower ESS. Note also that all coefficients except one has an ESS above 8000 making practically every sample independent.

The difference between the two runs appear to be caused by a difference in step-size and therefore also the mass matrix. The ideal mass matrix, as discussed in section 2.3, will decorrelate phase space resulting in uniform Hamiltonian trajectories that are easily traversed. Given that the average step-size, between the four chains, approximately is 0.275 for the GMM prior and 0.3 for the Gaussian prior emphasizes this.

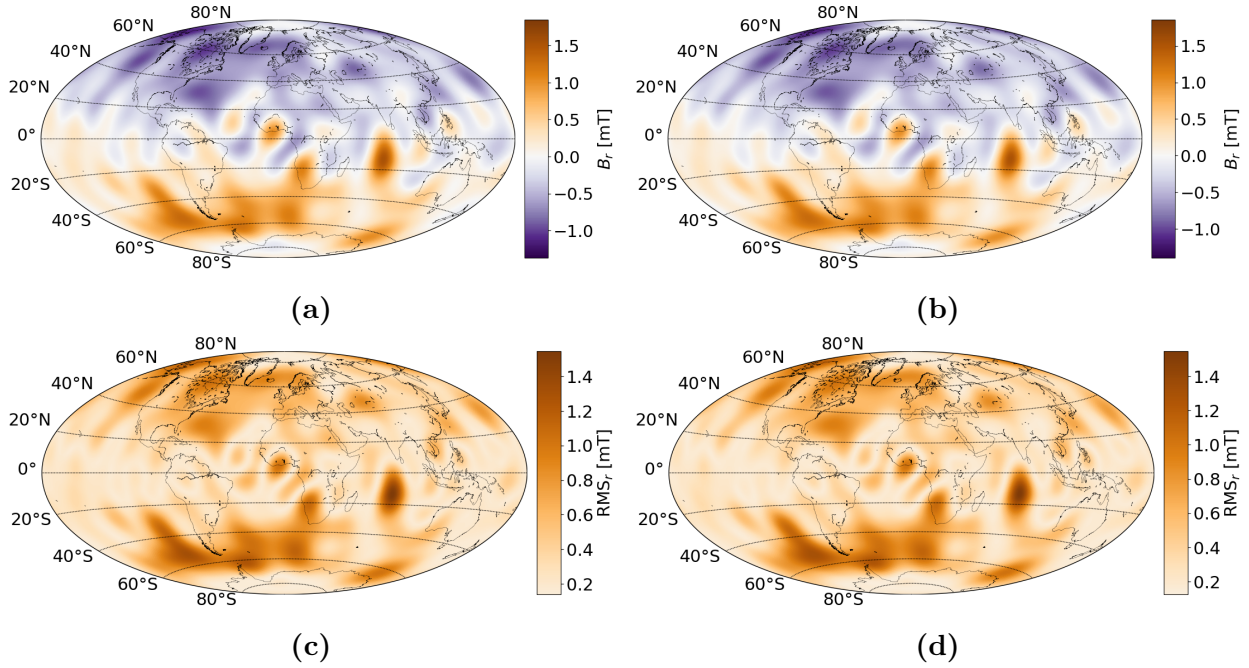



**Figure 5.2:** Various maps of the radial component, at the CMB, from two models; Left: Results of an independent Gaussian prior. Right: Results of an independent prior with distribution given from a two component GMM. First row is the mean model of the posterior. The second is RMS of the radial component revealing that the flux patches in 5.2a and 5.2b are well constrained.

Visualizing the mean models at the CMB, figure 5.2, shows similar features on both maps. The dipole pattern is clear with an intensity increase in the southern Atlantic hemisphere. A few flux patches lie along the equator and if consulting RMS maps, defined in section 3.3.7, the flux patches are well constrained.

With regards to computational time the two approaches are equal, table 5.1. This would change if the amount of warm-up iterations were to be increased with respect to the GMM prior to achieve a larger step-size, higher ESS and lower MCSE. But the main question needed to be answered was whether or not a Gaussian prior was a good approximation of the posterior distribution. When applying the GMM prior all marginal posterior distributions are fitted nicely by a single Gaussian, exemplified in appendix figure 8.1. Based on these findings a Gaussian prior is assumed a valid choice.

### 5.1.2 Influence of multivariate prior

In the previous section it was shown that an independent Gaussian prior performs equally as well as, or better than, to a more complex GMM prior. The question addressed in this section is whether or not the coefficients should be assumed independent.

A run with a multivariate Gaussian prior, see section 3.2.3, is made with similar hyperparameters as applied in the previous two runs. Its power spectrum, figure 5.3a, is very similar to the independent case. The spread is very low until SH-degree 13 where after it widens and follows the prior, although it is slightly more constrained.

More important is the mean model which no longer is extremely low in energy due to the inferred correlation between coefficients. This creates a correlation structure in the posterior distribution, figure 5.4a. In both independent cases there was no off-diagonal structure in the posterior correlation matrix. Note that it first begins after SH-degree 12 (168 coefficients) suggesting that the data constraint is sufficient up to this degree, after which the prior kicks in. Because of the reduced off-diagonal structure only three of the five correlation lines are visible, see figure 4.2b for comparison. Additionally, the radial component of the mean model, figure 5.3b, is more detailed. The higher harmonics now contribute to the small scale structure because their distributions no longer are centered around zero. This is exemplified by marginal posterior distributions in appendix figure 8.2.

The RMS map, figure 5.3c, is very similar to the mean model indicating that the random realizations all contain the same flux patches.

Applying the multivariate Gaussian prior does not come with additional expenses to the computational time, table 5.1.

Overall, one can conclude that the multivariate Gaussian prior is a significant improvement. And should thus be preferred over the independent Gaussian prior.

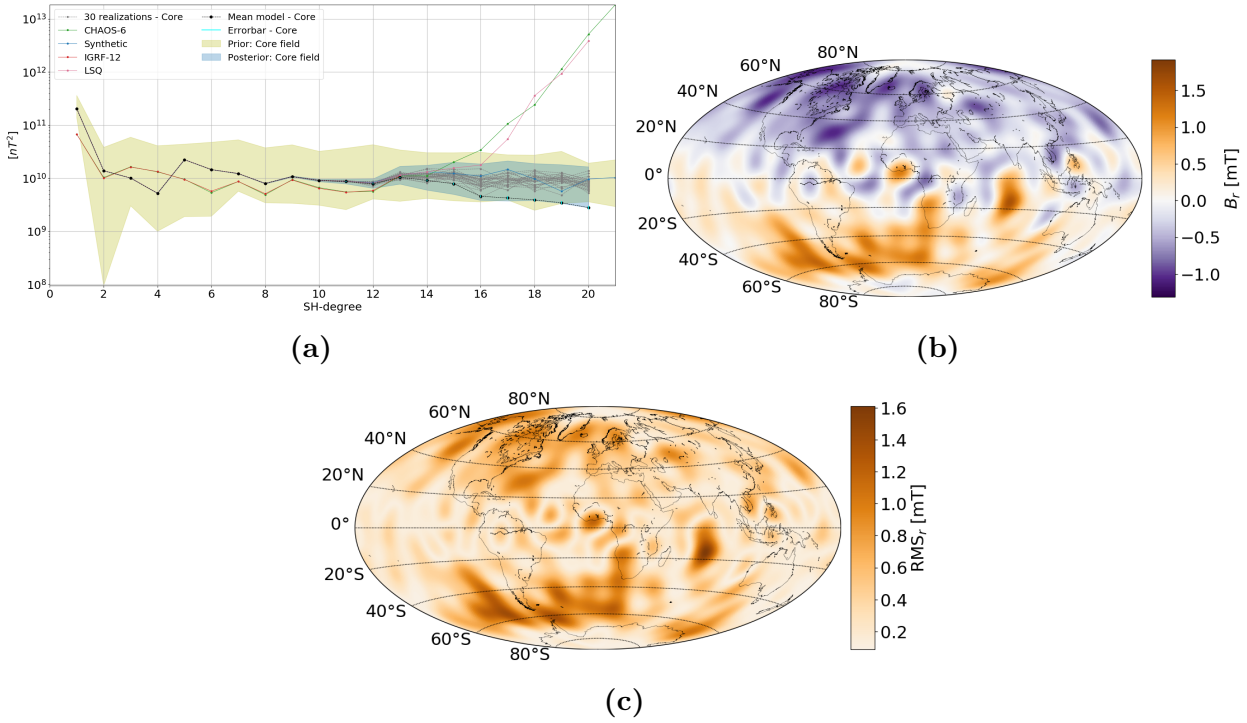



**Figure 5.3:** Power spectrum and mean model of the test case using a multivariate Gaussian prior, illustrated at the CMB. 5.3a: The power spectrum is slightly more constrained than in the independent cases. Additionally, the posterior mean model's power is higher because the marginal distributions of the posterior no longer have mean zero. 5.3b: As a result of the increased power the projection of the mean models radial component onto a map shows greater detail. 5.3c: RMS of the radial component shows that the flux patches in 5.3b are well constrained.

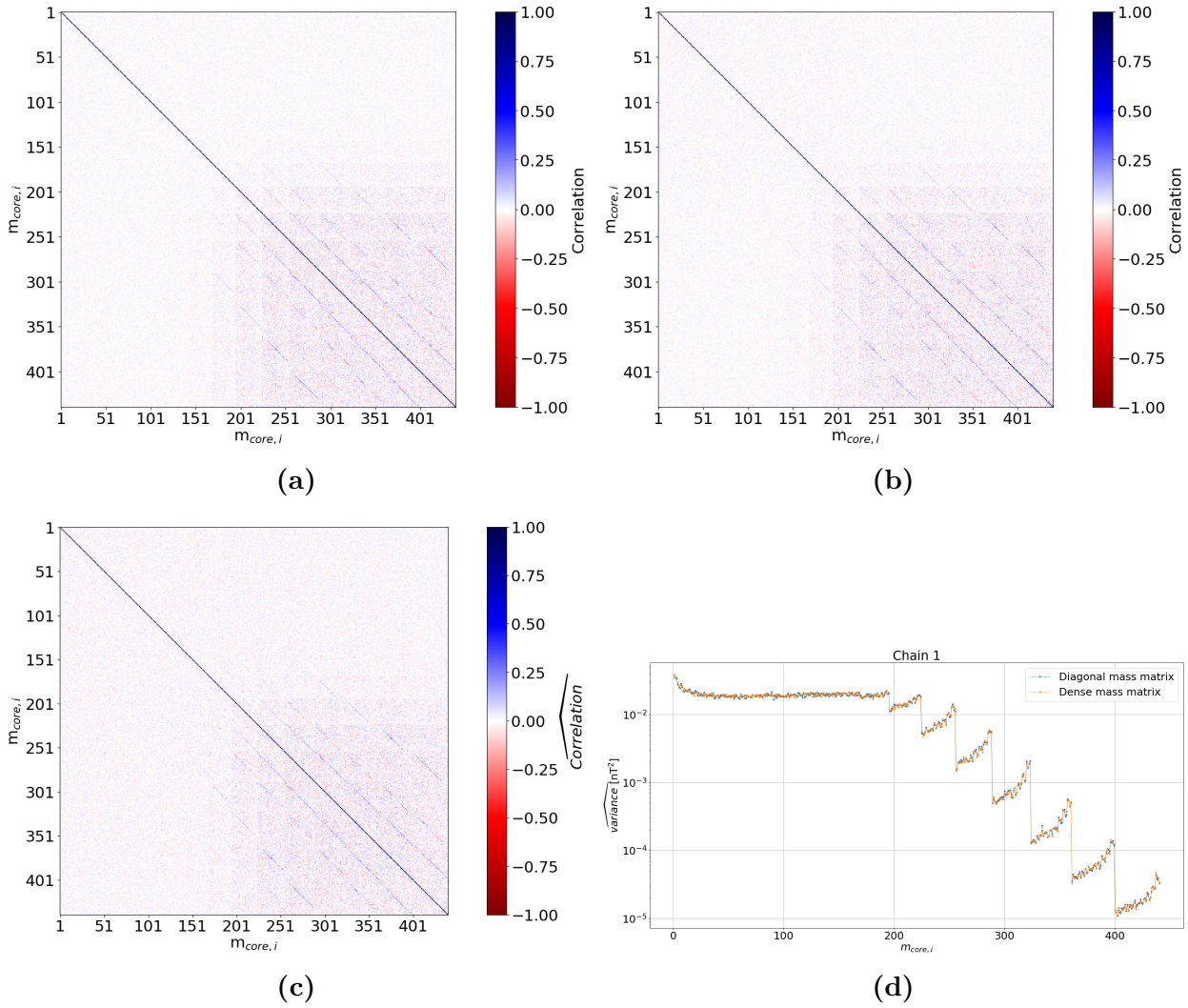



**Figure 5.4:** Comparison of correlation and mass matrices when using a dense or diagonal mass matrix. 5.4a: The posterior correlation matrix when using a dense mass matrix. Note how the structure first begins around SH-degree 12. Before which the data appears to constrain it well. 5.4b: The posterior correlation matrix when using a diagonal mass matrix. Note how the off-diagonal elements before SH-degree 12 are less noisy than in 5.4a. 5.4c: The mass matrix from the first chain when using a dense mass matrix. The structure is similar to the correlation matrices, but the off-diagonal elements are even more noisy. 5.4d: Comparison of the diagonal elements, in the mass matrices, when using dense or diagonal mass matrices. There is no significant difference and thus the performance difference must be due to the noisy off-diagonal elements when using the dense mass matrix.

### 5.1.3 Use of mass matrix

Until now the mass matrix has been assumed dense. Having a well approximated dense mass matrix can improve the sampling speed, but estimating it during the warm-up period takes time. If the off-diagonal correlation structure within the posterior is minor it can be a waste of time to compute the full matrix. In other words, if the model coefficients are close to independent. The alternative is to approximate the diagonal letting all off-diagonal elements be zero. In the following the trade-off between diagonal and dense mass matrices is examined. Results using a dense mass matrix and a multivariate Gaussian prior were already shown in section 5.1.2.

The run using a diagonal mass matrix was given the same hyperparameters as all previous runs, but in the initialization only the diagonal of the prior covariance matrix was passed. Again the diagnostics native to Stan issued no warnings. In fact there appears to be no performance difference between using a dense or diagonal mass matrix when comparing their power spectra, mean model and RMS maps, figure 5.5 and 5.3. The power spectrum widens after SH-degree 13 and follows the prior. The mean model in both cases appear to have equal power. Identical patterns are found in both the mean model and RMS. Neither does the assumption of the mass matrix change the posterior correlation structure, which can be compared in figure 5.4a and 5.4b. A keen eye might notice that the noise is smaller when using a diagonal mass matrix.

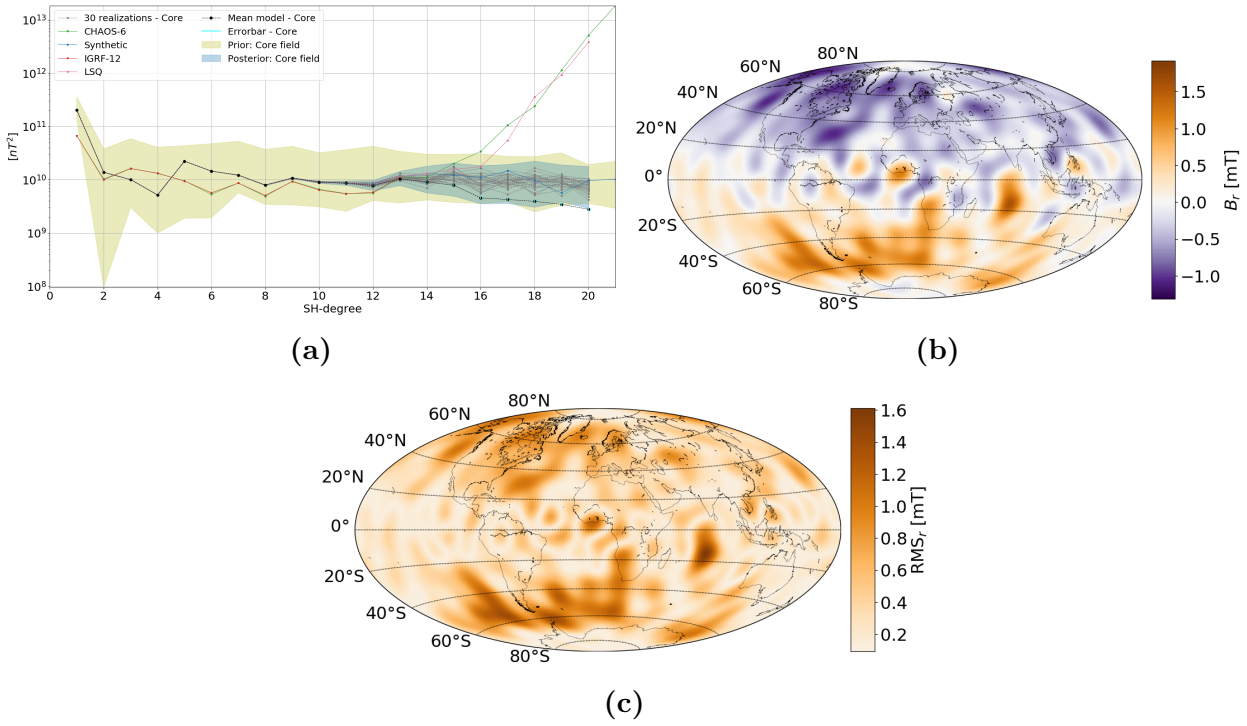
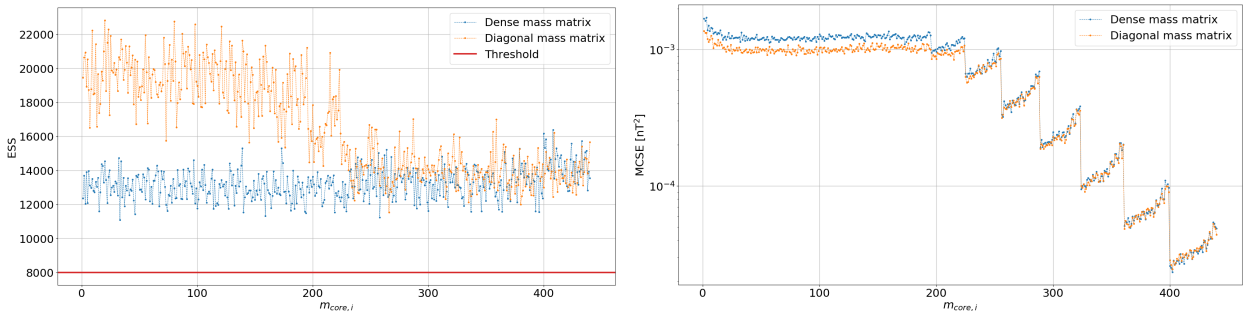



**Figure 5.5:** Power spectrum, mean model and RMS, at the CMB, of the test case using a multivariate Gaussian prior and a diagonal mass matrix. The performance when using a diagonal mass matrix appear to equal that of a dense mass matrix, figure 5.3. 5.5a: The power spectrum is close to identical to the run with a dense mass matrix. 5.5b and 5.5c: The mean model and RMS, of the radial component, have the exact same structure and detail as the run with a dense mass matrix.

The mass matrix is approximated throughout the warm-up period. The variance of each coefficient is easier to approximate, as is clear from figure 5.4d comparing the diagonal of a dense and diagonal mass matrix. The off-diagonal elements, on the other hand, are more difficult to estimate. Without a sufficient amount of samples belonging to the typical set or enough data constraint the approximation of off-diagonal elements can become noisy, figure 5.4c. Note the

off-diagonal elements after SH-degree 12 generally are larger than the corresponding elements in the posterior correlation matrices. Additionally, if the warm-up period is too short the mass matrix might not have converged.

It is believed that, as a consequence of the noise introduced by the mass matrix the ESS is reduced between SH-degree 1-15 (224 coefficients). The lowered ESS cause the MCSE to increase, figure 5.6b.




**(a)** **(b)**

**Figure 5.6:** Comparison of the ESS and MCSE when using a dense or diagonal mass matrix. 5.6a: The ESS in both cases is above 8000, which means that all samples are independent. Between SH-degree 1-15 the diagonal mass matrix results in a higher ESS. 5.6b: The MCSE follows the shape of the posterior variance. In case of the dense mass matrix the MCSE is slightly higher between SH-degree 1-15 due to the ESS difference, recall equation 3.25.

Before criticizing the use of the dense mass matrix too much it should be pointed out that the correct off-diagonal correlation structure does significantly help by creating a uniform phase space. The average step-size when using the dense mass matrix is 0.3 and only 0.18 with the diagonal mass matrix. As a consequence twice as many leapfrog steps are taken in the post warm-up period when using a diagonal mass matrix. If the complexity of the problem was increased, the SH truncation degree, the dense mass matrix might prove very valuable to keep the amount of computations down. Note also that the synthetic data is very well behaved and that the real satellite data does not necessarily constrain the lower harmonics as much as the synthetic data. Taking this into account as well as the time spend on the runs being equal, table 5.1, the dense mass matrix is the preferred choice.

### 5.1.4 Length of (post) warm-up and tree-depth

In the previous section the trade-off between using a dense or diagonal mass matrix was examined. The dense mass matrix was chosen due to its potential performance improvements with increased model complexity and less constraining data. Hyperparameters such as maximum tree-depth, the amount of warm-up and post warm-up iterations has not been taken into account. Here the consequences of using suboptimal values of these parameters are examined.

The post warm-up samples are fundamental in the comparison of performance. It is therefore important that enough samples are taken. Ideally the amount of sampled used to estimate the posterior should be infinite, but that is not possible. Instead the amount is chosen so that the posterior distribution is believed to have converged. With an insufficient amount of samples the posterior can be wrongfully estimated and the basis for comparison is gone. In all previous runs 2000 post warm-up samples were taken and every time the diagnostics native to Stan were

passed. The $\widehat{R}$ statistic, see section 3.3.1 for definition, specifically describes convergence of the posterior distribution. Thus 2000 post warm-up samples are enough to assume that the posterior distribution is stationary.

In the warm-up period the mass matrix and step-size are estimated and it can take time before they converge. If the period is too short they will not converge. Similarly if the maximum tree-depth is too low the sampler will continuously make suboptimal choices which can slow down convergence. It is therefore well advised to examine the warm-up period so to avoid counterproductive behaviour. By initially looking at the tree-depth throughout the warm-up, for the two previous cases; dense or diagonal mass matrices, figures 5.7a and 5.7b. It is apparent that using the diagonal mass matrix will make the sampler converge faster toward a low tree-depth and that the maximum tree-depth is initially exceeded in both cases.

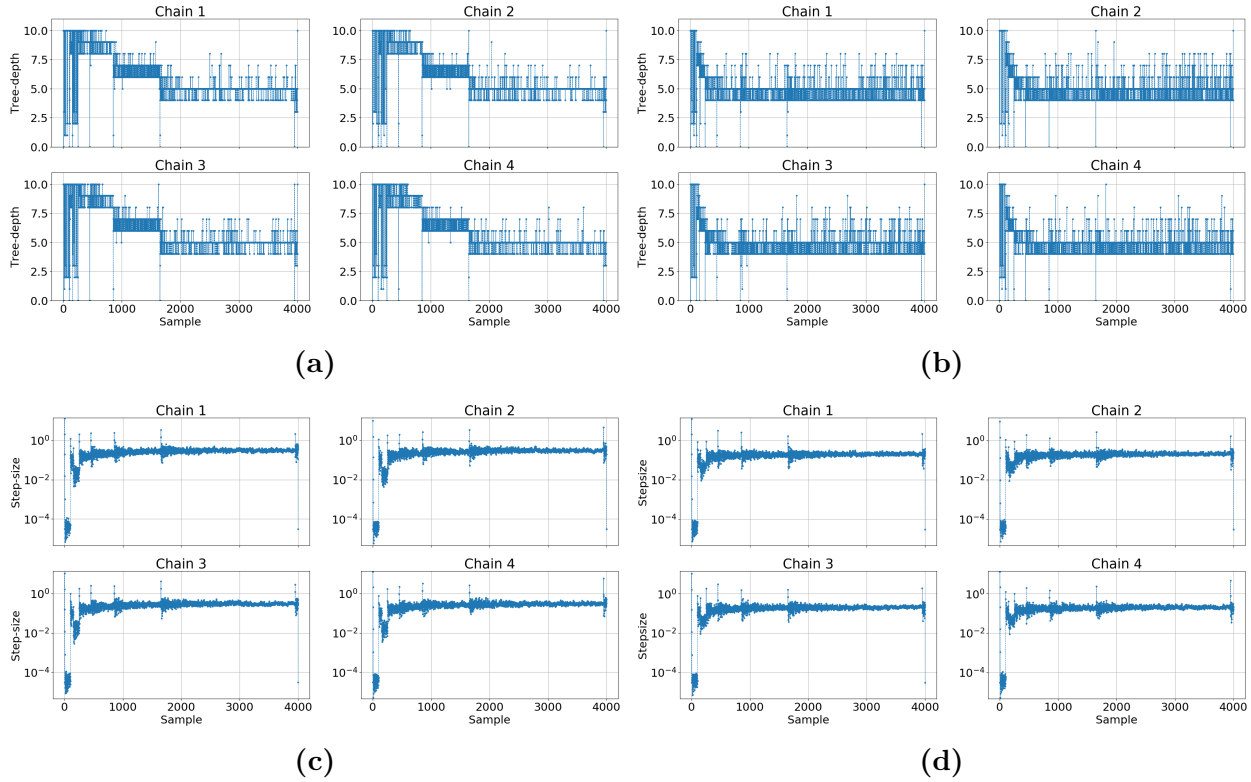



**Figure 5.7:** Comparison of tree-depth and step-size in the warm-up period when using a dense (left) or diagonal (right) mass matrix. Note that all chains have similar behaviour. 5.7a and 5.7b: Tree-depth in the warm-up period. The run using a diagonal mass matrix converges quickly towards a tree-depth of five, but the time series continues to be noisy. The run with the dense mass matrix converges slower, but ends up being more stable. 5.7c and 5.7d: Step-size in the warm-up period. The step-size initially drops to a magnitude of $10^{-4}$, because the initial mass matrix does not represent the posterior covariance well. It quickly bounces back as the mass matrix becomes well estimated. Note that the third phase in the warm-up is too short and the step-size does not converge. This affects the post warm-up acceptance rate, appendix figure 8.5.

The quick convergence rate when using a diagonal mass matrix confirms the previous statement that the diagonal of the posterior covariance is easily estimated, while the off-diagonal elements are more tricky. While the dense mass matrix is more difficult to estimate it does result in a more stable convergence of the tree-depth. This shows that the off-diagonal correlation structure is important for the samplers performance given the problem presented in this thesis.

The fact that the tree-depth is exceeded in both cases indicate that the initial mass matrix is not as good a representation of the posterior covariance. Given that it also occurs in the diagonal case suggests that it is not specifically the off-diagonal structure, but rather the size of the (co)variance that is the problem. This is only emphasized by time series of the step-size throughout the warm-up, figures 5.7c and 5.7d. The default step-size of one is quickly corrected to a magnitude of $10^{-4}$ to achieve the desired acceptance rate of 80 % given the initial mass matrix. This is a very small step-size and many leapfrog steps are needed to meet the NUTS termination criteria, see section 2.4 for definition. For this reason the initial iterations are terminated prematurely.

Increasing the maximum tree-depth would be a natural reaction, but doing so would not help, figure 5.8a. Here the maximum tree-depth has been increased to 12, allowing four times the amount of leapfrog steps. The first ∼300 iterations still exceed the maximum tree-depth and the warm-up time has increased by 75 %. It is possible to further increase the maximum tree-depth, but at a greater cost to the computational time. Note that the outcome of both warm-up periods are equivalent, there is no performance improvements. Additionally, a natural plateau is found at around 10 after the first few hundred iterations. For this reason the maximum tree-depth will be kept at 10.

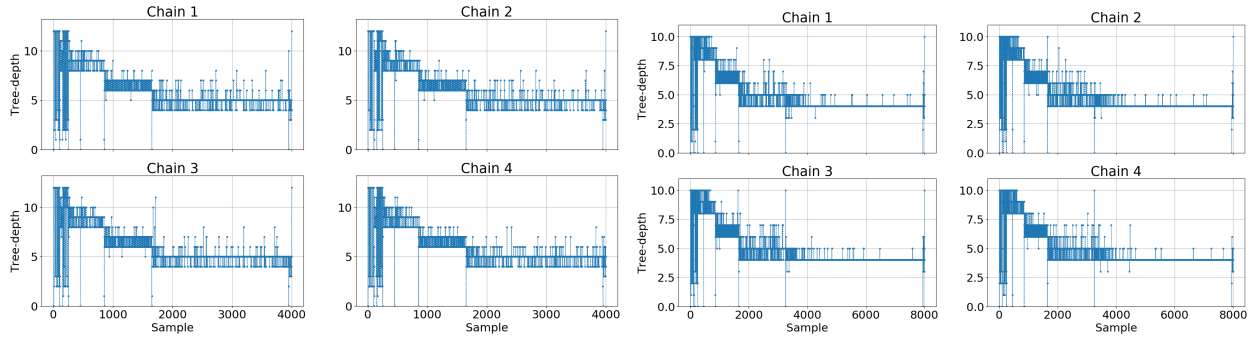



**Figure 5.8:** Time series of the tree-depth during the warm-up period when increasing the maximum tree-depth from 10 to 12, 5.8a, or increasing the amount of warm-up iterations from 4000 to 8000, 5.8b.

Instead the length of the warm-up period could be extended allowing the mass matrix and step-size to be fine tuned. In figure 5.8b the maximum tree-depth is kept at 10, but the amount of warm-up iterations has been doubled, 8000. By doing so an additional mass matrix adaption is achieved in the second warm-up phase. Now the tree-depth is holding steady at five and the computational time has only increased by 20 %. The extra warm-up does not help post warm-up performance. The average post warm-up step-size is still 0.3 even though the final mass matrix has a lower noise level, appendix figure 8.3. As a result the power spectrum is no more constrained than before, appendix figure 8.4. Thus 4000 warm-up iterations is believed to be suitable.

In section 3.1.3 it was mentioned that the first and third warm-up phases had lengths 75 and 50 by default, respectively. The first phase will be discussed further in the following section. The third phase has the purpose of fine tuning the step-size after the last mass matrix adaptation. It is clearly visible from figure 5.7c that every time a mass matrix adaptation takes place the steps-size is very unstable in the first ∼100 iterations and after ∼400 iterations it seems to have converged. Therefore the third warm-up phase is increased to 400 iterations. Recreating

the run with a dense mass matrix and similar hyperparameters except for the increased length of phase three yields an average step-size of 0.33 which causes the acceptance rate in the post warm-up period to oscillate around 80 % instead of 85 %, appendix figure 8.5. This is a small improvement and did not change the outcome of the particular run, but it does not cost any additional computations and it is silly not to take full advantage of the decorrelated phase space.

The length of the second phase will not be altered as it has not been tested. It is possible it should be increased to allow the step-size to converge more in the initial adaptations.

### 5.1.5 Role of having a warm guess

Until now all models have been initialized with the LSQ solution as starting point and the prior covariance as initial mass matrix. In this section the effects of using these initialization arguments are tested by making three different test runs.

**Test 1.** Without initializing the mass matrix.

**Test 2.** Without initializing the starting point.

**Test 3.** Without any initialization.

The three runs are otherwise identical to the one presented in section 5.1.2, which will be referred to as **test 0**.

**Test 1:**
Without initializing the mass matrix it is immediately apparent that the warm-up period is different when comparing its tree-depth, figure 5.9a, to that of **test 0**, figure 5.7a. The tree-depth is initially keeping fairly stable at 10, which can be explained by the corresponding step-size being a factor 10 larger than that of **test 0**, appendix figure 8.6. From this quick inspection the default mass matrix, the identity matrix, appears promising.

The first phase of the warm-up was mentioned in the previous section. By comparing the absolute log posterior from the first 900 iterations, figure 5.9e, it is clear that **test 0** does not reach the typical set within the first warm-up phase. Instead it happens after the first mass matrix adaptation. Oppositely, **test 1** reaches it in the first eight iterations. Note that all tests converge to an absolute log posterior value of $3\times10^4$ which is assumed to be the typical set.

It would be natural to think that using the default initialization, of the mass matrix, is the better option. Although in the remaining warm-up period **test 0** and **test 1** are similar to each other. Finally, looking at the post warm-up performance **test 0** has the highest ESS and lowest MCSE, figure 5.9c and 5.9d. Considering that they have identical posterior variance and average step-size this can only be caused by differences in the estimated mass matrices. Within the 4000 warm-up iterations a total of seven mass matrix adaptations occur. Comparing each step in the mass matrix evolution, appendix figures 8.7 and 8.8, as well as the correlation structure within, appendix figure 8.9 and 8.10, gives great insight. It would appear that only the prior correlation is a good representation, while the size of the covariance is far off. For that reason **test 0** appears to always be behind **test 1**. The final mass matrix of **test 0** is more noisy, but it can not be argued that the ESS and MCSE is in favor of **test 0**. The underlying reasons for this has to be examined further, but as of now runs will continue to be initialized using the prior covariance, so to encourage the use of custom mass matrices.

Initializing the mass matrix as the identity matrix appeared to be more successful based on its performance during the first few hundred warm-up iterations, but when comparing ESS and MCSE **test 0** proved to be better. Additionally, it was revealed that **test 0** did not reach the typical set within the first warm-up phase. In light of that several possibilities presents themselves.
First, increase the length of phase one allowing the sampler to reach the typical set. Because of the low step-size this would not be practical due to a significantly increased computational time.
Second, **test 0** could be initialized with a better starting point, preferably within the typical set. If the LSQ solution is to provide better initialization more data is required, but with increasing model complexity the amount of data would also have to increase, likely beyond what

is available. More robust methods such as regularization could be employed, but this was not done.
Third, shorten phase one such that the step-size is allowed to converge toward the target acceptance rate. Additionally, accept that the first one or two sections of phase two are used to reach the typical set.
Fourth and last, improve the initial mass matrix guess.
A mixture of the third and fourth proposal will be employed. It is accepted that the sampler may not reach the typical set within the first warm-up phase, although it would be preferable. Models with SH truncation degree larger than 20 can be initialized with a mixture of the prior covariance and the posterior covariance matrix from a previous less complex run.

**Test 2 and 3:**
It is unclear why initializing the mass matrix as the prior covariance performs better, but it is only confirmed when comparing **test 2** and **3**.
It is clear from the tree-depth time series of **test 2**, figure 5.9b, that not initializing the starting point is a large setback to the warm-up. Despite it the ESS and MCSE are better than **test 1**, figure 5.9c and 5.9d. Keep in mind that the sampler in **test 2** first reaches the typical set after the fourth, out of a total of seven, mass matrix adaptations. As a result it is much slower apparent from table 5.1.
**Test 3** is not illustrated anywhere because it did not pass the native Stan diagnostics. Suggesting that the prior covariance indeed does have a positive effect on the warm-up period.

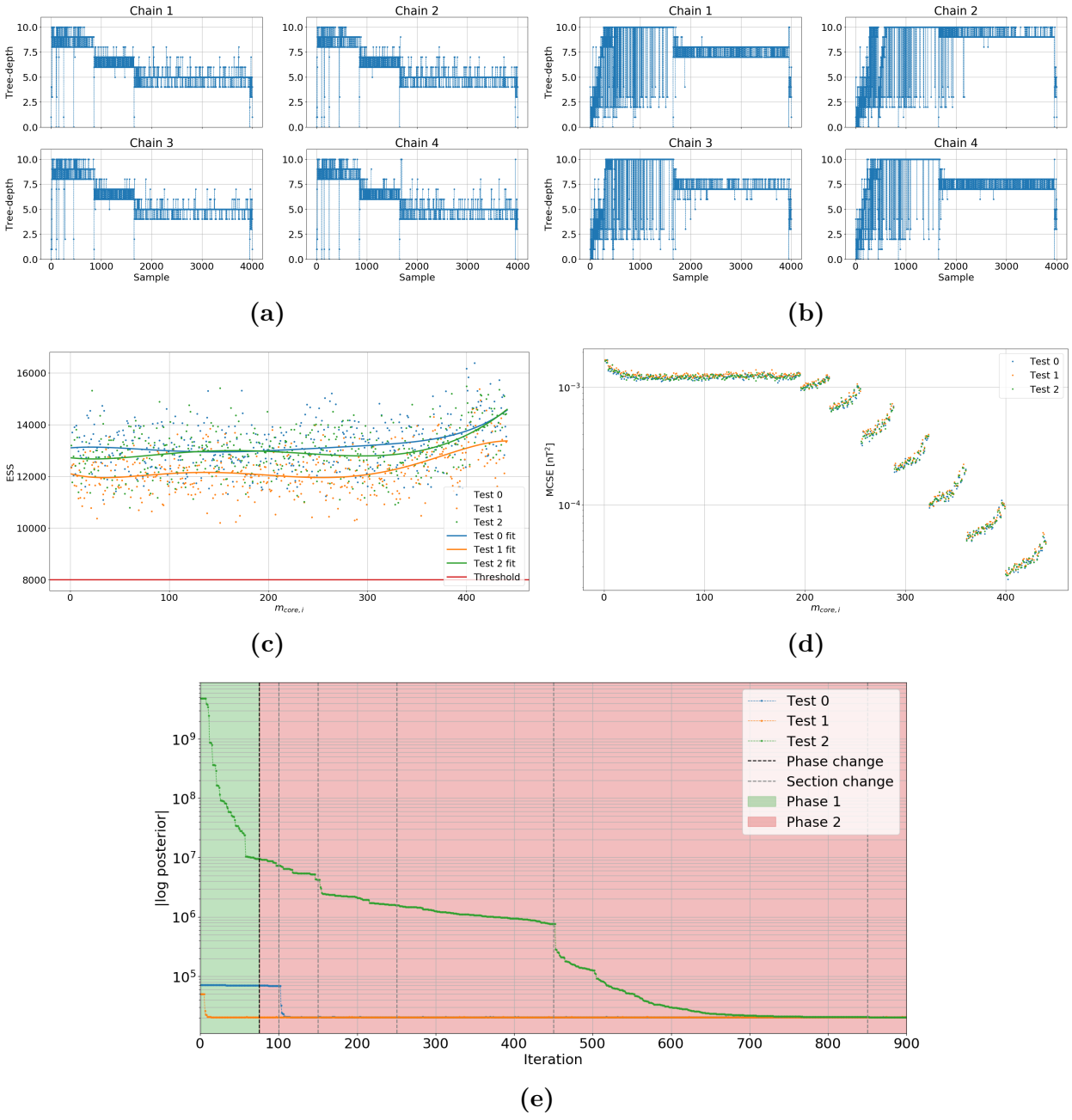



**Figure 5.9:** Comparison of the performance between **test 0-2**. 5.9a: The tree-depth of **test 1** in its warm-up period. 5.9b: The tree-depth of **test 2** in its warm-up period. 5.9c: A comparison of ESS between **test 0-2**. To help the reader a five order polynomial has been fitted to each time series. 5.9d: A comparison of MCSE between **test 0-2**. 5.9e: Visualization of how fast the sampler, chain 1, reaches the typical set. The green and red shaded area marks phase one and two of the warm-up, respectively. The grey horizontal lines indicate transitions between phase two sections, a mass matrix adaptation.

## 5.2 Tests with synthetic data

Synthetic data is typically used when the focus is entirely on understanding the sampling method and on testing the accuracy with which the chosen method can recover the truth. In that way the uncertainties and assumptions that comes with real data are eliminated. This section will show the capabilities of the HMC sampler under good conditions and the results will give an expectation of the performance with real data.

### 5.2.1 Data constraint

Although all runs until now, except one, have successfully sampled to SH-degree 20, it does not mean that they can not be improved. Immediately after SH-degree 12 the power spectrum widens and fills almost the entire prior. At this point the prior is the main constraint and no stronger information is available. The only way to increase the constraint is through data, simply by increasing the amount. This is illustrated very nicely in figure 5.10 where four power spectra are shown. The first power spectrum, figure 5.10a is made using 5000 data-points which is 2.5 times more than previously used. The power spectrum remains well constrained until SH-degree 14 and then gradually widens. Increasing to 10,000 data-points results in a constrained fit to SH-degree 15, figure 5.10b. Further increasing to 20,000 data-points is a minor improvement, figure 5.10c. Going to 100,000 data-points constrains the fit to SH-degree 16-17, figure 5.10d, and the last part up to degree 20 is slightly improved. Note how the LSQ solution improves with increasing amounts of data. This is also evident from the absolute log posterior of the initial warm-up iterations, figure 5.11a, where the jump in power gets smaller and smaller. Additionally, the magnitude of the posterior in the typical set increases with the amount of data.

The amount of data could be increased even further at a continually increasing cost to the computational time. Sampling with 5,000 data-points takes approximately 10 hours. Sampling with 10,000 and 20,000 data-points takes 18 and 40 hours, respectively. At this rate sampling with 100,000 data-points would take around 8 days, but as it was initialized with information from the posterior of a previous run the amount of warm-up iterations were reduced from 4000 to 1000 resulting in it being complete in two days.
Future runs will be restricted to a maximum of 20,000 data-points.
Besides a more constrained power spectrum increasing the amount of data decreases uncertainties, clear from the MCSE in figure 5.11b.

### 5.2.2 Increasing the SH truncation degree

SH-degree 20 was used as a reference point, but the prior information allows for models up to SH-degree 25. Attempts at increasing the truncation degree beyond 20 have been successful, but the computational time increases again due to higher model complexity resulting in a low step-size and high tree-depth if the mass matrix is not well estimated.

When sampling to SH truncation degree 21, using 2000 data-points, it takes seven hours. The time, approximately, doubles for every SH-degree. Continuing through to SH-degree 24 it takes $\sim$2.5 days. Thus more resources are spent sampling to SH-degree 24 with 2000 data-points than SH-degree 20 with 20,000 data points.

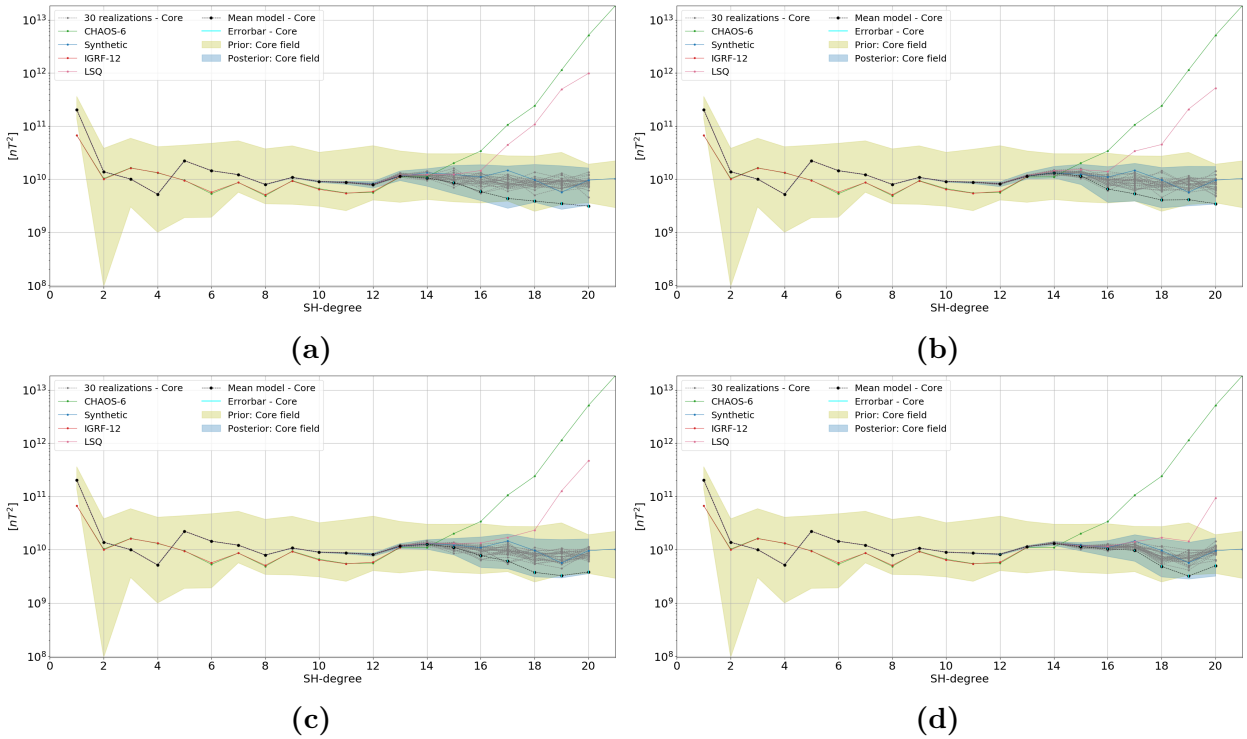



**Figure 5.10:** Power spectra at the CMB of four runs using a multivariate Gaussian prior and varying amounts of data-points. 5.10a-5.10c: Power spectra when using 5000, 10,000 and 20,000 data-points, respectively. It is clear that the spectra is more constrained with an increasing amount of data. 5.10d: This run was made with 100,000 data-points, but initialized with random realizations and the posterior covariance matrix from 5.10a resulting in a significant reduction in computational time.

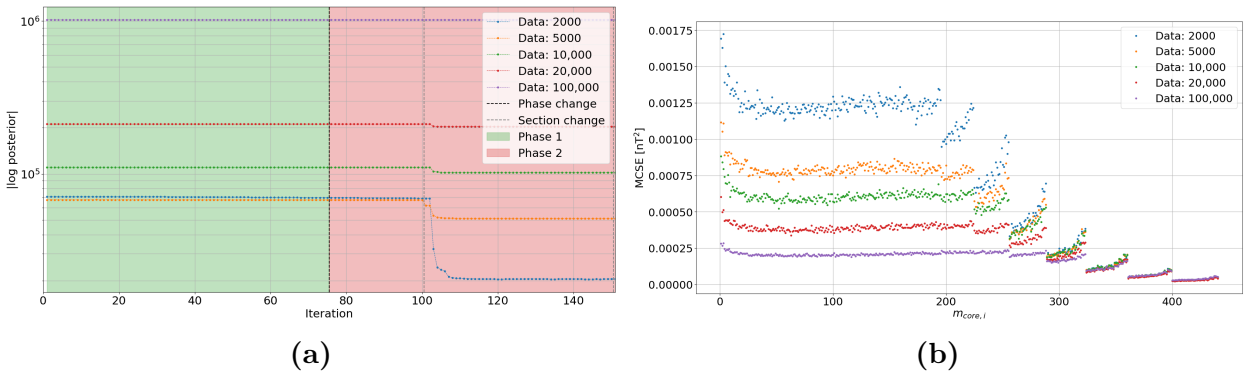



**Figure 5.11:** Comparisons of log posterior and MCSE given different amounts of data-points. 5.11a: Absolute log posterior in the first 150 warm-up iterations. The LSQ solution improves significantly with the amount of data resulting in a smaller change in power when reaching the typical set. 5.11b: MCSE comparison clearly shows a reduction in uncertainty when increasing the amount of data used during sampling.

Figure 5.12 show the results of sampling to SH-degree 24. From the power spectrum, figure 5.12a, it is clear that the posterior is equally as wide as the prior distribution above Sh-degree 15. It is evident that more data is needed if a more constrained posterior is wanted.
From the radial component of the mean model, figure 5.12b, it is clear that the higher SH truncation degree contributes to the small scale structure when compared to figure 5.3b where the truncation degree is 20. Although the spread of the posterior is large the flux patches observed are still well constrained as evident from the RMS of the radial component, figure 5.12c.

In section 5.1.4 it was stated that the maximum tree-depth was set to 10. In light of the large amount of prematurely terminated warm-up iterations, figure 5.12d, it is necessary to increase the maximum tree-depth if there should be any hope of extending the SH truncation degree beyond 24.
With the doubling of computational time when increasing the SH truncation degree it would take approximately 5 days to sample to SH-degree 25. Additionally, increasing the tree-depth to 12 would allow four times the amount of leapfrog steps per iteration which again would increase the computational time. For this specifically reason a successful run has not yet been made to SH-degree 25.

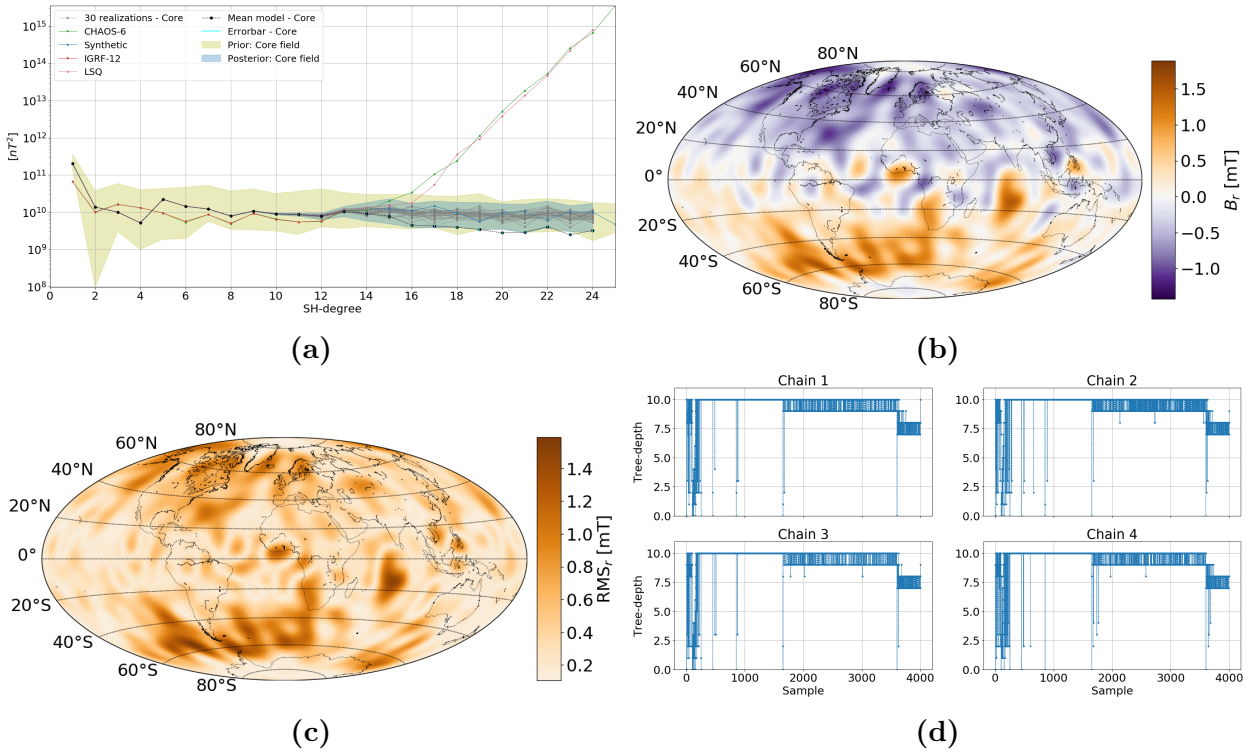



**Figure 5.12:** Results of sampling to SH truncation degree 24 using 2000 synthetic data-points. 5.12a: The power spectrum illustrated at the CMB. 5.12b: The radial component of the mean model at the CMB. 5.12c: RMS of the radial component at the CMB. 5.12d: The tree-depth during the warm-up. A majority of the warm-up iterations have been terminated prematurely due to a too low maximum tree-depth.

## 5.3 Satellite data

In the previous section it was possible to sample to SH truncation degree 24 using synthetic data. Repeating the success using real satellite data, see section 4.10, should not be a problem. Especially considering that 95% of the estimated noise, in the real data, is below 5 nT while the white noise added to the synthetic data has a standard deviation of 10 nT.

By using the same hyperparameters applied in section 5.2 a run was successfully made to SH-degree 20 with 2000 data-points. Its power spectrum is shown in figure 5.13a. The posterior follows both IGRF-12 and CHAOS-6 well, until SH-degree 13. Afterwards the power makes a small dip only to increase until it reaches the upper boundary of the prior. It is expected that the power of the core field continues horizontally and it is therefore unexpected to see the upper boundary of the prior constrain the spectrum after SH-degree 17. Although the spectrum itself is well constrained the amount of data is increased to 20,000. The associated power spectrum, figure 5.13b, is more constrained, but the peculiar behaviour has only worsened. Note how the LSQ solution has a similar dip after the introduction of more data. It must therefore be the data that force this behaviour.

This conclusion is only confirmed when rerunning the attempt with 2000 data-points, but without removing the lithospheric field, figure 5.13c. The spectrum no longer drops in power, but still tries to push beyond the priors upper boundary. It must thus be concluded that using a lithospheric field model to correct for the lithospheric contributions leads to spurious behaviour at SH-degree 14 and above that is not in accord with the prior information concerning the core field. Clearly a better approach of dealing with the lithospheric field, in a more correct way, is needed.

Despite the unsatisfactory results in figure 5.13b the posterior reveals a much larger spatial correlation structure than previously seen, figure 5.14. The posterior correlation, figure 5.14a, is very strong in the lower harmonics where in the synthetic case there had been no structure. Additionally, unlike in the synthetic tests the structure is weak after SH-degree 17 (323 coefficients). It would suggest that the prior is playing a much bigger role in constraining the lower harmonics, where in the case with synthetic data it did not. After SH-degree 17 the prior can not explain the data and the correlation weakens.

The correlation structure in the lower harmonics is much clearer than in any previous case. In section 4.1 five sequences described as lines in the off-diagonal elements were defined. The first line can only be seen in the higher harmonics of the correlation matrix while the rest are very clear. Additionally, two new sequences have emerged. From a close-up of the first six SH-degrees it is clear that coefficients within a SH-degree correlate strongly with each other. This propagates out into the before mentioned lines creating a new structure around them.

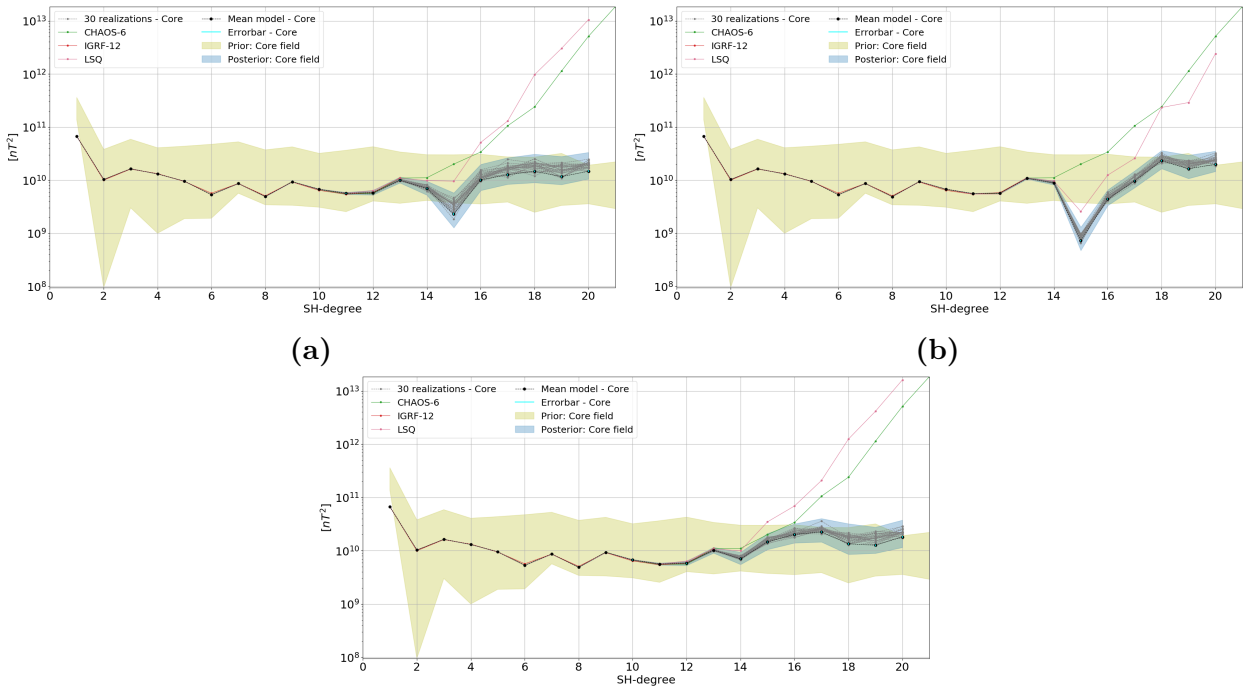



**Figure 5.13:** Power spectra at the CMB using real satellite data, with and without corrections for the lithospheric field. 5.13a: Result of sampling to SH-degree 20 with satellite data of the core field, 2000 data-points. 5.13b: Increasing the amount of data ten time, 20,000 data-points, only emphasize the already strange behaviour. It is expected of the core field to continue horizontally. 5.13c: Using satellite data without the lithospheric contribution removed, 2000 data-points, seems to have removed the initial dip in power.

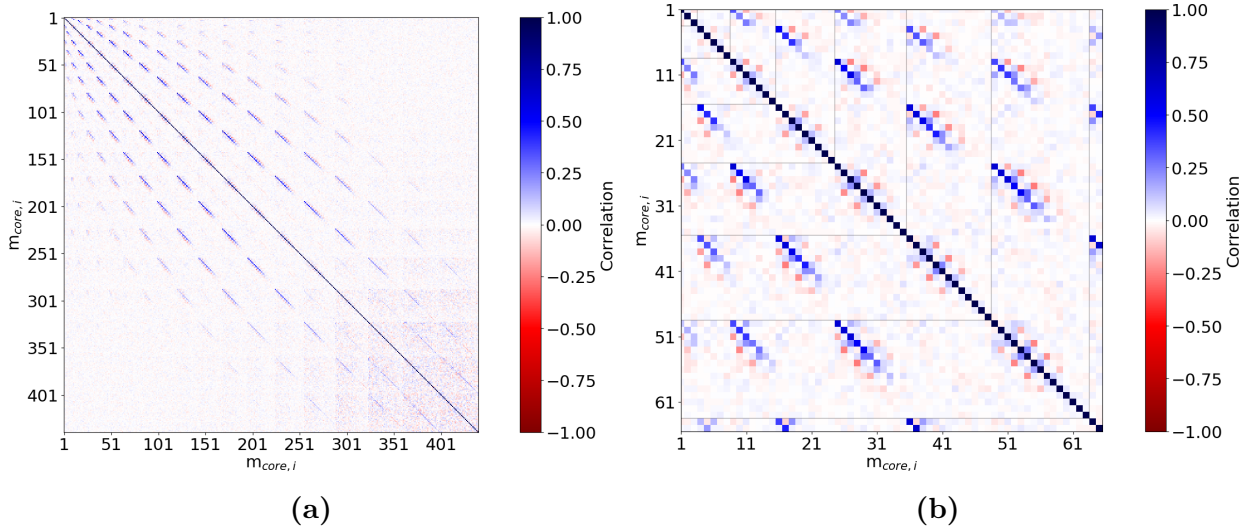



**Figure 5.14:** Posterior correlation matrix of figure 5.13b. 5.14a: The full correlation matrix showing strong off-diagonal structure. 5.14b: A close-up of the first six SH-degrees in 5.14a. New correlation structure not seen before has emerged.

## 5.4 Towards co-estimation of core and lithosphere models

It would appear that in order to move forward the core and lithospheric fields have to be co-estimated. In doing so it is possible to separate the two contributions such that they can be examined independently.
As a proof of concept the method will first be tested on a synthetic data set, a superposition of the core field previously used and a lithospheric field generated from the Masterton model as presented in section 4.2. Afterwards an attempt at co-estimating with real satellite data is presented.

### 5.4.1 Test with synthetic data

The synthetic test presented here was carried out using synthetic data constructed by superposition of internal sources with 2000 data-points. The model is as explained in section 3.2.4. The core field is given as a multivariate Gaussian prior while the lithospheric field is given as an independent Gaussian prior.

With regards to the hyperparameters they are as determined in section 5.1. The SH truncation degree is 20, high enough such that the power spectrum, figure 5.15a, of each model intersects. The core field behaves as has been observed previously, nicely constrained until degree 14. The lithospheric posterior, on the other hand, fills the entire prior, but random realizations are mostly located close to the Masterton model. The intersect of the two spectra occurs between SH-degree 15 and 16. This is made more clear in a close-up, figure 5.15b. The core field posterior has a similar spread as seen when fitting synthetic data without a lithosphere contribution, figure 5.10a. At SH-degree 16 the spectra are of equal power. Note how the MCSE continues to be low even as the mean models cross and none of the random realizations exhibit spurious behaviour. It would appear that co-estimating is not a problem for the sampler, but keep in mind that doing so will double the amount of model parameters. Thus the size of model space when co-estimating to SH-degree 20 is equivalent to SH-degree 28-29 when not co-estimating. Despite the large model space the computational time was only 2.5 days. Because of this observation it must be concluded that the difficulties of extending the SH truncation degree beyond 24, section 5.2.2, has more to do with the prior information and less to do with the dimensionality of the problem.

Figures 5.15c and 5.15d shows the radial component of the core and lithospheric field at the CMB and Earth's surface, respectively. The core field has the classical dipolar pattern and by consulting the RMS, figure 5.15e, all features seem well constrained. The lithospheric field is weak as is expected and its features are quite large due to the low truncation degree, but the flux patches are interesting. Recalling the Masterton model, figure 4.6a, there are strong magnetic anomalies under central Africa, the West Coast of USA, eastern Europe, Australia and east thereof. All these anomalies appear present and from the RMS, figure 5.15f, they are well constrained. This retrieval of lithospheric structure is interesting since the lithospheric prior was diagonal and contained no information on spatial correlation.

The correlation of the entire posterior, figure 5.15g, is a combination of the correlation within the core and lithospheric models and between them, indicated by two green lines. The squares drawn by the green lines will be referred to as quadrants similar to that of a Cartesian plane. The correlation matrix of the core, second quadrant, is close to diagonal. This is in good

agreement with what has already been observed. The data constrains the problem just fine and thus there is no need for the prior until SH-degree 12. For similar reasons, and because its prior is diagonal, the correlation matrix of the lithospheric posterior is diagonal, fourth quadrant. In the first and second quadrant the inter-model correlation structure is found. It is diagonal and has a strong negative correlation. This does make a lot of sense since the two models combine to make the full model of the internal sources, equation 3.9.

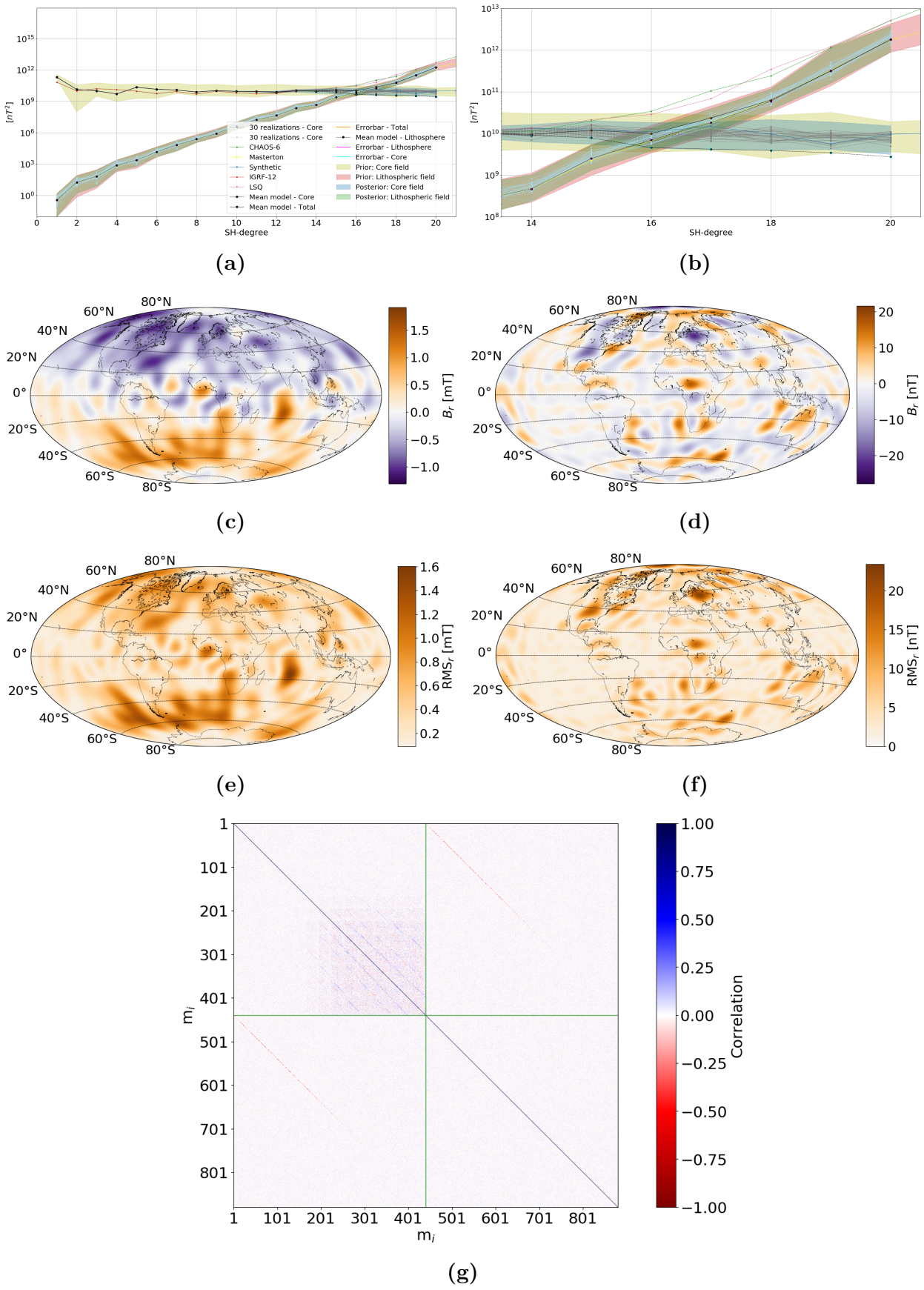



**Figure 5.15:** Results from co-estimation to SH-degree 20 using 2000 synthetic data-points. 5.15a: Power spectrum, at the CMB, looks very good there is no unnatural behaviour when the two models intersect as could be expected. 5.15b: Close-up of the intersection in the power spectrum. 5.15c and 5.15e: Radial components of the sampled core model illustrated on the CMB along with RMS. 5.15d and 5.15f: Radial components of the sampled lithospheric model illustrated at Earth's surface along with RMS. 5.15g: Correlation matrix of the full posterior. Green lines represent the division between the two models.

### 5.4.2 Attempt at co-estimation with real satellite data

In light of what appears to be a successful attempt at co-estimation using synthetic data the natural next step is to use real satellite data.

Using a similar approach as with the synthetic data a successful run was made to SH-degree 22 using 2000 data-points. The power spectrum is very well behaved, figure 5.16a. When the two spectra intersect at SH-degree 15, they remain parallel to each other until SH-degree 16 after which the lithospheric spectrum diverges to accommodate the necessary power increase needed to fit the data. If the combined prior had too much power and did not contain the true model then the lithospheric posterior would be expected to lie in the bottom of the prior. Seeing that it does not reveal that the posterior to some extent are correct.

The spectrum of the core field continues to push against the upper boundary of the prior, after SH-degree 16. This does make sense since the combined prior has no information about the correlation between the core and lithospheric model. There is therefore no reason for it not to maximize the power in the core field. Note also how the lithopsheric and total mean model follows the CHAOS-6 model.

The radial core field, figure 5.16c, looks very nice with reversed flux patches in the southern Atlantic hemisphere corresponding well with the weak field observed in the real observation, figure 4.10a. From the RMS, figure 5.16e, all structures are well defined and thus likely true or at least well constrained.

The mean lithopsheric model show strong anomalies at key locations, figure 5.16d, just as observed in the synthetic case. Additionally, there is a vague zonal pattern. Notice how the RMS map, figure 5.16f, shows the strong anomalies, but also a very weak small scale structure, specifically noticeable over the Pacific ocean. This can only be made if the random realizations, used to create the RMS maps, are not alike. In other words random realizations from the lithospheric posterior all show the strong anomalies, but also a non constrained zonal pattern, appendix figure 8.11. Note how this problem does not exist in the posterior of the core. This could be a matter of too few data-points compared to the size of model space, and if the SH truncation degree is reduced to 16 and the amount of data-points increased to 5000 the problem persists, appendix figure 8.12. This problem did not exist in the synthetic case, the observations were generated from models and although the noise added was large no other sources could contaminate it. Without a more informative, multivariate, lithospheric prior it will be difficult to get rid of this behaviour.

Although the lithospheric prior assumes the model coefficients independent the posterior correlation reveals some very nice structure, figure 5.17. The correlation between the coefficients within the core model is very strong even in the lower harmonics as previously observed in section 5.3. The off-diagonal correlation pattern has propagated into the fourth quadrant, the lithospheric posterior. This first happens in the last few SH-degrees, at the same time the inter-model correlation, quadrants one and three, show almost no correlation between coefficients after SH-degree 17.

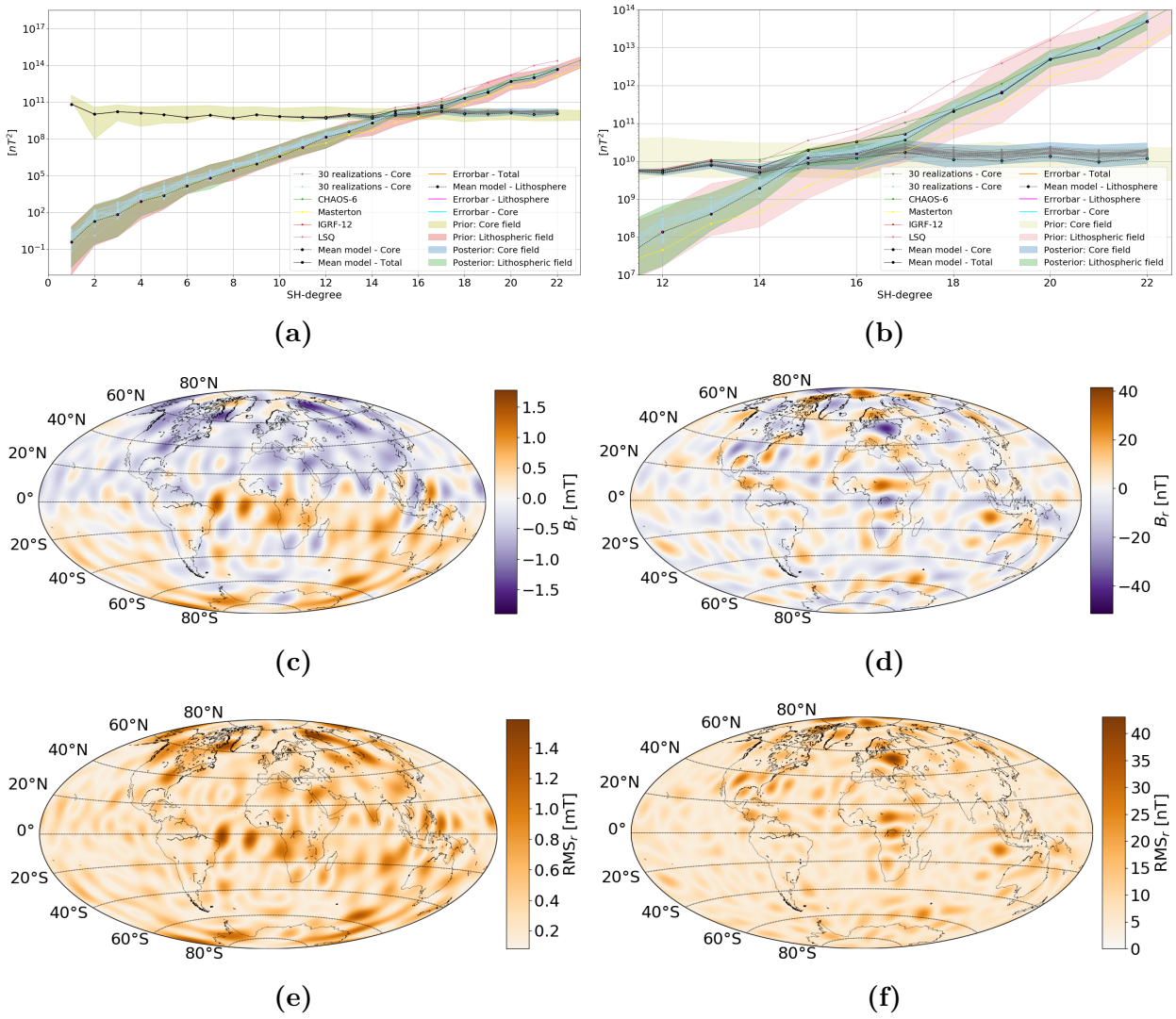



**Figure 5.16:** Results from co-estimation to SH-degree 20 using 2000 data-points from real satellite observations. 5.16a: Power spectrum, at the CMB, looks very good there is no unnatural behaviour when the two models intersect as could be expected. 5.16c and 5.16e: Radial components of the sampled core model illustrated on the CMB along with the RMS. 5.16d and 5.16f: Radial components of the sampled lithospheric model illustrated at Earth's surface along with the RMS.

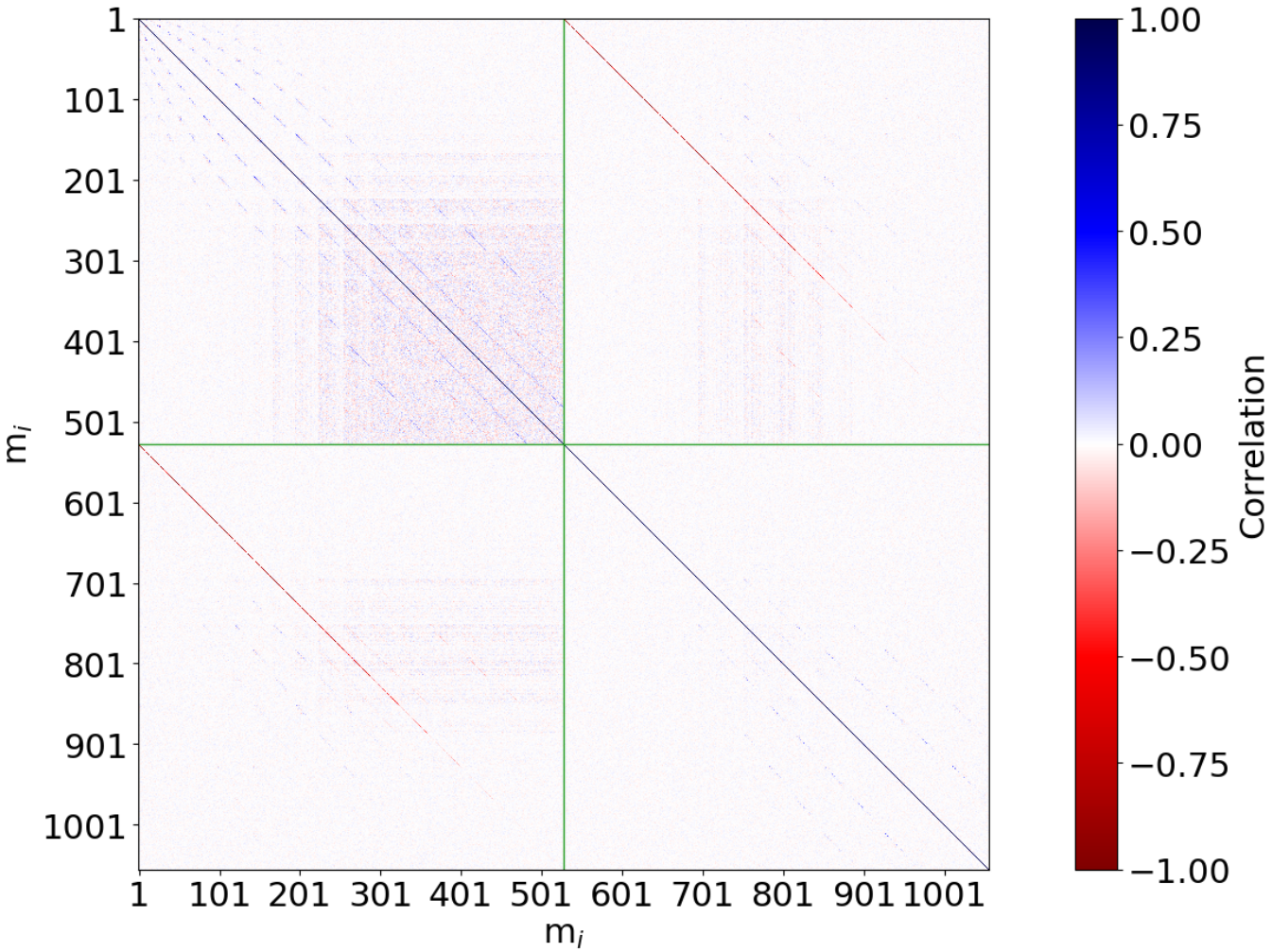



**Figure 5.17:** Correlation matrix of the posterior distribution

# CHAPTER 6

# Discussion

In the previous chapters the results of this thesis were presented along with the data and methods used to create them. This chapter will start by discussing the validity of the results in comparison with other known geomagnetic models. The chapter will continue with comments on limiting factors that have been discovered. Finally, a series of suggestions for improvements will be made in relation to both the sampling technique and the geomagnetic prior information introduced into the system.

## 6.1 Comparison with previous models

The main result of this thesis is the co-estimation of the core and lithospheric fields with real satellite data to SH truncation degree 22. It is possible to evaluate the sampled models and determine if the power contained within them is unrealistic or if magnetic flux patches are ill-constrained. But it is difficult to determine the correctness of the sampled models since the truth is not known. The next best option is to make comparisons to other well known models. Figures 6.1 and 6.2 compares the radial component of the sampled core and lithospheric fields with the CHAOS-6 and LCS-1 models.
Figures 6.1c-6.1d are truncated versions of the mean models presented in figures 6.1a-6.1b. The radial component of the core field, figure 6.1c, is illustrated on the CMB and truncated at SH-degree 13. The radial component of the lithospheric field, figure 6.1d, is illustrated at Earth's surface and truncated between SH-degree 15-20. They are truncated at these levels to make a fair comparison with the previous models. CHAOS-6 and LCS-1 are truncated similarly.

CHAOS-6 as presented in figure 6.1e is assumed to be a good representation of the core field up to SH-degree 13 since the lithospheric contribution is quite low at this truncation degree. By a visual comparison of the posterior mean model and CHAOS-6 they appear to be almost identical. The flux patches have very similar size, shape and intensity. It is only when subtracting the mean model from CHAOS-6 that the difference can be seen, figure 6.2a. Generally the residuals are very low, but they are relatively stronger over landmasses where there are known lithospheric anomalies. Calculating the RMS of the residuals from 300 random realizations, figure 6.2c, reveals that the patterns in figure 6.2a are well constrained. With the specific locations of the more intense residuals it is hard not to assume that they are caused by CHAOS-6 containing both core and lithospheric contributions. Calculating the RMS of the residuals between CHAOS-6 and 300 realizations of the combined core and lithospheric posterior, figure 6.2e, while keeping the SH truncation degree of 13, shows flux patches of equal intensity and the maximum RMS has decreased by approximately 50 %. This is a good indication that the probabilistic approach indeed was able to separate the two sources.

The truncated lithospheric mean model shown in figure 6.1d should be compared with the LCS-1 model in figure 6.1f. Immediately it is noticed that the mean posterior is weaker than LCS-1 and that the flux patches do not match as neatly as in the case of the core field. This is also apparent from figure 6.2b showing the difference between the two. In general the residuals

are in the region of five nT except for the field over Australia where there is a clear disagreement in that the flux patches are of opposite polarity, see figures 6.1d and 6.1f. This result is not exclusive to the mean model as is clear from the RMS of the residuals between 300 realizations of the lithospheric posterior and the LCS-1 model, figure 6.2d. This is possibly correlated with the core field power spectrum, figure 5.16b, being pushed against the upper boundary of the core field prior. The independent lithospheric prior is likely so weak that some of the lithospheric field is being deposited in the core field and as the power of the lithospheric field grows so does the misplaced contribution in the core field.

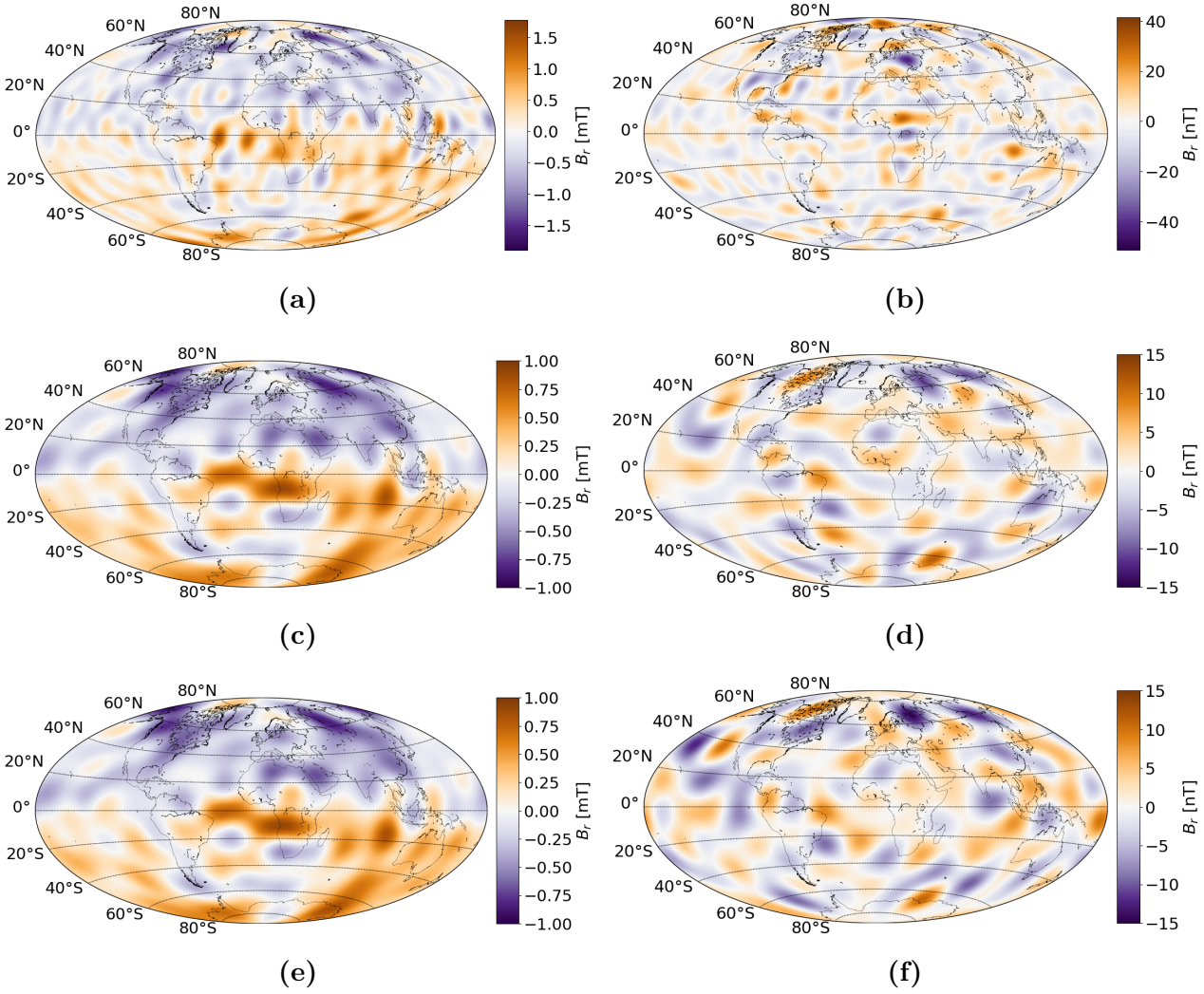



**Figure 6.1:** Comparison of the radial component when co-estimating with real satellite data, section 5.4.2, and known models. Left: The sampled core field and CHAOS-6 illustrated at the CMB. Right: The sampled lithospheric field and LCS-1 at the Earth's surface. 6.1a-6.1b: The mean model of both the core and lithospheric fields, respectively. These are identical to figures 5.16c and 5.16d and are repeated here for easier comparison. 6.1c-6.1d: The mean model of both the core and lithospheric fields, respectively. The core field is truncated at SH-degree 13 while the lithospheric field is truncated between SH-degree 15-20. 6.1e-6.1f: CHAOS-6 and LCS-1 truncated at SH-degree 13 and between SH-degree 15 and 20, respectively.

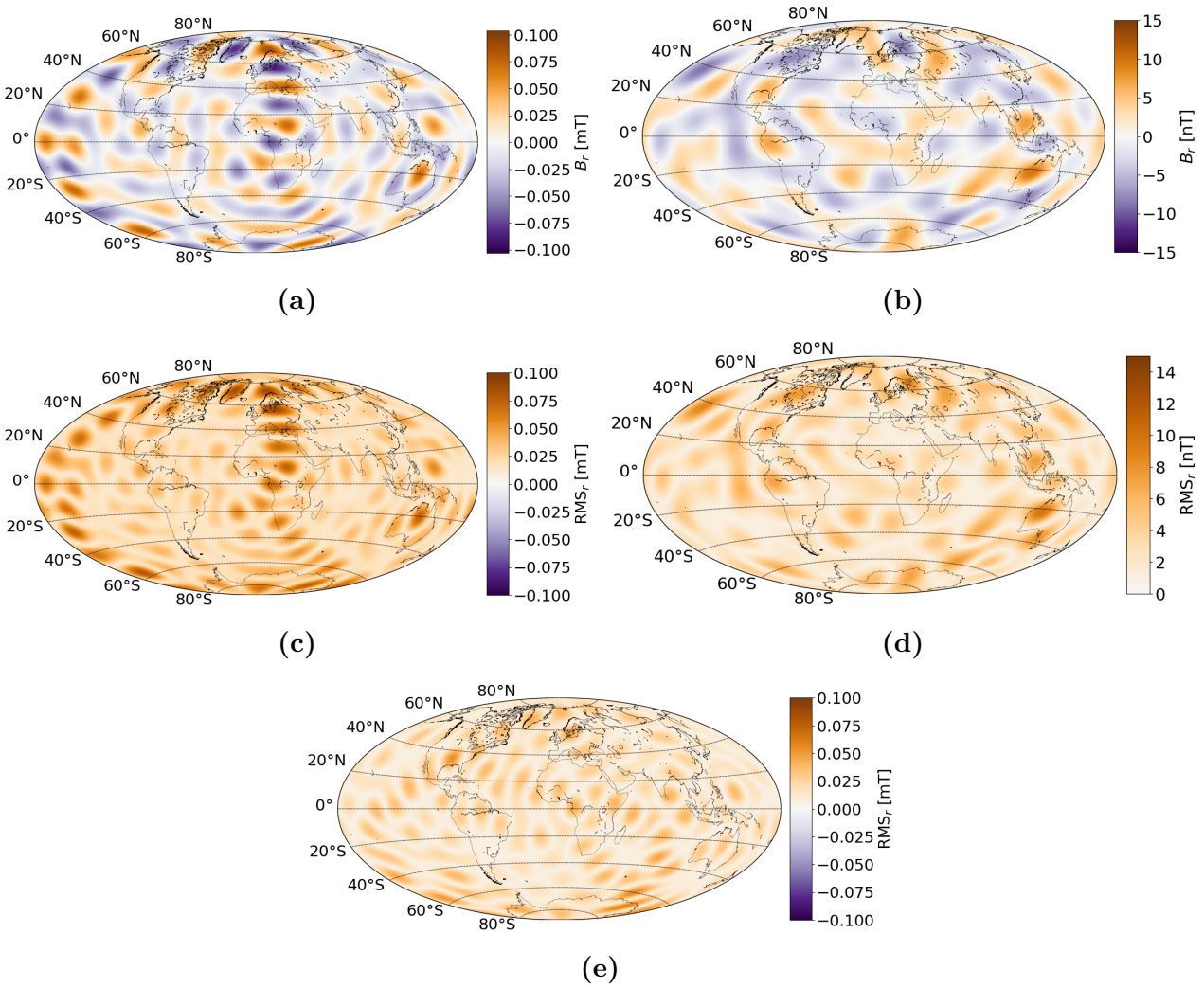



**Figure 6.2:** Continuation of the comparison in figure 6.1. 6.2a-6.2b: Difference between 6.1c-6.1d and 6.1e-6.1f. 6.2c-6.2d: RMS of the difference between 6.1e-6.1f and 300 random realizations of the core and lithospheric posteriors. 6.2e: RMS of the difference between CHAOS-6 and the combined core and lithospheric posterior truncated at SH-degree 13 and shown on the CMB.

## 6.2 Limitations

In this thesis it was shown to be possible to co-estimate the core and lithospheric magnetic fields. In the process of doing so two limiting factors were noticed; these concern the core and lithospheric prior information.

The core field prior was determined from a core dynamo simulation as presented in section 4.1. Here it was discovered that some of the coefficients in the 687 realizations had a very high autocorrelation. Thus the covariance matrix defining the multivariate prior was not based on completely independent samples. Additionally, the amount of realizations determines the maximum SH truncation degree. In order to provide the logarithm of the prior probability it is necessary for the covariance matrix to be positive definite so as to ensure a unique solution to the LDLT decomposition. For that reason the amount of realizations needs to be equal to or larger than the amount of model parameters described by the prior. The prior information about the core field was taken from a core dynamo simulation (Aubert et al. 2017). Alternatively a longer time series from Aubert et al. 2013 could be used since it is morphologically the

same core field. This time series is, to the authors knowledge, also truncated at SH-degree 30, which then would be the next natural limitation. If it is proven possible to sample to degrees higher than 30 additional prior information such as suggested in Jackson and Sambridge 2005 could be valuable.

The second limiting factor is the lithospheric prior. The one applied in this thesis is very weak because it assumes the coefficients to be independent. In section 5.1.2 it was made clear that assuming the SH coefficients of the core field independent instead of multivariate impaired performance significantly. It was shown in figure 5.17 that some of the off-diagonal correlation structure of the core field was repeated by the lithospheric field, which is perhaps concerning. But it is also the belief that the lithospheric field has some off-diagonal correlation structure of its own. And as discussed briefly in section 4.2 in the future it might be possible to generate information about this structure by varying the crustal thickness and magnetic susceptibility parameters used to create the model presented in Masterton et al. 2013.

## 6.3 Future work

In this section suggestions for future work are made. These include both improvements to the current method, and also how the probabilistic approach could be extended to include time dependence.

### 6.3.1 Improved initialization

In section 5.1 several hyperparameters were systematically tested in the attempt to tune them for the specific problem presented in this thesis. It is the belief that by fine tuning the hyperparameters the performance during warm-up will be improved. Especially the initialization of the starting point and mass matrix influence the performance of the sampler. It is therefore recommended that the LSQ solution used as starting point, see section 3.1.3, be replaced with a more robust solution such as a regularized LSQ approach that can provide a more realistic fit to the higher harmonics when the amount of data is restricted.
Additionally, the mass matrix, which was initialized as the prior covariance matrix, was shown to be a bad representation of the posterior covariance. This was clearly seen in section 5.1.5 where the use of the prior covariance as mass matrix lead the sampler to not converge in the first warm-up phase, which is the intention.

It is possible that the information gathered on the size of the posterior covariance and its correlation structure, from the prior information in section 4.1 and the posterior correlation matrix in section 5.3, can be used to generate initial mass matrices to the SH truncation degree chosen.

It should also be emphasized that the maximum tree-depth was found to be dependent on how well estimated the starting point and initial mass matrix were along with the dimensionality of the sampled space. This was particularly clear when sampling the core field with synthetic data to SH-degree 24, figure 5.12d. It is therefore suggested that a few hundred test warm-up iterations are made and analyzed before committing to the entire run.

### 6.3.2 Investigate correlation structure in the co-estimated posterior

Part of the information gathered on the posterior correlation structures regards the co-estimated model. Specifically, the inter-model structure and how the core field structure appears to propagate to the lithospheric correlation matrix, see figure 5.17. It is unknown how the inter-model structure should look and how well behaved co-estimated models should interact. It is possible that the duplicate structures in the core and lithospheric field, quadrant two and four of figure 5.17, could be a sign of incorrect source separation. If this is investigated further and understood it could be a major help in the initialization of future co-estimated models, an excellent diagnostic to evaluate the correctness of the output and generally further the understanding of this techniques ability to separate sources of the geomagnetic field.

### 6.3.3 Riemannian-Gaussian kinetic energies

In the HMC implementation used for this thesis the kinetic energy is assumed to be of the EGKE family. This means that the phase space is decorrelated globally. For this reason local areas with very steep gradients can be problematic and cause the sampler to diverge or simply increase the needed integration time. Riemannian-Gaussian kinetic energies can decorrelate phase space locally decreasing the integration time by avoiding large gradients. This type of kinetic energy is actually implemented in Stan, but not fully tested, and will be part of future releases (Stan Development Team 2019a). This is especially interesting if custom distributions are implemented, which may be necessary for the lithospheric field or with the introduction of time dependence.

### 6.3.4 Time dependent field

For reasons made clear in sections 4.4 the real satellite data was restricted to a three months period in order to avoid significant secular variations such that the assumption of a stationary geomagnetic field would hold. A natural extension to the current approach would be to introduce a time dependent core field. In doing so prior information about the core field evolution would be needed to constrain and keep the evolution physically feasible. A probabilistic approach to model a time dependent core field was presented in Gillet et al. 2013 and it is possible that similar prior information could be applied in the present context. This would significantly increase the amount of model parameters and data. Preliminary tests to see how well Stan handles such large models should be made. But from the fact that sampling the core field, with synthetic data, is difficult beyond SH-degree 24 (624 coefficients). While co-estimating to SH-degree 22 (1056 coefficients) is fairly easy, suggests that the high dimensionality is not the problem, but rather the amount of constraint that can be applied to keep integration time down.

# CHAPTER 7

# Conclusion

In this thesis the geomagnetic field has been modelled using a Bayesian approach called Hamiltonian Monte Carlo that utilizes a mixture between MCMC and gradient information. The motivation for solving the inverse problem presented in this thesis using a probabilistic approach comes form the possibility of obtaining uncertainties of model parameters in the form of a posterior distribution. Typically, this problem would be solved with a regularized least squares approach even though probabilistic methods for solving inverse problems have been around for some time. The classical MCMC approaches do not work efficiently in high dimensional spaces. But because Hamiltonian Monte Carlo uses gradient information, a high dimensionality is not a problem. The downside is the need for suitable prior information on the field sources as well as user-defined parameters that require experience to determine. With the most challenging of these parameters determined by the No-U-Turn algorithm from Hoffman and Gelman 2014 the probabilistic approach is closer to becoming a real alternative to the classical techniques.

From the offset, the goal was to model the core field. The prior probability of the core was assumed multivariate Gaussian and the prior information came from 687 realizations of a core dynamo simulation truncated at SH-degree 30 that stem from Aubert et al. 2017.
Benchmark tests using synthetic data generated from the midpath dynamo, truncated at SH-degree 60, of Aubert et al. 2017 yielded promising results. When introducing real satellite data similar to that of Hammer and Finlay 2019 spurious behaviour was observed at SH-degree 14 and above. It is believed to be caused by the attempted removal of the lithospheric contribution using the LCS-1 model.
Without the ability to provide real data of the core magnetic field alone, co-estimation of both the core and lithospheric fields was the natural next step. The lithospheric coefficients were assumed independent and prior information taken from Masterton et al. 2013. The lithospheric field superimposed on the synthetic data was generated from the same source.
Again benchmark tests were promising. It was possible to separate the core and lithospheric fields leading to the recreation of lithospheric anomalies, despite the prior having no information on the correlation structure thereof.
Attempts at co-estimating using real satellite data was successful at sampling to a SH truncation degree of 22. The power spectra of the two sources intersect between SH-degree 14-15 and show no sign of unnatural behaviour that might be expected when the sources are of equal strength. The generated radial core field is nearly identical to the CHAOS-6 model when truncating them both at SH-degree 13. They both exhibit strong flux patches stretched in the east/west direction over the equator in the Atlantic hemisphere. Increasing the truncation degree of the sampled core field to 22 gives the flux patches a more sectoral pattern instead.
The sampled lithospheric field manages to recreate major lithospheric anomalies and when comparing with random realizations of the posterior these patterns are well constrained. However there was generally a zonal pattern over areas of low field strength, such as the Pacific ocean. Additionally, when comparing with LCS-1, truncating both between SH-degree 15-20, the two flux patches that cover Australia have opposite polarity and LCS-1 is generally a few nT stronger.
The power spectrum of the core field is found to push against the upper boundary of the core prior distribution. It is possible that the core prior is too weak at higher SH-degrees causing

some the core field to be deposited in the lithospheric field. Likewise this could be caused by using an independent lithospheric prior that is insufficient. Given a stronger lithospheric prior this could be further tested.
In addition to co-estimating the core and lithospheric fields a prior distribution based on a two component Gaussian mixture model was successfully implemented. This proves the possibility of using custom prior distributions which can be useful when implementing a multivariate lithospheric prior that is not necessarily Gaussian.

The approach presented in this thesis can be extended and future work could include time dependence of the core field. Although it was shown to be difficult to sample the core field with a SH truncation degree above 24 (624 coefficients) the HMC sampler should have no trouble handling the increased dimensionality that comes with time dependency, since co-estimating to SH-degree 22 with 1056 model parameters presented no serious difficulties.

# Bibliography


Aster, Richard C., Brian Borchers, and Clifford H. Thurber (2013). *Parameter Estimation and Inverse Problems*. Elsevier Inc. ISBN: 9780123850485, 9780123850492, 0123850495, 0123850487. DOI: 10.1016/C2009-0-61134-X.

Aubert, Julien, Chris Finlay, and Alexandre Fournier (2013). "Bottom-up control of geomagnetic secular variation by the earth's inner core". In: *Living on a Magnetic Planet - Abstract Volume*.

Aubert, Julien, Thomas Gastine, and Alexandre Fournier (2017). "Spherical convective dynamos in the rapidly rotating asymptotic regime". In: *Journal of Fluid Mechanics* 813, pp. 558–593. ISSN: 14697645, 00221120. DOI: 10.1017/jfm.2016.789.

Baratchart, L. and C. Gerhards (2017). "On the recovery of core and crustal components of geomagnetic potential fields". In: *Siam Journal on Applied Mathematics* 77.5, pp. 1756–1780. ISSN: 1095712x, 00361399. DOI: 10.1137/17M1121640.

Betancourt, Michael (2017). *Robust Statistical Workflow with RStan*. URL: https://mc-stan.org/users/documentation/case-studies/rstan_workflow.html (visited on 08/03/2019).

— (2018). "A Conceptual Introduction to Hamiltonian Monte Carlo". In:

Bloxham, J, D Gubbins, and A Jackson (1989). "Geomagnetic Secular Variation". In: *Philosophical Transactions of the Royal Society A-mathematical Physical and Engineering Sciences* 329.1606, pp. 415–502. ISSN: 14712962, 1364503x, 00804614. DOI: 10.1098/rsta.1989.0087.

Brooks, Steve, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall.

Cain, Joseph C., Zhigang Wang, Dave R. Schmitz, and J. Meyer (1989). "The geomagnetic spectrum for 1980 and core-crustal separation". In: *Geophysical Journal-oxford* 97.3, pp. 443–447. ISSN: 1365246x, 0956540x, 09524592. DOI: 10.1111/j.1365-246X.1989.tb00514.x.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus A. Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell (2017). "Stan: A probabilistic programming language". In: *Journal of Statistical Software* 76.1, pp. 1–29. ISSN: 15487660. DOI: 10.18637/jss.v076.i01.

Carpenter, Bob, Matthew D. Hoffman, Marcus Brubaker, Daniel Lee, Peter Li, and Michael Betancourt (2015). "The Stan Math Library: Reverse-Mode Automatic Differentiation in C++". In:

Fichtner, Andreas, Andrea Zunino, and Lars Gebraad (2019). "Hamiltonian Monte Carlo solution of tomographic inverse problems". In: *Geophysical Journal International* 216.2, pp. 1344–1363. ISSN: 1365246x, 0956540x. DOI: 10.1093/gji/ggy496.

Finlay, Chris, Nils Olsen, Stavros Kotsiaros, Nicolas Gillet, and Lars Tøffner-Clausen (2016). "Recent geomagnetic secular variation from Swarm and ground observatories as estimated in the CHAOS-6 geomagnetic field model". In: *Earth, Planets and Space* 68.1. ISSN: 1880-5981. DOI: 10.1186/s40623-016-0486-1. URL: http://dx.doi.org/10.1186/s40623-016-0486-1.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin (2013). *Bayesian data analysis, third edition*. CRC Press. ISBN: 9781439840955, 9781439898208.

Gillet, N., D. Jault, Chris Finlay, and Nils Olsen (2013). "Stochastic modeling of the Earth's magnetic field: Inversion for covariances over the observatory era". In: *Geochemistry, Geophysics, Geosystems* 14.4, pp. 766–786. ISSN: 15252027. DOI: 10.1002/ggge.20041.

Hammer, Magnus Danel and Christopher C. Finlay (2019). "Local Averages of the Core-mantle Boundary Magnetic Field from Satellite Observations". In: *Geophysical Journal International* 216.3, pp. 1901–1918. ISSN: 1365246x, 0956540x. DOI: 10.1093/gji/ggy515.

Hoffman, Matthew D and Andrew Gelman (2014). "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." In: *Journal of Machine Learning Research* 15.1, pp. 1593–1623.

Hollerbach, Rainer and David Gubbins (2007). "Inner Core Tangent Cylinder". In: *Encyclopedia of Geomagnetism and Paleomagnetism*. Ed. by David Gubbins and Emilio Herrero-Bervera. Dordrecht: Springer Netherlands, pp. 430–433. ISBN: 978-1-4020-4423-6. DOI: 10.1007/978-1-4020-4423-6_154. URL: https://doi.org/10.1007/978-1-4020-4423-6_154.

Holschneider, Matthias, Vincent Lesur, Stefan Mauerberger, and Julien Baerenzung (2016). "Correlation-based modeling and separation of geomagnetic field components". In: *Journal of Geophysical Research: Solid Earth* 121.5, pp. 3142–3160. ISSN: 21699356, 21699313. DOI: 10.1002/2015JB012629.

Jackson, A and M Sambridge (2005). "Softening a hard quadratic bound to a prior pdf - an example from geomagnetism". In: *Aip Conference Proceedings* 803, pp. 499–506. ISSN: 15517616, 0094243x.

Jones, Eric, Travis Oliphant, and Pearu Peterson (2001). *SciPy: Open Source Scientific Tools for Python*.

Langel, R. A. and R. H. Estes (1982). "A Geomagnetic Field Spectrum". In: *Geophysical Research Letters* 9.4, pp. 250–253. ISSN: 19448007, 00948276. DOI: 10.1029/GL009i004p00250.

Lowes, F. J. (1974). "Spatial power spectrum of the main geomagnetic field, and extrapolation to the core". In: *Geophysical Journal of the Royal Astronomical Society* 36.3, pp. 717–30, 717–730. ISSN: 00168009.

Lunn, David, David Spiegelhalter, Andrew Thomas, and Nicky Best (2009). "The BUGS project: Evolution, critique and future directions". In: *Statistics in Medicine* 28.25, pp. 3049–3067. ISSN: 10970258, 02776715. DOI: 10.1002/sim.3680.

Masterton, S. M., D. Gubbins, R. D. Muller, and K. H. Singh (2013). "Forward modelling of oceanic lithospheric magnetization". In: *Geophysical Journal International* 192.3, pp. 951–962. ISSN: 1365246x, 0956540x. DOI: 10.1093/gji/ggs063.

Monnahan, Cole C., James T. Thorson, and Trevor A. Branch (2017). "Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo". In: *Methods in Ecology and Evolution* 8.3, pp. 339–348. ISSN: 2041210x. DOI: 10.1111/2041-210X.12681.

Mosegaard, K and C Rygaard-Hjalsted (1999). "Probabilistic analysis of implicit inverse problems". In: *Inverse Problems* 15.2, pp. 573–583. ISSN: 13616420, 02665611. DOI: 10.1088/0266-5611/15/2/015.

Nilsson, Andreas, Neil Suttie, and Mimi J. Hill (2018). "Short-term magnetic field variations from the post-depositional remanence of lake sediments". In: *Frontiers in Earth Science* 6, p. 39. ISSN: 22966463. DOI: 10.3389/feart.2018.00039.

Olsen, Nils, G. Hulot, and T. J. Sabaka (2010). "Measuring the Earth's Magnetic Field from Space". In: *Space Science Reviews* 155.1-4, pp. 65–93. ISSN: 15729672, 00386308. DOI: 10.1007/s11214-010-9676-5.

Olsen, Nils, Gauthier Hulot, Vincent Lesur, Chris Finlay, Ciaran Beggan, Amaud Chulliat, Terence J. Sabaka, Rune Floberghagen, Eigil Friis-Christensen, Stavros Kotsiaros, and Lars Tøffner-Clausen (2015). "The Swarm Initial Field Model for the 2014 geomagnetic field". In:

*Geophysical Research Letters* 42.4, pp. 1092–1098. ISSN: 19448007, 00948276. DOI: 10.1002/2014GL062659.

Olsen, Nils, H. Luhr, T.J. Sabaka, M. Mandea, M. Rother, L. Tofner-Clausen, and S. Choi (2006). "CHAOS-a model of the Earth's magnetic field derived from CHAMP, Orsted, and SAC-C magnetic satellite data". In: *Geophysical Journal International* 166.1, pp. 67–75. ISSN: 0956-540X.

Olsen, Nils, Dhananjay Ravat, Chris Finlay, and Livia Kathleen Kother (2017). "LCS-1: A high-resolution global model of the lithospheric magnetic field derived from CHAMP and Swarm satellite observations". In: *Geophysical Journal International* 211.3, pp. 1461–1477. ISSN: 1365246x, 0956540x. DOI: 10.1093/gji/ggx381.

Olsen, Nils and Claudia Stolle (2012). "Satellite Geomagnetism". In: *Annual Review of Earth and Planetary Sciences* 40, pp. 441–465. ISSN: 0084-6597. DOI: 10.1146/annurev-earth-042711-105540.

Plummer, Martyn (2009). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*.

Al-Rfou, Rami, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, Alexander Belopolsky, et al. (2016). "Theano: A Python framework for fast computation of mathematical expressions". In: *arXiv preprint arXiv:1605.02688*.

Sabaka, Terence J., Lars Tøffner-Clausen, Nils Olsen, and Christopher C. Finlay (2018). "A comprehensive model of Earth's magnetic field determined from 4 years of Swarm satellite observations". In: *Earth, Planets and Space* 70.1, p. 130. ISSN: 18805981, 13438832. DOI: 10.1186/s40623-018-0896-3.

Saff, E. B. and A. B.J. Kuijlaars (1997). "Distributing many points on a sphere". In: *Mathematical Intelligencer* 19.1, pp. 5–11. ISSN: 18667414, 03436993. DOI: 10.1007/BF03024331.

Salvatier, John, Thomas V. Wiecki, and Christopher Fonnesbeck (2016). "Probabilistic programming in Python using PyMC3". In: *Peerj Computer Science* 2016.4, e55. ISSN: 23765992. DOI: 10.7717/peerj-cs.55.

Stan Development Team (2019a). *CmdStan Interface: User's Guide 2.19*.

— (2019b). "Stan Reference Manual 2.19". In:

— (2019c). *Stan User's Guide 2.19*.

Statisticat, LLC (2016). "LaplacesDemon: A Complete Environment for Bayesian Inference within R". In: *R Package version* 17.

Tran, Dustin, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei (2017). *Edward: A library for probabilistic modeling, inference, and criticism*.

Voorhies, C. V., T. J. Sabaka, and M. Purucker (2002). "On magnetic spectra of Earth and Mars". In: *Journal of Geophysical Research: Planets* 107.E6.

# CHAPTER 8

# Appendix

## 8.1 Equidistant grid for synthetic data

```
#%% Import
import numpy as np
import matplotlib.pyplot as plt
from mpl_toolkits.basemap import Basemap

#%% Figure properties
font = {'family' : 'sans-serif',
        'weight' : 'normal',
        'size'   : 22}
plt.rc('font', **font)
plt.ioff()

#%% Create grid
radius = 1
N=10000

grid = np.zeros((N,3))
s = 3.6/np.sqrt(N)
dz = 2.0/N
long = 0
z = 1-dz/2
for i in range(0,len(grid)):
    r = np.sqrt(1-z*z)
    grid[i, :] = [np.cos(long)*r, np.sin(long)*r, z]
    z = z - dz
    long = long + s/r

lat = np.arccos(grid[:,2]/ np.sqrt(grid[:, 0]**2 + grid[:, 1]**2 + grid[:, 2]**2))
lon = np.arctan2(grid[:,1],grid[:,0])
grid[:,0] = np.ones((N))*radius
grid[:,1] = -1*(lat*180/np.pi)+90
grid[:,2] = lon*180/np.pi

#%% Plot grid
fig = plt.figure(figsize=(12, 10))
ax = plt.gca()
map = Basemap(projection='hammer', resolution='l', area_thresh=1000.0,
              lat_0=0, lon_0=0)
x,y = map(grid[:,2], grid[:,1])
map.plot(x,y, marker='.', color='tab:red', markersize = 1, linewidth=0)
```

```
map.drawcoastlines(linewidth=0.25)
map.drawparallels(np.arange(-80, 81, 20), labels=[1, 0, 0, 0])
fig.savefig('/home/micmad/cmdstan/grid/grid.png', bbox_inches='tight')
plt.show()
```

# 8.2 Independent Gaussian prior

```
data {
  // Observations and forward problem
  int N;                                    // Length of data vectors
  int M;                                    // Amount of model coefficients
  vector[3*N] B;                            // Observations
  matrix[3*N, M] G;                         // Data kernel

  // Likelihood
  vector[3*N] varLike;                      // Error estimate on observations

  // Prior
  vector[M] muCore;                         // Mean of prior
  vector[M] varCore;                        // Variance of prior
}

transformed data {

  // Likelihood constants
  real kL1 = -(3*N)/2*log(2*pi());          // Log of initial Gaussian term
  real kL2 = -0.5*sum(log(varLike));        // Log-determinant

  // Prior constants
  real kP1 = -(M)/2*log(2*pi());
  real kP2 = -0.5*sum(log(varCore));
}

parameters {
  vector[M] mCore;                          // Initialize model coefficients
}

model {
  vector[3*N] res = G*mCore-y;                    // Calculate residuals

  // Prior
  target += kP1;
  target += kP2;
  target += -0.5*((mCore-muCore)'*(mCore-muCore)./varCore);

  // Likelihood
  target += kL1;
  target += kL2;
  target += -0.5*(res'*res./varLike);
}
```

# 8.3 Multivariate Gaussian prior

```
data {
  // Observations and forward problem
  int N;                                    // Length of data vectors
  int M;                                    // Amount of model coefficients
  vector[3*N] B;                            // Observations
  matrix[3*N, M] G;                         // Data kernel

  // Likelihood
  vector[3*N] varLike;                      // Error estimate on observations

  // Prior
  matrix[M, M] inv_cov;                     // Inverse co-variance matrix
  real log_det_cov;                         // Log of co-variance determinant
  vector[M] muCore;                         // Mean of prior
}

transformed data {

  // Likelihood constants
  real kL1 = -(3*N)/2*log(2*pi());          // Log of initial Gaussian term
  real kL2 = -0.5*sum(log(varLike));        // Log-determinant

  // Prior constants
  real kP1 = -(M)/2*log(2*pi());
  real kP2 = -0.5*log_det_cov;
}

parameters {
  vector[M] mCore;                          // Initialize model coefficients
}

model {
  vector[3*N] res = G*mCore-y;                    // Calculate residuals

  // Prior
  target += kP1;
  target += kP2;
  target += -0.5*((mCore-muCore)'*inv_cov*(mCore-muCore));

  // Likelihood
  target += kL1;
  target += kL2;
  target += -0.5*(res'*res./varLike);
}
```

# 8.4 Appendix - Co-estimation prior

```
data {
  // Observations and forward problem
  int N;                                    // Length of data vectors
  int M;                                    // Amount of model coefficients
  vector[3*N] B;                            // Observations
  matrix[3*N, M] G;                         // Data kernel

  // Likelihood
  vector[3*N] varLike;                      // Error estimate on observations

  // Prior - Core
  matrix[M, M] inv_cov;                     // Inverse co-variance matrix
  real log_det_cov;                         // log of co-variance determinant
  vector[M] muCore;                         // Mean of prior

  // Prior - Lithosphere
  vector[M] muLitho;                        // Mean of prior
  vector[M] varLitho;                       // Variance of prior
}

transformed data {

  // Likelihood constants
  real kL1 = -(3*N)/2*log(2*pi());          // Log of initial Gaussian term
  real kL2 = -0.5*sum(log(varLike));        // Log-determinant

  // Prior constants - Core
  real kP1 = -(M)/2*log(2*pi());
  real kP2 = -0.5*log_det_cov;

  // Prior constants - Lithosphere
  real kP3 = -0.5*sum(log(varLitho));
}

parameters {
  vector[M] mCore;                          // Initialize model coefficients
  vector[M] mLitho;                         // Initialize model coefficients
}

model {
  vector[3*N] res = G*(mCore+mLitho)-y;                  // Calculate residuals

  // Prior - Lithosphere
  target += kP1;
  target += kP3;
  target += -0.5*((mLitho-muLitho)'*(mLitho-muLitho)./varLitho);

  // Prior - Core
  target += kP1;
  target += kP2;
  target += -0.5*((mCore-muCore)'*inv_cov*(mCore-muCore));


  // Likelihood
```

```
  target += kL1;
  target += kL2;
  target += -0.5*(res'*res./varLike);
}
```

# 8.5 Appendix - GMM prior

```
data {
  // Observations and forward problem
  int N;                                    // Length of data vectors
  int M;                                    // Amount of model coefficients
  vector[3*N] B;                            // Observations
  matrix[3*N, M] G;                         // Data kernel

  // Likelihood
  vector[3*N] varLike;                      // Error estimate on observations

  // Prior
  vector[M] muCore;                         // Mean of prior
  vector[M] sigmaCore;                      // Standard deviation of prior
  vector[M] weightCore;                     // weight of prior
}

transformed data {

  // Likelihood constants
  real kL1 = -(3*N)/2*log(2*pi());          // Log of initial Gaussian term
  real kL2 = -0.5*sum(log(varLike));        // Log-determinant
}

parameters {
  vector[M] mCore;                          // Initialize model coefficients
}

model {
  vector[3*N] res = G*mCore-y;                    // Calculate residuals

  // Prior
  for (i in 1:mMax) {
    target += log_sum_exp(log(weightCore[i,1]) + normal_lpdf(mCore[i] | muCore[i,1],
sigma[i,1]), log(weightCore[i,2]) + normal_lpdf(mCore[i] | muCore[i,2], sigma[i,2]));
  }

  // Likelihood
  target += kL1;
  target += kL2;
  target += -0.5*(res'*res./varLike);
}
```

# 8.6 Complimentary figures for chapter 5

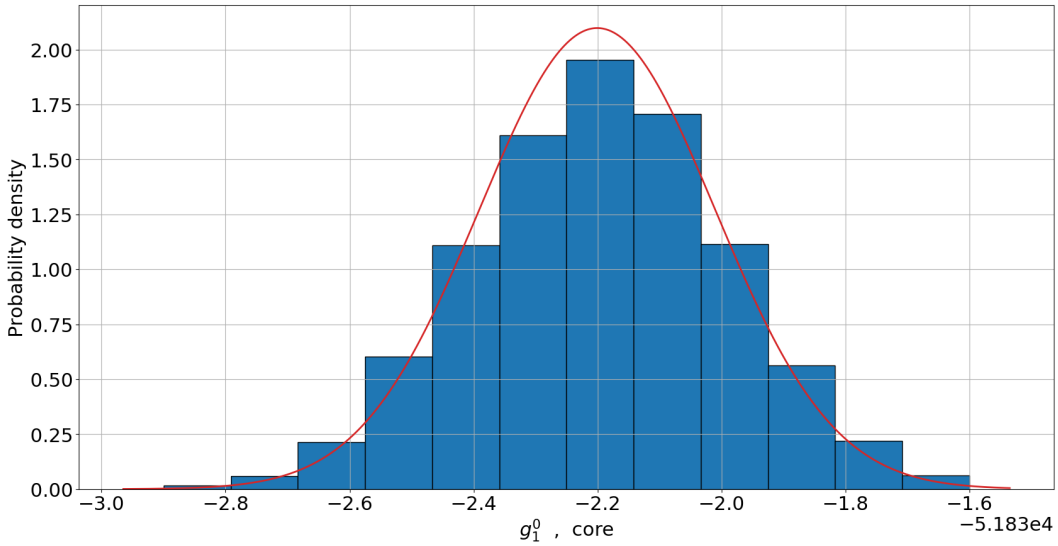



**Figure 8.1:** Marginal posterior distribution of the axial dipole component when applying a GMM prior. Note that the distribution is fitted well be a single Gaussian, as illustrated by the superimposed red line.

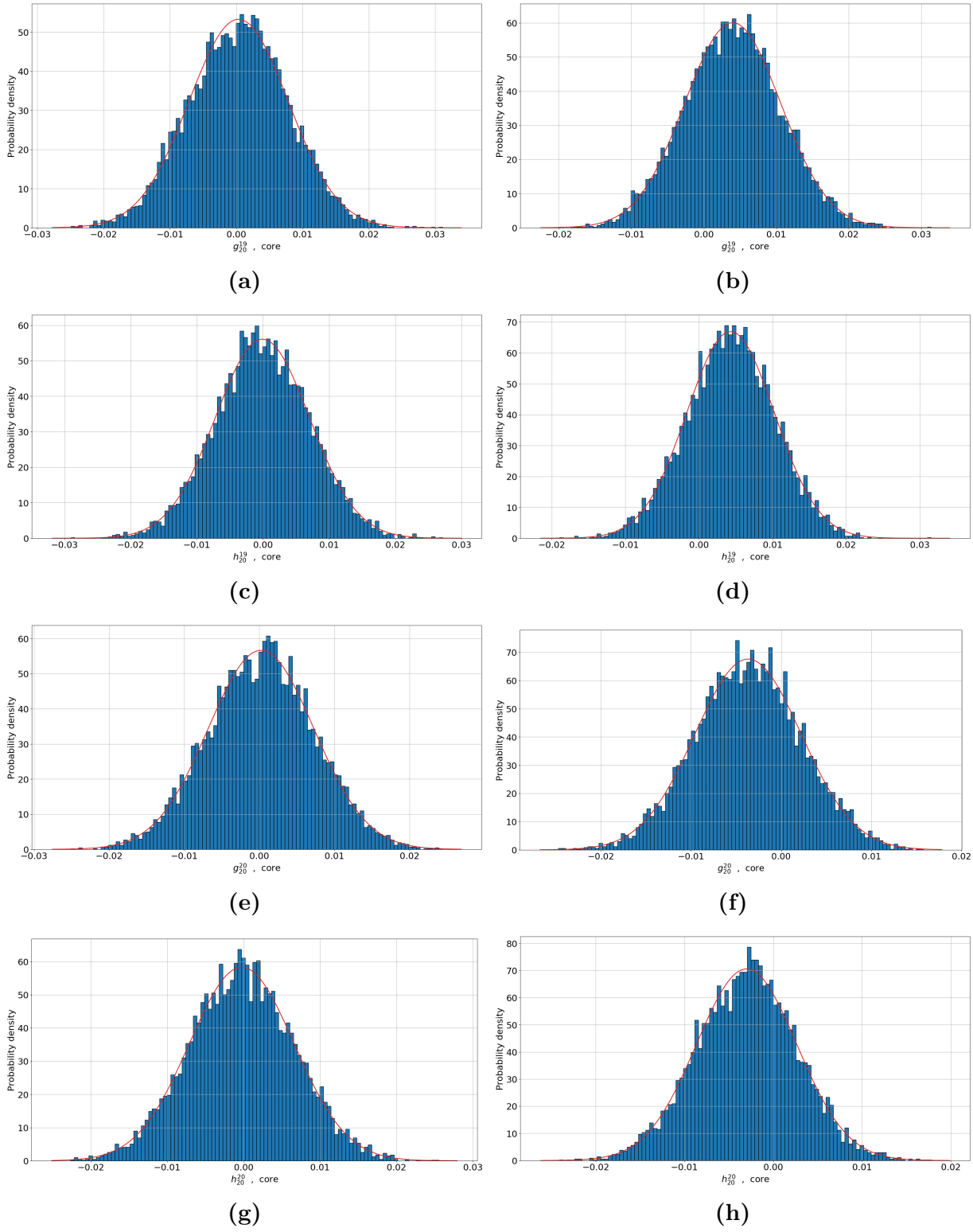



**Figure 8.2:** Comparison of marginal posterior distributions between independent and multivariate Gaussian priors. Superimposed is a Gaussian fit. Left: Independent Gaussian prior. Right: Multivariate Gaussian prior. Note how the multivariate prior cause the mean of the distributions not to be zero by inferring a forcing a correlation between the SH coefficients.

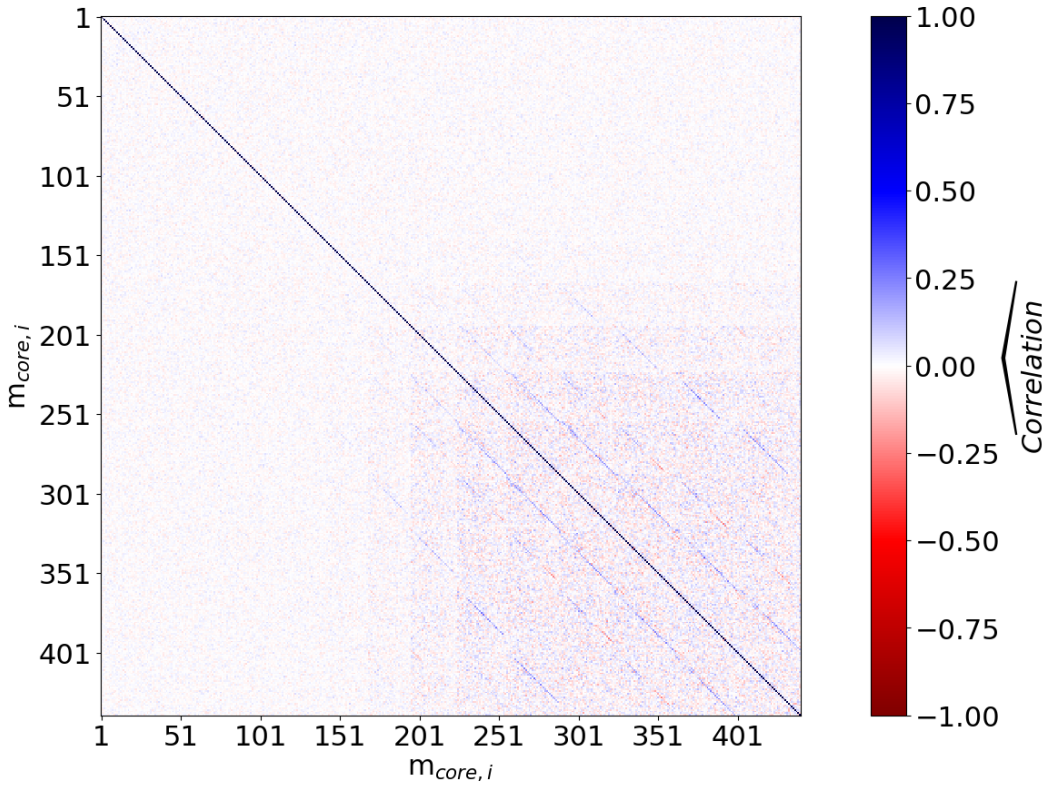



**Figure 8.3:** Mass matrix, chain 1, of test run to SH truncation degree 20 with 8000 warm-up iterations, 2000 post warm-up iterations and a maximum tree-depth of 10. Initialized with the least squares solution and a mass matrix equal the covariance of the core dynamo simulation. The prior is a multivariate Gaussian. Is to be compared with figure 5.4c.

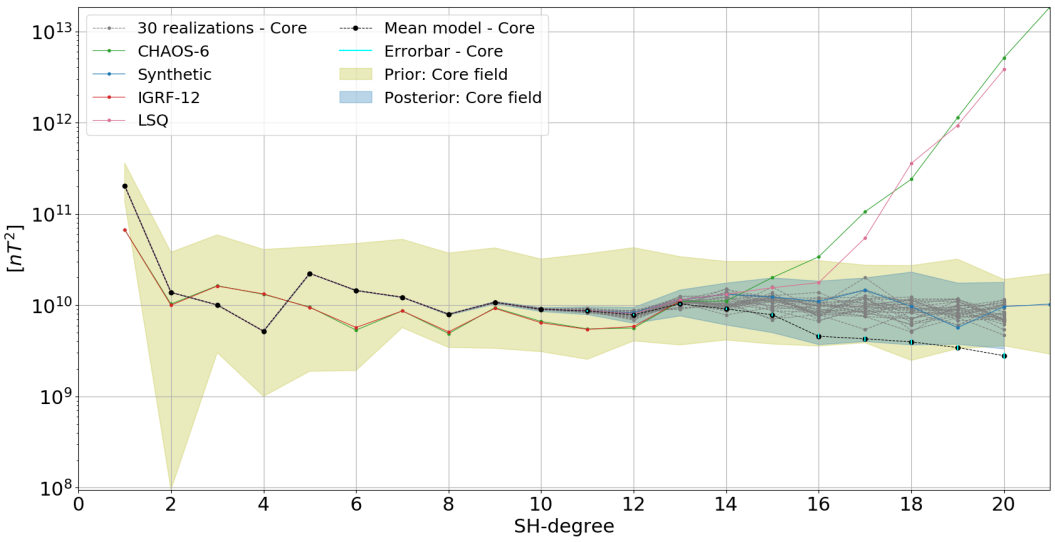



**Figure 8.4:** Power spectrum of test run to SH truncation degree 20 with 8000 warm-up iterations, 2000 post warm-up iterations and a maximum tree-depth of 10. Initialized with the least squares solution and a mass matrix equal the covariance of the core dynamo simulation. The prior is a multivariate Gaussian. Is to be compared with figure 5.3a.

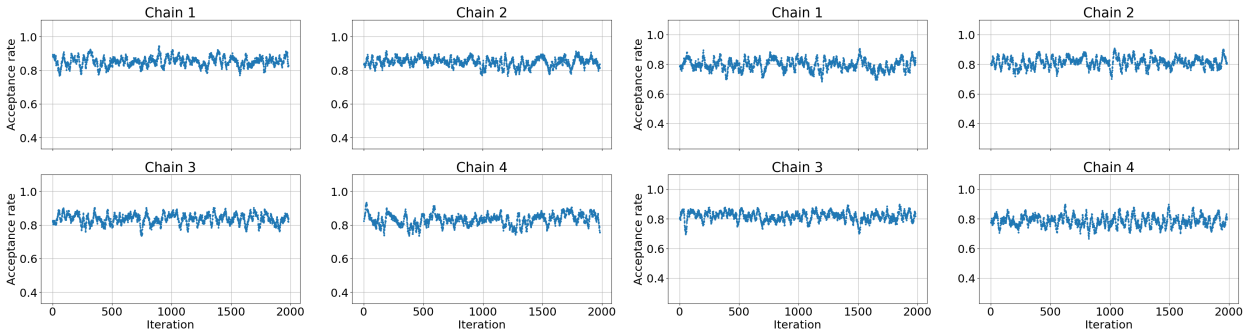



**Figure 8.5:** Comparison of the acceptance rate in the post warm-up period when varying the length of the third warm-up phase. Left: 50 iterations (default). Right: 400 iterations. Note that this a running average, window size 20, of the true acceptance rate.

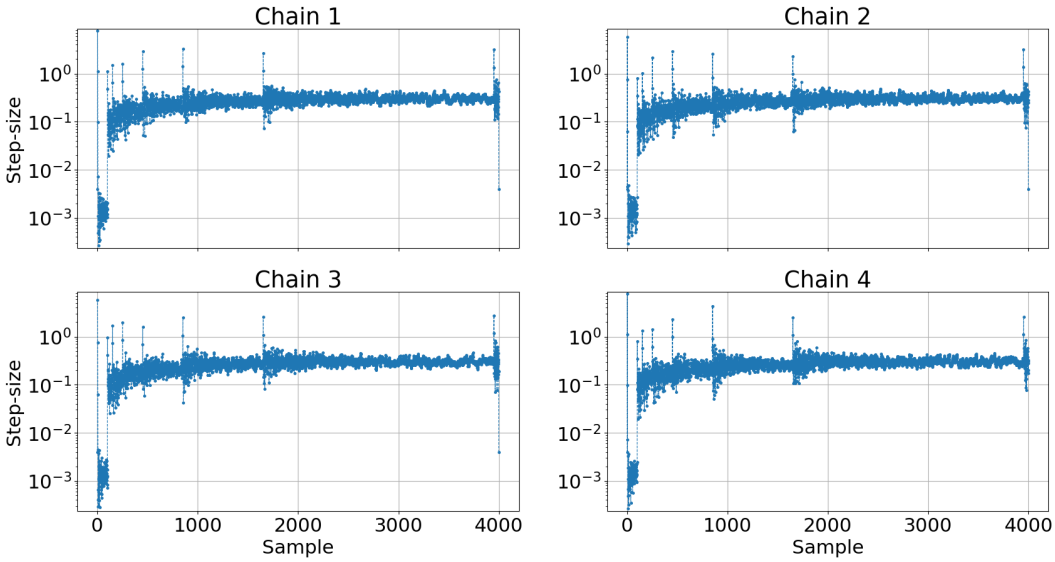



**Figure 8.6:** Time series of step-size in the warm-up period from **test 1**.

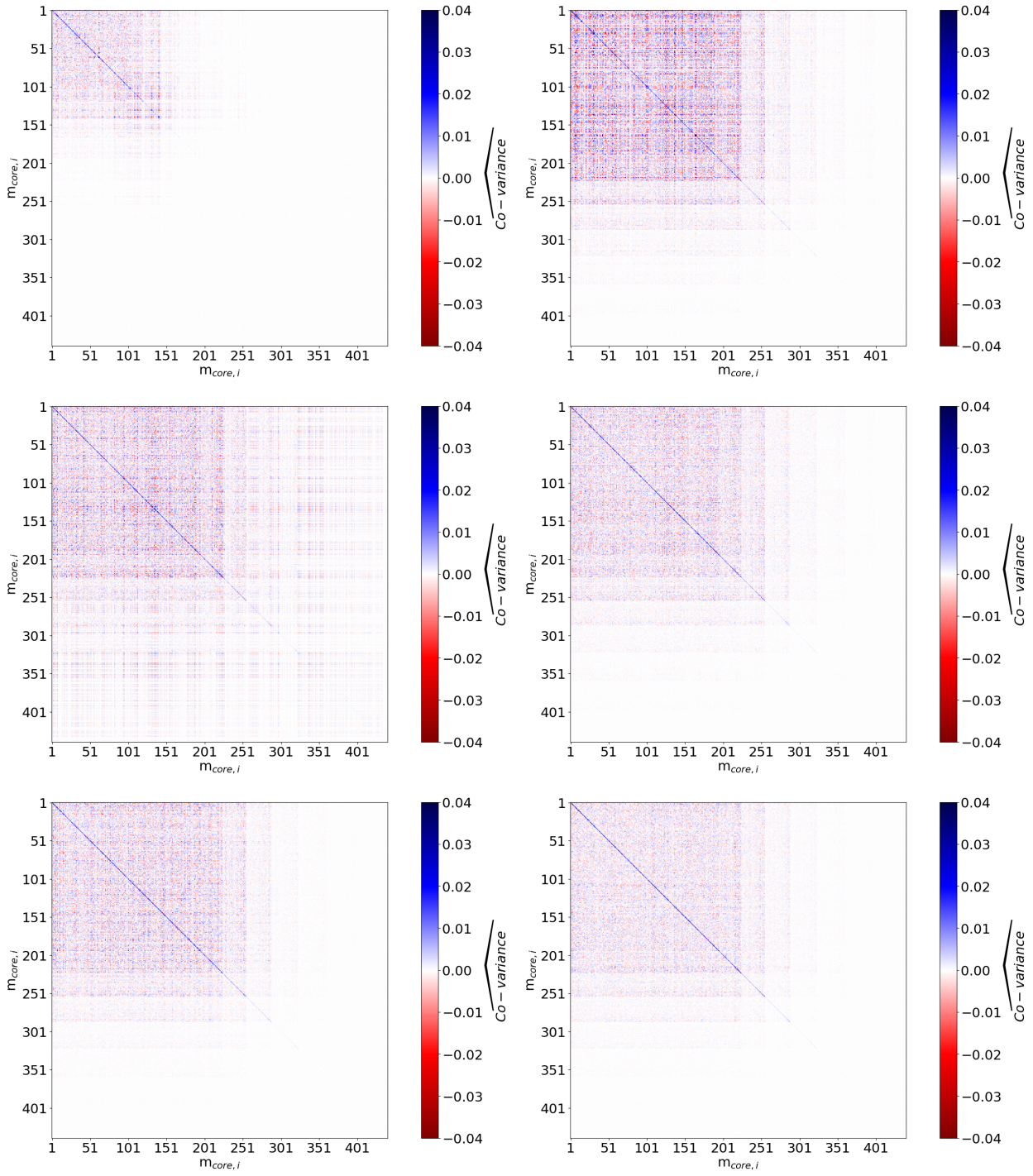



**Figure 8.7:** Comparison of the mass matrix as it evolves through the seven mass matrix adaptations when using 4000 warm-up iterations. On the left is **test 0** on the right **test 1**. Each row represents a mass matrix adaptation. The comparison if continued in figure 8.8.

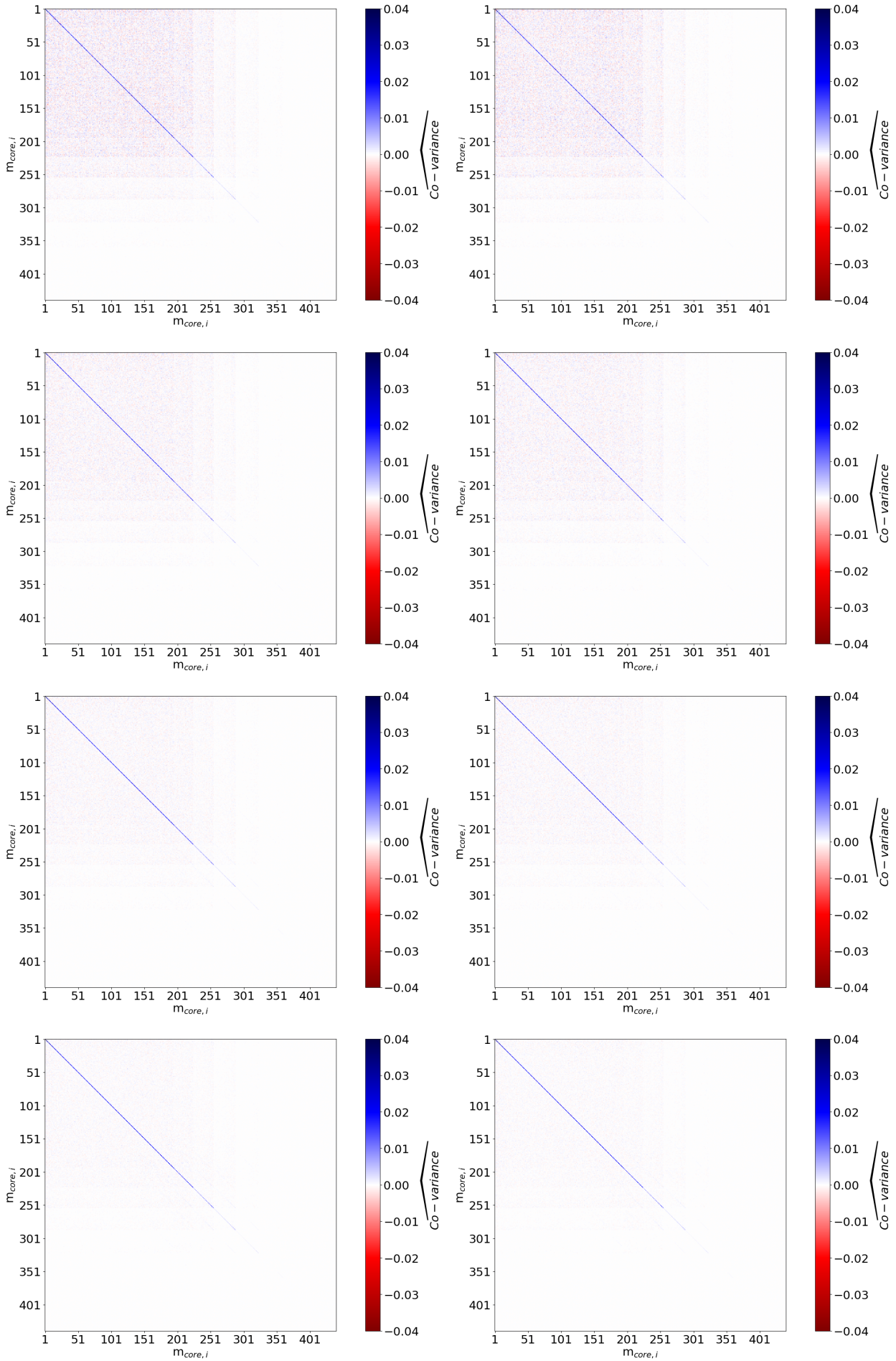



**Figure 8.8:** Continuation of figure 8.7

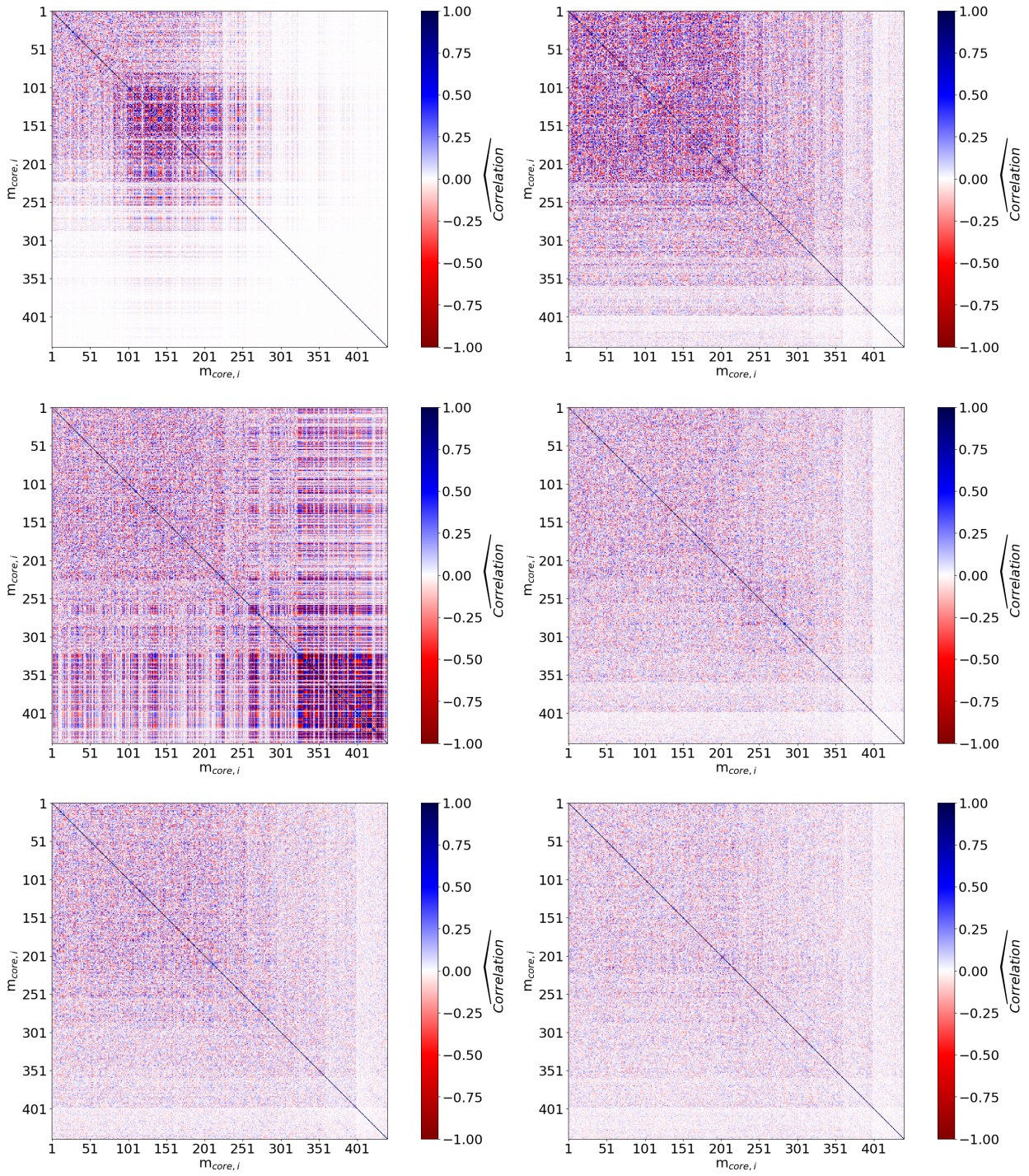



**Figure 8.9:** Comparison of the correlation structure in the mass matrix as it evolves through the seven mass matrix adaptations when using 4000 warm-up iterations. On the left is **test 0** on the right **test 1**. Each row represents a mass matrix adaptation. The comparison if continued in figure 8.10.

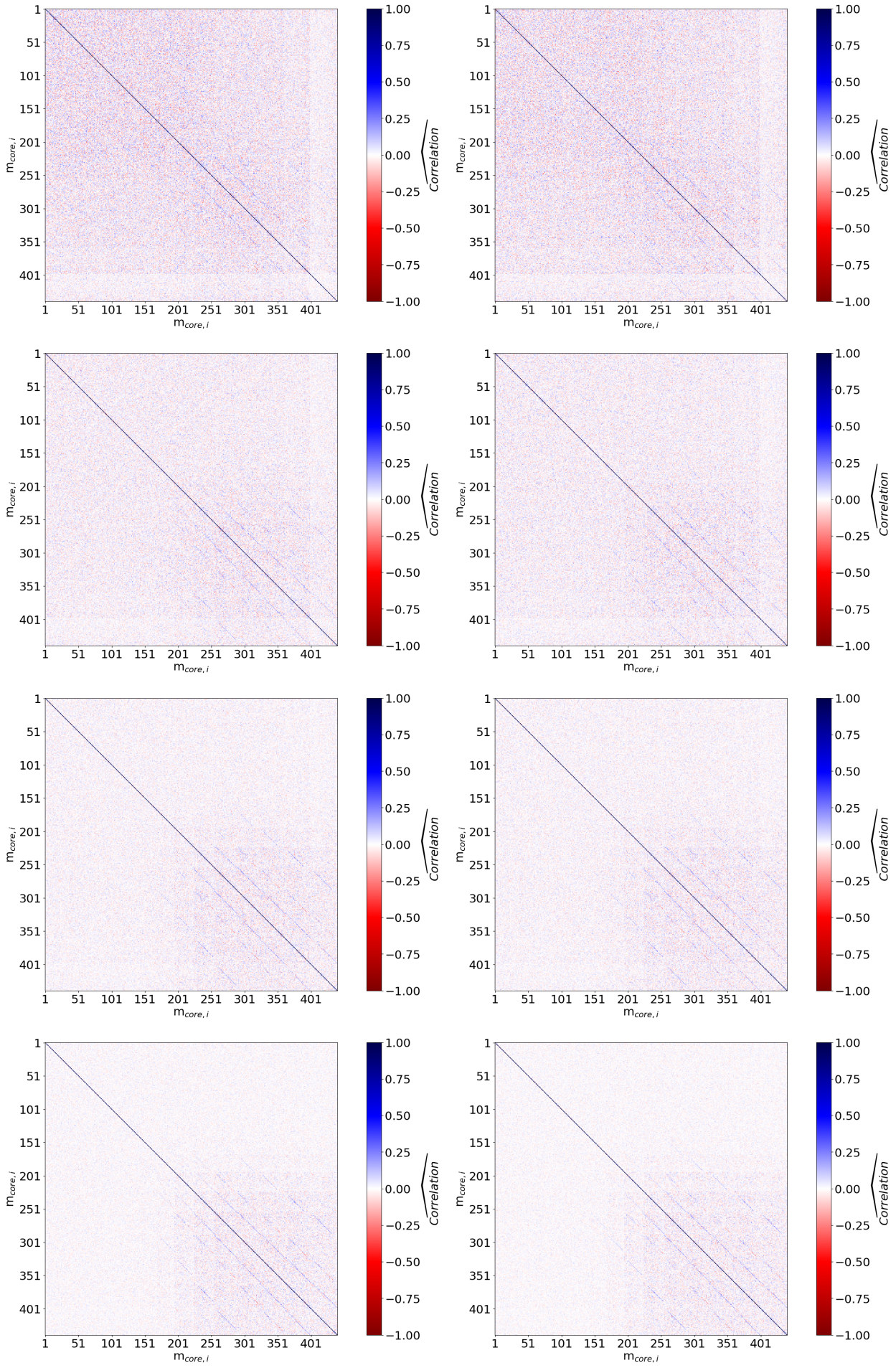



**Figure 8.10:** Continuation of figure 8.9

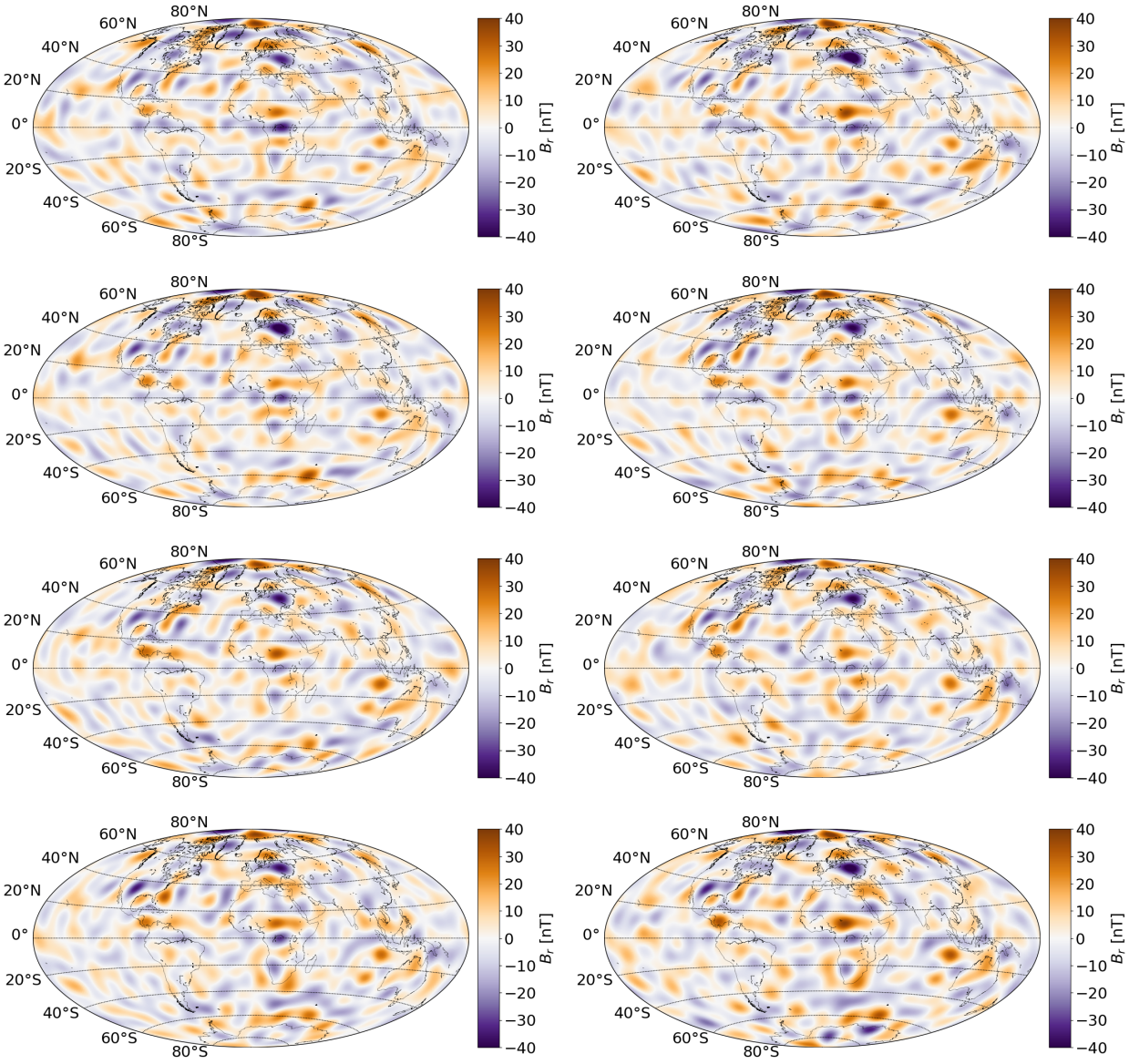



**Figure 8.11:** Eight, random realizations from the posterior lithospheric distribution. These come from co-estimating using 2000 data-points from a data set of real satellite data and sampled to a SH truncation degree of 22.

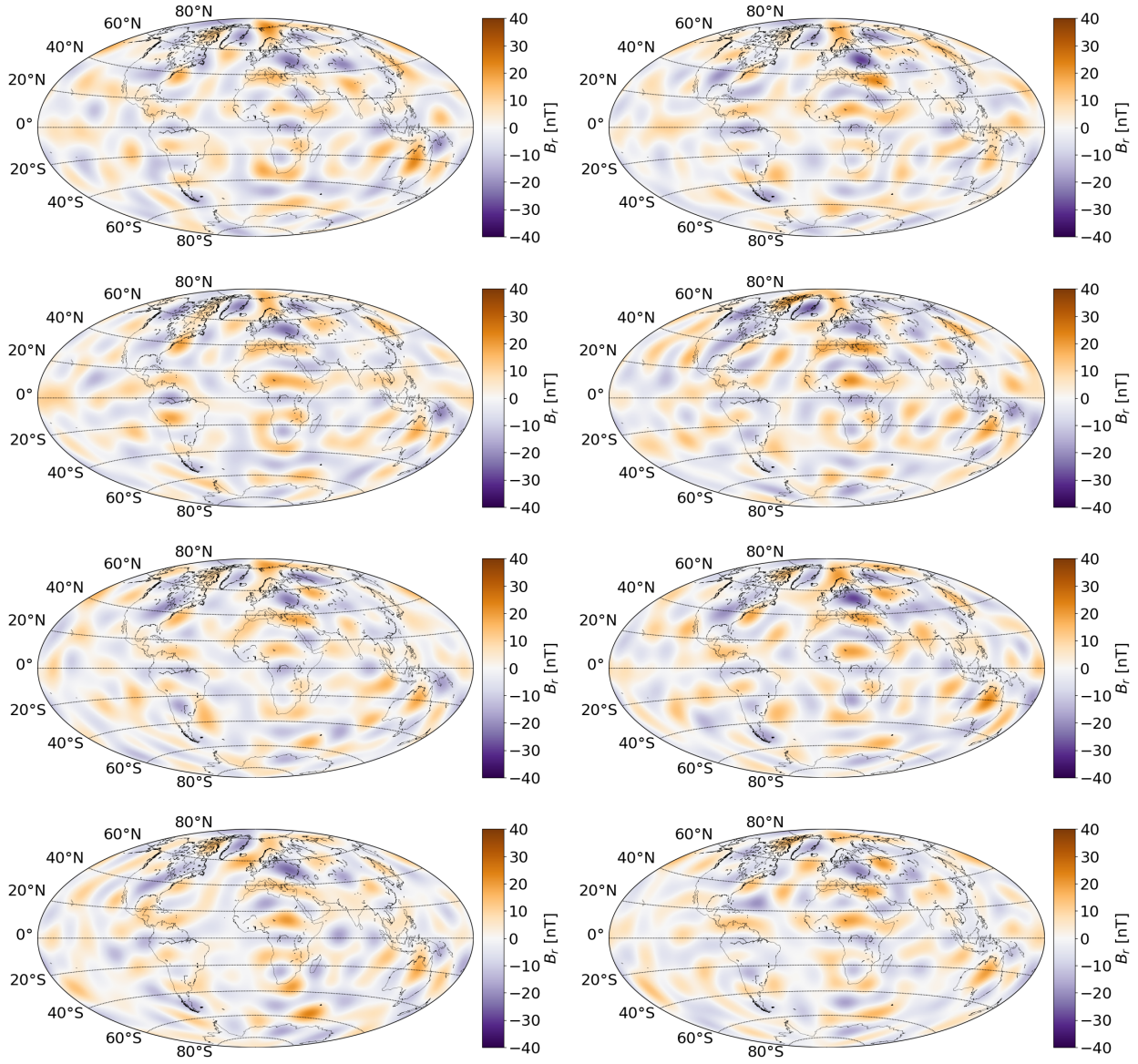



**Figure 8.12:** Eight, random realizations from the posterior lithospheric distribution. These come from co-estimating using 5000 data-points from a data set of real satellite data and sampled to a SH truncation degree of 16.

**DTU Space**
National Space Institute
Technical University of Denmark

Elektrovej, building 327
DK - 2800 Kgs. Lyngby
Tlf. (+45) 4525 9500
Fax (+45) 4525 9575

www.space.dtu.dk