Finding the patterns in mantle convection

Suzanne Atkins

UTRECHT STUDIES IN EARTH SCIENCES

No. 130

Members of the dissertation committee:

Prof. dr. Patrick Cordier
Unité Matériaux et Transformations
Université Lille 1, France

Prof. dr. Nicolas Coltice
Laboratoire de Géologie de Lyon: Terre, Planètes, Environnement
Université Claude Bernard Lyon 1, France

Prof. dr. Boris Kaus
Institut für Geowissenschaften
Johannes Gutenberg Universität Mainz, Germany

Prof. dr. Malcolm Sambridge
Research School of Earth Sciences
The Australian National University, Australia

Prof. dr. Paul Tackley
Institut für Geophysik
Eidgenössische Technische Hochschule Zürich, Switzerland

Printed in the Netherlands by IPSKAMP Printing, Amsterdam.

ISBN: 978-90-6266-470-2

Finding the patterns in mantle convection

Patroonherkenning voor aardmantelconvectie

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
woensdag 24 mei 2017 des middags te 12.45 uur

Suzanne Atkins

Promotor: Prof. dr. J. Trampert

Copromotor: Dr. A.P. Valentine

This thesis was accomplished with financial support from the Netherlands Research Centre for Integrated Solid Earth Science (ISES 2012-81) and the European Research Council Project, iGEO (FP/2007- 2013/ERC Grant Agreement n. 320639).

Contents

Contents

1

Introduction

Humans live on the outer surface of our planet, which we have observed scientifically for a few thousand years. There then remains a vast volume of the Earth and expanse of its history which have never been directly accessible to the scientific community. In this thesis, I explore new ways to study the interior of the Earth.

The mantle lies immediately below the rocky crust and makes up over 80% of the Earth by volume. The mantle is convecting, thereby removing heat from Earth. Figure 1.1 shows a schematic representation of the radial structure of the Earth. Whilst the mantle comprises the bulk of the Earth by volume, there exist a great many fundamental questions about its composition, rheology, dynamics and history. Many of these questions remain unanswered because of the difficulty in making even indirect observations of the mantle, and the many

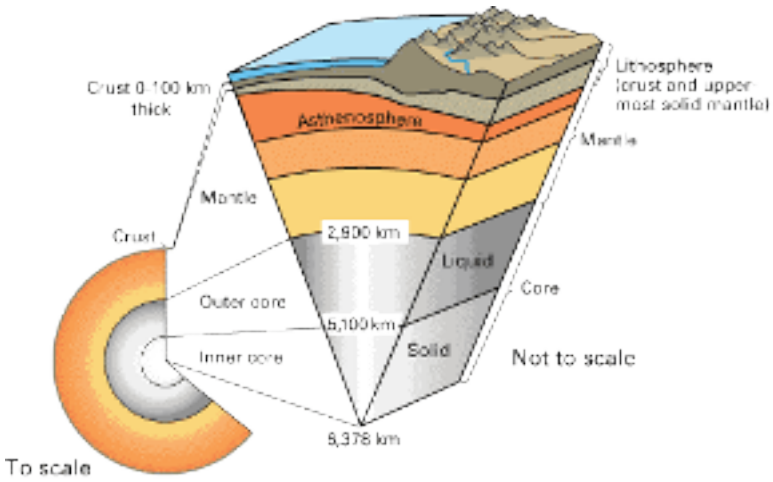


Figure 1.1: The radial structure of the Earth. Figure from the US Geological Survey

possible interpretations of those observations which do exist. There are three main observation types which allow us to make inferences about the mantle: the geological record preserved in the crust, and geophysical and geochemical observations made at the surface. Together these observations are used to constrain models and theories about the Earth. These models and theories can be encapsulated in computer codes which simulate the movements of the mantle, helping us to understand its dynamics and to interpret our observations.

The crust provides information about the mantle in two ways: in the geological record and through the movement of tectonic plates. The two are intrinsically linked, with the former recording the latter. The geological record forms a history of the surface of the Earth preserved in rocks. These have been deposited through sedimentary processes or have cooled and crystallised from molten material. The rocks record information about the changes in environment, with locations becoming cooler or hotter, and wetter or drier, and even experiencing flooding by the sea. These changes can be caused by global climate change, but also by much more drastic regional events caused by movement of the underlying tectonic plates. The plates are rigid or semi-rigid rafts lying on top of the mantle, which fit together to make up the lithosphere, of

which the crust is the uppermost layer. They move continuously, and have probably done so for at least three quarters of the Earth's history. As they collide and move apart, they can massively change a region's landscape, throwing up mountain belts or opening new oceans. These vertical movements are also recorded in the rocks. Oceanic rocks are now found kilometres above sea level in the Alps and Himalaya, whilst metamorphic rocks record a process of burial, heating and subsequent exhumation. With time, mountains erode, but traces of them are found in the sediments which fill basins at the bases of the mountains. These basins also provide evidence of smaller scale uplift and subsidence, with each vertical movement recorded by erosion surfaces. As well as displacing rocks vertically, the plate movements across the surface of the Earth can take a rock far from its latitude of formation. This movement is recorded as the sediments change their environmental signature: desert sandstones formed near the equator may be topped with more temperate riverine deposits, followed by glacial rocks as the plate moves towards the poles. The rocks can also preserve evidence of their latitude of formation in a more quantifiable form, through preserved palaeomagnetism. A rock section's palaeomagnetic history can then be used to trace its movement across the surface of the Earth, as each successive layer in a sedimentary sequence will record changes in the section's position with respect to the Earth's magnetic field.

The geological record therefore constrains how the plates and the rocks on top of them have drifted across the surface of the Earth. This is of great importance, because the tectonic movement of the crust has long been recognised as an expression of the deeper dynamics of the mantle (Hager and O'Connell, 1981). The geological record therefore gives us one way to attempt to access the history of the mantle.

Unfortunately, there remain many puzzles in this relationship. The geological record is incomplete. Rocks tend to erode, so much of the record was lost millions of years before we evolved sufficiently to be able to study it. When rocks do not erode, they run the risk of being subducted. This is a particular problem when trying to piece together the history of the Earth using the magnetic record. Oceanic basalts preserve the history of changes to the Earth's magnetic field in the most accessible manner, but these are the most likely rocks to be subducted. There is little oceanic crust older than 150 Myr on the planet (Torsvik et al., 2010). Palaeomagnetic plate reconstructions rely heavily on the oceanic palaeomagnetic record, making reconstructions which attempt to go further back in time than 150 Myr fraught with uncertainty. For time periods earlier in Earth history than 150 Myr ago, the latitude of some continents can

be constrained, but their longitudinal relationship to each other is much more uncertain (e.g. Austermann et al., 2014; Torsvik et al., 2014).

The other puzzle is the exact nature of the relationship between the crust and the mantle. The coupling between the mantle and the crust depends on the rheology of the mantle, as does the force exerted by the sinking of subducting slabs (e.g. Lowman et al., 2011; Höink et al., 2012). Clues to the rheology of the mantle are also preserved in the geological record. Continents are depressed by loading from ice sheets, which force the continent down into the softer mantle until they reach isostatic equilibrium. When the ice melts, the load is removed and the continents rise up again (e.g. Peltier, 1998). The rebound is recorded by relative changes in sea level. These are fast enough to be recorded over decades. This gives geophysicists a method to find the viscosity of the mantle, albeit with a rather non-unique solution (e.g. Thoraval and Richards, 1997; Forte and Mitrovica, 2001; Rudolph et al., 2015).

The crust preserves a few direct samples of the mantle in the geological record (e.g. Pearson et al., 2014) and indirect samples that are derived by melting. The composition of erupted material provides clues about the temperature of the mantle and how it varies both spatially and temporally (e.g. Klein and Langmuir, 1987; McKenzie and Bickle, 1988; Lee et al., 2009), with hotter mid-ocean ridges producing different basalts to cooler ones. It also provides hints about the composition and history of the source rock for these melts, with some ocean island basalts including geochemical traces of both subducted crustal material and potentially very ancient reservoirs of rock with unusual compositions, which may have existed in isolation since the Earth formed (e.g. Coltice and Ricard, 2002). These ocean island basalts may come from deep in the mantle, while mid-ocean ridge basalts generally come from the upper mantle. Theories about how the mantle works must therefore be reconciled with these samples of mantle rocks, and what they imply about how the mantle mixes and produces melt.

The information contained in the geological record is complemented by geophysical observations, which whilst still made at the surface, give us access to deeper features in the mantle. Seismic waves are one such observation. The waves, originating from earthquakes or explosions, penetrate the mantle and core before returning to the surface to be recorded in seismograms. When the data from the seismograms are inverted for velocity variations in the mantle, they give us a snapshot of the processes happening within our planet today. However, it is difficult to use velocity variations to make interpretations about the dynamics of the mantle, because these variations can be caused by temper-

ature and/or compositional effects (e.g. Trampert et al., 2004). Both chemistry and temperature contribute to density variations that drive convection, but the dynamics are different depending on the cause of the contrast. Besides the uncertainty about what they show, tomographic images have limited resolution. This is due to uneven spacing of earthquake sources and seismic receivers, but also due to the long wavelength of the seismic waves that sample the mantle. Some dynamic processes will therefore simply not be observed because they are too fine scale.

A tomographic image is an instantaneous snapshot of the current state of the interior of the mantle. It therefore does not include direct information about the dynamics of much of the mantle and how it is flowing. Anisotropy in seismic waves may be due to crystal alignment caused by flow in the mantle (e.g. Wookey et al., 2002). However, most of the lower mantle is isotropic, limiting the regions in which this can be used to find the flow patterns. The seismic velocity structure of the mantle does still include other clues about Earth history, besides anisotropy. For example, we can see dense slabs sinking through the mantle from which some dynamical information can be inferred, such as relative density and viscosity contrasts (e.g. van der Hilst et al., 1997).

Together, the instantaneous snapshots taken from geophysical observations and the more temporally extensive geological record can be used to constrain geodynamical models. Geodynamical models are simulations that model the flow of the mantle, subject to what we know about its properties and the choices made during model setup. With these models we can see how features in the mantle evolve, such as the shapes made by rising plumes of hot buoyant material, or how subducting slabs sink into the deep mantle. By changing simulation parameters, such as the rheological conditions, geodynamicists can investigate what conditions are necessary to produce Earth-like behaviour in the mantle. The convection simulations are constrained by the geophysical observations, in that we want the models to resemble the observations, but they also help us to interpret the observations by showing what interpretations are geodynamically feasible. The geological record provides a constraint on the upper boundary layer of the mantle and provides timings for major tectonic events such as the initiation of subduction of plates into the mantle.

When trying to use models to unravel the history of mantle convection, most geodynamicists start with present-day geophysical observations, then push back the movement of mantle material under the condition that it must produce the plate tectonic configuration seen in the geological record, either by running the simulations with the boundary constraints (e.g. Bower et al., 2015)

or by using adjoint inversion methods (e.g. Liu and Gurnis, 2008; Horbach et al., 2014). This sort of study is therefore limited to periods of Earth history with reliable plate reconstructions (Bocher et al., 2016), which exist for around 150 Myr, or up to around 500 Myr with rapidly increasing uncertainty. Beyond this, only very general studies can be conducted, because the mantle is a highly non-linear system. Precise predictions of the location of features therefore cannot be made without imposed boundary conditions that are simply not known for most of Earth history.

We therefore have two tools with which to study the mantle: observations made at the surface, which are both spatially and temporally limited; and convection simulations which model what is happening inside, given assumptions about how the mantle works. Given their limitations, we need to find ways to extract as much information from them as possible.

In this thesis I present a new method for studying the mantle which links convection simulations with observations through machine learning. I use a large data set of convection simulations that allows an artificial neural network to find the relationship between patterns seen in the convection simulations and characteristics of the mantle or its history about which we are interested. Subject to certain caveats, which I will discuss later, I can theoretically use the neural networks to interpret the patterns seen in geophysical observations of the Earth, giving me a new way to access its history and present-day characteristics. Using geophysical observations is currently beyond the scope of this thesis, but is a very promising possibility for further work.

A major strength of my method is that it is fully probabilistic and therefore gives me a way to assess the uncertainty of any inferences I make about the mantle. Without uncertainty estimates, inferences about the state of the mantle are not particularly useful because it is impossible to tell how many other different inferences may also fit the observations or the extent to which these inferences can be relied upon in future studies. This probabilistic approach also allows me to tackle the entire history of the Earth, because I can treat it in a statistical manner, removing some of the problems caused by the mantle's non-linearity.

I can then begin to study problems that have plagued geophysicists for years. Because I work in a probabilistic framework, I can tackle highly non-unique problems, such as the viscosity structure or the nature of compositional variations in the mantle. The non-uniqueness of the problem may simply mean that my solutions are very uncertain, but at least I can quantify the uncertainty. I can also investigate historical features in the mantle, such as the evolution of

a distinct layer at the base of the mantle. In the future, I may be able to use this method to address problems such as the evolution of plate tectonics. Whilst not explored here, the applications of this method extend beyond the Earth. Planetary science is plagued by many of the same problems as deep-Earth geophysics, but often amplified. Very few observations of extraterrestrial bodies exist, compared to the number of observations that we have for the Earth, and obtaining more observations is incredibly expensive and limited by available technology.

The results presented in this thesis are presented as a proof of concept and use only synthetic cases that include too many simplification and assumptions to be used at present for making inferences from real observations. However, by showing some of what is possible by applying this method to synthetic cases, I demonstrate that this approach is feasible and may be worth developing so that it can be used for the Earth.

1.1 Thesis Outline

In the following two chapters, I outline the theory and method behind my approach. Chapter 2 discusses how I set up a geodynamic investigation in a probabilistic framework and how pattern recognition can be used in this context. Chapter 3 then describes my particular geodynamic setup. Chapter 4 shows the extent of the problems geodynamicists face when studying mantle convection. Chapter 5 demonstrates that it is possible to resolve some of these limitations by using a probabilistic approach based around prior sampling and pattern recognition. I apply this method to make inferences about the input parameters to convection simulations. The work in chapter 5 has already been published (Atkins et al., 2016). Chapters 6 and 7 develop and extend this method, exploring other geophysical observations (chapter 6) and investigating the use of this method to make inferences about the bulk composition of the mantle (chapter 7, in preparation for publication). Chapter 8 then presents a possible future application of this method within geodynamical simulations, and shows how pattern recognition can be used to solve problems in convection simulations, as well as to make inferences from observations.

2

Probabilistic methods to find patterns in convection simulations

The biggest challenge when studying mantle convection is uncertainty. Uncertainty impacts every step of the process: how should we set up our geodynamical simulations; how well do they model the Earth; how much difference would changing things make and what do we know about the Earth anyway? It affects the accuracy of any study and means that a range of options may potentially be compatible with extant data.

By using probability theory, I can take into account these uncertainties when trying to find answers to questions about how the mantle works. In this chapter, I explain why a probabilistic approach is best for geodynamical problems and how exactly my probabilistic approach works.

2.1 What is a probabilistic approach to a problem?

A probabilistic approach to a problem provides an answer to that problem not as a single outcome, but as a probability density function. A probability density function is a mathematical description of the probability of the answer taking any value, or, when integrated, the probability of the answer falling within a range of values. In contrast, a deterministic solution produces a single answer or outcome, with no reference to any other possible solutions. The major advantage of describing an answer with a probability density function is that the probability density function also describes all of the uncertainty associated with that answer.

For example, I might want to try to find the mean temperature at a depth of 1000 km in the mantle. A deterministic approach would chose a composition at 1000 km, take an observation of seismic velocity and use a mineral physics relationship to find the temperature, given the chosen composition and observed velocity. This yields a single value for the temperature. However, there are many assumptions and uncertainties in this process. A probabilistic approach would take into account a range of possible compositions, the uncertainty in the relationship between seismic velocity and temperature, and the uncertainty with which the velocity is known. This would then yield a distribution of possible temperatures. The distribution takes into account all of the uncertainty about the process, and gives me an answer that there is X% probability that temperature is between values y and z. It may also show that there is more than one possible temperature range which fits the observed seismic velocity.

2.2 Why use a probabilistic approach in geodynamics?

Geodynamics is the study of the convection and dynamics of the Earth. As with any problem in science, knowing the uncertainty of a solution in geodynamics is important. Without knowing the uncertainty, a solution cannot be reliably used as a foundation when building up more advanced investigations into a system. However, the consideration of uncertainty in geodynamics is particularly important because geodynamicists work with non-linear systems when trying to simulate and model the flow of the mantle.

A non-linear system is one where changing the inputs does not lead to a linear change in the outputs. Figure 2.1 shows a non-linear function, $y = \exp(x)$, to act as an example. A change to the inputs can produce a very large change in the outputs (e.g. when $x \gg 1$) or almost no change in the output (e.g. when

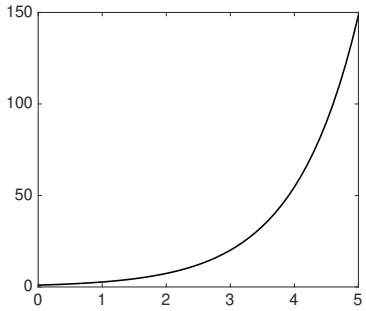


Figure 2.1: The non-linear function $y = \exp(x)$

$x \ll 1$). This system is not random, and the result can be calculated exactly from the input. The problems arise when the desired input is not known precisely. In this example, if I use $x = 4.4$ rather than $x = 4.5$, the difference in the outcome is 8.45. If this exponential function feeds its result into another non-linear function, then another, the small original error could force this complex system to take a course which looks completely different from the one it would have followed if I had used $x = 4.5$ in place of $x = 4.4$.

Such a highly non-linear system is chaotic, where a small change to the inputs leads to outcomes which are very different and the evolution of the system is unpredictable. In climate science, this is often called the butterfly effect, where a small perturbation, such as the beat of a butterfly's wings, can send the system off on a completely different course (Lorenz, 1963). Non-linear chaotic systems can still be fully deterministic, where the path the system takes is dependent only on the input and boundary conditions, as with the example in figure 2.1. However, the margin for error on the inputs, before the effects of the non-linearity overwhelm the process, is often very small, and can be below the accuracy with which it is realistically possible to know the inputs.

Despite their unpredictability, the behaviour of a chaotic system can generally be described statistically using probability distributions. This provides a way to describe the system in terms of the probability of an outcome, taking into account the uncertainty with which the conditions are known. Given a set of conditions which fall within an expected range, an expected range of outcomes can be predicted.

Mantle convection simulations are such a non-linear systems. The outcome is fully dependent on the initial and boundary conditions and the input parameters to the simulation. However, small perturbations to these can generate

very different mantle structures. Unfortunately, many of the necessary parameters for mantle convection simulation are associated with large uncertainties. To demonstrate the limits of the predictability in convection simulations, Bello et al. (2014) calculated the Lyapunov time for 3-D incompressible cases of the mantle convection simulation code StagYY. The Lyapunov time is a measure of the time that a dynamic system takes to diverge unrecognisably from its original path following a perturbation. It is also described as the predictability limit of the system, within which the exact path the system takes can be predicted. Given a 5% uncertainty in the initial temperature structure of their simulations, they estimate that the predictability limit for the Earth is around 95 Myr. Their results demonstrate the impossibility of allowing a simulation which is freely evolving over billions of years to reach the exact thermal structure of the Earth today.

The Lyapunov time is an important limit, because it imposes a constraint on most methods for studying the evolution of mantle convection, and shows why geodynamists struggle to unravel the history of the mantle, especially when other sources of uncertainty besides the initial temperature structure are considered. Most methods for studying the recent history of mantle convection start with observations of the structure of the mantle today, and then attempt to rewind it. The structure is generally taken from seismic tomography, which is converted into thermal maps of the mantle. The thermal heterogeneity in the mantle causes buoyancy difference which drive convective flow. With an approach such as this, several sources of uncertainty have already been added before any inversion has been attempted: those associated with the seismic data; the seismic inversion process necessary to produce a tomographic model; conversion from seismic wave speed to temperature; the relationship between temperature and buoyancy; and the equations used to describe mantle flow which require assumptions about rheology. All of these uncertainties, and many more, are what limit the predictability of the mantle, because the initial state cannot be adequately described.

Once an initial mantle state has been established, with all the uncertainties therein, there are several methods for rewinding mantle flow. The time-dependent flow equations can be solved in reverse, by stepping time backwards through history instead of forwards. However, thermal diffusion is time-irreversible and must thus be neglected, meaning that eventually backwards convection produces a stable, stratified mantle configuration (Kaus and Podladchikov, 2001). The adjoint inversion method used by Bunge et al. (2003), Ismail-Zadeh et al. (2004), and Liu and Gurnis (2008) addresses the effects of

thermal diffusion by linearising the relationship between initial model conditions and the misfit between the final flow pattern and the observations, having assumed lengths in the parameter space over which it is reasonable to linearise this relationship. The method can be used to investigate both present-day (e.g. Worthen et al., 2014; Ratnaswamy et al., 2015) and historical mantle structure (e.g. Liu and Gurnis, 2008; Bocher et al., 2016). Both the time-reversal method and the adjoint method are fundamentally time-limited by the predictability of mantle flow since the initial mantle structure is not perfectly known (Conrad and Gurnis, 2003; Bello et al., 2014). By assimilating geological observations, such as plate reconstructions (e.g. Bower et al., 2013; Shephard et al., 2014), this time limit may be extended somewhat, because it reduces the number of possible paths that the chaotic system can take. The timescale is then determined by the resolution of the data coverage, both spatially and temporally, and is limited to periods with reliable plate tectonic reconstructions (Bocher et al., 2016). The data which are assimilated into the system add a further source of uncertainty.

The expense of running these convection simulations adds one more source of uncertainty. Because of the expense, most studies only allow a few parameters to vary during the simulations (e.g. temperature and viscosity variations, and the location of subducting slabs and tectonic plates). This means that many other features, such as the bulk composition of the mantle or the relationship between viscosity and pressure are fixed in the simulations and their contribution to the evolution of mantle flow are not considered. In an ideal case, many convection simulations would be run, varying all possible parameters in all possible combinations. However, this is generally not feasible.

2.3 Finding stable statistics to study

The exact locations of heterogeneities in the mantle become unpredictable within a few million years out of the Earth's 4.5 Gyr history. However, if some stable statistics describing the probabilistic evolution of mantle convection patterns can be found, which do not rely on knowing the exact structure of the mantle at any time, I can begin to investigate the longer term evolution of the mantle in a probabilistic way. Such a stable statistic might be the 1-D mantle temperature profile. This does not describe the precise location of heterogeneities in the mantle, but does constrain the general statistical path the simulation has taken.

In figure 2.2, I show examples of how the non-linearity affects two example convection simulations. The input parameters for these simulations are given

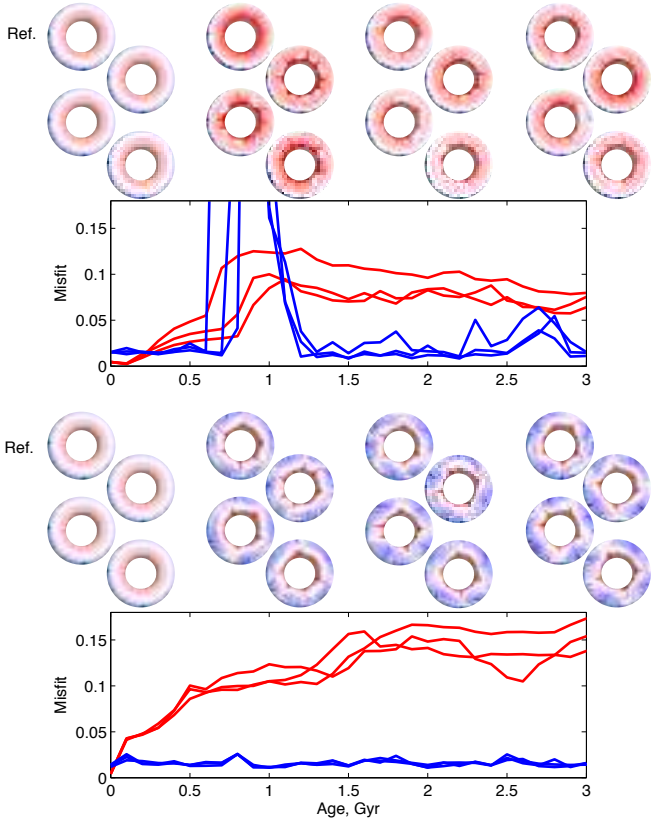


Figure 2.2: The misfit between the full temperature field (red) and amplitude spectra (blue) when four simulations are started with identical model parameters, but initial 20 K perturbations located in different places. Two different comparison sets are run with different model parameters. The misfit is calculated according to the method of Bello et al. (2014), where the misfit $= 1/N \sum |t_i - t_i^{\text{ref}}|/t_i$, where t^{ref} is the reference simulation, shown in the top row of annuli. The annuli show the four simulations at various time steps. In general, the misfit for the amplitude spectra are much lower than for the temperature in the spatial domain, indicating that the amplitude spectra are stable with respect to input parameters despite the small initial differences. Figure previously published in Atkins et al. (2016).

in table 3.5. The upper model is number 16 and the lower is number 7 in the table in section 3.3. In each case, I start four simulations using the convection code StagYY (Hernlund and Tackley, 2008; Tackley, 2008) with the same input parameters, but each has differently placed initial 20 K random perturbations. I then allow each simulation to run for 3 Gyr. The annuli show the temperature field of each simulation at 1 Gyr time steps, all plotted with the same colour scale. For each set of input parameters, I take one simulation, plotted on the top row, to be a reference simulation and calculate the misfit between this reference and the other three simulations. The misfit in the spatial domain is the same as used by Bello et al. (2014):

$$\text{misfit} = \frac{1}{N}\sum_{i=1}^{N}\frac{|t_i - t_i^{\text{ref}}|}{t_i} \tag{2.1}$$

where t_i is the temperature of each cell in the perturbed and reference case, summed over the total number of cells in the simulation grid. The red line in figure 2.2 shows the development of this misfit function with time, as the perturbations cause the simulations to diverge. I then repeat the same calculation, but instead of using the temperature field, I use the full amplitude spectra of the temperature field, following a Fourier transformation into the frequency domain. This process is described in more detail in section 3.1. The evolution of the spectral misfit is plotted in blue. By comparing the spatial and spectral misfits, the advantage of working in the frequency domain is immediately apparent. Whilst the simulations diverge spatially, with upwellings and downwellings in different locations, their heterogeneities all have very similar wavelengths, making the difference between them much smaller in the spectral domain. The amplitude spectral representation for these simulations is stable with respect to the small initial perturbations to the temperature field, unlike the spatial representation. This stability holds for many millions of years, potentially allowing me to study the evolution of the mantle over periods much longer than its Lyapunov time.

The large peak in the misfit for the upper set of simulations (case 16) in figure 2.2 is because subduction begins last in the reference case, as can be seen from the annuli. This causes the amplitude spectra to diverge, but they later converge again as convection stabilises. This difference in onset time demonstrates the necessity of using a probabilistic approach: at this time step, the same input parameters can produce two very different observations.

By using the amplitude, I remove the phase information, which contains the the spatial variations brought about by the non-linearity of convection, leaving

just the relative strength of each wavelength of heterogeneities. This is a standard analytical approach for analysing geodynamical models (e.g. Becker and Boschi, 2002; Deschamps and Tackley, 2008) because the absolute position of a heterogeneity such as an upwelling plume is only a function of the random initialisation. However, by using only the amplitude, I lose how features are located relative to each other, and whether this is affected by the choice of simulation input parameters. For example, the position of plumes with respect to thermochemical piles at the base of the mantle may provide information about the likely dynamics within the piles. Whether plumes rise from the edges or the middle of the piles may be a function of the heating rate and composition of the piles, and may also indicate how stable they are likely to be and how much anomalous material can be expected to be entrained in the plume (e.g. Steinberger and Torsvik, 2012).

2.4 Bayes' Theorem

The ultimate aim is to be able to take some observation or feature of the Earth and to use it to make an inference about something geodynamical. Once I have found a suitable observation which is stable with respect to very small perturbations, such as the initial temperature perturbations, I can use it to make inferences. Such an observation might be the amplitude spectra of the temperature field in the frequency domain, as shown in figure 2.2. This observation could then be used to make inferences about the input parameters to the convection simulation, some mantle structure other than the temperature such as the viscosity, or a point in the simulation's evolution, for example the time at which subduction began.

All of the inferences I make are represented using probability density functions. These take into account all of my knowledge about the system and its associated uncertainties using Bayes' theorem. Each of the uncertainties, including those associated with the input parameters, boundary conditions and the choice of the mathematical description of the geodynamical processes can be described by a probability density function. These distributions combine to give a posterior probability density function for whatever feature I am considering. This can be described using Bayes' theorem (Bayes, 1763), which is written as:

$$P(m_i|d_j) = \frac{P(m_i)P(d_j|m_i)}{P(d_j)} \tag{2.2}$$

Bayes' theorem contains three probability density functions: $P(d_j)$, $P(m_i)$ and $P(m_i|d_j)$. The first probability density function, $P(d_j)$, describes the probability of an observation taking a particular value. This describes how informative a sample drawn from the distribution $P(d_j)$ is. If the probability of observing a particular sample is low, it indicates an unusual set of conditions which created that sample, making it more informative. For example, $P(d_j)$ might describe the probable mean temperature of a convection simulation at a depth of 1000 km. If I draw a sample from this distribution, it is much more probable that I get a simulation with a temperature of 2000 K than 200 K. However, a sample observation of 200 K suggests something unusual is happening in the simulation, therefore it is more informative because there are very few combinations of input parameter values which would lead to such an observation.

The probable distribution of the feature about which I want to make an inference is described by the probability density function $P(m_i)$. This is the prior distribution of feature m_i, and includes all of the information I already know about the feature m_i before I start the experiment. This feature could be, for example, the input parameters to the convection code.

The third probability density function is the posterior probability density function $P(m_i|d_j)$ for feature m_i given that I already have a sample from the distribution $P(d_j)$. The sample from $P(d_j)$ could be a temperature observation, for which I want to find the probability that a particular set of convection simulation input parameters were responsible for that temperature observation. For this, I need one more piece of information. This is a mathematical description about how samples from $P(m_i)$ are related to the observation from distribution $P(d_j)$. The samples from these two distributions lie in two different parameter spaces, d_j in the data or observational parameter space, and m_i in the model space. The mathematical relationship can then be said to describe a mapping between the two spaces (Tarantola, 2005). For convection simulations, the simulation code contains this mathematical relationship. The relationship between the two parameter spaces enters the posterior distribution through the likelihood function, $P(d_j|m_i)$. This finds the likelihood that a sample m_i from the model space is compatible with observation sample d_j. It is evaluated by using the model space sample as the input to the mathematical forward relationship and calculating the outcome. The difference between the outcome of this forward calculation and the observation sample d_j gives a misfit, the reciprocal of which gives the likelihood.

There are several ways in which I want to use Bayes' theorem to study mantle convection. In some cases, I take observations of the end-state structure

of my convection simulations, and want to make inferences about the input parameters to the convection code. The structure of the mantle is then the observation d_j, be it thermal or chemical. The model parameters are any or all of the input parameters to my convection code, StagYY. The prior distribution of these model parameters is described in chapter 3. I can also reverse the process by using Bayes' theorem to predict the probability of a particular end-state of a convection simulation, given a set of StagYY input parameters. In both cases the likelihood function is calculated using the convection code StagYY.

2.5 Sampling to find the posterior distribution

Bayes' theorem allows me to find the posterior probability distribution for parameters which lie in the model space, given an observation. It is particularly useful for problems where a forward theory exists linking model parameters to observations, but no mathematical inverse solution exists. Using Bayes' theory, an inverse relationship between model parameters and observations can be found where only the known forward problem has to be solved. The distribution describing this inverse relation is found by sampling. Samples are drawn from the model space prior distribution and the mathematical forward problem is evaluated which maps these model space samples into the data space. The likelihood function then gives them a weight, turning the distribution of samples into a posterior probability density function given the observation. This probability density function can then be used to describe the probable distribution of model parameters which would be needed to generate the observed sample from the data space.

To construct the posterior for any observation, there are two approaches to deciding where to place my samples: prior and posterior sampling. The difference between these two approaches is the point at which my observation d_j, for which I want to make an inference about its unknown model parameters, is introduced, and therefore at what point the likelihood function is evaluated.

Prior sampling selects model parameter values randomly according to the prior distribution of model parameters $P(m_i)$, with no reference to d_j. The samples in the data space then cover all of the regions which are possible according to the prior distribution of m_i. The forward simulations are run for all of these model parameters, mapping them into the data space. Only then is the observation d_j introduced. The posterior is found by interpolating between these prior samples. This relies on the posterior varying smoothly between samples (Käufl et al., 2016). It is also possible to use the prior samples to con-

struct a marginal posterior probability density function $P(m_i|d_j, m_{k\neq i})$ which only considers one dimension of the model parameter space, but includes all of the effects of the other dimensions, $m_{k\neq i}$, which are integrated out.

Posterior sampling works in the opposite direction. I begin with the observation d_j for which I do not know the corresponding model parameters. I then draw a sample in the model space from the prior distribution of possible model parameter values, calculate the forward problem and immediately evaluate the likelihood function. Depending on the sampling algorithm used, the model space samples are preferentially selected to maximise the likelihood function. There is then a higher sample density in regions of model space which have a higher likelihood of explaining the observation. The sample density follows the posterior distribution for the model parameters given the observation. Monte Carlo sampling methods work by posterior sampling (e.g. Sambridge and Mosegaard, 2002; Tarantola, 2005). Posterior sampling considers all dimensions at once, producing a full conditional probability density function. Because the samples are clustered in regions of high likelihood, they are much more closely spaced so interpolation distances between samples are much smaller (Käufl et al., 2016).

There are advantages and disadvantages to both sampling methods. The main differences are in the uncertainty of the posterior distribution and the reusability of the samples. These differences arise simply because posterior sampling is entirely tuned to one single datum, whilst prior sampling is a much more general approach. This is covered in more detail in Käufl et al. (2016).

The posterior distribution produced by prior sampling is the result of interpolation between samples. Because the samples are not selected to maximise the likelihood function, there may not be many samples in the high likelihood region of the model space for a given observation. The interpolation distances between samples are therefore potentially greater than in posterior sampling, where the samples are focused in this high likelihood region. This means that the posterior is generally less certain, and a more conservative estimate for the posterior distribution is produced than would be the case with a posterior sampling approach. Posterior sampling therefore generally produces a more detailed representation of the posterior probability density function, although in the case of infinite samples, the posterior distribution found by the two sampling approaches should be identical.

Constructing the posterior by interpolation also assumes that the distribution varies smoothly between samples. If this is not the case, the posterior distribution produced will be unrepresentative of the true posterior. I test this as-

sumption by using independent test samples. These are model space samples drawn from the same prior distributions for which the data space observation is calculated using the same forward calculation. These test samples are not used to construct the posterior distribution, so are a completely independent test of the interpolation. If the interpolated posterior distribution produces an adequate representation of the known model parameters given the test observation for the test sample, I assume that the sampling distribution is sufficient to capture the posterior probability density function in that region of model space.

The major strength of the prior sampling approach lies in the reusability of the samples. In posterior sampling, the samples are clustered in a region of model space that maximises the likelihood of a single observation. If the posterior distribution for the model parameters given a different observation is required, a whole new set of samples must be drawn and the forward problem calculated for the new samples. For computationally expensive forward problems, it is prohibitively expensive to run multiple sampling campaigns for different observations. The advantage of posterior sampling is that the samples are distributed throughout the model space. The likelihood function can then be calculated for any observation very rapidly and the posterior found by interpolating between the samples, provided that the prior distributions for the model parameters are adequate. The same samples can therefore be used repeatedly with multiple different observations. The evaluation of the posterior is a very rapid and computationally inexpensive process. For forward simulations which produce multiple observations, I can use the same set of samples to find the likelihood function and posterior distribution for different types of observation. For example, I get observations including seismic velocity of the mantle, plate motion velocity and the location of temperature heterogeneities from a single forward simulation. I can use these different observations from the same sample set to make inferences about each parameter, giving me an *a priori* method to test to which observations the parameters affect the most, without further sampling.

The rapidity and reusability of the prior approach has been shown to be advantageous for repetitive geophysical problems, such as inversion on a point-by-point basis, e.g for the 2-D inversion for the depth of the Moho discontinuity, as done by Meier et al. (2007); or where the speed of the inversion is crucial, such as in earthquake early warning applications (e.g. Käufl et al., 2014). My application benefits in a third way from this reusability. Observations of the mantle have high levels of uncertainties and vary depending on the data and

method used (for a recent example see Schaeffer and Lebedev (2015) for comparisons of mantle tomographic models). I can therefore easily compare the posterior distribution for mantle features and model parameters given different observations of the Earth, without having to commit an expensive sampling campaign to a single observations. This gives me a method of testing the robustness of any inferences I make. For example, if I get a very different inference from two different seismic tomographic images, then the causes of the differences between the models must be carefully considered and it may be that the inference is not robust given the uncertainty in the tomographies.

2.6 Mixture density neural networks

Having drawn samples from the prior space, I need a reliable way to interpolate between them in order to use them to make inferences about new data space samples. To do this, I use mixture density neural networks. These are a type of pattern recognition learning algorithm which are also capable of approximating complicated mathematical functions and providing an output in the form of a parameterised probability density function. Once set up and trained, they can rapidly find the posterior probability density function for the model parameters associated with any sample observation taken from the same data space distributions as the training samples.

They do this by learning an approximation to the mapping between probability distributions in the data space and probability distributions in the model space. Because they learn this relationship, they can interpolate between samples to make inferences when presented with new samples. This is one way in which neural network representations of the posterior differ from Monte Carlo representations. Monte Carlo sampling leads to a collection of sample pairs in the model and data space. However, no information on the mapping function between the two parameters is learnt, so interpolation between samples is not easy (e.g. Sambridge, 1999). The best way to find out what happens between samples is to take another sample.

Network setup and architecture

Neural networks are complicated non-linear functions formed by a network of interlinking functions. The books by Bishop (1995) and MacKay (2003) both provide a good introduction to neural networks. I want to train a network to find the relationship between model and data space, such that when presented

with a data space sample, it can give me posterior probability density function for a model space parameter.

The basic building block of a network is a neuron, which contains a weighting factor and a function, illustrated in figure 2.3. The neuron receives an input value, d, which is multiplied by a weighting factor, w. The weighted input is then fed into (or activates) the function. The function is generally simple, and could be linear, hyperbolic, step or another such function. I use a variety of functions in my network. The output of the neuron may be the result of the network, or can form the input to a subsequent neuron. The neuron input, output and function are often described as the nodes of the neuron respectively.

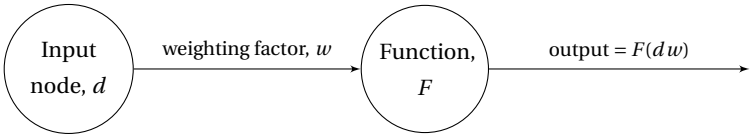


Figure 2.3: A neuron, the building block of a neural network

A neural network is built up of interlinking neurons. A simple two-layer feed-forward network is shown in figure 2.4. The network is given an observation as an input, which is fed forward through all of the functions to produce an output. Each set of weights and function nodes make up a layer. The middle layer of weights and functions (w^1_{ij} and F^1_j) is called the hidden layer. The function nodes in the hidden and output layers are all connected to multiple nodes in the preceding layers. These input nodes can be the network inputs (d_i), or the output from the previous layer (e.g. $F^1_j(d_i\,w^1_{ij})$).

The input nodes introduce an observation to the network. Each input node could, for example, be one pixel of a seismic tomographic image of the mantle. Using more input nodes, and therefore more details from the observation, will allow for more detail to be included in the output. The number of input nodes therefore imposes a constraint on the complexity of the outcome. The same logic applies to the number of hidden nodes used. However, the number of nodes and weights in the network are constrained by the number of samples available to train the network (Baum and Haussler, 1989).

A network can also have multiple hidden layers. With enough hidden nodes, a network such as the one in figure 2.4 is capable of approximating any function (e.g. Hornik et al., 1989). However, especially if few samples are available, it is not always possible to train a network which is big enough to approximate a very complex function with a high enough degree of accuracy (Baum and Haussler, 1989), and the inputs must be preprocessed to extract relevant features

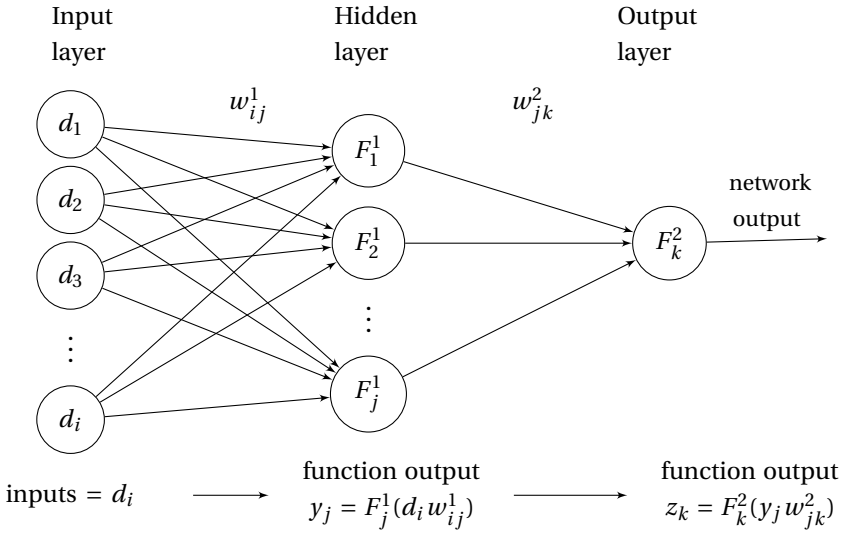


Figure 2.4: A simple multi-layered neural network

from the raw data. This reduces the dimensionality of the input, and therefore the size of the network. A rapidly developing field of machine learning is deep learning, where networks have many hidden layers to extract complex patterns from high-dimensional raw inputs (e.g. LeCun et al., 2015). In deep learning, much less preprocessing of the inputs needs to be done, because the networks extract interesting features automatically, allowing the networks to solve much more complex problems (Schmidhuber, 2015).

I train my networks using sample convection simulations. The simulations are very computationally expensive, therefore I only have a small set. I therefore use a relatively small and simple network architecture which is similar to the one shown in 2.4. However, the network in figure 2.4 only produces a single-valued outcome, whereas mine produces a probability density function, parameterised by a set of Gaussian distributions in a Gaussian mixture model. Each of these Gaussian distributions is characterised by a mean, a standard deviation and a weighting factor, which determines its contribution to the probability density function. I therefore have three network outputs for each Gaussian distribution. Figure 2.5 shows a mixture density network.

The algorithm underlying the mixture density network is relatively straight-

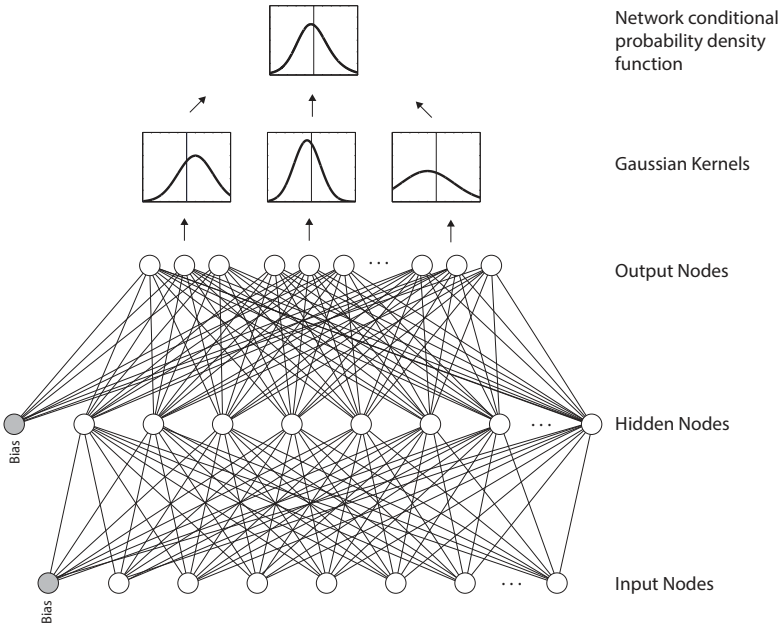


Figure 2.5: A mixture density network, with the number of nodes I use in each layer. Figure modified from Atkins et al. (2016).

forward, and does not vary significantly from that given in figure 2.4. In my networks I use hyperbolic tangents in the hidden layer, F_j^1. The output, y_j, from the hidden nodes is then:

$$y_j = \tanh(d_i w_{ij}^1) \tag{2.3}$$

where d_i are the inputs to the network and w_{ij}^1 are the weights joining the input nodes to the hidden function nodes. The outputs from the hidden layer then act as the inputs to the next layer.

The network output is a marginal probability density function, $P(m_i|\mathbf{d},\mathbf{w},m_{j\neq i})$. It is parameterised by a set of K Gaussian kernels which are added together according to:

$$P(m_i|\mathbf{d},\mathbf{w},m_{j\neq i}) = \sum_{k=1}^{K} \alpha_k(\mathbf{d},\mathbf{w}) \frac{1}{\sqrt{2\pi}\sigma_k(\mathbf{d},\mathbf{w})} \exp\left\{ -\frac{||m_i - \mu_k(\mathbf{d},\mathbf{w})||^2}{2\sigma_k^2(\mathbf{d},\mathbf{w})} \right\} \tag{2.4}$$

In this case, the marginal PDF for model parameter dimension m_i is dependent on the input vector to the network $\mathbf{d}$, the two layers of network weights $\mathbf{w}$

and the other dimensions in the model space $m_{j\neq i}$, which are integrated out in the marginal distribution. Each Gaussian kernel, k, contributes to the PDF according to its weighting factor, $\alpha_k(\mathbf{d},\mathbf{w})$, which is dependent on the input vector and network weights. Each kernel has a mean $\mu_k(\mathbf{d},\mathbf{w})$ and standard deviation $\sigma_k(\mathbf{d},\mathbf{w})$. Each of μ_k, σ_k and α_k have their own output function node. There are $3k$ output nodes from the network. The functions in the output nodes are different depending on whether they are producing a mean, a standard deviation or a weighting factor.

The nodes for the kernel means have a simple linear function:

$$\mu_k = F_k^{\mu}(y_j w_{jk}^{\mu}) = y_j w_{jk}^{\mu} \tag{2.5}$$

All of the hidden nodes connect to all of the output nodes, regardless of the output node type. The connections between hidden nodes and output nodes which produce the mean are denoted w_{jk}^{μ}.

The output nodes for the standard deviation of the kernels, σ_k, use an exponential function to prevent the standard deviation from going to zero:

$$\sigma_k = F_k^{\sigma}(y_j w_{jk}^{\sigma}) = \exp(y_j w_{jk}^{\sigma}) \tag{2.6}$$

The weighting factor for each kernel ensures that, when added together, the final probability density function given in equation 2.4 integrates to 1. The weighting factor nodes, α_k, therefore use a softmax function:

$$\alpha_k = F_k^{\alpha}(y_j w_{jk}^{\alpha}) = \frac{\exp(y_j w_{jk}^{\alpha})}{\sum_{l=1}^{l=K} \exp(y_j w_{jl}^{\alpha})} \tag{2.7}$$

One of the strengths of a neural network is that each neuron can deal with a vector of inputs, rather than just a single value. The network then produces a vector output for every element of the vector at the same time. In the notation used in figure 2.4, that would mean that I could show the network n input patterns at once, in a matrix of size d_{in}, where each input node deals with n inputs. The network then produces z_{kn} outputs, one for each of the n input patterns. This feature is important for the generalised performance of the network, because it means that it can be trained to perform optimally on many different input patterns at once.

The number of nodes and weights in the network is a trade-off between the complexity and the ease of training the network. As general rule of thumb the number of training sets needed is approximately ten times the number of weights in the network (Baum and Haussler, 1989). The number of training sets

therefore limits the size of the network, but the optimal number of nodes is still unknown. Different networks may also perform differently with the same data due to the complexity of network training (e.g. Blum and Rivest, 1992). I attempt to overcome this optimisation problem by treating the network size as an extra variable during the inference process. For every problem, where I want to find posterior distributions for a particular parameter m_i given observations from some distribution $P(d_j)$, I train 100 neural networks. These 100 networks form a committee, each of which contributes to the final result. The size of each network in the committee is a random variable. The final result from the committee is a weighted average from all of the committee members, weighted by generalised performance on a test set. In this way, the most optimal network architectures dominate for each problem without being specified in advance. The range of sizes varies slightly between the applications I present in this thesis, but in general the networks have between 25 and 50 hidden nodes and between 3 and 8 Gaussian kernels. Using a committee also has a regularising effect and generally improves performance (Peronne and Cooper, 1993), as I will discuss later in this chapter.

Training the network

My networks are initialised with the weights taking random numbers. They do not initially find a mapping between the data and model spaces and must be tuned to do so through training. To train them, each network is shown many sample pairs, where both the observations and the target model parameters are known. The observations are given to the network as an input, which then produces an estimate for the model parameter values. The difference between the true model parameter value and the network estimate can then be found. The network weights are updated to reduce this error. At each training iteration, the mean error from all of the samples in the training set is used. This improves the general performance over a wide region of parameter space, because the mean error takes into account all of the prior space. By showing the network may sample pairs drawn from the prior distributions for the model parameters and data at once, the network will hopefully find an underlying relationship between data and model space which is independent of the sample location, allowing it to predict, with a low error, the model parameters for samples covering a wide region of parameter space. By finding this underlying function, the network is able to interpolate between samples and make inferences about the model parameters associated with new data space samples.

At the start of the training process, the weights in the network are initialised

randomly. The weights in the first layer of the network are initialised using random numbers within the range $-1/N$ to $1/N$, where N is the number of inputs to each layer. The weights in the second layer of the network are initialised so that the network initially outputs the prior distribution of model parameter values. This is necessary for me to be able to consider the results in a Bayesian sense, because the network then begins with the specified prior distribution. I do this using k-means clustering, with the same number of clusters as Gaussian kernels (McLachlan and Chang, 2004). The mean of each Gaussian kernels is then set to the mean of each cluster. If network training fails, it will continue to output the prior distribution regardless of the input values.

To train the network, I update the weights iteratively based on their contribution to the error function:

$$E = -\ln P(m_i|\mathbf{d}, \mathbf{w}, m_{j\neq i}) \tag{2.8}$$

which is the negative logarithm of the posterior probability density function for the known model parameter, given in equation 2.4 (Bishop, 1995). In this case, m_i is the known model space sample location associated with the the data space sample, $\mathbf{d}$, which was the input to the network. The error function is high when the network thinks that the likelihood of the target value being the model space counterpart of the observation is low. The likelihood can be low because the estimate is inaccurate or if the PDF is very broad with a large uncertainty. The network tries to update the weights to maximise both the accuracy and the certainty. The resulting PDF therefore includes a measure of the uncertainty of the inversion process.

I train the network using a suite or batch of training samples to calculate the error at each iteration. The number of training samples are given in table 2.1. At each training iteration, I show the network all of the samples and calculate the error function for each sample. The update for each weight is then calculated by finding its contribution to the integrated error of the entire batch. By training using a batch, it helps to maximise the networks performance over the entire range of the sampled data space. I use the Rprop+ algorithm to update the weights, which is a type of back-propagation gradient descent algorithm (Igel and Hüsken, 2000). At each training step, it records the error from the previous iteration. If the error increases, it undoes the weight update. If there is no fundamental relationship between the data and model spaces being considered, the update will only improve the likelihood of the inference for a subset of the samples and will worsen the inference for the rest. The error will therefore increase, and the weights will be reverted back to their original state, which

was set to the prior. This is also why the network returns the prior model parameter distribution if training fails. This can also happen if the relationship between model and data spaces is too complicated for the size of the network or if there are not enough samples in the data set. Small networks and widely spaced samples mean small scale variations in the relationship between the model and data spaces are missed.

I stop the training process after a set number of training iterations. For some applications, I also use a monitoring set of simulations to check how the generalised error function is evolving. These samples are drawn from the same distributions as used in the training set, but are not used to update the network. This independent monitoring is necessary because the networks are subject to overfitting. The overfitting occurs as the network fits the target model parameters to the inputs using an artificially over-complicated function. Because the function is over-complicated and does not describe the true function well, the generalised performance suffers, which can be identified using the monitoring set of simulations. As the overtraining progresses, the error for the monitoring set will increase, despite the training set error decreasing.

In any sort of function fitting problem, overfitting can be constrained using regularisation (Tarantola, 2005). The regularisation specifies *a priori* characteristics of the fitted function, such as its smoothness. There are many ways to regularise neural networks. Generally, smaller networks suffer less from overtraining (e.g. Baum and Haussler, 1989), and there are many ways to constrain the complexity of neural networks during training (Schmidhuber, 2015). The effects of overfitting can also be reduced by using a committee of networks. The random initialisation means that two different networks with identical architecture and training data overfit in different directions. Using multiple networks generally means that the overtraining cancels out, often giving a better approximation to the true function than any single network (Peronne and Cooper, 1993). The committee is also an approximation to an integration over all possible neural networks which could be used to attempt to solve the problems. By using a committee I can capitalise on the different results produced by networks of different sizes. There is no way to predict the optimal number of nodes for any network. I therefore vary the number of nodes within my committee of networks, increasing the chances of finding an optimal network structure.

The final result from the committee of networks is a weighted mixture of all the outputs from all of the networks. The weighting of the networks depends on their performance on a third, independent set of simulations, as described in

Age (Gy)	Training	Monitoring	Committee Assembly	Test	Total
0.4	800	250	250	250	1550
1	553	200	200	200	1153
2	408	150	150	150	858
3	334	130	130	130	724
4.5	457	150	-	150	757

Table 2.1: Number of convection simulations at each age. The monitoring set is used to monitor the error of the network and to stop training, the committee assembly set to weight each committee member and the test set is used to assess the performance of the committee of networks. All results presented in this thesis are for inferences made given observations from the test set. The networks trained at 4.5 Gyr use slightly different training approach and use the monitoring set to weight the committee. These networks are stopped after a set number of iterations rather than when they reach a stable error. The code to train the networks for the 4.5 Gyr cases was developed by A. P. Valentine to run in parallel. The networks trained for earlier time steps used a code written by S. Atkins run on a desktop computer. There was no difference between results.

Käufl et al. (2014). The best performing networks in the committee have most influence on the results. I use two different neural network codes, one developed by A. P. Valentine to run in parallel and the other written by me to run on a desktop computer. They give the same results when tested with the same data. Depending on the neural network program I use, the committee is assembled based on either the monitoring set of samples or a separate committee assembly set. Since the networks are stopped after a set number of iterations, the monitoring set can be reused, increasing the number of samples available for the training set.

2.7 Using neural networks with real data

Having hopefully trained my networks to make inferences about geodynamic processes using observations taken from convection simulations, the ultimate end-goal is to use them to make inferences about the values of parameters which are responsible for a real observation taken from the Earth. There are three significant hurdles to overcome before this can even be considered:

- Is it actually fundamentally possible to make an inference about a particular parameter using any type of observation? And if so, is it possible in all regions of data space? (Section 2.8)

- Do the forward simulations that were used to train the neural networks adequately capture the processes happening in the Earth such that the

inference is likely to be reasonable? (Section 2.9)

- Can I make reasonable inferences given the uncertainties that are associated with real observations? How sensitive are the inferences to these uncertainties? (Sections 2.10 and 2.11)

Using real data is currently a long way outside the scope of this thesis. I still consider these points, because otherwise studying this approach is an entirely academic exercise, and they are important even when only synthetic cases are considered.

2.8 Have the networks learnt anything?

After the network has been trained, I need to establish that it has actually learnt something. To do this, I use a separate test set of samples. All of the training and test sets of samples are produced by StagYY and are drawn from the same distributions. They therefore have the same mathematical relationship linking the data and model spaces as the training set. They have not been used to train the networks or assemble the committee and are therefore completely independent of the training process.

If the network failed to find a fundamental relationship between data and model spaces during training, it will not be able to predict the parameters associated with the test set. In this case, because of the way the networks are initialised and trained, it will return the prior distribution. The network could potentially update itself so that it manages to fit all of the training samples without finding a true relationship, although this is unlikely using batch training. If it has done this, it will be highly specialised with regard to those samples, but will be unable to generalise with a test set. False positives showing successful training are therefore highly unlikely, especially when using a large test set. If the network has instead found some real relationship in the data, it should be able to repeat this for a test set which has the same underlying physical relationship. The network is highly unlikely to find a relationship which finds the model parameters perfectly for all test sets (Baum and Haussler, 1989; Blum and Rivest, 1992), but a cut-off point for success for a given number of test sets can be established.

The test set of samples also tests the assumption that the function linking the data and model space varies smoothly between samples and that the networks are large enough to capture the complexities of this function. If the mapping function varies over much shorter length scales than the spacing between

my samples in either data or model parameter space, the inferences for the test set will be highly inaccurate, because they are made assuming smooth variations. The true location of the model space counterpart for each data space sample, for a rapidly varying posterior, will lie a long way from this smooth surface. Unfortunately, the results in the case that the posterior is not smooth are identical to the case where there is no information present in the observation about the parameter of interest. In the former case, the inference could be improved by using more samples and therefore decreasing interpolation distances and by increasing the network size. In the latter case, increasing the sample size will not help the inference at all.

For each investigation, I present the marginal posterior probability density functions for the test sets to show the quality of the network inferences. They are mostly plotted in the same way. For every test data sample presented to the network, a posterior PDF for the model parameter in produced. Figure 2.6 gives an example of how these are plotted. In this case it is for yield stress, given observations of the temperature structure of the simulations. The six boxes on the left show six posterior PDFs for six different simulations. Some of them are quite complicated with multiple peaks. The cyan line shows the true yield stress for each simulation. I shade the PDFs, so that the maximum amplitude is black. I can then stack multiple PDFs for many different simulations together in a grid such as the one on the right. Each PDF (each vertical stripe) is placed along the x-axis according to the true value for the simulation. Ideally, the maximum amplitude of the PDF should correspond to the true value. A well-performing network would produce a grid with a slash of black across the diagonal, indicating that the network predicts the parameter accurately for the majority of simulations.

As well as checking the accuracy of the predictions, I can assess how much information I have gained in comparison to the prior. For this, I use a measure called the Kullback-Leibler distance to measure the change in entropy in bits between the marginal posterior probability distribution for the input parameter and the prior distribution for that input parameter (Johnson and Sinanović, 2000):

$$D_{KL} = \int P(m_i) \log_2 \frac{P(m_i)}{P(m_i \mid \mathbf{d}, m_{j \neq i})} dm_i \tag{2.9}$$

where $P(m_i)$ is the prior distribution of the model parameter within the training set, and $P(m_i \mid \mathbf{d}, m_{j \neq i})$ is the network inference for the marginal posterior pdf for the same model parameter, given an observation $\mathbf{d}$. This also provides a measure of how certain the network is about the result.

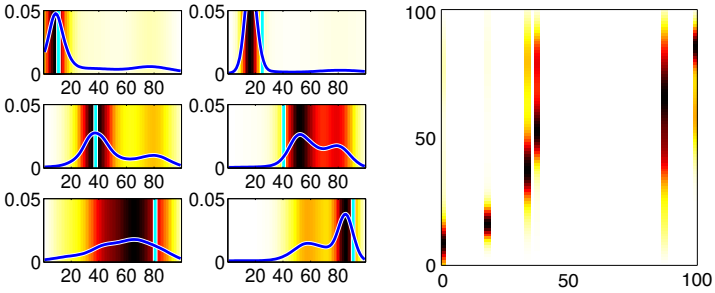


Figure 2.6: Some randomly selected examples for posterior PDFs inferring yield stress after 3 Gyr, taken from the test set of simulations. The six PDFs to the left are the committee output, coloured with the same colour scale as in the right-hand panel and figure 5.7. The colour scale is black at the maximum regardless of amplitude. The pale blue line indicates the true target value of yield stress for each simulation. Ideally, the maximum of the PDF should correspond to the target value. The target value is then used to align the coloured representations of these PDFs along the x-axis of a grid such as on the right-hand side. If the PDF maximum is close to the target values for all the PDFs in the test set, there will be a diagonal stripe of high amplitude across the grid. Figure previously published in Atkins et al. (2016).

It is important to remember that these results are probability density functions. They do not say that the most likely value is the true value, merely that it is most likely. All the other values are also possible, just less likely. For some samples, the maximum likely value is a long way from the known target. This does not mean that network training has failed, merely that the target value lies in an unlikely area of model space and, given the observation, a different value is more expected. Any inference for real data must to be presented as a probability rather than a statement of fact.

2.9 Modelling assumptions

In section 2.7, I listed three considerations which must be taken into account before inferences can be made from real data. The first was discussed in the previous section. The second and third are assumptions about the relationship between the data and model space. The first of these concerns the uncertainties built into the forward simulations and how well these simulations represent the Earth. The second concerns the uncertainties inherent in the observations themselves, as discussed in the next section. The way in which these uncertainties are accounted for are similar, therefore I discuss their treatment together in section 2.11.

Any inferences made by the networks are subject to the caveat that it is assumed that the Earth behaves in the same way as the forward simulations used to train the network. These assumptions take many forms. In my case, the most obvious one is that I train the networks on 2-D convection simulations. In order to use them to with real data, I would have to establish how different an observation from a 2-D simulation is from a comparable observation taken from a 3-D calculation. Other more subtle assumptions include the viscosity calculations used. I use a simple viscosity law, which does not take into account phase changes. Again, the effect of this assumption can be quantified by running comparison simulations to see how much the results change when varying these parameters.

The biggest problem when trying to quantify the uncertainties is when the physics of a process are not completely understood. For example, there are crystal scale effects on rheology through processes such as recrystalisation and the healing of defects. These could significantly alter our understanding of the rheology and therefore the dynamical processes taking place in the mantle. Because we do not know yet precisely how they will affect the geodynamics, it is impossible to quantify how neglecting these effects will affect the simulations.

2.10 Data uncertainties

The other limiting factor listed in section 2.7 is the uncertainty associated with the observation. When training the networks, I know the value of my data space sample perfectly. For real data, each observation is normally associated with an uncertainty. For example, I might want to train my networks using seismic tomographic images of the mantle. A tomographic image uses seismic data, produced by an earthquake that has been recoded by a seismometer. The seismometer also records a lot of noise from other sources, which must be removed before the record of the earthquake can be used, introducing one source of uncertainty into the data. The seismograms are then turned into a tomographic image. This is an inversion process, which also has inbuilt assumptions and uncertainties which must be taken into account before the resulting tomographic image is used. Sources of uncertainty in seismic tomographic models include the treatment of the earthquake itself, where and when it happened and how the fault moved because this determines how the wave propagates. There are then errors introduced by the computational limitations imposed when calculating how waves propagate through the mantle and through the multiple ways to parameterise and regularise the inverse problem that produces the tomo-

graphic image.

2.11 Treating data and modelling uncertainties

I can include all of the uncertainties discussed in sections 2.10 and 2.11 by training the networks to map from a probability function in the data space to a probability function in the model space, rather than mapping from discrete points in the data space. This is again achieved through sampling. The basic set of data space samples are created by running the forward simulation for a set of model space samples. This creates discrete data space samples. However, I can assume that each discrete sample is actually the most likely point of a probability distribution. The width and shape of this distribution is determined by expected modelling and observational uncertainties. By sampling this distribution repeatedly, I produce a new set of sample pairs. There are now many data space samples, drawn from the uncertainty distribution paired with the same model space sample. When trained using this set, the width of posterior PDF produced by the network includes the data space uncertainty distribution.

Sampling from a data space uncertainty distribution acts to regularise the network. The differences in the data space samples which are linked to the same model space sample blur the relationship between the inputs and targets because the same input-target pair is shown to the networks multiple times with slight offsets. This prevents the networks from finding a fit between samples which is too complicated. The regularisation also gives me an *a priori* method to assess whether a given observation could realistically be used to make inferences about the Earth, given the expected uncertainty of that observation.

2.12 Preprocessing my inputs

Whilst in theory a large enough network would allow me to train a network to find a mapping between a data space of any dimension and any model parameter (e.g. Hornik et al., 1989), in practice I need to preprocess my input observations. This is for two main reasons. Firstly, as discussed in sections 2.2 and 2.3, mantle convection is highly non-linear and I must therefore find some stable statistics to study. Secondly, neural networks are much easier to train the smaller they are. Reducing the dimensionality of the data increases their stability and performance. This is especially important in my case because I have

relatively few simulations with which to train the networks (Baum and Haussler, 1989; Schmidhuber, 2015).

Smaller neural networks with fewer input nodes are generally much easier to train than larger ones (Baum and Haussler, 1989; Blum and Rivest, 1992). I therefore reduce the dimensionality of my observations as much as possible. Exactly how I do this varies between applications and is covered on a case-by-case basis in each subsequent chapter.

My common approach is to use an auto-encoding neural network. This was developed by Valentine and Trampert (2012), based on the work of Hinton and Salakhutdinov (2006). The auto-encoding neural network is similar in architecture to the network shown in figure 2.5, but with more layers. A schematic representation of an auto-encoder is given in figure 2.7. It takes an observation, for example the amplitude spectra of the temperature field, and reduces the dimensionality to give an encoded representation. The input layer takes the original observation and feeds it to successively smaller layers. Each layer has half the number of nodes of the previous layer, reducing the input down, layer by layer, to a predetermined number of discrete values. A second, inverted network then ensures that these discrete values are meaningful. The second network does this by taking the discrete the values and expanding them, layer by layer, back up to the original number of dimensions. The two networks are trained together on many different input patterns so that the difference between the original pattern and the decoded pattern is minimised. The process is not compression in the conventional sense, which generally aims to reduce data size for storage purposes (e.g. `gzip` Ziv and Lempel (1977); Valentine and Trampert (2012)), because it is not loss-less. The expanded pattern generally loses details which were present in the original and tends to have a smoothing effect. An example is shown in figure 2.8. However, one of the strengths of this method is that it compresses the input pattern by identifying common features between all of the patterns in the training set and removing them. It therefore works as a feature extractor, automatically identifies the features that differ between patterns, which are what I am most interested in.

I make inferences from observation of convection simulations in two stages with two different neural networks: a feature extracting dimensionality reduction stage; followed by an inference stage with mixture density networks, which use the encoded observations. A deep learning neural network with many layers could theoretically perform the whole task in a single stage, with lower layers extracting features and higher layers interpreting them and making the inferences which are currently done by my mixture density network (LeCun et al.,

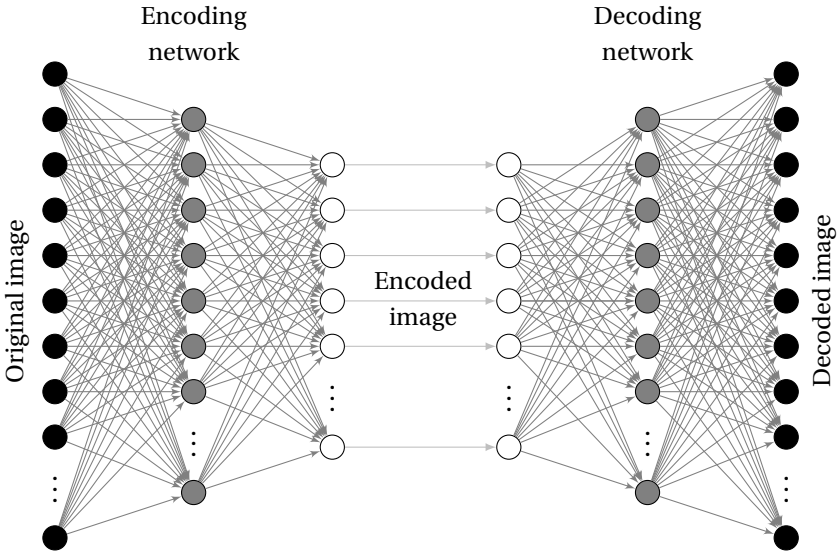


Figure 2.7: A simplified illustration of an auto-encoding neural network pair. There are two networks, the first of which reduces the dimensionality of the image down to a set number of discrete points (the encoded image). The second network then decodes these discrete points to produce an image of the same dimensionality as the original. The two networks are trained together so that the decoded image is as close as possible to the original. The similarity between the original and decoded image depends on the number of training images and the required degree of compression. I generally setup my encoding networks so that they have one layer for every factor of two dimensionality reduction. A network which reduces a 640 pixel image down to 20 dimensions will therefore have 6 hidden layers, including input and output. See Valentine and Trampert (2012) for more information on auto-encoding neural networks.

2015). However, networks with many layers are slow and difficult to train satisfactorily. Since the features extracted are broadly similar for every inference, I speed up the training process by using the auto-encoder separately from the inference-making networks, significantly reducing the required complexity and therefore the required training time. This may reduce the quality of the inference somewhat, as different features may be interesting for different inferences. However, the increased stability of the smaller networks makes this division of the process a very practical solution for a proof of concept investigation such as presented here.

For most patterns I use the spectral representation in the frequency domain, as discussed in section 2.3. I outline precisely how I undertake the transforma-

tion in section 3.1. I divide the spectra up into their degree 0 component and higher order components. I then use two different auto-encoders to encode them. I do this for practicality because by separating the degrees, I can easily use the encoded data for multiple applications (such as in chapter 7) without having to train a new auto-encoder each time. In general the result are not significantly different when using one auto-encoder or two. Figure 2.8 shows one example for the temperature spectra for a single simulation. The lower grids show the spectra after they have been encoded and decoded. The one on the left is the result using two auto-encoders, one for degree 0 and one for degrees 1–10. The one on the right used a single encoder for all degrees. There is very little difference between the results.

There is a trade-off between the degree of compression and the information loss, with greater compression leading to more information loss. I have relatively few simulations to train my networks with, requiring a high degree of compression, in order to keep the networks small and stable. Figure 2.9 shows the change in variance between the original and encoded spectra at different compression levels for the temperature observation. The variance is:

$$\text{var} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{x_i^{\text{orig}} - x_i^{\text{enc}}}{x_i^{\text{orig}}}\right)^2 \tag{2.10}$$

where x_i^{orig} is the original spectral representation and x_i^{enc} is the pattern after encoding. There are N pixels in both patterns. The black line is the average variance for all simulations, and the coloured lines show individual simulations to indicate how the loss varies for different patterns. The original pattern had 704 dimensions. Picking a compression point is a trade-off between the amount of detail retained in the patterns and the stability of the network. With more samples a larger network can be used, requiring a lower degree of compression. The auto-encoder training also improves with a larger number of samples. In general, I use a compression with 28 data points. Whilst this results in a significant amount of data loss, especially in the finer details of the spectra, the inferences made by the MDNs are not particularly worse than when a lower degree of compression is used. This is because of the trade-off between network size and compression accuracy. Most of the data loss is in the fine scale variations in the amplitude spectra but the relative dominance of each wavelength with depth is generally maintained, which seems to be the most important information.

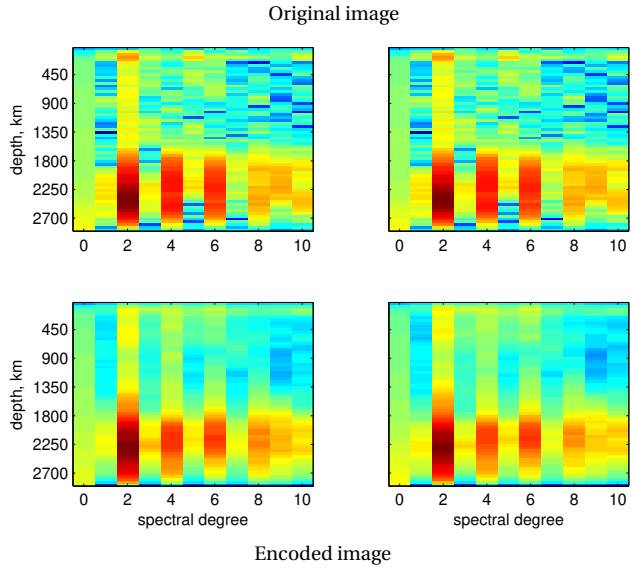


Figure 2.8: Comparing two different approaches to encoding the temperature spectra. The top row are the inputs to the auto-encoder, and are identical. The bottom row shows how the input pattern changes when it has been encoded then decoded. The spectra on the left have been split and encoded using two different encoders, one for degree 0 and one for degree 1–10. The one on the right used just one encoder for all degrees. All of the grids are plotted on the same colour scale. They have already been subject to some preprocessing, including the removing the mean of degree, the

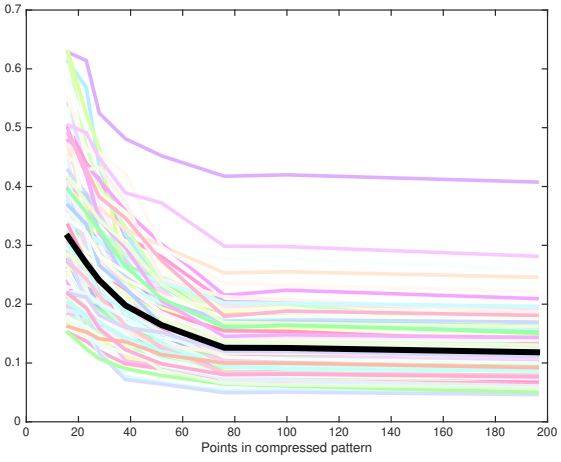


Figure 2.9: Variance between the original and encoded temperature spectra at different compression levels, as calculated using equation 2.10. The black line is the average for all the patterns in my training set of simulations, the coloured lines are for individual patterns to show the range of information loss when different patterns are encoded.

3

Convection Simulations

The long term goal of the research in this project is to be able to take an observation of the mantle and use them to make an inference about characteristics of the Earth, with an estimate of the associated uncertainties. In chapter 2, I outline a method by which this may be achievable, using a sampling based approach. I shall use forward simulations to calculate synthetic observations that can then be used to find an approximation to the inverse relationship between observations and mantle properties. The forward simulations I shall use are run with the mantle convection code StagYY (Hernlund and Tackley, 2008; Tackley, 2008). This code takes a set of simulation input parameters, solves various flow laws and gives me resulting observations such as the temperature and density structure of the mantle after millions of years of simulated convection.

In this chapter, I discuss some general features of the code; the observa-

tions it produces and how they can be preprocessed before auto-encoding; and which parameters I am studying, why I have chosen them and why the prior distributions for these parameters were chosen. There are a selection of simulations included at the end of this chapter for illustrative purposes.

3.1 The convection simulation code, StagYY

Governing equations for convection simulations

Over long time scales, the mantle behaves like a creeping fluid. As it flows, mass, momentum and energy must all be conserved. The continuity equations for these quantities set the rules for the mantle convection simulations. The first equation is for continuity of mass:

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho v_i}{\partial x_i} = 0 \tag{3.1}$$

where ρ is density, t time, v_i and x_i are particle velocity and location. I adopt the Einstein summation convention over repeated indices. Equation 3.1 is in the Eulerian form, which describes the movement of mass with respect to a fixed frame, as opposed to the Lagrangian form which considers a moving point. StagYY models the mantle as a compressible material, meaning that density changes with pressure and temperature (Tackley, 2008). To include changes in density, the code uses an anelastic approximation. This assumes that density changes with position, but only very slowly or not at all with time. The time-derivative of density in equation 3.1 is therefore assumed to be zero, leaving:

$$\frac{\partial \rho v_i}{\partial x_i} = 0. \tag{3.2}$$

The second conservation equation is for momentum, which describes the relationship between forces and deformation in a continuous medium. The Eulerian form is:

$$\frac{\partial \sigma_{ij}}{\partial x_j} + \rho g = \rho \left(\frac{\partial v_i}{\partial t} + v_j \frac{\partial v_i}{\partial x_j} \right) \tag{3.3}$$

where σ_{ij} is the stress tensor:

$$\sigma_{ij} = -P\delta_{ij} + \eta \left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} - \frac{2}{3}\delta_{ij} \frac{\partial v_k}{\partial x_k} \right) \tag{3.4}$$

where P is pressure and η is viscosity. Together, 3.4 and give:

$$-\frac{\partial P}{\partial x_i}+\frac{\partial}{\partial x_j}\left\{\eta\left(\frac{\partial v_i}{\partial x_j}+\frac{\partial v_j}{\partial x_i}-\frac{2}{3}\delta_{ij}\frac{\partial v_k}{\partial x_k}\right)\right\}+\rho g=\rho\left(\frac{\partial v_i}{\partial t}+v_j\frac{\partial v_i}{\partial x_j}\right) \tag{3.5}$$

In highly viscous flows, the inertial force on the right-hand side of equation 3.5 is very small with respect to the viscous resistance and the gravitational force and can therefore be set to zero, giving the Stoke's equation (Gerya, 2010). The buoyancy term, ρg, includes variations in density due to temperature and composition. My simulations are run using dimensional units for these calculations.

The third conservation equation is for energy, which includes heating through viscous dissipation and adiabatic heating because the material is compressible:

$$\rho C_P\left(\frac{\partial T}{\partial t}+v_i\frac{\partial T}{\partial x_i}\right)=\alpha T\left(\frac{\partial P}{\partial t}+v_i\frac{\partial P}{\partial x_i}\right)+\frac{\partial}{\partial x_i}\left(k\frac{\partial T}{\partial x_i}\right)+\rho H+\sigma_{ij}\frac{\partial v_i}{\partial x_j} \tag{3.6}$$

where C_P is heat capacity, T temperature, α thermal expansion, k thermal conductivity, and H is radiogenic heating per unit mass. The first term on the right-hand side describes the adiabatic heating, and the last term is the viscous dissipation function.

The final constraint is that whilst the composition of individual cells can change, the bulk composition of the simulation is constant. These equations and the underlying theory are described in more detail in the books by Gerya (2010) and Ismail-Zadeh and Tackley (2010).

Geometry, gridding and boundary conditions

At each time step StagYY solves the governing equations so that the mantle can flow under the stresses caused by the buoyancy forces and viscous resistance. The equations are discretised using the finite volume method. The mantle is divided into control volumes using a staggered grid. The pressure, temperature and other scalar properties are defined at the centre of the cell on such a grid, whilst velocities are defined at the centre of the cell faces. The equations are solved using a multi-grid solver.

My simulations are run in two dimensions, using a spherical annulus geometry. This geometry allows a 2-D simulation to approximate the patterns expected in a 3-D simulation because the ratio of the crustal surface to the core-mantle boundary surface is much greater in a sphere than in a cylinder with the same radius. The size of the core cannot simply be reduced to keep these

proportions constant in a cylindrical model setup as that can lead to structures at the lower boundary being forced closer together than they should be. Spherical annulus geometry includes a radius term in the governing equations that scales the volume integrals in the governing equations. The virtual thickness of the annulus therefore increases outwards from the centre, keeping the ratio of volume-dependent parameters, such as heat flux, the same as they would be in a spherical system without having to reduce the lateral spacing of cells by reducing the inner core size. For a detailed description of how this works and affects the governing equations, see Hernlund and Tackley (2008).

Both top and bottom boundaries are iso-thermal and free-slip, meaning that the surfaces have a fixed temperature and that no vertical motion is allowed. The temperature within the top layer of cells can vary.

The models are initiated with an adiabat determined by the initial core mantle boundary temperature and initial mantle potential temperature. The initial boundary layer thickness is 30 km. Random perturbations of 20 K are then added to cause density differences and instabilities. Once convection has begun, the signal of these perturbations will be lost, but changing the location of perturbations causes thermal and chemical heterogeneities to evolve in completely different places, as shown in figure 2.2 (Bello et al., 2014; Atkins et al., 2016). This initialisation ignores magma ocean stages in the evolution of Earth. Other initialisation methods, such as adding initial long wavelength temperature variations can also be used, and can change the time it takes for convection to begin. The differences between initialisation methods are yet another source of uncertainty in the modelling process, which could be investigated in future studies.

Composition and physical properties

StagYY allows the mantle to be made up of several end-member compositions. In my simulations, the majority of the mantle is composed of a mechanical mixture of basalt and harzburgite (a sort of peridotite) end-member rock types. The simulations are set up in this way to model the melting that forms mid-ocean ridge basalts. Mid-ocean ridge basalts (MORBs) are formed by partial decompression melting of a source rock, which then rise to the surface to form oceanic crust. The MORB composition is mainly controlled by the melt fraction, which is generally a function of temperature (e.g. Klein and Langmuir, 1987; McKenzie and Bickle, 1988). When the basaltic melt has been extracted, it leaves a residual peridotite (Johnson et al., 1990). This gives one way to estimate the composition of the original source rock, by combining the compositions of basalt

and peridotite, given an assumed fraction of melting based on mineralogy (e.g. Workman and Hart, 2005; Lyubetskaya and Korenga, 2007). Some of my simulations also have a third end-member that I describe as primordial material. Upon initialisation, this is a continuous layer at the base of the mantle, which is designed to model possible origins of deep mantle heterogeneities seen in seismic tomographic images (e.g. Garnero, 2000; Lekić et al., 2012). The origin and composition of this primordial end-member is discussed later in this chapter.

By setting up my simulations as an initial fertile mixture of basalt and harzburgite, I can model the melting processes that form basaltic crust. During convection, the basaltic end-member can melt if at any point the cell temperature exceeds a set melting temperature. The solidus function is based on the results of Herzberg et al. (2000) in the upper mantle and Zerr et al. (1998) in the lower mantle. The cell produces enough melt to either bring the cell temperature to the solidus, given the latent heat of melting, or until all of the basaltic end-member has melted, in which case the cell is entirely depleted. If the cell is shallow enough (<300 km), the melt is considered buoyant and is removed instantly to form basaltic crust. Below 300 km, the melt stays in the cell, where it is treated as an extra phase, affecting the viscosity and density. When the melt has been extracted, it leaves the cell more harzburgitic in composition.

The composition is tracked using Lagrangian tracer particles (Gerya, 2010). By using Lagrangian points that move through an Eulerian grid, sharp differences in properties, such as chemical contrasts, are maintained because they are not subject to numerical diffusion. These tracers are also used to track the amount of radiogenic material in each cell and the melting history of any basaltic end-member present in the cell. The time at which any basaltic melt was produced can be saved onto the tracers, which allows the tracking of the formation and subduction of basaltic crust through time.

The composition of each cell determines the density; thermal expansivity and enthalpy for the continuity equation for energy (equation 3.6); viscosity; fertility (the amount of melt the cell can produce); heating rate (the basalt and primordial end-members are enriched in heat producing elements compared to harzburgite, and these elements move with the end-member); and bulk and shear modulus which are needed to calculate the seismic velocity. The mineral physics properties (density, expansivity, enthalpy and bulk and shear moduli) are calculated using the Perple_X package (Connolly, 2009) that finds the phase assemblage with the minimum free-energy for any pressure, temperature and composition, given suitable equations of state. I use equations of state formulated by Stixrude and Lithgow-Bertelloni (2005) and Stixrude and

Lithgow-Bertelloni (2011) that use thermodynamic relationships to calculate various material properties in a fully self consistent manner, without needing extra sources of information. For each of my end-members, these properties are stored in look-up tables, one for each end-member, which are accessed by StagYY at each time step. The look-up tables contain all of the phase changes, and associated density and enthalpy changes that would occur given any particular composition. I therefore do not need to impose phase boundaries in my simulations. Whilst the properties for each end-member are calculated individually in a fully self-consistent manner, they are used in StagYY in a mechanical mixture. This means that the properties for each cell are not fully self-consistent. For density, viscosity and heating, the bulk cell property is the arithmetic mean of the end-members, according to how much of each lies in the cell. The seismic velocity uses a Voigt-Reuss-Hill average. There are two reasons for doing this, one scientific and one practical. Scientifically, it is not clear to what extent the mantle is in fully mixed chemical equilibrium. After millions of years of convection, lots of basalt has been subducted into the mantle. The short length scales of diffusion make it highly unlikely that subducted basaltic slabs ever fully remix (e.g. Hofmann and Hart, 1978). The coexistence of distinctly different lithologies in the mantle is supported by seismic observations of subducting slabs (e.g. van der Hilst et al., 1997) and small scale scatterers in the mantle (e.g. Hedlin et al., 1997). These lithologies can exist in full thermodynamic equilibrium with each other (Tirone et al., 2016), making it worthwhile to calculate their individual end-member equilibrium properties. From a practical point of view, using a mechanical mixture model makes most sense. It is straightforward to trace the movement of the three end members and inexpensive to run Perple_X for three different compositions. Using a fully equilibrated mantle (assuming this is scientifically reasonable) would require much more complexity. A stepped look-up table can be used (e.g. Zunino et al., 2011) which includes the equilibrium assemblage properties with a stepped mixture of end-members. However, for a three component system, stepped at 10% intervals, this would require 100 different look-up tables to be stored whilst running StagYY, the memory requirements for which are impractical. This also glosses over another assumption in the mechanical mixture method. In my simulations, it is assumed that source rock melting produces basalt of the same composition regardless of melt fraction, leaving a basalt-harzburgite remnant. Using an equilibrated mixture is therefore more thermodynamically sound, but not necessarily any more realistic. Tracking the movement of elements through the system and recalculating the equilibrium at each point would therefore be

preferable, but this adds in other computational limitations and requires assumptions about melting processes to be included in the code which in some cases are not yet fully understood.

The viscosity of the cell can depend on composition in several ways. Viscosity jumps can be imposed in a manner similar to a Clapeyron slope for a phase change, with different boundaries for different compositions. The cell viscosity is calculated based on the percentage of each composition end-member in the cell. In my simulations, I use the same viscosity for basaltic and harzburgitic end-members, with no viscosity jumps. I do however use compositionally dependent viscosity when considering the 'primordial' material. The primordial material has a viscosity contrast, which is a factor of the viscosity of the harzburgitic end-member at the same temperature and depth. In a cell with mixed composition, this increased viscosity contributes to the total cell viscosity proportional to the amount of primordial material present. In the future, it would be possible to locate viscosity jumps at the same place as phase changes as calculated in Perple_X, but this has not yet been implemented.

The internal heating of the cell also depends on composition. The end-members can be preferentially enriched in heat-producing elements, which are also tracked using the tracers. Upon melting, the partitioning of heat producing elements into the melt can also be varied, leaving the cell more or less depleted in heat producing elements.

Outputs

StagYY saves various fields at either set time intervals or after a set number of time steps. My simulations save the state of the mantle at least every 100 Myr. For each time interval, I have observations including temperature, density, composition, viscosity, the melt fraction in each cell, dynamic pressure, the flow velocity in each cell and the thickness of the basaltic crust. These can be post-processed to give seismic velocity, using the elastic moduli calculated using Perple_X. In this thesis, I mostly concentrate on temperature and density as first order observations of the state of the mantle system.

As discussed in section 2.3, the observations must be preprocessed to remove their sensitivity to small initial variations that result in the unpredictable spatial distributions of heterogeneities. Stable statistics include the mean temperate integrated over the whole mantle, the mean 1-D structure of the mantle where the properties are averaged azimuthally but not radially, or the spectral representation in the frequency domain.

Most of my applications use a spectral representation of the mantle struc-

ture. The process is illustrated in figure 3.1. To calculate this representation, I consider the azimuthal variation in the field at each depth slice. This means that the lateral variation can be considered a modulo 2π function because the radial variation is neglected. I then take the fast Fourier transformation of each slice. The phase information in the frequency domain is discarded by taking the amplitude of the frequency representation. This removes information about the relative spacing of the lateral variations, which are sensitive to the initial perturbations in the system, leaving only information about relative amplitude of variations at different wavelengths. I then generally discard the higher wavelengths information, keeping up to degree 20 (wavelength $\pi/10$) depending on the application. This is to reduce the dimensionality of the input, but also because this resolution is commonly used for reference model seismic models of the mantle (S20RTS by Ritsema et al. (1999)).

3.2 Parameters investigated

I chose to vary 12 StagYY input parameters, plus the composition of the end-member rock types. This gives me a 29 dimension model space. The StagYY input parameters include initialisation conditions, such as initial temperature, that are used for the setup at the start of the simulation, and constant parameters, such as reference viscosity, that are fixed throughout the simulation and appear in the flow laws at every time step. The parameters that I vary are given in table 3.1, with their ranges, which are their prior distributions. This particular set were selected to investigate a few unknowns in mantle convection studies. There are many more parameters I could investigate, but the number chosen were limited for practicality. The common parameters which do not vary between simulations are given in section 3.4.

I chose the ranges of the parameters according to previous studies to try to produce a wide range of convection styles. This means that many simulations are not Earth-like in any way. For some of the parameters, the ranges do not encompass those that are likely for Earth. I shall go through each varying parameter and discuss the reasoning behind the values chosen. For each prior distribution I use a uniform or logarithmic distribution, depending on the number of orders of magnitude that the range spans. I do this so that there is no initial bias towards any region of model space.

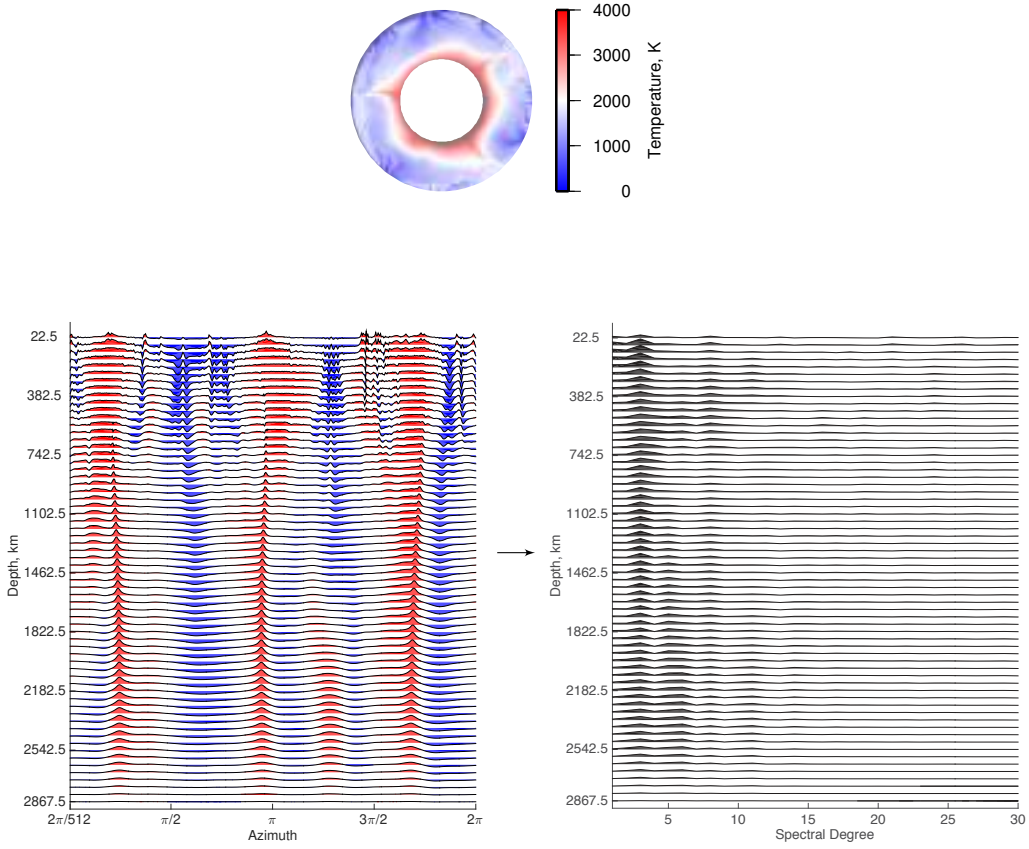


Figure 3.1: Creating a spectral representation for the temperature field of a simulation. The StagYY output is an annulus. Dividing the annulus radially gives 64 modulo 2π functions (left) that can be transformed into the frequency domain. I then get 64 amplitude spectra (right) that can be stacked up into a spectral representation of the temperature structure. This contains information about magnitude each wavelength (or spectral degree) of lateral variations in the temperature structure, but contains no information about the precise spatial variations in the temperature structure. I have removed the mean 2-D temperature from the both representations so that the lateral variations are clearer in this diagram.

Initial mantle potential temperature

Initial mantle potential temperature at the surface is the temperature that a package of mantle material would have were it to rise adiabatically through the mantle. It is one way to define mantle temperature, in the absence of anomalous temperature changes, caused, for example, by high radiogenic element composition or the subduction of a cold slab.

In my simulations, initial mantle temperature is used to define the adiabat at the initiation of the convection code. As convection begins, the signal of this parameter is expected to be lost as the simulation stabilises and the temperature becomes dependent on the efficiency of heat transfer by advection and diffusion through the mantle, which is dependent on other parameters (e.g. Sharpe and Peltier, 1978; Turcotte et al., 1979; Solomatov, 2001). By changing the initial mantle temperature, I can test if this is indeed true, or whether there remains a signal of initial mantle temperature even after the simulation has entered a stable mode of convection. If there is indeed a signal, many assumptions about the initialisation of mantle convection simulations would need to be revisited. The present day mantle potential temperature is in the range 1300–1400 °C beneath mid-ocean ridges (Lee et al., 2009, e.g.), therefore I set the lower bound at 1400 K. The maximum current mantle potential temperature is estimated to be around 1830 K in mantle plumes under hotspots such as Hawaii (Schubert et al., 2001), therefore I set the upper bound at 1900 K.

Mantle heating and distribution of radioactivity

The mantle is heated from the base as the core cools and from within by the decay of radioactive elements. Mantle heating is almost entirely from the decay of potassium, uranium and thorium. The relative amount of basal to internal heating determines the vigour of convection, with increased internal heating reducing the vigour of convection. High initial concentrations of radioactive elements also make the early stages much hotter and more vigorous, increasing the amount of melt and producing more crust early in the Earth's life.

The heat flux at the surface of the Earth is 47 ± 2 TW (Davies and Davies, 2010), but the relative proportion contributed by the core, crust and mantle internal heating are very uncertain. Depending on the geochemical and geophysical models used, the modern mantle may have heat producing capacity between about 3 and 25 TW (Huang et al., 2013; McDonough, 2016), with lower values assuming loss of lithophile elements due to collisional erosion early in the Earth's history (e.g. O'Neill and Palme, 2008) or non-chondritic

Earth models with low heating (e.g. Javoy et al., 2010; Campbell and O'Neill, 2012) and higher values coming from energy balance considerations in geodynamical modelling (e.g. Davies, 2010).

In my simulations, I average the half lives of potassium, uranium and thorium, based on their ratios of abundance to give a single radioactive half-life of 2.43×10^9 yr. The heating rate is set at the beginning of the simulation and decays with time. I start my simulations with 4.5–27.0 pW kg^{-1}, which is equivalent to between 18–108 TW for the whole mantle. After 4.5 Gyr, this decays to give a total mantle heat production of 5–30 TW, in line with the range of estimates for the mantle heating rate.

I vary the partition coefficient with which radioactive elements enter basaltic melt. This changes the rate of depletion of radioactive elements in the mantle and the heat production of the basaltic crust and any subsequently subducted material. Upon initialisation, the basalt component in each cell is preferentially enriched in heat-producing elements (HPEs) by a factor of 10, and basalt is uniformly distributed throughout the pyrolitic mantle. The HPEs preferentially enter the melt, with the partition coefficient varying between 10^{-5} and 1. The smaller the partition coefficient, the more preferentially the HPEs enter the melt. This range is used simply to see if any difference between partitioning coefficients can be detected, which is why I allow it to vary by five orders of magnitude.

80% of my simulations have an initial basal layer of material with composition distinct to that of the rest of the mantle. I refer to this as primordial material. The origins of this material are discussed later in this section. The heating rate of this material changes the Rayleigh number because it acts as a basal heat source, complementing that from the core. If I can find evidence suggesting that a chemically distinct layer at the base of the mantle has a low or high heating rate, it puts constraints on its likely origin. It would help to determine whether high ^{3}He/^{4}He ratios seen in ocean island basalts are due to the source region being undegassed and primitive (lots of ^{3}He) or having a low radioactivity (low ^{4}He) (Coltice and Ricard, 1999). The stability of such a layer is a trade-off between chemical and thermal buoyancy (e.g. Deschamps et al., 2011), so by varying the heat production I can investigate this trade-off. The minimum heating rate is the same as the pyrolitic mantle. The maximum is 500 times greater, with samples distributed logarithmically between these limits. These limits were chosen to cover a huge range simply to see what happens in the simulations.

Core heating rate and initial core-mantle boundary temperature

The rate and evolution of heat flow across the core mantle boundary (CMB) determines the existence and history of the dynamo in the core. The heat flux across the core-mantle boundary is dependent on mantle convection (Nakagawa and Tackley, 2010) and rate of cooling of the core, which is controlled by the core's conductivity and heat production (Labrosse, 2015). The heat flux also determines how long a basal magma ocean could exist in the Earth, which in turn delays the onset of dynamo activity (Labrosse et al., 2007). However, the presence of heat producing elements in the core, and therefore the initial temperature of the core, is debated on both thermodynamic and geochemical grounds (e.g. McDonough, 2016). High-pressure, high-temperature experiments suggest a concentration of 60–130 ppm of potassium into the core (Murthy et al., 2003). Increased concentration of radiogenic elements requires a lower initial CMB temperature to maintain a dynamo because the core cools more slowly. However, the required temperature for the dynamo has been found to be more strongly dependent on the conductivity of the core than on the radioactivity. Depending on conductivity and radioactivity, the core-mantle boundary temperature may have started at between around 4400 and 7150 K in order to cool to approximate present day conditions (Labrosse, 2015). However, Nakagawa and Tackley (2010) found that while the initial core temperature and heating rate determined the growth rate of the inner core, it made little difference to the evolution of the mantle. By changing the heating rate of core cooling, I can investigate whether the mantle contains clues about the core's thermal history.

In StagYY, I set the initial concentration of ^{40}K in the core to be between 0 and 800 ppm. The initial CMB temperature is between 3000 and 7000 K, which are in the range required for a dynamo, dependent on heating rate (Labrosse, 2015). Depending on the mantle solidus used, these temperatures may result in lower mantle melting (Fiquet et al., 2010; Andrault et al., 2011; Nomura et al., 2014). My simulations use a CMB solidus of 4350 K from Zerr et al. (1998), but only the basalt portion of the cells is allowed to melt. The initial CMB temperature and heating rate are selected independently, therefore there is the possibility of ending up with a very high core mantle boundary heat flow. The conductivity of the core in my simulations is 46 W m^{-1} K^{-1} after Stacey and Anderson (2001), which is somewhat lower than results from more recent studies (e.g. Gomi et al., 2013; Seagle et al., 2013; Pozzo et al., 2014). This low value of thermal conductivity means the core cools more slowly, so the initial temperature needs to be less high and a lower concentration of potassium is needed

to maintain the CMB heat flow necessary for a dynamo (Labrosse, 2015). The core cools according to the model of Buffett et al. (1992).

Viscosity

Viscosity determines the vigour with which the mantle convects and therefore the rate at which it loses heat. The original estimate of 10^{21} Pa s by Haskell (1935) is still considered valid as an approximate average mantle value, although newer studies, (e.g. Whitehouse et al., 2012; Argus et al., 2014) include much more complex lateral and radial variations. There is expected to be a viscosity jump of at least an order of magnitude around the transition zone, although the size and location of the jump varies between studies. The depth is often co-located with seismic discontinuities such as the 660 km on the basis that these are possibly caused by phase changes, although recent studies suggest the jump may be deeper than the expected phase change (e.g. Rudolph et al., 2015). These studies use a mixture of observations from glacio-isostatic adjustment, estimates of temperature taken from seismic tomography, and assumed scaling relationships between viscosity and temperature, all of which come with large uncertainties and are not necessarily self-consistent. Previous studies have attempted to invert for viscosity at the same time as mantle flow by using surface observations (Liu and Gurnis, 2008), making the problem more self-consistent, but this is severely time constrained by the availability of plate motion data (Conrad and Gurnis, 2003; Bello et al., 2014). If just the shape of seismic heterogeneities in the mantle can provide information on the viscosity, it gives me a new and independent method to constrain the viscosity structure which is complementary to existing methods with better constrained uncertainties.

The viscosity in my calculations is continuous with depth and calculated according to:

$$\eta(T,p) = \eta_0 \exp\left(\frac{E_\eta + pV_\eta}{RT} - \frac{E_\eta}{RT_{\eta_0}}\right) \tag{3.7}$$

where η_0 is the surface reference viscosity for zero pressure and reference temperature T_{η_0}, which is 1600 K. The viscosity reference temperature is different from the initial mantle potential temperature and is constant between simulations so that the effects of changing reference viscosity can be compared more easily. The reference viscosity varies between 10^{18} to 10^{21} Pa s. The activation energy is E_η (= 162 kJ mol^{-3}, a lower than usual value, thus reducing both viscosity and the temperature sensitivity of viscosity), R is the ideal gas constant,

and V_η is the activation volume, which decays with pressure as:

$$V_\eta(p) = V_{\eta_0} \exp\left(-\frac{p}{p_{\text{decay}}}\right) \tag{3.8}$$

I vary V_{η_0} between simulations, with range 1×10^{-6} to 3×10^{-6} m^3 mol^{-1}. The viscosity decay constant, p_{decay}, is 1610 GPa. Generally, different values of p_{decay} are used for upper and lower mantle (e.g. Lourenço et al., 2016), but I keep them the same for simplicity. For the upper mantle, most studies have no pressure dependence for the viscosity activation volume, so that $V_\eta = V_{\eta_0}$. A value of around 200 GPa for p_{decay} is often used for the lower mantle. Whilst my simulations use a higher than normal value for p_{decay}, they all use this value, so I can still make comparisons between simulations to see the effects of changing V_{η_0}. Increasing the pressure dependent viscosity by increasing V_{η_0} could promote mobile lid development in simulations, because higher viscosity increases the convective stresses exerted by the mantle (Stein et al., 2013). The low value for E_η may be responsible for some computational instability whilst running the simulations as convection becomes to vigorous, but because all of the simulations share this value, I can still compare them, even though they are not Earth-like.

I use a smoothly increasing viscosity profile so that I do not have to chose a depth for a viscosity jump. I do not impose a viscosity jump because I do not want to impose one that was inconsistent with the Perple_X calculated phase change locations. It is uncertain if a phase change is responsible for the inferred viscosity jump at the transition zone, especially with recent evidence of another viscosity jump at 1000 km, for which no major phase transition is known (Rudolph et al., 2015). However, it is also entirely possible that the ringwoodite phase boundary does change the rheology, causing a viscosity jump. In the future, it would be possible to use Perple_X to find the phase change and impose a viscosity jump at this location. In my setup, the viscosity does not depend on the relative proportions of harzburgite or basalt (although this would also be possible), but it is affected by the presence of primordial material in a cell. The primordial material has an associated viscosity contrast relative to the viscosity of the pyrolitic material at the same pressure-temperature conditions.

Yield stress

The yield stress parameter determines how much stress the material can withstand before it begins to undergo plastic or brittle deformation. If the lithosphere is weak enough relative to convective stresses it will yield, forming a

mobile-lid simulation regime. This is analogous to plate tectonics in simulations that can only produce basaltic oceanic-style crust, which can undergo subduction. The yield stress has been observed in many previous studies to be the major factor in determining whether a planet is stagnant or evolves a mobile lid (e.g. Moresi and Solomatov, 1998; Valencia et al., 2007; van Heck and Tackley, 2011; Lenardic and Crowley, 2012), and when continents are present, the strength is a factor in determining the wave length of convective flow (e.g. Zhong et al., 2007; Rolf et al., 2014). The yield stress required for a mobile-lid in convection simulations is much lower than those expected from rock deformation experiments. This may be because subduction initiates in highly fractured rocks with high fluid pressure, giving a much lower local yield stress within a high yield stress crust (e.g. Dymkova and Gerya, 2013). The simulations cannot model the small scale local variations, so use a lower average value. The lithosphere has the same strength throughout my simulations and it is not taken into account that yield stress is probably a time dependent, as well as spatially dependent parameter. For example, on Earth, the evolution of the atmosphere, and therefore the addition of water to the crust may reduce yield stress (Valencia et al., 2007).

? found that, when simulations were allowed to produce crust by melting, yield stresses below around 50 MPa produced continuously mobile-lid regimes. Above this until 120–180 (depending on viscosity) the simulations went through episodic periods of stagnation and mobile lid regimes. They found that the critical yield stress could be predicted based on melt eruption rate and internal temperature gradients. This study only varied a few parameters, so with more simulations with more variables, I can investigate if yield stress is still a dominant driver for the style of mantle convection. I use yield stresses between 0 and 100 MPa. This means that most of my simulations should spend at least some of their time in a mobile-lid regime.

Pyrolite composition

In my simulations, the bulk of the mantle is initially of a pyrolite composition, modelled using a mechanical mixture of basalt and harzburgite, as discussed earlier in this chapter. I assume that the entire mantle is homogeneous at the start of the simulation on the basis that the mantle was likely to have been relatively well mixed after the solidification of a low-viscosity magma ocean. The ranges for the major element oxides used to calculate the mineral physical properties are based on, but wider than those explored by Nakagawa et al. (2010), which in turn are based on those of Xu et al. (2008). I add an extra vari-

able: the initial fraction of basaltic material in pyrolite can be between 0.2 and 0.3 of the total. The compositional ranges for harzburgite and basalt are given in table 3.2. The maximum ranges of the composition of the combined mechanical mixture of harzburgite and basalt are given in table 3.3.

When the cells reach the solidus, a basaltic melt can be formed, then, if the cell is shallow enough, this will be removed to form crust. The changing fraction of basalt in each cell therefore determines the density and physical properties of the cell, the amount of melt a cell can produce, and its heating rate because heat producing elements partition preferentially into basalt upon melting. The initial basalt fraction will therefore also determine the crustal thickness, because it determines the fertility of the upper mantle.

Primordial material

At the base of the lower mantle, seismologists image anomalous regions with low shear velocities. These have long been observed to correspond to geoid anomalies (Le Pichon and Huchon, 1984) and the locations of large igneous provinces (e.g. Austermann et al., 2014). The interpretation of what these anomalies are is debated. They are possibly just thermal anomalies (e.g. Schuberth et al., 2009), but seismic observation (e.g. Trampert et al., 2004), combined with their probable stability through time suggest that they are chemical in origin (e.g. Deschamps et al., 2011; Torsvik et al., 2014). If they are a chemically distinct region, they are often invoked as the source of ocean island basalts (OIB), which have much more varied isotopic compositions than mid-ocean ridge basalts. From helium isotopes, some geochemists argue that the source regions for the OIBs may have been isolated since very early in the Earth's history (e.g. Allègre et al., 1983; Hofmann, 1997). Many models have been proposed for how an isolated reservoir could form and persist.

In my simulations, I include the possibility of a primordial chemical heterogeneous layer by starting 80% of the simulations with a third compositional end-member which forms a distinct layer at the CMB. I can then see if this end-member disperses and how different the simulations with and without primordial material end up. The composition of this primordial end-member is again set as a function of the major element oxides and its properties calculated using Perple_X. The compositions are chosen to model three hypotheses for the origin of a primordial layer, and are given in table 3.4.

The first model for the primordial material is that it is the result of subduction of an iron-rich early crust, after Tolstikhin and Hofmann (2005). This model suggests that after the giant moon-forming impact, the Earth formed a

crust. This collected smaller pieces of solar-system debris which were not large enough to cause significant damage to this early crust. Some of this debris had chondritic composition and was therefore enriched in iron. When subduction initiated on Earth, this iron-rich crust was significantly denser and subducted to the bottom of the mantle, where it remains. 10% of my simulations follow this hypothesis, with the primordial material having a basaltic composition with an extra chondritic component.

A second group of theories suggest that the anomalous layer at the base of the mantle is the result of a crystallising magma ocean. Peridotite melt is enriched in iron and is denser than solid peridotite under some temperature and pressure conditions, making in negatively buoyant. It would stop rising as it becomes neutrally buoyant and would crystallise, upon which it becomes denser than the surrounding rocks and sinks, forming an iron-rich layer at the base of the mantle (Lee et al., 2010). An iron rich layer can also be formed by fractional crystallisation of a basal magma ocean (Labrosse et al., 2007). 35% of my simulations assume this origin for the primordial material, with pyrolitic oxide composition plus varying amounts of extra FeO and SiO_2

The remaining 35% of models have a primordial component with basaltic composition to model segregated subducted crust, since basalt will produce a high density layer at such pressures and when partially melted and recycled may help to explain the composition of some ocean island basalts (e.g. Coltice and Ricard, 1999). These have compositions drawn from the same range as in table 3.2, but the composition may not be the same as that of the true basalt component in the same simulation.

The thickness of the primordial layer is also allowed to vary. Nakagawa et al. (2010) observed that this is the only way in which composition significantly influences core-mantle boundary heat flow, because a very thick dense layer blankets the core preventing heat from escaping. Some, but not too much, primordial material may help to preserve the geodynamo action by regulating the core-mantle boundary heat flux (Nakagawa and Tackley, 2014).

In previous studies, the viscosity contrast between a chemically heterogeneous basal layer and the overlying mantle has been seen to determine the shapes into which the basal layer was pushed during convection (e.g. Becker et al., 1999; Davaille, 1999; McNamara and Zhong, 2004; Deschamps et al., 2011). I therefore vary the viscosity of the primordial material relative to the bulk of the mantle to investigate if I can see any effect when all the other parameters vary at the same time.

Parameter	
Initial mantle potential temperature at surface	1400 – 1900 K
Initial mantle heating	4.5 – 27.0 pW kg^{-1}
Basalt heating with HPE partition coefficient	factor 10^{-5} – 1
Primordial heating by HPE enrichment	factor 1 – 500
Initial CMB temperature	3000 – 7000 K
Core heating by initial potassium concentration	0 – 800 ppm
Surface reference viscosity (η_0 in eq. 3.7)	10^{18} – 10^{21} Pa s
Primordial viscosity contrast	factor 10^{-2} – 10^{2}
Viscosity activation volume (V_eta in eq. 3.7)	10^{-6} – 3×10^{-6} m^3 mol^{-1}
Yield stress	1 – 100 MPa
Basalt fraction	0.2 – 0.3
Initial primordial layer thickness	0 – 800 km

Table 3.1: Input parameter ranges to StagYY. All input parameters are drawn independently from uniform distributions between these ranges.

Basalt	molar %
Al_2O_3	9 – 10.5
CaO	11 – 15
FeO	6 – 8.5
MgO	14.5 – 18.5
Na_2O	0 – 2.5
SiO_2	45 – 59.5
Harzburgite	
Al_2O_3	0.2 – 0.8
CaO	0.05 – 1
FeO	4.5 – 6.5
MgO	53.7 – 61.25
SiO_2	34 – 38

Table 3.2: Basalt and harzburgite major element composition ranges used to calculate properties in Perple_X (Connolly, 2009). For basalt, NCFMA are drawn randomly, with the remainder being SiO_2. For harzburgite, CFAS are drawn randomly, brought to 100% by MgO.

3.3 Some example convection simulations

To illustrate the variety in my convection simulations, I provide a selection in figures 3.2 to 3.12. They can all be considered samples drawn from the prior distribution of observations with which I train my networks. All of these simu-

Oxide	molar %
Al_2O_3	1.96 – 3.71
CaO	2.24 – 5.2
FeO	4.8 – 7.1
MgO	41.94 – 52.7
Na_2O	0 – 0.75
SiO_2	36.2 – 44.45

Table 3.3: Maximum possible ranges for bulk pyrolitic mantle composition when basalt and harzburgite end-members have been mixed mechanically. This is not an equilibrium assemblage, but a mechanical mixture of two equilibrium assemblages.

Basalt + Chondritic	(e.g. Tolstikhin and Hofmann, 2005)
10% of models	
	molar %
Al_2O_3	8.16
CaO	10.59
FeO	11.28
MgO	20
Na_2O	1.5
SiO_2	48.47
Pyrolite + FeO + SiO_2	(e.g. Lee et al., 2010)
35% of models	
Al_2O_3	1.26 – 2.59
CaO	1.84 – 3.79
FeO	5.8 – 20
MgO	27.45 – 56.51
Na_2O	0.15 – 0.32
SiO_2	30.99 – 49.30
Basalt	as in table 3.2
35% of models	
No primordial	
20% of models	

Table 3.4: Primordial material major element composition ranges used to calculate properties in Perple_X (Connolly, 2009). The model for primordial material is selected randomly, then the composition is drawn from the ranges given.

lations are 4.5 Gyr old. The input parameters for each simulation are given in table 3.5. For each simulation, I show the temperature field in Kelvin; the lateral density anomaly for each simulation, after the 2-D profile is subtracted; the viscosity of each simulation, which is capped in the range 10^{18} to 10^{25} by StagYY; the fraction of basalt in each cell and the fraction of primordial material in each cell. The primordial material replaces pyrolite, therefore in the basalt plot, the white (low basalt) regions may be harzburgitic or primordial. In table 3.5, I also

give the mean surface plate velocity of the simulations after 4.5 Gyr. This gives an indication of how tectonically active they are. I generally consider any simulation with mean surface velocity > 1 cm/yr to be tectonically active.

The simulation numbers are not consecutive because I take the first twelve which completed 4.5 Gyr of evolution out of all the simulations started. The simulation numbers not included here crashed before reaching 4.5 Gyr. The high initial core temperatures used in some simulations lead to excessively high lower mantle temperatures and therefore lower viscosities than the code could handle. Simulations with low viscosity activation volume also tended to fail. Other simulations ran, but with very small time-step intervals and had still not reached 4.5 Gyr after 18 months of calculations. The simulations took so long because I chose to run them on one processor each. This was because the code does not scale linearly, so running the simulations on one processor meant I got more partially run cases faster than running fewer cases to completion in parallel. These crashing simulations mean that the distribution of input parameters for the observations produced by the simulations is different from the distribution with which the simulations started. The neural networks are initialised with the prior distribution for parameters associated with the observations, therefore this is not a problem when making inferences.

Two thirds of the simulations have mobile lids, although only a few have clear evidence of subduction. Simulation 7 (figure 3.2) has basalt slabs going down to the transition zone which then get swept up into a marble cake structure in the lower mantle. There are also small drips coming down from the crust in simulations 16 and 61, and intact detached slabs in the lower mantle. Simulations 19, 42 and 70 also have a marble cake structure. In simulation 42, the basalt seems to be concentrated in three places above the primordial layer, which are low-viscosity, high-density, high-temperature regions and are the base of the upwellings. The basalt seems to be being swept into these regions by the upwellings and possibly entrained into the upper mantle. This simulation has very low yield stress, which is probably why there is so much subducted basalt mixed into the mantle. Simulation 69 has the clearest subduction, with attached slabs penetrating the lower mantle. This simulation has a strong degree-one structure with subduction happening on one side. This has pushed the originally thick layer of primordial material into a heap on the opposite side of the mantle.

Simulation 9 (figure 3.3) has a lot of basalt in the lower mantle suggesting an overturning event has occurred. This simulation has a very rapidly moving surface. Simulations 45, 50 and 67 show a similar basaltic layer at the base of

the mantle, but the simulations are stagnant with almost no movement. Simulation 45 shows an interesting degree-one structure, but the cause of this structure is not clear from the snapshot. The basaltic layer in 67 has ended up underneath the primordial layer.

Only simulations 19, 42 and 69 show any evidence of the primordial material being entrained into the originally pyrolitic mantle. In the other simulations, the primordial material gets swept around, but remains in intact piles or layers.

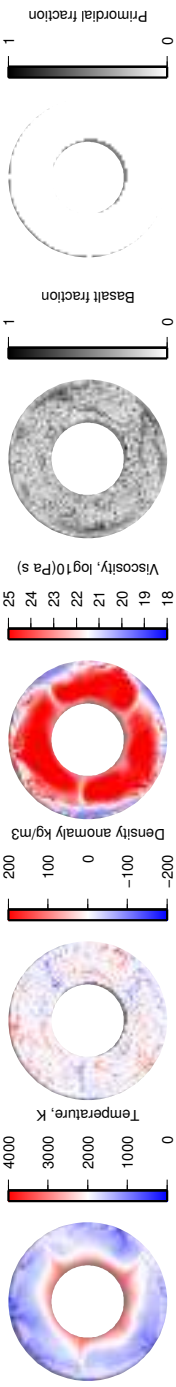


Figure 3.2: Simulation 7 after 4.5 Gyr.

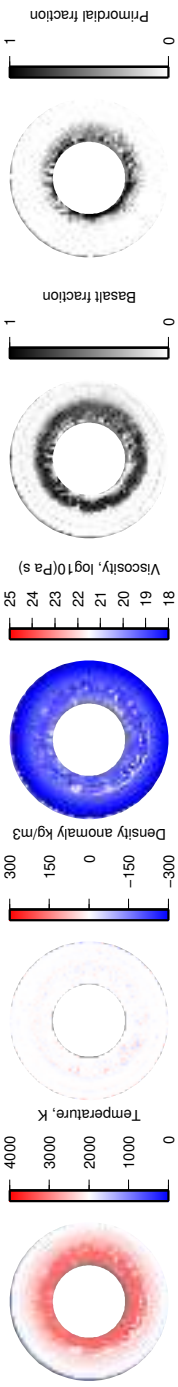


Figure 3.3: Simulation 9 after 4.5 Gyr.

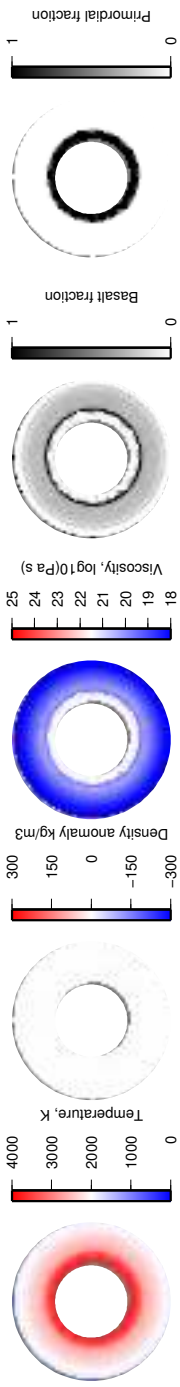


Figure 3.4: Simulation 14 after 4.5 Gyr.

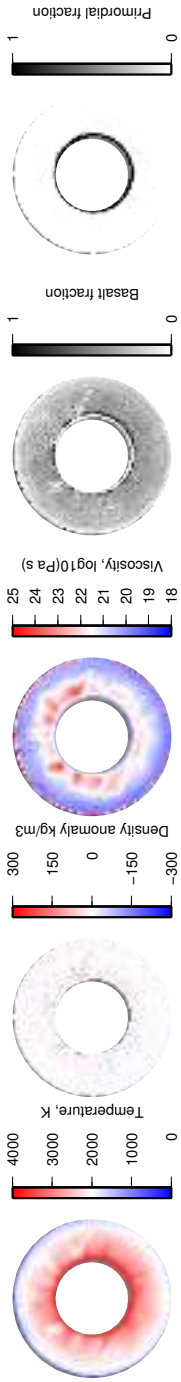


Figure 3.5: Simulation 16 after 4.5 Gyr.

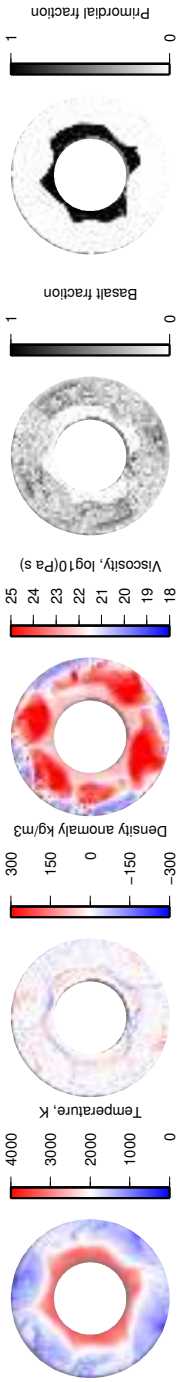


Figure 3.6: Simulation 19 after 4.5 Gyr.

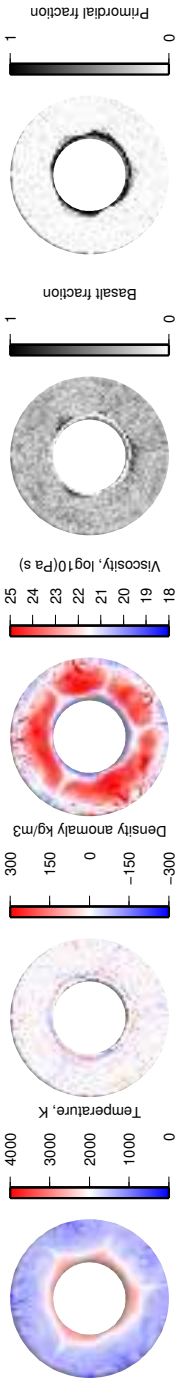


Figure 3.7: Simulation 42 after 4.5 Gyr.

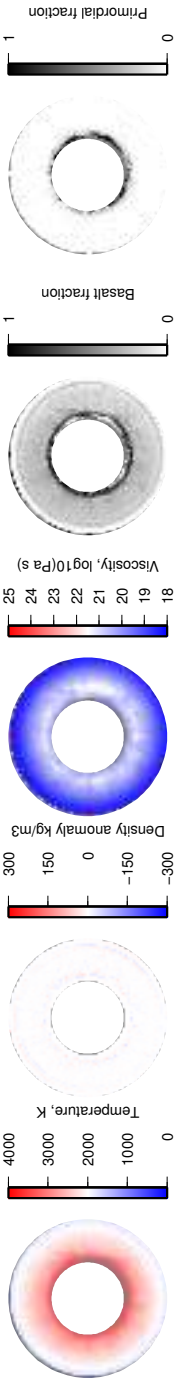


Figure 3.8: Simulation 45 after 4.5 Gyr.

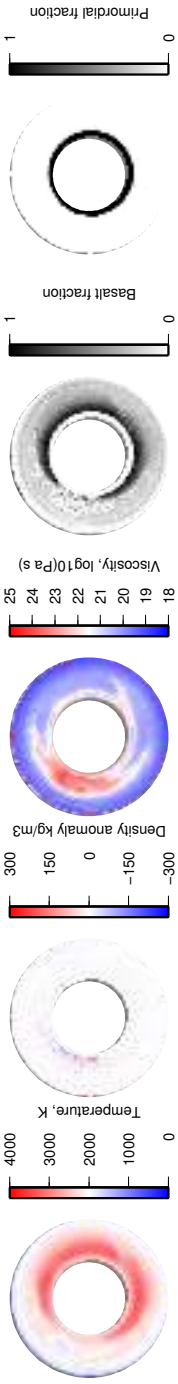


Figure 3.9: Simulation 50 after 4.5 Gyr.

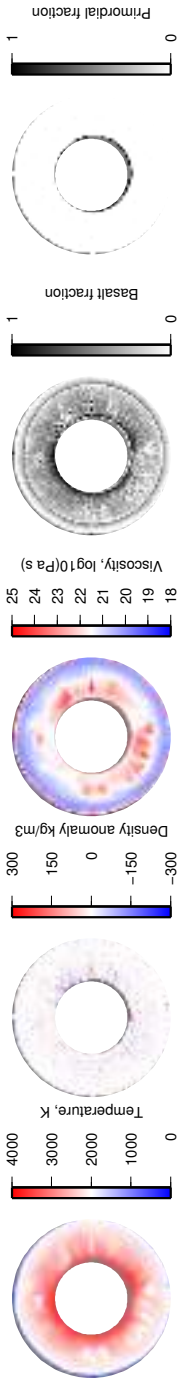


Figure 3.10: Simulation 61 after 4.5 Gyr.

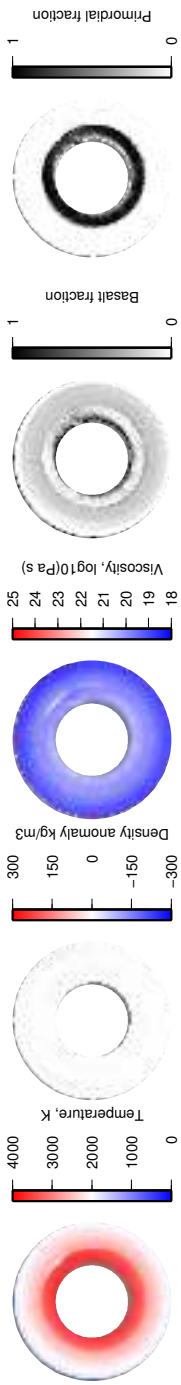


Figure 3.11: Simulation 67 after 4.5 Gyr.

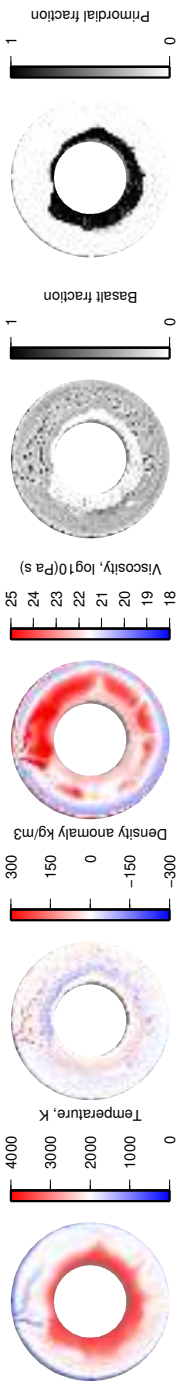


Figure 3.12: Simulation 69 after 4.5 Gyr.

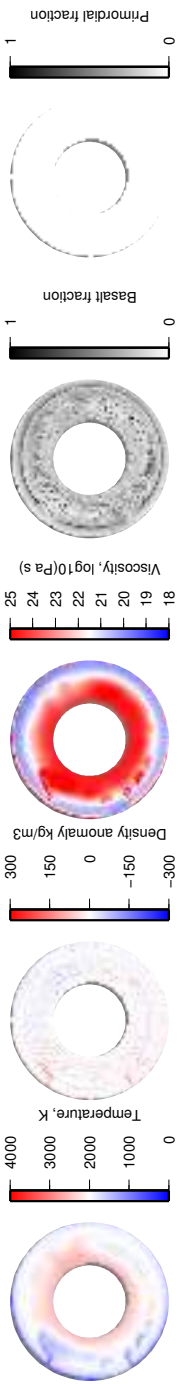


Figure 3.13: Simulation 70 after 4.5 Gyr.

Simulation	StagYY input parameters																												Surface velocity after 4.5 Gyr
	Initial mantle potential temperature	Initial mantle heating	Basalt HPE partition coef.	Primordial HPE enhancement	Initial CMB temperature	Initial core ^{40}K	Reference viscosity	Primordial viscosity contrast	Viscosity activation volume	Yield stress	Pyrolite basalt fraction	Initial primordial thickness	Basalt molar %						Harzburgite molar %						Primordial molar %						
	K	W kg^{-1}	factor	factor	K	ppm	Pa s	factor	m^3 mol^{-1}	M Pa		km	Al_2O_3	CaO	FeO	MgO	Na_2O	SiO_2	Al_2O_3	CaO	FeO	MgO	SiO_2	Al_2O_3	CaO	FeO	MgO	Na_2O	SiO_2	cm yr^{-1}
7	1611	1.08e-11	1.53e-05	5.45e+01	3634	650	4.86e+19	6.46e+01	2.97e-06	9	0.29	0	9.06	12.43	7.21	18.04	2.06	51.19	0.32	0.26	4.50	37.61	57.31	-	-	-	-	-	-	51.70
9	1702	2.60e-11	4.20e-04	4.98e+00	3209	216	3.06e+18	1.31e-01	2.24e-06	62	0.28	397	9.79	14.60	6.20	15.61	0.03	53.77	0.71	0.97	6.41	35.86	56.06	9.23	13.88	7.10	15.48	1.62	52.69	353.27
14	1404	1.54e-11	2.79e-03	3.49e+00	4124	271	1.95e+18	9.94e+01	2.60e-06	81	0.26	547	10.23	13.13	6.65	18.26	1.03	50.69	0.51	0.09	5.14	35.89	58.37	9.07	14.78	7.69	18.12	0.53	49.81	0.09
16	1550	7.69e-12	1.08e-02	6.97e+01	3261	611	1.23e+20	7.70e-02	2.42e-06	73	0.30	165	9.65	13.08	8.14	15.95	1.67	51.52	0.80	0.76	5.19	35.16	58.09	1.68	2.46	19.99	36.63	0.20	39.03	34.58
19	1716	2.23e-11	8.79e-04	9.31e+00	3647	176	5.31e+19	1.16e+00	2.75e-06	10	0.23	645	10.23	14.11	6.32	18.19	0.24	50.90	0.66	0.29	5.70	35.49	57.86	9.03	14.25	6.55	17.30	2.24	50.63	61.58
42	1693	1.31e-11	2.78e-01	3.24e+01	3606	268	4.25e+19	1.65e-02	1.93e-06	7	0.30	347	10.16	11.54	8.24	14.87	1.97	53.21	0.74	0.51	5.20	37.13	56.43	8.16	10.59	11.28	20.00	1.50	48.47	105.91
45	1556	8.70e-12	2.12e-03	3.45e+00	4478	190	6.52e+18	2.69e-01	2.38e-06	59	0.24	146	10.46	11.88	7.36	15.53	2.25	52.53	0.56	0.92	6.22	37.71	54.60	9.41	13.91	7.56	15.79	1.67	51.66	0.05
50	1694	2.15e-11	1.03e-05	4.54e+02	3061	267	5.68e+19	2.69e-02	2.34e-06	52	0.28	335	9.42	12.56	6.06	14.58	2.35	55.03	0.57	0.58	5.12	36.08	57.65	1.93	2.83	14.36	42.12	0.24	38.52	0.15
61	1672	2.14e-11	6.05e-04	3.55e+00	4297	31	6.50e+20	5.39e+01	2.09e-06	48	0.30	81	9.55	12.23	6.29	15.57	1.89	54.47	0.24	0.53	4.86	36.86	57.51	1.49	2.18	16.03	32.55	0.18	47.56	8.28
67	1509	1.68e-11	8.82e-02	1.53e+01	3670	609	3.62e+19	4.51e+00	1.94e-06	67	0.22	752	10.16	12.04	8.39	15.91	0.86	52.64	0.70	0.33	4.86	36.73	57.37	9.81	13.86	6.30	15.72	0.47	53.85	0.14
69	1729	7.86e-12	7.01e-04	3.44e+00	4156	7	5.46e+20	4.17e-01	2.79e-06	83	0.28	736	9.42	13.83	8.03	15.36	2.49	50.86	0.75	1.00	5.04	35.83	57.38	2.04	2.99	10.22	44.59	0.25	39.91	6.88
70	1675	1.03e-11	2.52e-03	3.44e+02	3340	153	1.22e+20	1.76e+00	2.64e-06	32	0.26	0	9.71	14.49	7.41	15.20	1.14	52.06	0.37	0.85	6.34	37.06	55.38	-	-	-	-	-	-	10.81

Table 3.5: Input parameters for the example simulations shown in figures 3.2 to 3.13

3.4 Other StagYY parameters which are not varied

The parameters listed in this section are common to all of my simulations. Many of them could also be varied in future investigations. There are also a selection of parameters not listed which are associated with phase change boundaries. These are not active in my simulations, but could be used in future studies to study, for example, phase dependent viscosity changes. There are also many parameters associated with the solvers which are not listed here. These have an effect on the final state of convection, but are mathematical effects rather than physical ones.

Parameter	Default	Comments & notes
Geometry		
Cells in x direction	1	3-D simulations would add a whole new layer of complexity
Cells in y direction	512	With greater resolution we may see more interesting features
Cells in z direction	64	
Boundaries		
Top thermal boundary layer mode	isothermal	
Bottom thermal boundary layer mode	isothermal	
Top boundary layer velocity mode	free-slip	Zero vertical velocity
Amplitude of initial perturbations	20 K	
Initial boundary layer thickness	30 km	
Reference state		Used to setup initial state of the mantle at first time step
Radioactive half life	2.43×10^{9} yr	Based on expected abundances of U, Th and K
Reference geotherm potential temperature	1600 K	For reference viscosity and density
Radiogenic heating compositionally dependent	true	
Basalt heating enhancement	10	
Core cooling model		Based on Buffett et al. (1992)
Core melting temperature	5600 K	At centre of the Earth to calculate inner core growth and core cooling
Thermal conductivity	3 W m^{-1} K^{-1}	
Reference temperature change	2500 K	Between surface and CMB for initial adiabat. The value is modified by changing CMB temperature
Reference density	3300 kg m^{-3}	Used as reference to set up initial mantle state
Specific heat capacity	1200 J kg^{-1} K^{-1}	
Thermal expansivity	5×10^{-5} K^{-1}	
Reference surface dissipation number	1.18	

Parameter	Default	Comments & notes
Rheology		
Reference viscosity temperature	1600 K	Temperature at which reference viscosity η_0 applies (eq. 3.7)
Reference viscosity depth	0 km	Depth at which reference viscosity η_0 applies (eq. 3.7)
Stress of transition between between diffusion and dislocation creep	1 MPa	
Viscosity activation energy	162.0×10^3 J mol^{-1}	
Viscosity pressure dependent decay constant	1610 GPa	See eq. 3.8
Max viscosity	1×10^{25} Pa s	Viscosity is capped at these values
Min viscosity	1×10^{18} Pa s	
Depth dependence of yield stress	0.005 Pa m^{-1}	
Byerlee's law friction coefficient	0.5	
Viscosity contrast with phase change	1	No phase changes included because I use Perple_X
Viscosity contrast of basaltic end-member	1	No viscosity difference between harzburgite and basalt
Melt		
Harzburgite can melt	false	Only basalt is allowed to melt
Permeability constant for Darcy's law	1×10^{-9}	
Viscosity of liquid	10 Pa	For Darcy's law
Density contrast solid-liquid	500 kg m^{-3}	
Latent heat of melting	600×10^3 J mol^{-1}	
Compositional dependent solidus	true	
Pressure dependent solidus	false	
Eruption depth	300×10^3 m	Above this depth, melt is erupted
Tracers		
Primordial buoyancy contrast	0	Perple_X density calculations cover this
Primordial initialisation mode	layer	Could change this so that starts with a more complex structure
Buoyancy contrast of basalt	0	Perple_X density calculations cover this

4

Investigating the effects of StagYY input parameters

I begin my study by looking for simple relationships between the StagYY input parameters and the resulting mantle states. In this chapter, I investigate to what extent the resulting convection state is predictable from the input parameters, using both standard linear and non-linear methods, such as linear regression and cluster analysis. This gives me an initial indication of what may be achievable using neural networks, which are somewhat more expensive and complex to train, but it also demonstrates the limits of conventional approaches and why a fully probabilistic method is necessary. It potentially gives me an insight into what the networks may find. Neural networks, used in subsequent chapters, are such non-linear functions that it is difficult to decipher what features they are using to make their inferences, therefore using simpler methods gives

me an initial representation which is somewhat easier to comprehend.

4.1 Looking for correlations between temperature and input parameters

Linear correlations

I begin by using standard linear statistical techniques to try to establish a simple relationship between observations from my convection simulations and input parameters.

I first consider the integrated mean mantle temperature of 757 simulations which have run for 4.5 Gyr. Plotting mean mantle temperature as a function of each input parameter to StagYY produces no clear dependence on any individual parameter (figure 4.1) apart from lower values of yield stress. At low yield stresses there is a correlation, but this relationship decays as yield stress increases. I expected to see at least a slight correlation with some other parameters, for example higher viscosity simulations might be expected to lose heat less efficiently and therefore be hotter, as would highly radioactive simulations. However, the trade-offs between parameters swamp the signal of individual parameters. The effects might be observable if only one parameter varies at a time. It may also be the case that the mean mantle temperature is simply too crude a measure and the effects of each parameter can be better detected when considering the whole mantle structure.

The study of the mean mantle temperature introduces four possibilities for why I cannot see any correlation between the observation and the StagYY model parameters. These return repeatedly in slightly different forms throughout this thesis. The possibilities are:

1. The trade-offs between parameters mean that it is not possible to identify the effects of any parameter independently;

2. The observation does not include enough detail to find the parameter;

3. The parameters simply do not affect that observation;

4. I do not have enough samples to be able to find a relationship.

Trade-offs between parameters occur when a pair (or more) of parameters interact such that if both change together the end-state of the simulation remains

constant. It is therefore difficult to tell what combination of parameter values were used. For example, one parameter may act to cool the mantle (e.g. a low yield stress means the mantle loses heat more efficiently) whilst another heats it (e.g. radioactive element content). If these two parameters trade-off, a low-yield stress, highly-radioactive simulation may end up with the same mean mantle temperature as a high-yield stress, low-heating simulation because the rate of heat production is balanced by the insulation provided by the crust.

Possibility 2 concerns the observation chosen. Some parameters may affect small scale variations, for example the deflection of a phase change by a few kilometres. If the observation does not have high enough resolution to resolve these fine-scale structures, I cannot use it to make inferences about the mantle, regardless of the analytical technique I employ. This is linked to point 3. The observation may simply not include any information about a parameter. This could be because it is not high enough resolution, or that the particular observation is not affected. For example, short wavelength gravity variations are dominated by near surface variations in the Earth's density, so are not useful for studying deep mantle anomalies. For some parameters, it may not be immediately obvious whether there will be effects which are present in any observation. For example, compositional variations directly affect density and seismic velocity, but unless they affect the distribution of heat producing elements or the evolution of the shape of convective cells, they may not affect the thermal structure of the mantle.

The fourth reason why I may be unable to find a relationship between a parameter and an observation is a sampling problem. As discussed in chapter 2, low sample density leads to great distances between data points. If the samples are too far apart in either the data or model space, it is difficult to interpolate between them to find any meaningful relationship. This problem becomes more apparent when I use neural networks to interpolate between samples later in this thesis. However, it may also be a problem in a simple analysis, such as figure 4.1. Although the plots in figure 4.1 look densely populated, it must be remembered that each is a projection from a 29 dimensional space onto a single dimension. The points may therefore actually lie a long way away from each other. If I had more points which lay much closer together in the other dimensions, patterns may begin to emerge which are not visible with this number of samples.

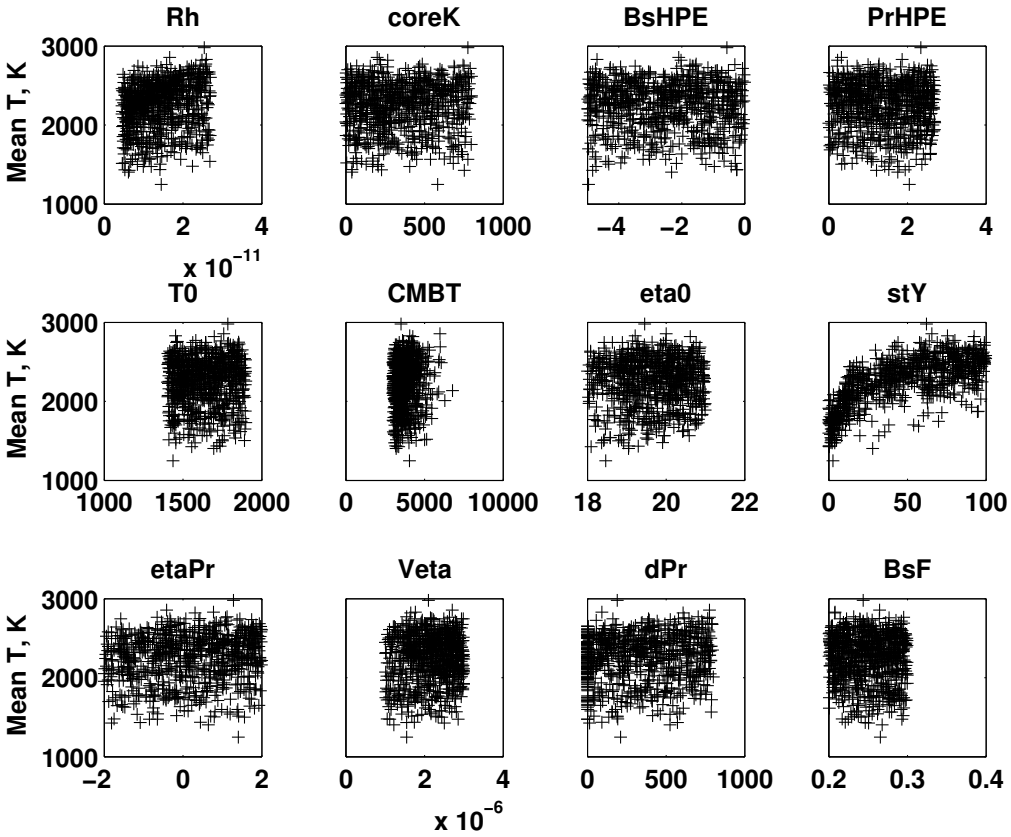


Figure 4.1: Mean mantle temperature for 757 simulations after 4.5 Gy as a function of each input parameter to StagYY.

Cluster analysis

I can begin to investigate possibilities 1 and 2, (trade-offs and details in the observation) by using a using cluster analysis. I categorise the mean 1-D temperature profile using k-means cluster analysis. This measure how similar a set of simulations are to each other, and clusters them accordingly. By using the 1-D profile, I include far more information about the temperature, so can investigate if finer details contain more information.

Figure 4.4 show the results of clustering the 1-D temperature profiles into six clusters. Some patterns emerge in the temperature structures, which are shown in the upper panels in figure 4.4. The mean profiles for the simulations

in each cluster are plotted in red, and are then repeated in figure 4.2 to aid comparison. Cluster 1 contains the coolest simulations, which are generally nearly isothermal away from the boundaries. Simulations in cluster 2 are much hotter, and have a smooth increase of temperature with depth, with a increase at in the lowermost mantle, from about 2000 km depth. The magnitude of the increase varies, with some simulations showing a large increase of up to around 1500 K. Cluster 3 contains still hotter simulations, which show a large jump in temperature in the lowermost mantle, but starting at a slightly greater depth than those in cluster 2, generally around 2200 km. Cluster 4 only contains 19 simulations, out of 757, all of which have an exceptionally low temperature at the base of the mantle, probably due to computational instability in the convection code. These simulations should therefore be considered outliers. Cluster 5 is another cluster with a cold mid-mantle simulations, but they have a very strong temperature increase in the deep mantle with temperatures increasing by as much as 2000 K over 1000 km, starting below around 2400 km depth. Cluster 6 simulations generally have smoothly increasing profiles, with a slight increase towards the base. They are very similar in shape to cluster 2, but are on average around 200 K cooler throughout.

Having clustered the temperature profiles, I can then study how the StagYY input parameters vary between clusters. If the distributions of parameters vary significantly between clusters, then they may be producing effects in the temperature structure which are distinctive enough for the clustering algorithm to identify. In figure 4.4, the middle panel for each cluster shows the distribution of StagYY input parameter values within that cluster of simulations. The values for each parameter within the cluster are binned and coloured according to the number of simulations which fall into that bin, black being highest. For example, in cluster 1, almost all of the simulations have very low yield stress. To show how this distribution varies from the prior distribution for the full un-clustered data set, I plot the prior mean for the full data set in cyan. This shows whether the full distribution has a bias, as seen for example in the initial core temperature. If a parameter makes very little difference to the 1-D temperature structure, the values of that parameter will approximately match the distribution of the parameter in the full data set. The simulations will not be concentrated in one bin but spread across the entire range (e.g. primordial viscosity contrast in cluster 1). Figure 4.3 gives a example to show how to read these histograms.

Several parameters show a skew within the clusters. Cluster 1 has low mantle heating and very low yield stress, both of which help to explain why these simulations have cool temperature profiles. Low yield stress allows the man-

tle to lose heat easily. They also have low initial temperatures, low initial core temperatures and almost no simulations have any primordial material. The lack of primordial material allows the core to lose heat into the mantle, keeping CMB temperatures low. No other parameters show a noteworthy pattern in their values. Working though the clusters in order of increasing mid-mantle temperatures, cluster 5 is the next coolest, with low mid-mantle temperatures, but high core-mantle boundary (CMB) temperatures. These simulations also have low yield stress, explaining the cool mid-mantle, but have thicker than average primordial layers, which keeps the lowermost mantle hot. That low yield stress simulations produce cool mantles and thick dense layers at the base of the mantle keep the core hot is consistent with many previous studies. Cluster 6 has a near isothermal mean profile, with low mantle heating and no primordial material. The simulations in this cluster have a middling yield stress, which is probably what is responsible for the middling mid-mantle temperatures. Cluster 2 is very similar to cluster 6, with a mean profile which is hotter but quite smooth. Again, the simulations have very little primordial material. but have a high yield stress. Finally, the hottest simulations are in cluster 3, with a large lower-mantle temperature jumps. They have high initial mantle heating rates, very high yield stress, a range of primordial layer thicknesses, and the primordial material has high viscosity. When all these factors are combined, it is clear why these simulations are the hottest: they are blanketed top and bottom and produce a lot of heat from within.

Whilst I can see that some clusters tend to contain simulations with particular characteristics, I cannot see how parameters interact by considering one at a time. It may be that some of the parameters which have broad ranges are actually trading-off with other parameters. A broad range of values may then produce similar temperature structures if that parameter is being cancelled out by another parameter which moves to oppose it. In an attempt to identify tradeoffs between the parameters, I also calculate the Pearson correlation coefficient between parameters within each cluster. This is defined as:

$$C(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{(A_i-\bar{A})}{\sigma_A}\frac{(B_i-\bar{B})}{\sigma_B}\right) \tag{4.1}$$

where A and B are vectors of the values of two StagYY variables belonging to the N simulations in each cluster. $\bar{A}$ and $\bar{B}$ are the mean of variable A and B within each cluster, and σ_A and σ_B their standard deviation. The correlations coefficient, C, is plotted in a grid in the lower panel for each cluster in figure 4.4. The colour scale saturates at the relatively low value of $\pm$ 0.5, and values between

± 0.2 are white. The lack of colour across most of the grids suggests that there is almost no correlation between parameters, apart from cluster 4, which has too few simulations to draw any meaningful conclusions. Any colour, red or blue, suggests that if one parameter changes, the other must move in order to balance it and maintain a similar temperature structure. However, it must be emphasised that none of the correlations are particularly strong.

Yield stress is generally inversely correlated with mantle heating. Both internal heating rate and high yield stress reduce the likelihood of entering a mobile lid regime. However, higher yield stress generally increases the the lateral size of convection cells whilst higher internal heating increases the vigour of convection and therefore decreases the wavelength of convection. These affects may therefore be separable when the length scale of thermal heterogeneity is considered. The only cluster where there is no correlation between yield stress and heating is cluster 3, which has both high heating and yield stress and contains the hottest simulations. These two parameters are amplifying each other and so do not have to change together to maintain a hot profile. Reference viscosity and activation volume are correlated in 4 out of 5 relevant clusters. Both parameters increase viscosity and therefore decrease the vigour of convection and thus the rate of heat loss. In all four, the correlation is negative: as one increases, the other must decrease to maintain the temperature structure appropriate for each cluster.

Initial core mantle boundary temperature is very weakly correlated with a lot of different parameters. The trade-offs between this initial bottom boundary layer temperature and other heating parameters such as initial mantle temperature and radioactive heating rate in the core and mantle are intuitive, but still somewhat unexpected after 4.5 Gyr of convection. No linear relationship for either initial core temperature or mantle heating with mean mantle temperature are seen in figure 4.1, suggesting a more subtle effect, if this correlation is more than chance. The correlation with the primordial material parameters is also intuitive. If the lower boundary layer is initially hot, it will heat the primordial material, decreasing its density and viscosity and thus increasing its chance of entrainment into the rest of the mantle, depending on the initial thickness and viscosity.

I have repeated this exercise using different numbers of clusters and preprocessing the data in different ways, including the deviation of the temperature from an adiabat and the amplitude of the non-degree 0 components of the amplitude spectra of the temperature field, which gives a measure of the lateral heterogeneity at each depth. The results are all very similar, with yield stress

consistently showing the biggest variation away from the prior mean between clusters.

It should be highlighted that there are very few simulations in each cluster (between 100 and 200). The conclusions drawn here, especially the correlation between parameters, are therefore not particularly statistically reliable. This investigation was conducted simply to see if any very obvious trends appeared.

4.2 Emulator modelling

Mean mantle temperature clearly does not depend linearly on any one parameter, as shown in figure 4.1, but there are definitely trade-offs which affect the temperature, as shown by the cluster analysis. By considering all of the parameters together, I can use the pattern recognition powers of a neural network to identify these trade-offs and predict the mean mantle temperature.

I do this by training a neural network to predict the mean mantle temperature given all the StagYY parameters. The networks emulate the processes within StagYY by finding relationships between the input parameters and resulting temperature without the convection processes ever being explicitly described to the networks. In this case, the networks are still operating within a Bayesian framework, as described in chapter 2, but I am using Bayes' theorem to predict an outcome, rather than to perform an inverse calculation. The known observation is now a vector containing the StagYY parameters, and the target space has one dimension, which is the mean mantle temperature of the simulations.

Figure 4.5 shows the probability density functions for the network prediction for the mean mantle temperature for a test set of simulations. Figure 2.6 gives an example demonstrating how to read these figure. I find that higher mean mantle temperatures are predicted with more accuracy than lower temperatures.

Mantle convection is fully deterministic, therefore the mean mantle temperature is dependent only on the input parameters and boundary conditions. There is therefore no reason why a sufficiently complex network should not be able to emulate the convection process (e.g. Hornik et al., 1989). However, it is still nice to see that such a small emulator neural network can approximate the complex non-linear functions built into StagYY, especially given the hundreds of hours of computational time it takes to calculate the forward simulation code. Whilst mean mantle temperature is a single very general parameter, this demonstrates that neural networks are a potentially very powerful tool for

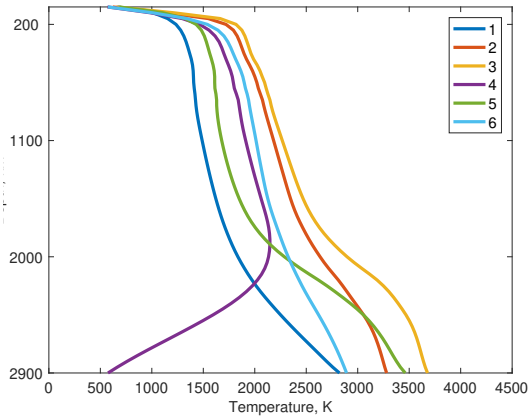


Figure 4.2: The mean temperature profile from each cluster, to aid comparison between clusters in figure 4.4. All the profiles are relatively smooth because there is no viscosity jump at the transition zone. This, combined with very low yields stresses means that most simulations are colder than expected for Earth. Cluster 4 only contains 19 simulations, out of 757, all of which have an exceptionally low temperature at the base of the mantle, probably due to computational instability in the convection code

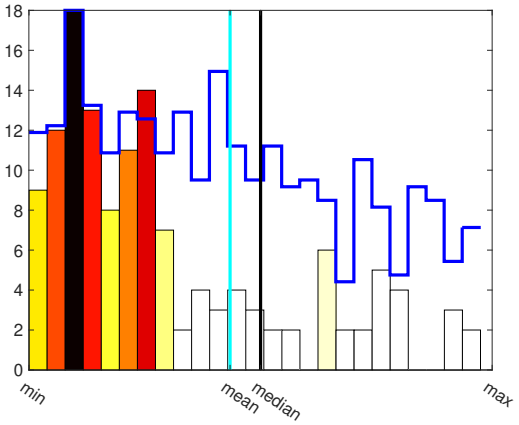
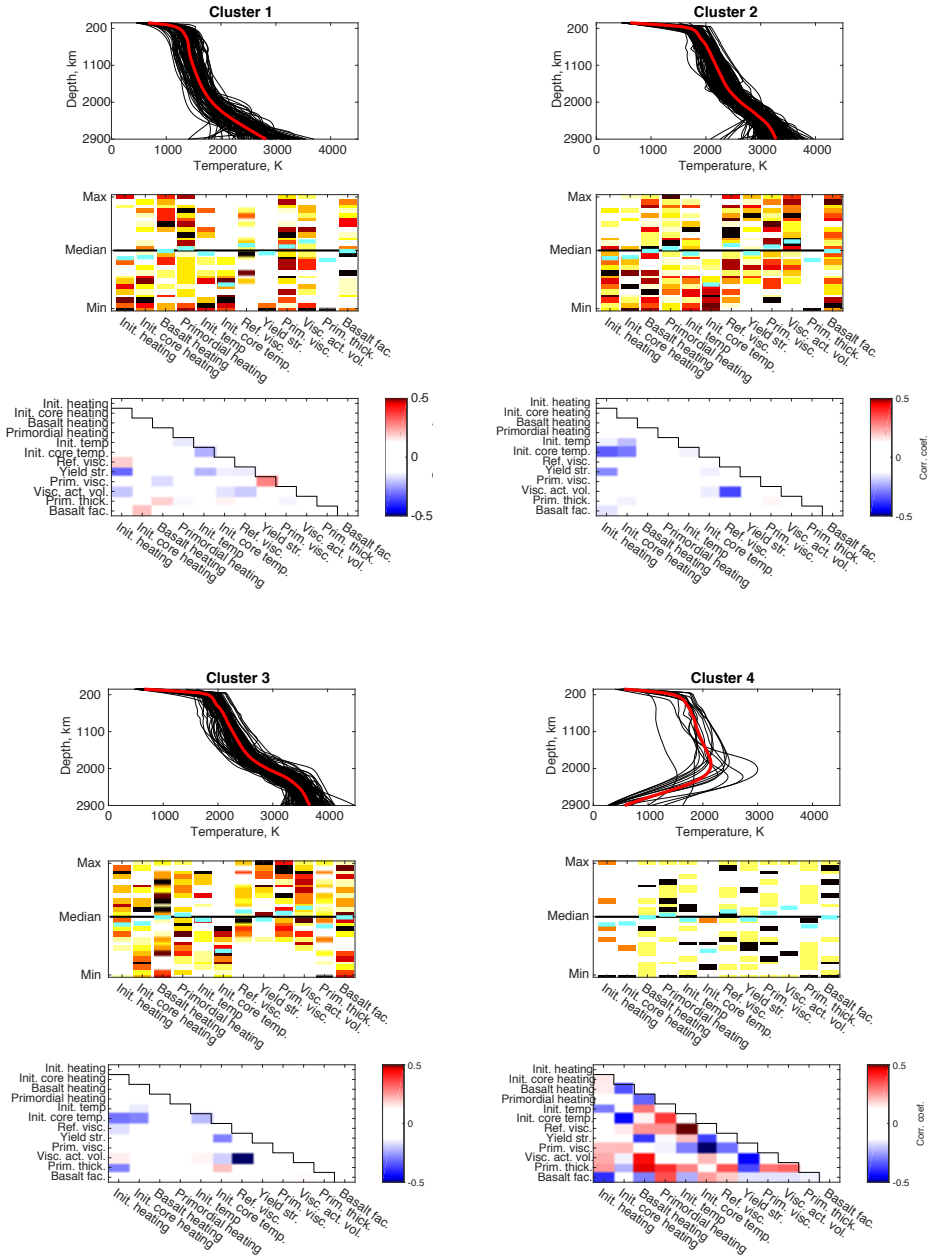


Figure 4.3: Example histogram to demonstrate how to read the middle panels in figure 4.4. The coloured columns in figure 4.4 are the histogram for the values of a particular parameter within a cluster. They are coloured relative to the bin count, with a cut-off so that bins with less than 40% of the counts of the maximum bin are white. The prior for the entire suite of simulations prior to clustering is plotted in blue. The mean for the prior is plotted in cyan. This shows the skew of the prior, so that any skew in the cluster can be compared to that of the prior. Each column in the middle panels in figure 4.4 is the top view of such a histogram.


Cluster 1
Cluster 2
Cluster 3
Cluster 4
Depth, km
Temperature, K
Max
Median
Min
Corr. coef.
Init. heating
Init. core heating
Basalt heating
Primordial heating
Init. temp
Init. core temp.
Ref. visc.
Yield str.
Prim. visc.
Visc. act. vol.
Prim. thick.
Basalt fac.

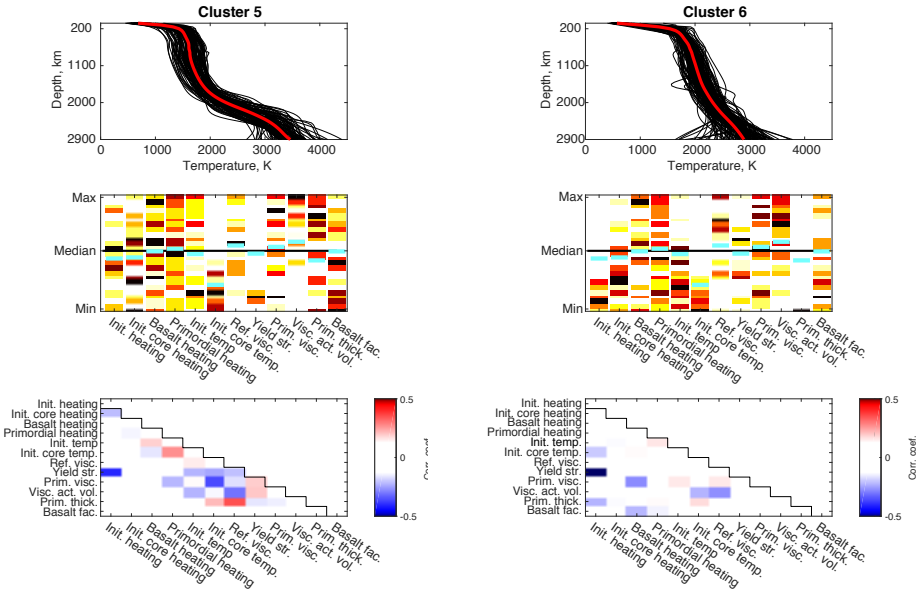


Figure 4.4: Clustering simulations according to 1D temperature profile. For each cluster, the top panel shows all the profiles in the cluster. The middle panel shows the distribution of the values of each input parameter within the cluster. Each vertical stack of bins is for one input parameter. The colour indicates the number of simulations per bin, black being the largest. The horizontal black line is the median value for each parameter and the minimum and maximum ranges are the same as in table 3.1. The cyan line shows the mean value for each parameter across the whole data set, to show the skew of the mean away from the median value. The distribution of each parameter within the cluster therefore gives an indication of which parameters are independently affecting the profiles enough to affect the cluster analysis. See figure 4.3 for more explanation. The bottom panel shows the correlation between the input parameters for the simulations within the cluster, calculated according to equation 4.1. This gives an indication of the existence of trade-off between parameters. Correlations suggest trade-offs between parameters, showing that if one parameter changes, the other must move to balance the effect to maintain a similar temperature structure.

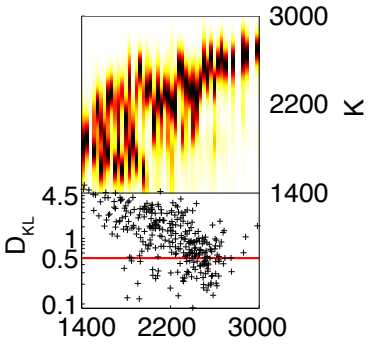


Figure 4.5: Predicting mean mantle temperature after 4.5 Gyr using a MDN given all StagYY input parameters.

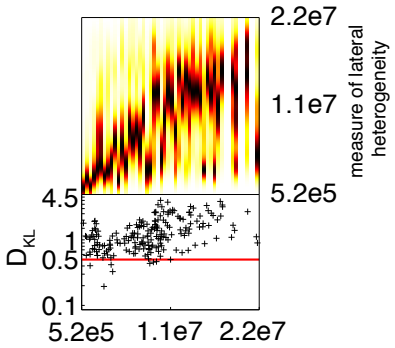


Figure 4.6: Neural network trained to predict amplitude contained in degree 1–20 in the amplitude spectrum of the temperature field at all depths. This works as a proxy for the magnitude of lateral variation in temperature. It is calculated according to equation 4.2.

studying mantle convection.

I am not limited to calculating mean mantle temperature. I can also use proxies for the lateral heterogeneities in mantle structure. In figure 4.6, I show the PDFs from a test set of simulations trained on the amplitude of the temperature spectra contained in degrees 1 to 20, given by:

$$\text{measure of lateral heterogeneity} = \sum_{d=1}^{d=64} \sum_{f=1}^{f=20} |T_{df}| \tag{4.2}$$

where T is the amplitude spectra of the temperature field, f is each spectral degree and d is each radial interval in the simulation grid. This gives a measure of how the temperature varies laterally, including both the amplitude of the

variation and the complexity of these variations. Figure 4.6 shows that I can predict this measure using a MDN give all the inputs to StagYY. Using neural networks, I can therefore predict two measures for the final state of the mantle, despite the non-linearity of the system.

I can also use an emulator to approximate the relationship between the StagYY input parameters and the density of the mantle at a given depth. I train the networks to find variations in density at a constant pressure. It should be theoretically possible to train a single network to predict the density at any depth, however then the variations caused by StagYY input parameters would be overwhelmed by the strong pressure dependence of density. This could be overcome with larger networks, but with such a small training set, it is much easier to train the networks at constant depth. I switch from using temperature (as in the previous section) to density simply to demonstrate that both work for my simulations. I train 64 constant pressure networks, one for each radial layer of cells in my simulations. Figure 4.7 shows some test results from these networks. Each profile is the prediction of the 1-D density profile for 4 different test simulations using the constant pressure emulating networks. The grey shading shows the amplitude of the posterior probability density function provided by the networks at each depth for the probable density. The red line is the true simulation profile. The cyan line shows the extreme values and prior range of the training set, providing an indication of the information learnt by the networks and shows that the networks have actually made inferences which are significantly better than the prior. It also aids comparison between the simulations.

The example profile produced by the surrogate networks shown in figure 4.7 perform very well between 1000 and 2000 km, with narrow, high PDFs, peaking at the correct density. Between 100 and 1000 km, the networks perform nearly as well, picking the correct density jumps with phase changes around 660 and 1000 km, but with slightly greater uncertainty, indicated by wider PDFs. They are somewhat less accurate at 660 km, but recover on either side. In the crust and lithosphere, the estimates are less accurate. The estimates for the deep mantle, below 2000 km, are much less certain, in some cases entirely inconclusive. This is due to the much greater variation in composition and therefore density because of the option for a layer of dense primordial material in the prior distributions. The PDFs for the full test set at various depths are given at the end of this chapter, which demonstrate that the cases given in figure 4.7 are representative of the general performance.

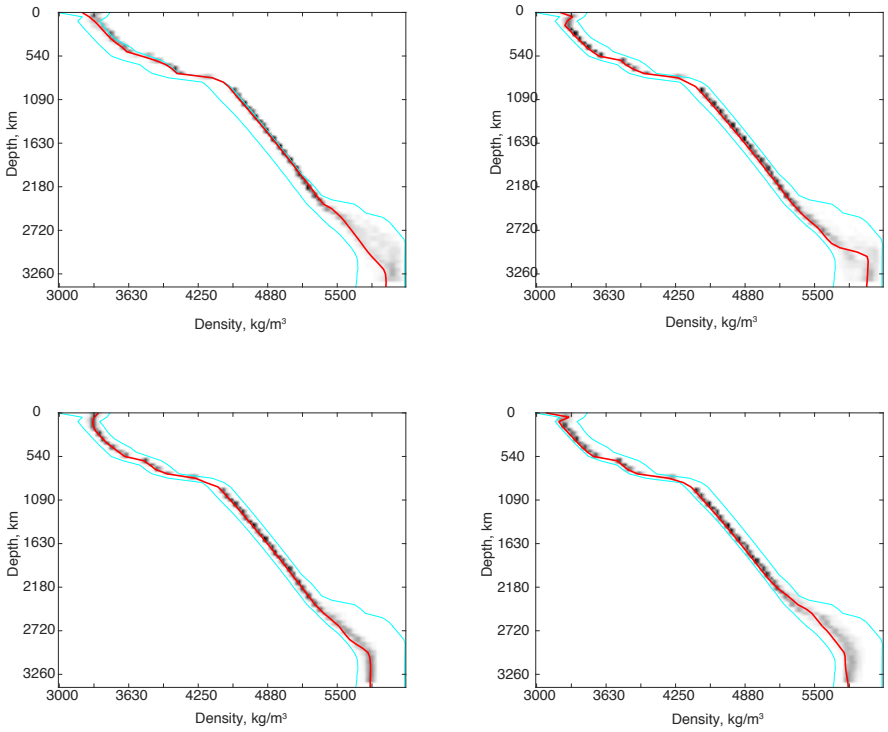


Figure 4.7: 1D mantle density profile predicted by a suite of surrogate networks for 4 different test simulations. The red line is the true profile for each simulation. The cyan lines give the extreme values present in the training set, to provide a reference for comparison.

4.3 Emulator modelling to investigate simulation sensitivity

Having successfully trained networks to predict the outcome of convection simulations, it is extremely rapid to evaluate the probable outcome of a simulation after 4.5 Gyr of convection, given the simulation input parameters. This gives me a rapid way to investigate the sensitivity of the end state of the simulation to changes in the StagYY input parameters. I can then investigate the uncertainties of the results of a single simulation, by showing how much the result would change as the input parameters change, without running more convection simulations. This approach is already used to test model sensitivity in some other

fields, such as volcano hazard monitoring (e.g. Spiller et al., 2014) and climate sensitivity analysis (e.g. Olson et al., 2012).

Having established that my emulating networks provide a reasonable inference for density, as shown in figure 4.7, I can use them to investigate the sensitivity of the density structure to each parameter. Here, I am not trying to find the uncertainty on any particular simulation, but am simply investigating how much each parameter affects the resulting density structure. I use the 64 constant-pressure density-predicting networks demonstrated in the previous section to find out how much each parameter influences the density at each depth. I therefore also get an indication of the depth dependent sensitivity of the simulations to each parameter. These results would allow me to target a full inversion to the most sensitive regions of the mantle.

I investigate the effects each of the 12 StagYY input parameters independently. For each parameter, I hold all of the other parameters constant, then change the parameter of interest, within its prior range, to see what the emulating network expects to happen to the density structure. Figure 4.8 shows an example where I change the yield stress, holding all other parameters constant. In this case, the density prediction is clearly sensitive to yield stress, varying by up to 150 kg m^{-3}. The sensitivity of the density structure to each parameter is also determined by trade-offs between the other non-varying parameters. I therefore repeat this exercise with 200 sets of model parameters, where the other 11 parameters are different, but drawn from the same prior distributions as the training set. I therefore hope to capture some of the effects of the trade-offs.

For each of the 200 simulations, the network calculates 20 PDFs for which the parameter of interest is varied randomly, holding the other 11 parameters constant. This produces a set of PDFs as in figure 4.8. I take the maximum probability value of these 20 posterior PDFs and calculate the variance between these maxima. This provides a measure of the variation caused by that parameter. I can then plot the variances taken from each simulation. The variance between the density estimate for each simulation at each depth, one parameter at a time is shown in figure 4.9. This gives an indication of where the emulating networks think that each parameter has a significant impact on the density.

Again, yield stress has a significant impact, mostly in the upper mantle. The viscosity parameters: reference viscosity and viscosity activation volume; make more difference to the density estimate in the lower mantle. The effects of the initial thickness of the primordial layer are most pronounced in the lower mantle. For the other parameters, the effects are smaller. The initial mantle heating

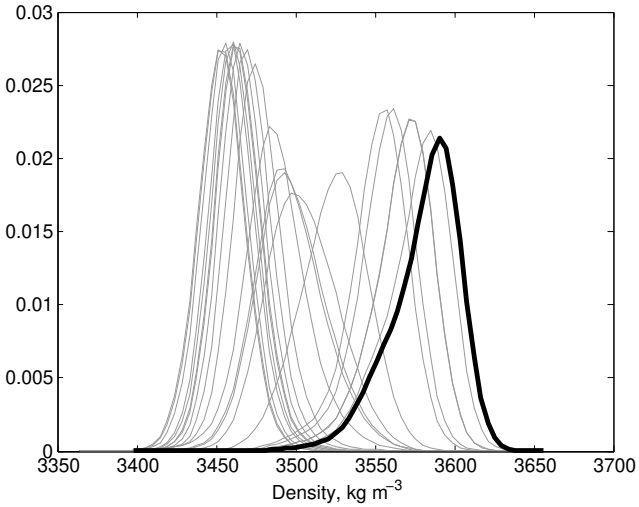


Figure 4.8: Effect of changing yield stress when using a neural network emulator to predict density at 360 km depth. The thick pdf is the prediction when using the correct yield stress for this simulation. By changing the yield stress randomly, but keeping other parameters constant, I can alter this prediction by up to 150 kg m^{-3}.

rate makes a small difference in the lower mantle and upper mantle. The magnitude of the signal from the other parameters is very small, but almost all produce some signal in the lowermost mantle and around the transition zone. Unfortunately, these are the regions where the emulating networks perform least well. The poor performance of the networks may be because all of the parameters have stronger effects here, increasing the number of relevant dimensions and making the networks harder to train.

The estimates for the transition zone demonstrate one of the disadvantages of prior sampling. Large variations are expected in the transition zone, which are strongly sensitive to the StagYY input parameters. Interpolation between samples is therefore much harder because the variation between samples is greater and the results are less satisfactory. This is reflected in the attempts to use the prior-sample-trained emulator in posterior sampling. With more training samples, the emulator performance would improve, but the cost of generating sufficient prior samples may no longer balance the cost saved by using

emulators in place of conventional posterior sampling.

4.4 Using a neural network emulator in a Monte Carlo inversion

Having trained my neural network StagYY emulators using a set of prior samples, I can then use them to search the model space in an attempt to match an observation. This combines both prior and posterior sampling approaches: the networks are trained on prior samples and are then used to invert an observation by posterior sampling.

As an initial investigation, I try to find the posterior PDFs for each of the 12 StagYY input parameters for three StagYY simulations, by using the neural network emulator to predict the mantle density after a 4.5 Gyr simulation. For each simulation, I know the resulting density profile, and pretend that I do not know what StagYY model parameters created it. I chose a particular depth interval and create a probability density function for the mean density at that depth interval. The probability density function is Gaussian, with a variance of 400 kg m^{-3}. I draw samples from the prior ranges of StagYY model parameters and use the emulating neural networks to give me a posterior probability density for the mean density at that depth interval given those input parameters. I can then compare the network-produced posterior PDF with the PDF for the actual density for that simulation to produce a likelihood function for a Metropolis Hastings Monte Carlo posterior sampling approach. In this case, I use a very simple likelihood function, taking the dot product between the two PDFs. Figures 4.10 to 4.12 show the result of this Monte Carlo posterior-neural network emulator sampling approach to finding the input parameters for three different test simulations. For each simulation, I vary the 12 StagYY parameters, but assume that the composition is perfectly known. I use networks trained for the depth intervals 316–361 km, 541–587 km and 1445–1490 km, because these networks perform well across the test sets (see figure 4.7). In these figures, the blue line shows the target value for the simulation. The red histogram is the prior distribution for each parameter, with which the emulator networks were trained. The grey histograms are the outcome of the Monte Carlo simulations looking for each parameter at the three different depths, with the darkest coloured histogram being the for the network trained at the deepest mantle depth.

For the simulation in figure 4.10, the Monte Carlo sampling using the network emulator for most of the parameters simply returns the prior distribution.

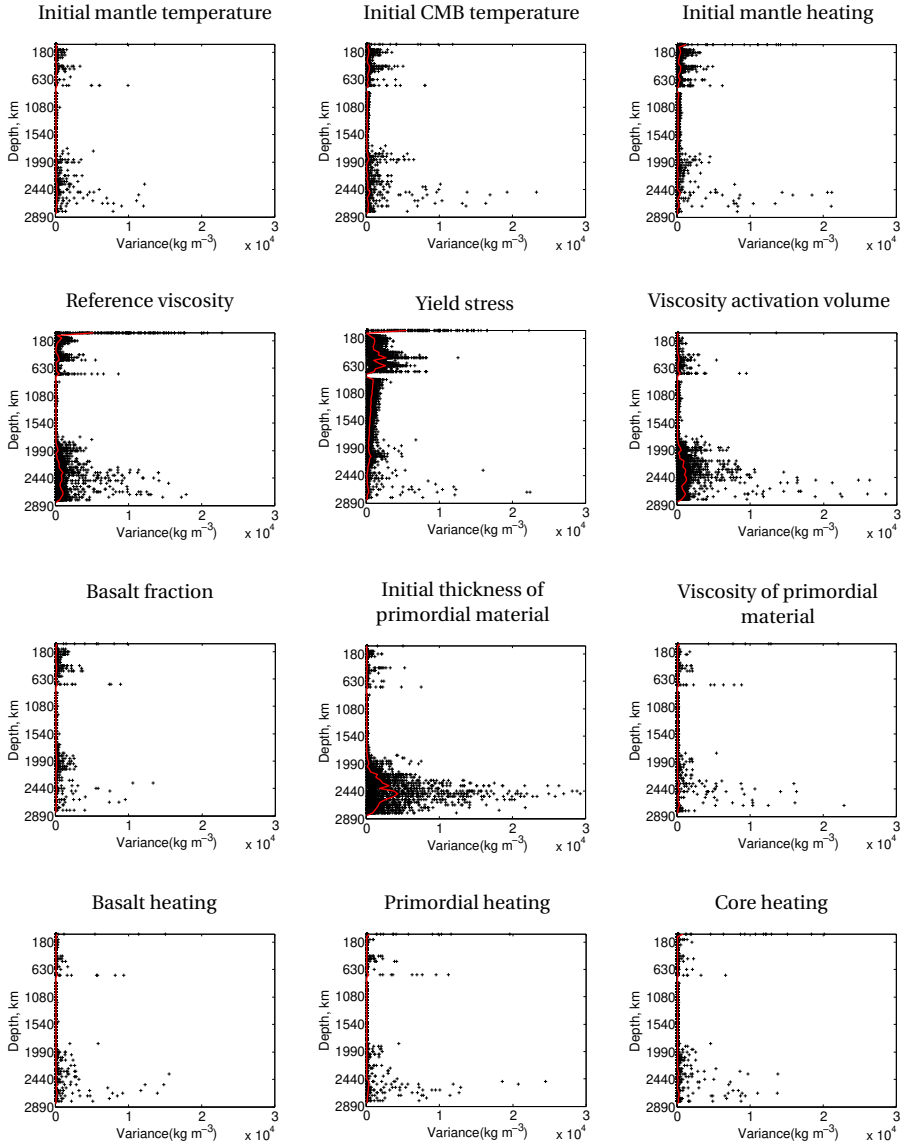


Figure 4.9: The variance in density estimate caused by varying one parameter at a time as a function of depth, when using a neural network to emulate the density produced by StagYY after 4.5 Gyr of convection. The red line is the average variance and indicated the sensitivity of the density to each parameter as a function of depth. The larger the variance, the more sensitive the density estimate is.

Initial mantle heating rate has an improvement over the prior, with a slightly better estimate when trying to match the deepest observation (darkest histogram). The deepest histogram for initial mantle density shows an improvement over the prior in the right location, but the other two are very poor, suggesting this may just be chance. Reference viscosity has a significantly sharper histogram than the prior, but misses the true value of the target. One histogram predicts yield stress perfectly, but the other two miss the value. The shallower networks perform better for basalt fraction, but the result is not particularly good. The results for the second test simulation in figure 4.11 are similar. Yield stress is found much better for this simulation. Initial mantle heating and core heating are also both found better. The initial thickness of primordial material is also found, but with much less certainty. For simulation 3 (figure 4.12) the results for core heating rate are similar to the previous two simulations. However, all three simulations have similar true values for this parameter, so it may be that the networks always return such distributions, regardless of the true value.

These examples are not particularly successful. However, they demonstrate how prior and posterior sampling methods can be combined. It has the advantage of very rapidly narrowing the range of possible model space parameters. There may be better ways to train the emulators so that they are more accurate and sensitive to the input parameters. This is also only a brief investigation and using a different posterior sampling algorithm may also improve the inferences made.

This approach is also limited in several ways. Most fundamentally, I never add any new information when I draw new samples. I simply make a new interpolation at a different location. The accuracy is therefore limited by the initial prior sampling approach. In this particular implementation, I can also only train the networks to give a single output. In this example, I used multiple networks together to increase the dimensionality and so that I can compare results. However, a method which could produce a multi-valued output (e.g. the full 1-D depth density profile)may be more powerful as some of the effects of certain parameters are going to be in the depth or lateral variations, rather than the absolute value at any depth.

4.5 Conclusion

In this chapter, I take a first step towards assessing the effects of the input parameters to the convection simulation code StagYY. I show that the most basic linear statistical methods are not satisfactory and produce no useful results,

which is not surprising given the non-linear nature of mantle convection. With cluster analysis, I begin to see the trade-offs between parameters. I then introduce neural networks as a method for studying convection simulations. These demonstrate that whilst crude observations such as the mean mantle temperature are too dependent on the highly non-linear interplay between parameters to be useful for inversion, they are easily predicted by relatively simple networks. This gives me a method with which to investigate the sensitivity of the simulations to input parameters in a fully Bayesian framework. This approach seems to have promise and demonstrates one possible application of neural networks in geodynamics.

4.6 1-D density posterior PDFs

For reference, I include the full test set pdfs for the surrogate networks trained to predict the 1D density structure of the simulations, given all the StagYY input parameters and composition. Four examples are included earlier in the chapter (figure 4.7), but since they were only single cases, it does not fully demonstrate that these networks perform well. One committee of networks is trained at each depth point. I do not include all of the results for the lower mantle as these do not vary significantly.

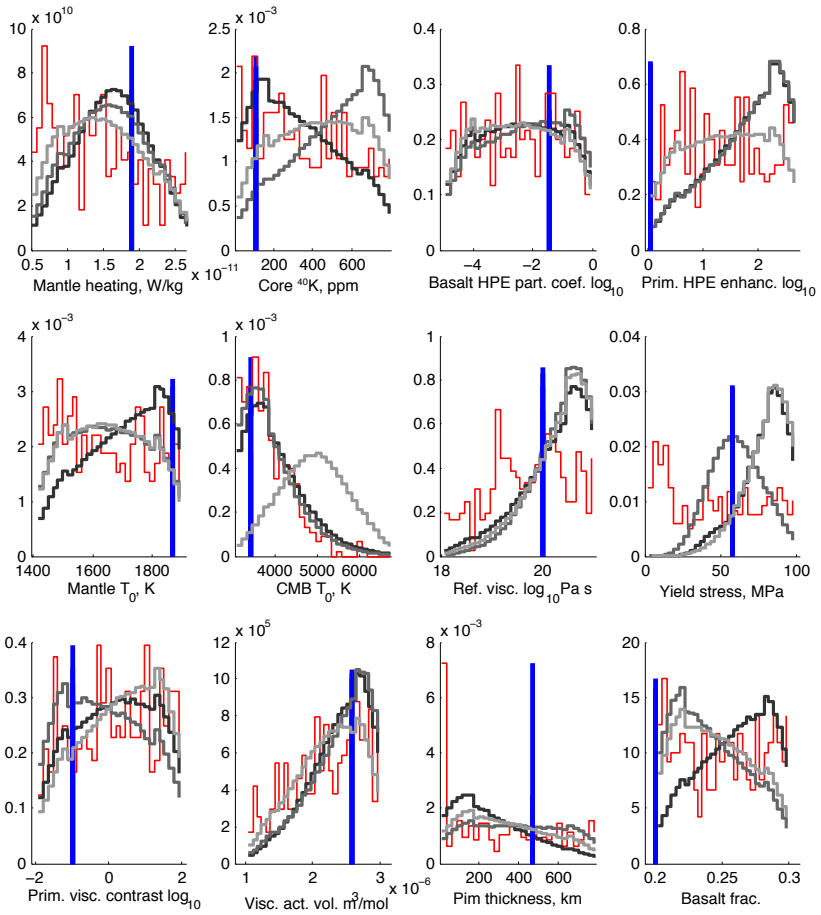


Figure 4.10: Using the density predicting surrogate networks to calculate samples for a Monte Carlo inversion for StagYY model parameters for test simulation 1. The red histogram is the prior distribution for each parameter; the blue line is the true parameter for each simulation; and the grey histograms are the distribution after sampling the model space looking for the best fit to density at three different depth intervals (316–361 km, 541–587 km and 1445–1490 km). The darkest grey is for the deepest.

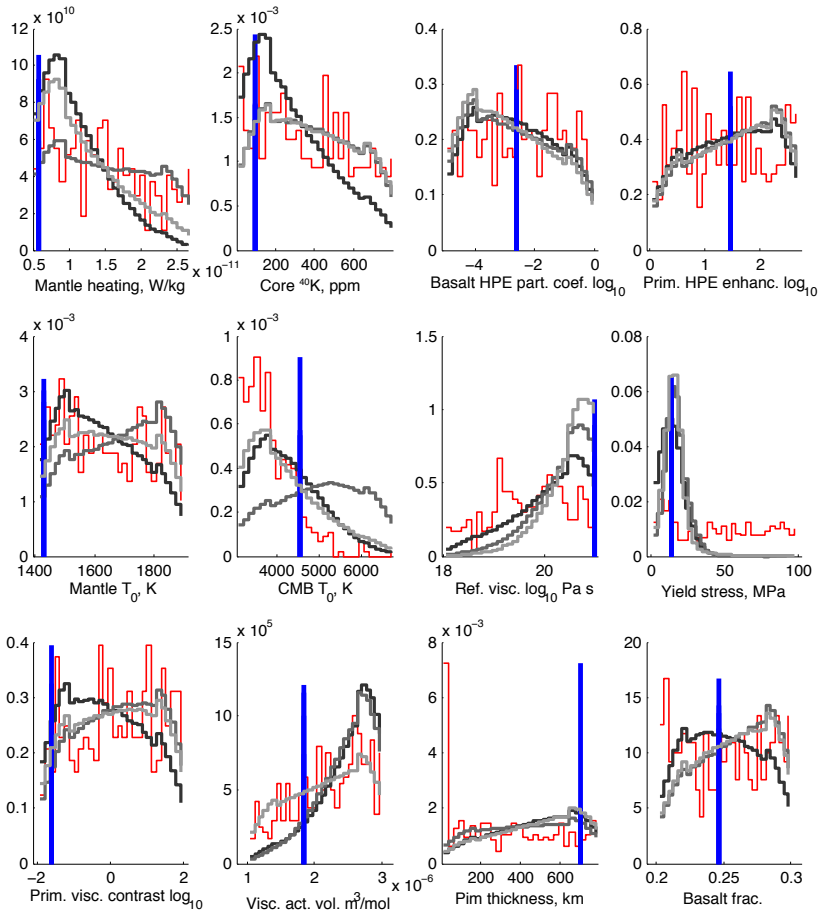


Figure 4.11: Using the density predicting surrogate networks to calculate samples for a Monte Carlo inversion for StagYY model parameters for test simulation 2. The red histogram is the prior distribution for each parameter; the blue line is the true parameter for each simulation; and the grey histograms are the distribution after sampling the model space looking for the best fit to density at three different depth intervals (316–361 km, 541–587 km and 1445–1490 km). The darkest grey is for the deepest.

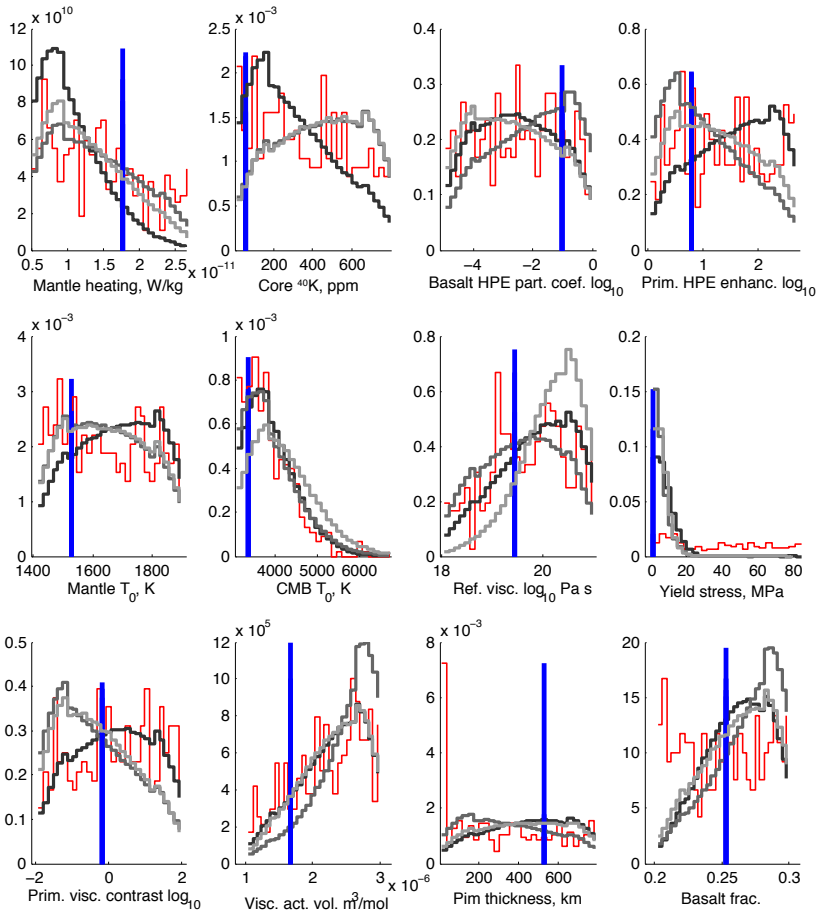
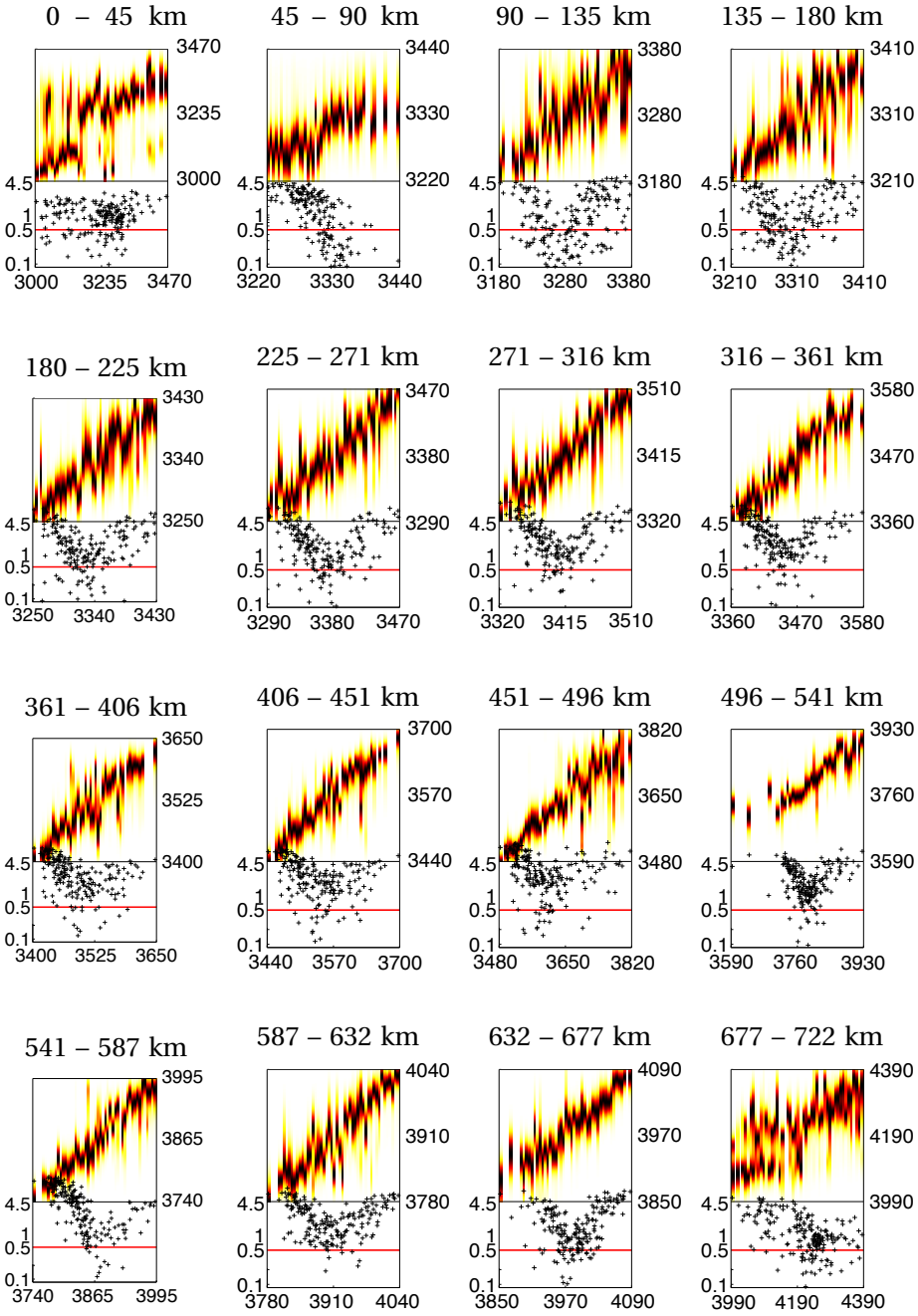
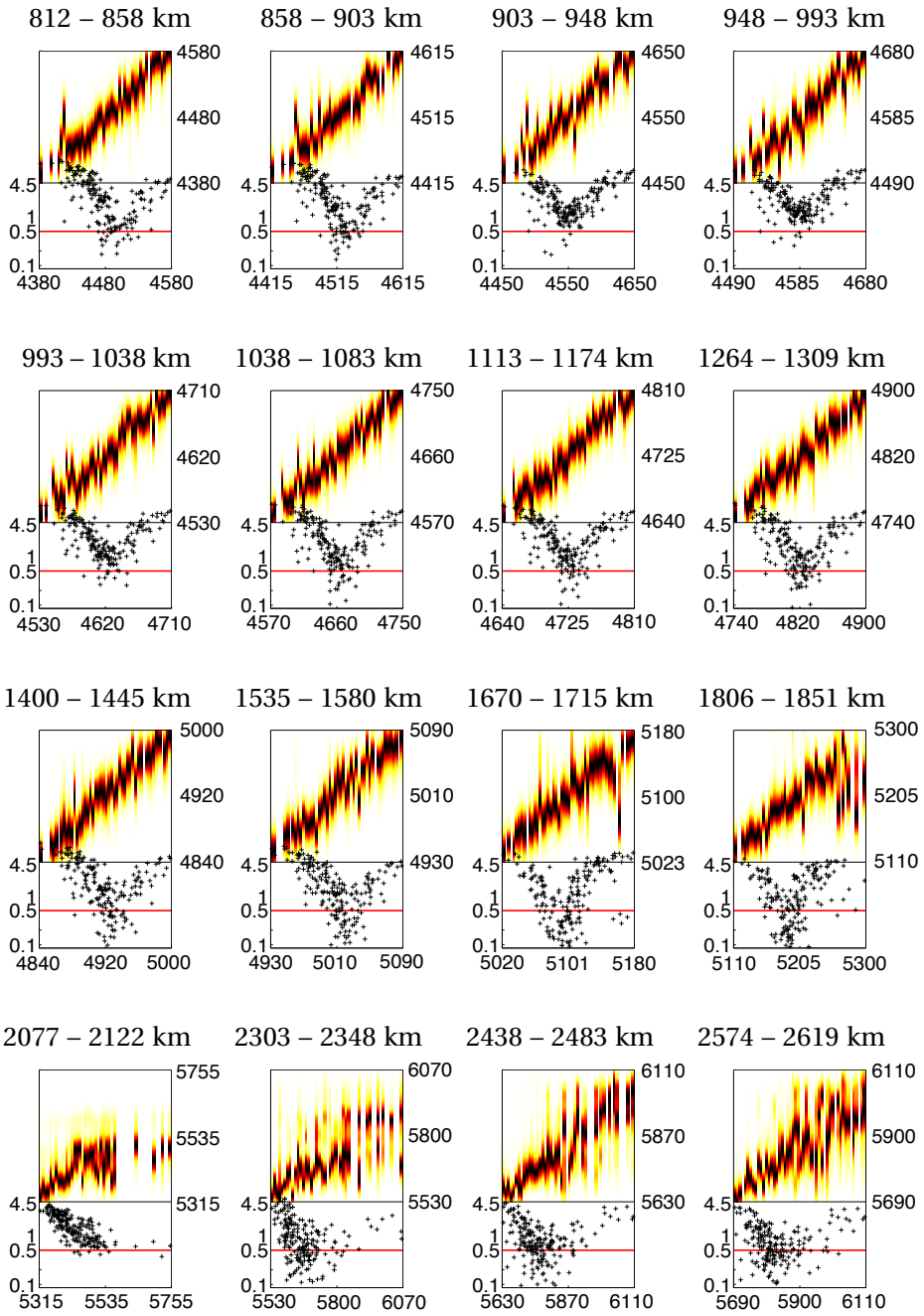


Figure 4.12: Using the density predicting surrogate networks to calculate samples for a Monte Carlo inversion for StagYY model parameters for test simulation 3. The red histogram is the prior distribution for each parameter; the blue line is the true parameter for each simulation; and the grey histograms are the distribution after sampling the model space looking for the best fit to density at three different depth intervals (316–361 km, 541–587 km and 1445–1490 km). The darkest grey is for the deepest.


0 – 45 km
3470
3235
3000
4.5
1
0.5
0.1
3000 3235 3470
45 – 90 km
3440
3330
3220
3220 3330 3440
90 – 135 km
3380
3280
3180
3180 3280 3380
135 – 180 km
3410
3310
3210
3210 3310 3410
180 – 225 km
3430
3340
3250
3250 3340 3430
225 – 271 km
3470
3380
3290
3290 3380 3470
271 – 316 km
3510
3415
3320
3320 3415 3510
316 – 361 km
3580
3470
3360
3360 3470 3580
361 – 406 km
3650
3525
3400
3400 3525 3650
406 – 451 km
3700
3570
3440
3440 3570 3700
451 – 496 km
3820
3650
3480
3480 3650 3820
496 – 541 km
3930
3760
3590
3590 3760 3930
541 – 587 km
3995
3865
3740
3740 3865 3995
587 – 632 km
4040
3910
3780
3780 3910 4040
632 – 677 km
4090
3970
3850
3850 3970 4090
677 – 722 km
4390
4190
3990
3990 4190 4390


812 – 858 km
858 – 903 km
903 – 948 km
948 – 993 km
993 – 1038 km
1038 – 1083 km
1113 – 1174 km
1264 – 1309 km
1400 – 1445 km
1535 – 1580 km
1670 – 1715 km
1806 – 1851 km
2077 – 2122 km
2303 – 2348 km
2438 – 2483 km
2574 – 2619 km

5

A proof of concept: Inferring mantle convection parameters from the temperature structure

Part of this chapter has been published as Using pattern recognition to infer parameters governing mantle convection*, by S. Atkins, A. P. Valentine, P. J. Tackley and J. Trampert, 2016, in Physics of the Earth and Planetary Interiors*

In the previous chapter, I showed that neural networks can be used to rapidly predict simple density and thermal observations, given all the convection simulation input parameters. The neural networks learn to approximate the non-linear processes happening over 4.5 Gyr of simulated convection. However, whilst the outcome of these non-linear processes are predictable, they are very difficult to invert. Due to trade-offs between parameters, there are simply too

many combinations of input parameters which could produce the same 1-D thermal or density profile. Linear statistical methods cannot begin to unravel these trade-offs. In chapter 4, I tried to use the neural network emulators, which predicted the outcomes of the simulations, to attempt a posterior sampling based inversion approach. This was not particularly successful, partly because I have too few samples which do not adequately capture the sensitivity of the simulations to the input parameters, and partly because the networks can only predict simple observations, such as depth-dependent mean mantle temperature, for which any solution is much too non-unique to be useful.

In this chapter I try a different approach to inversion. I train neural networks to approximate the inverse relationship between more complex observations of the mantle and StagYY input parameters. I use the full amplitude spectra of the temperature field at various time intervals during convection. I can then see whether the temperature contains sufficient information on these parameters to make inferences, given the number of simulations I have, and whether the influence of these parameters on the temperature structure changes with time. This chapter is designed as a proof of concept to show that using observations of the mantle structure it is indeed possible to make inferences about the governing parameters of mantle convection. As such, part of this chapter has been published as Atkins et al. (2016). Having shown that this approach is feasible, I then build upon it in later chapters.

There are several advantages to this neural network inversion approach, compared to the methods tried in chapter 4. To begin with, I use much more detailed observations of the structure of the mantle. I train the networks to make inferences from the amplitude spectra of the temperature field. This contains information about the degree of lateral heterogeneity at all depths, as well as the change in temperature with depth. By using so much more information to make inferences, I remove some of the non-uniqueness introduced by trade-offs between parameters which limited my previous attempts. The other advantage is that the inference is provided in the form of a full PDF, as described in chapter 2, including much more information on the uncertainties than the conclusions drawn from the cluster analysis in the previous chapter.

Using neural networks trained on thermal observations to make inferences does still have some limitations. Some of the more general limitations were covered in chapter 2, but there are also more specific limitations. The low number of samples and therefore the interpolation distance between samples is nearly as much of a problem in this method as for the ones in the previous chapter. However, with a more complex observations, the non-uniqueness of the

problem decreases, and therefore the observations cover a wider area of the data space and interpolation stands more chance of success. The other limitation is the observation used. The thermal structure of the Earth's mantle is not well constrained, so this is not the most practical observation for applying the method to the Earth. However, the temperature structure is a fundamental observation of convection and therefore works well for a proof of concept: the amplitude and spacing hot upwellings and cold downwellings indicate the vigour of convection. The temperature structure is simply the product of the convection and is not a direct function of composition, as is the case with density. This therefore removes some complexity, but means that the networks struggle to make inferences about compositional parameters.

5.1 Method particulars

I show here that neural networks can be used to make inferences about the input parameters to mantle convection simulations using observations of the temperature structure of these simulations.

As described in chapter 2, I use a prior sampling approach to find the posterior probability density functions for each model parameter, given an observation of the temperature structure. My prior samples are many different cases of StagYY, each run with different simulation input parameters. In order to investigate how the posterior varies with time for each StagYY input parameter, I take observations from StagYY after 0.4, 1, 2, 3 and 4.5 Gyr of simulated convection. However, when the initial investigation was being conducted in spring 2015, not enough simulations had completed 4.5 Gyr of run time. This was for a combination of reasons. Many simulations crashed due to computational instabilities, particularly associated with very high initial core temperatures. Some of these simulations completed a few million years before crashing, meaning there are more samples for which have run for 0.4 Gyr than 3 Gyr. Some combinations of input parameters also caused the simulations to run very slowly. Since the earlier simulations had to finish before new ones could be started, due to limited computing resources, I concentrated on getting a large number of simulations to a stable convective state rather than 4.5 Gyr. Subsequently, enough simulations have reached 4.5 Gyr so that I can use these simualtions to train my networks. Table 2.1 gives the number of simulations used at each time period.

The simulation input parameters are drawn from the prior ranges given in chapter 3. The temperature observations from these simulations are trans-

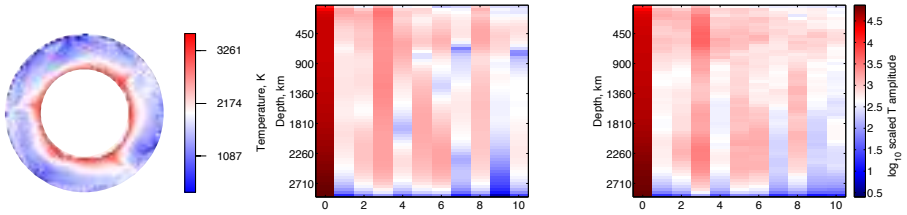
formed into the frequency domain, as in figure 3.1. At this point, after transformation, regularising noise is added, so that I can control its frequency. I vary the amplitude of the noise, to investigate its effects on the inferences and add it to both real and imaginary parts with the same amplitude at all wavelengths. I then find the amplitude spectra. I use the logarithm of the amplitude because there are large variations in the spectrum, particularly between deep and shallow mantle temperatures. Taking the logarithm prevents the networks from being overwhelmed by the higher magnitude parts of the spectrum. From the amplitude spectra, I keep degrees 0 to 10 in order to reduce the dimensionality of the network input. The dimensionality is then further reduced by using an auto-encoding neural network. The encoding process reduces the 64×11 elements of the amplitude spectrum to 28 discrete numbers in the encoded version. By trial and error I find that reduction to 28 dimensions retains enough information to preserve the original pattern, whilst being of sufficiently low dimensionality for the inversion process to succeed. The encoding is not lossless, with the loss being in the fine details of the amplitude spectrum, as can be seen in the example in figure 5.1. The spectrum is smoothed, but the broad patterns are retained.

The encoded temperature amplitude spectra are used to train mixture density networks, as described in chapter 2. Each network is trained to find the marginal posterior PDF for one model parameter from a given amplitude spectrum using a training set of observations and target simulation input parameters.

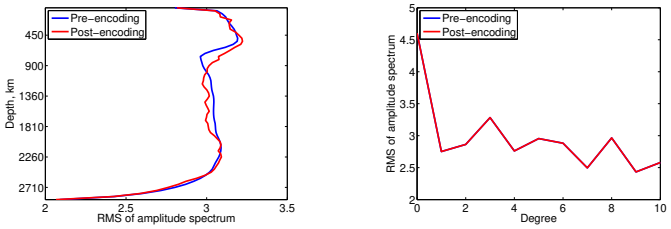
The details of the method are covered in more detail in chapter 2 and 3. They are summarised for the application in this chapter in figure 5.2.

5.2 Proof of concept

I assess the performance of each committee of networks after training by carrying out a series of synthetic tests. I use a separate test set of convection simulations, for which the simulation input parameters are known. These have not been used to update any part of the network at any point and are completely independent of training. Because they are independent, I can be confident that any positive results I see are derived from underlying physical relationships between observation and convection parameter, allowing us to test the generalised performance of the network. The convection input model parameters in the test set are all drawn independently and randomly from the same prior distributions as those in the training set. The number of simulations in



(a) Temperature (b) Amplitude spectra (c) Encoded amplitude spectra



(d) Smoothing effect with depth (e) Smoothing effect by degree

Figure 5.1: Example of the effects of using an auto-encoding neural network to reduce dimensionality. (a) the original temperature field; (b) the original amplitude spectrum for the temperature field for degree 0–10; (c) amplitude spectrum after encoding and decoding. Both spectra are on the same scale. The amplitude is scaled by the square root of number of samples. The encoding network is trained to reduce the original amplitude spectra from 64×11 points to a 28 dimensional representation. The same network can then decode the 28 dimensions back to a 64×11 spectrum, showing the possible loss of information in the encoding. The decoded amplitude spectrum is smoothed with respect to the original spectrum but retains all of the large scale features of the amplitude spectrum. The bottom row of figures shows the root-mean-square amplitude as a function of depth (d) and spectral degree (e), for the original and decoded spectra.

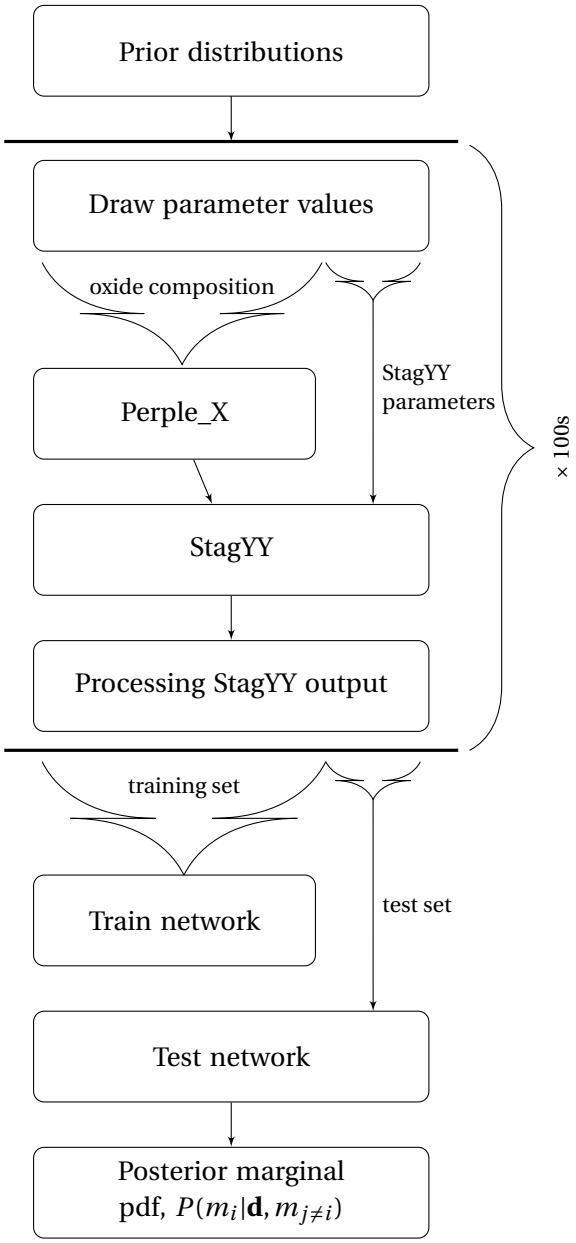


Figure 5.2: Workflow to train a network to infer a convection model parameter.

the training and test sets for each age group are given in table 2.1.

I use the Kullback-Leibler distance

$$D_{KL} = \int P(m_i) \log_2 \frac{P(m_i)}{P(m_i \mid \mathbf{d}, m_{j \neq i})} dm_i \tag{5.1}$$

to measure the change in entropy in bits between the marginal posterior probability distribution for the input parameter and the prior distribution for that input parameter (Johnson and Sinanović, 2000). If the network has learned to find patterns that can be used to infer the simulation input parameters, the network has gained information on the relationship between observation and input parameter. The more information the network has learned, the narrower the posterior PDF is, and therefore the smaller its entropy relative to the prior, giving a large D_{KL}. Figure 5.3 shows the Kullback-Leibler distance between a Gaussian mixture approximation to a uniform distribution ($P(m_i)$ in equation 5.1) and Gaussian distributions with decreasing standard deviation, $P(m_i \mid \mathbf{d}, m_{j \neq i})$. The networks are initialised to output a Gaussian mixture approximation to the prior distribution, which is uniform for most parameters. The D_{KL} between a Gaussian distribution with standard deviation of 0.62 and a standardised uniform distribution is 0.5.

I calculate the D_{KL} between the prior sample distribution of parameter values in the training set and the network calculated posterior distribution for each parameter at each time step. Figure 5.4 shows the mean D_{KL} across all the simulations in the test set. The convection simulation input parameters with the highest information gain is yield stress, the inference of which improves with time. The information gain for reference viscosity and the initial thickness of primordial material are also moderately high and stable with respect to run time. The information gain for the initial core mantle boundary and initial mantle temperature start high but decrease markedly with time.

I now return to the potential reasons why no signal for some simulation parameters is found, first raised in section 4.1. It may be that there is no information to be learnt for that input parameter from the observations shown to the network. Alternatively, a signal may be present in the temperature field, but the training set may not contain enough simulation samples to allow the network to find a mapping between observation and input parameter. Finally, there are an infinite number of ways in which to set up and train the neural networks, and I may simply have picked an approach which is not optimal for this case. It is impossible to distinguish these causes and I have no way to estimate how many samples I will need before training the network. I stress that although I do not find these parameters in this study they are not necessarily unknowable,

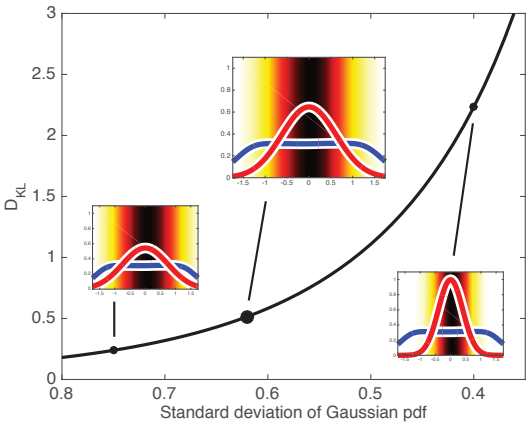


Figure 5.3: D_{KL} between a Gaussian mixture approximation to a uniform distribution and Gaussian distributions with varying standard deviation. The inset distributions show Gaussians with standard deviation of 0.75, 0.62 and 0.4 respectively (in red), with the Gaussian mixture distribution plotted in blue behind each. The D_{KL} is 0.24, 0.5 and 2.23 respectively. The colour in the background corresponds to those used in figure 5.7. The PDF maximum is coloured black in each case and the width of the transition from black to yellow shows the width of the PDF.

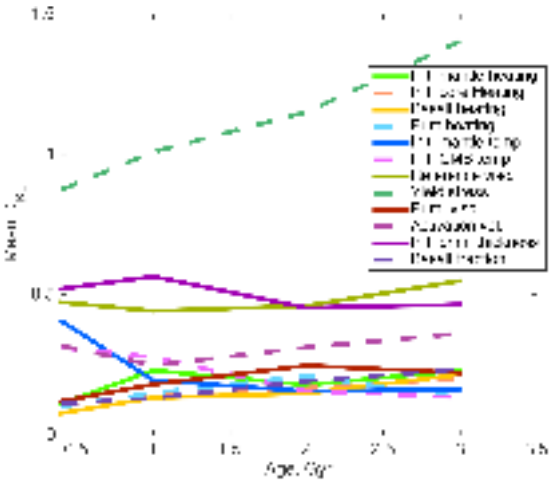


Figure 5.4: Mean D_{KL} between the prior distribution of parameter values in the training set and the network calculated posterior distribution for the test set simulations. Each point represents the outputs for one committee of networks, trained to find the particular model parameter.

and they may be recoverable with more training data or different observations. A null result in this study therefore can not be regarded as evidence that a particular parameter has no signature in present-day observables.

My neural network implementation produces a conservative estimate for the posterior PDF when compared to results produced by directly sampling from the posterior distribution by Monte Carlo methods. When the data points produced by sampling the prior model parameter space are not concentrated close to an observation that I am trying to invert, the interpolation between samples is over greater distances, increasing the uncertainty. With more samples, the D_{KL} would increase as the uncertainties introduced by interpolation decrease. In the case of too few samples, the inference simply returns the prior. More details on the comparison between inferences made by mixture density neural networks and Monte Carlo techniques can be found in Käufl et al. (2016). I have fewer samples at greater ages, therefore I would expect the D_{KL} to decrease with age, unless this is compensated by an increase of information in the data.

As discussed in chapter 2, there are other sources of uncertainty in the posterior probability density function. The uncertainty in the value of simulation input parameters is described by the prior and has a direct effect on the posterior PDF via Bayes' theorem. There are also uncertainties in the forward simulation process and in the observations. In this study, the training and test data are entirely synthetic, but to apply this method to real data, I would have to take into account the errors introduced by the assumptions implicit in StagYY in addition to shortcomings in our understanding of the physics of mantle convection, and errors and noise in the real data. If I can quantify these uncertainties, it is straightforward to include them in our method. During network training, noise can be added to the observations (Bishop, 1995; Käufl et al., 2014) encapsulating modelling and data uncertainties. Adding noise to the training data is similar to regularisation and has the effect of desensitising the networks (Bishop, 1995). With greater noise, the network is forced to find mappings using the features that vary the most between training observations. With smaller noise levels, the network is allowed to use smaller differences to distinguish between observations and is thus more sensitive to details in the data.

To investigate the influence of noise in the data, I train different networks with noise drawn from distributions with standard deviations of 10, 50 and 100 K. This noise is added when the temperature structure has been transformed into the frequency domain to both the real and imaginary parts of the structure, before any dimensionality reduction takes place. I do this so that I can

control the frequency of the noise and so that it does not get removed by the Fourier transformation. The noise is then included in the amplitude spectrum. The noise with the same amplitude is used at all depths. I do this because it reflects the amplitude of the lateral heterogeneities, although the magnitude of the noise therefore decreases as a percentage of the mean of the temperature with depth. If I were using real data, for instance seismic tomography, the noise could be varied both laterally and with depth to reflect different levels of knowledge in each region, as well as taking into account uncertainties in seismic tomographic modelling and the conversion of these models into temperature, density or composition.

The noise level that produces the highest mean information gain for the test set depends on the simulation input parameter of interest. Figure 5.5 shows the mean D_{KL} and an error measure as a function of different noise levels. There is very little difference in D_{KL} with noise level. The error measure is the mean difference between the maximum of the network inferred posterior PDF and the true model parameter value of each simulation in the test set in terms of the variance of prior distribution. This measure shows more variation with noise level. Whilst this measure gives an indication of success, it cannot be treated as more than an indication because the true value may still fall within the region of high likelihood, meaning the inference can be successful even if the peak of the PDF does not lie exactly at the desired point. In this paper all the presented results are for committees trained using Gaussian noise with a standard deviation of 50 K.

Whilst figure 5.4 gives the mean D_{KL} for all the simulations in the test set, examining the individual PDFs for the each member of the test set allows me to get more insight into the inversion, for example to look at how performance varies in different regions of the model space. Figure 5.7 show PDFs for the training sets for the best resolved parameters and one badly resolved parameter. The number of simulations in each test set are given in table 2.1. Each vertical line is a marginal posterior PDF for the input parameter of interest given the temperature amplitude spectra from one convection simulation. The vertical line is coloured according to the amplitude of the PDF. The y-axis is the value of the input parameter of interest, and therefore the colour of each point along the column gives $P(m_i|\mathbf{d})$ for $m_i = y$, normalised so that the maximum of each PDF is black, as shown in figure 5.6. The vertical line for the PDF is positioned along the x-axis according to the known value of this input parameter for that particular simulation. Figure 5.6 demonstrates how the PDFs for six test simulations are placed into the grids in figure 5.7. If the network effectively infers

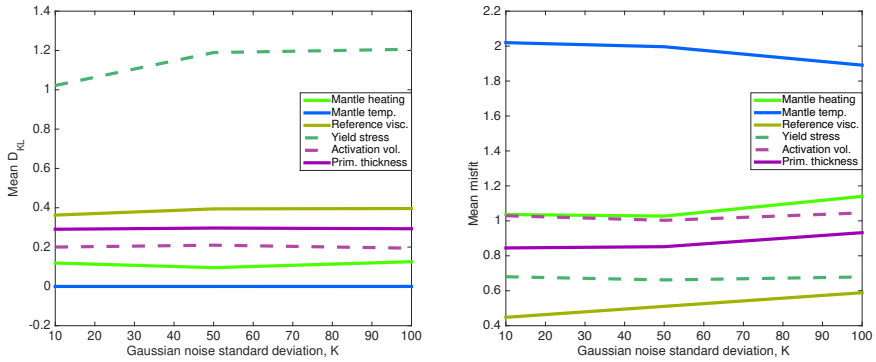


Figure 5.5: (a) mean D_{KL} with different levels of noise. (b) mean difference between PDF maximum and true simulation parameter for each simulation with different noise levels after 3 Gyr. The unit is the variance of the prior distribution of each parameter.

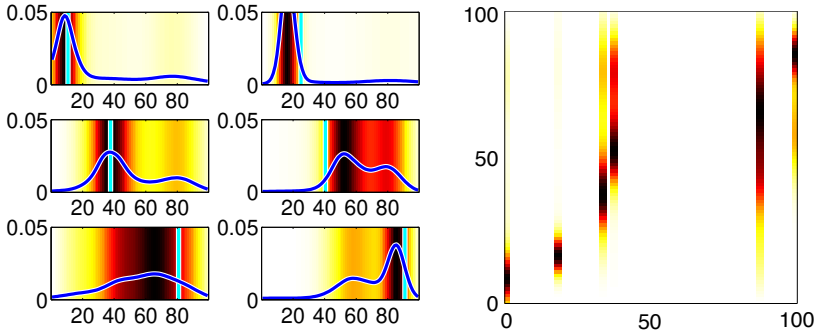


Figure 5.6: Some randomly selected examples for posterior PDFs inferring yield stress after 3 Gyr, taken from the test set of simulations. The six PDFs to the left are the committee output, coloured with the same colour scale as in the right-hand panel and figure 5.7. The colour scale is black at the maximum regardless of amplitude. The pale blue line indicates the true target value of yield stress for each simulation. Ideally, the maximum of the PDF should correspond to the target value. The target value is then used to align the coloured representations of these PDFs along the x-axis of a grid such as on the right-hand side. If the PDF maximum is close to the target values for all the PDFs in the test set, there will be a diagonal stripe of high amplitude across the grid.

the value of the input parameter for all the simulations in the test set, the diagram should have a diagonal trend of high PDF amplitudes running across it, as seen for instance in figure 5.7c. I also need to know how certain the networks are, therefore underneath each grid I plot the D_{KL} for each test set. The red line marks D_{KL} equal to 0.5, corresponding approximately to a posterior PDF with standard deviation of 0.62 compared to a standardised uniform distribution, as shown in figure 5.3. Cases with a D_{KL} of 0.5 or over show a significant improvement on the prior distribution. A D_{KL} below 0.5 does not mean that the network has learnt nothing, but simply that the uncertainty of the prediction is higher. The PDFs for such cases should be considered before rejection.

The values of the input parameters for the simulations which make up the test set are also drawn randomly from the prior distributions. They are therefore not evenly distributed across the prior space, leaving gaps in the diagrams. For some parameters (e.g. initial CMB temperature), the prior is skewed because some ranges of values cause the simulations to become computationally unstable or run very slowly and therefore the outputs from simulations occupying these regions of model space are missing.

Because all 29 parameters vary at once in all the test simulations, the marginal PDF includes all the trade-offs between model parameters in its width.

The most successful inference is for yield stress, particularly at low values. The PDFs produced by the networks, shown in figures 5.7a to 5.7d, are narrow and high with peaks that correspond to the true value of yield stress used to run each simulation. Networks inverting temperature patterns for reference viscosity also perform reasonably well. In general, there are few under- or overestimates for either yield stress or viscosity, and the differences between the maximum of the PDF and the true model parameter value are not large and certainly within one standard deviation. For yield stress, the majority of the simulations in the test sets are predicted with a D_{KL} over 0.5. The networks find yield stress with much lower uncertainty for low yield stress simulations, which have a much higher D_{KL}. About half of the inferences for reference viscosity show a D_{KL} over 0.5. The appearance of bi-modality at high yield stresses is probably an artefact resulting from the parameterisation of the posterior using Gaussian kernels. This is generally how distributions which are close to uniform over a particular range appear when parameterised in this way. Yield stress and viscosity determine whether tectonic plates form and the vigour of convection, therefore it is not surprising that I can make inferences about these parameters from the temperature field. I discuss this further in section 5.3.

The thickness of primordial material can be determined from the temper-

ature field after 0.4 Gyr (figure 5.7i). After 3 Gyr, the network still manages to categorise, mostly correctly, whether models have an initially thin, medium or thick layer, but the uncertainty is greater (figure 5.7l).

The inversions for initial mantle temperature are unsuccessful for any time step after 0.4 Gyr, but they demonstrate what happens when the network learns nothing about a parameter. For figure 5.7p, the networks return an approximation to the prior distribution using Gaussian kernels. The D_{KL} is non-zero here simply because the prior distribution of parameters in the training set is not perfectly smooth, but the difference between the prior and posterior distributions are very small.

Extending the investigation up to 4.5 Gyr

Due to limited computational resources, at the time of writing the paper Atkins et al. (2016), I only had enough simulations to investigate simulations which had run for up to 3 Gyr. We hypothesised in the paper that the results would be very similar for simulations which had completed 4.5 Gyr of run time. here I show that that is indeed the case, as well as investigating one extra parameter not covered in our paper.

After 4.5 Gyr, the signal for yield stress is somewhat reduced (figure 5.8, in comparison with 5.7d). For simulations with low yield stress, I am still able to find yield stress reliably. After 4.5 Gyr, 45% of the simulations have stagnated, with surface velocities $\leq$ 1 cm/yr. This is comparable to the number of stagnated models at 3 Gyr. However, the mean velocity at the surface has halved as the convection slows and more simulations have entered a stable, near stagnant convection state, making them indistinguishable. This is due to the extra 1.5 Gyr cooling time. For lower yield stress simulations, where mobile lids are maintained for longer, there is still enough variation in temperature structure to be able to precisely distinguish the value of yield stress used. For higher yield stress simulations, the networks identify that they are high yield stress, but merely provides a near uniform distribution in the range 50-100 MPa and does not differentiate within this range. This still provides some constraint, although it is less useful.

The signal of the initial thickness of primordial material was very weak after 3 Gyr (figure 5.7l). After 4.5 Gyr, I cannot infer the initial thickness of primordial material using observations of the temperature structure. The networks simply return the prior distribution when presented with the test set of observations (figure 5.9).

I have more success inferring the resulting thickness of the primordial ma-

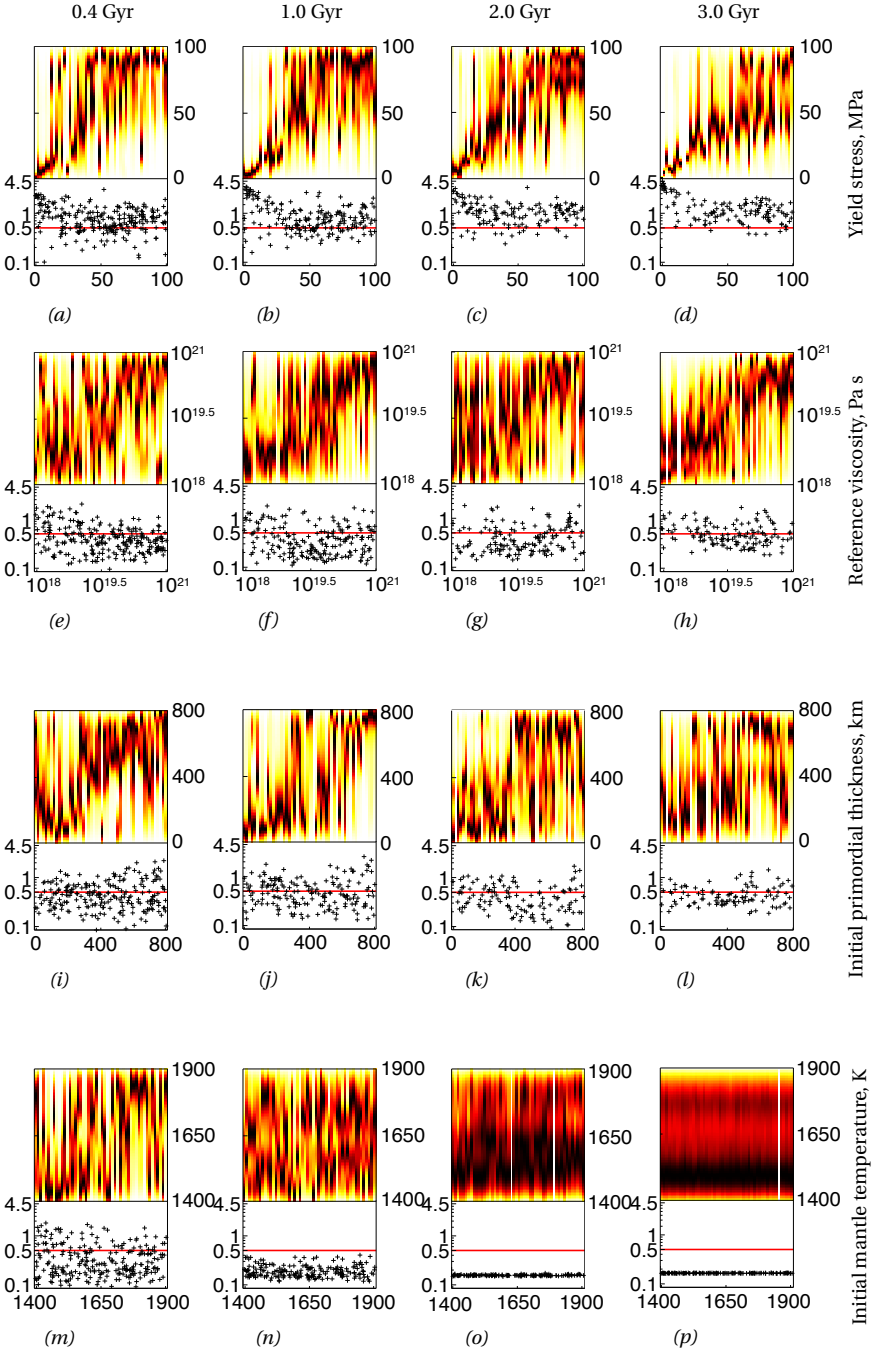

0.4 Gyr
1.0 Gyr
2.0 Gyr
3.0 Gyr
Yield stress, MPa
Reference viscosity, Pa s
Initial primordial thickness, km
Initial mantle temperature, K
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)
(o)
(p)

Figure 5.7

Figure 5.7: PDFs for the test set of simulations at each age which provide an independent demonstration of network performance. In the coloured grids, each vertical column is one posterior PDF for the relationship between the temperature structure of a single simulation and the model parameter given on the left. The column is positioned along the x-axis according to the true value of the model parameter. The colour scale gives $P(m_i|\mathbf{d})$, where $m_i = y$ for each value of the model parameter ranging along the y-axis. The colour scale is set so that the maximum of each PDF is black. See figure 5.6 for a demonstration of how to interpret these figures. The D_{KL} for each simulation is plotted below the coloured grid on a $\log_{10}$ scale. PDFs with a D_{KL} above 0.5 (red line) indicate that the network has learnt a significant amount of information on that model parameter, which corresponds to a Gaussian distribution with standard deviation of approximately 0.62, as shown in figure 5.3. A lower D_{KL} simply indicates greater uncertainty. The D_{KL} values for all the test simulations are plotted, but to improve clarity for the PDFs, the test simulations are binned according to input parameter value and one PDF from each bin is chosen randomly.

terial after it has been subjected to 4.5 Gyr of convection. This is a parameter which was not investigated for the earlier time steps. The primordial material begins as a homogeneous flat layer at the base of the mantle. During convection, it can be swept around into heaps or piles, or diffuse into the rest of the mantle. After running the simulation, I consider any cells which are more than 80% primordial material to be remnants of the original homogeneous layer. I then calculate mean depth over which cells contain 80% or more primordial material. The inference for this parameter is not particularly strong (figure 5.10), comparable to the inference of the initial thickness of primordial material after 3 Gyr (figure 5.7l). The network predicts zero remaining primordial material by default, giving PDFs with peaks at very low values. Simulations for which the network does make an inference generally have the remaining thickness inferred reasonably accurately, although with quite large uncertainty.

The primordial material cannot be lost or melted and therefore stays in the mantle. It can be dispersed and mixed in with the pyrolitic material above it. The example simulations in section 3.3 show how primordial material can be moved around by convection. In some cases, it does this, leaving very few cells with more than 80% primordial material. In most cases, the primordial is denser and more viscous than the overlying primordial material. This means that only a small amount diffuses into the rest of the mantle. The mean depth of cells which contain 80% or more primordial material is therefore strongly dependent on the initial depth of cells which contained 100% primordial (the initial thickness of the primordial layer).

The inference for reference viscosity does not decay further with time, maintaining a strong signal at all values (figure 5.11).

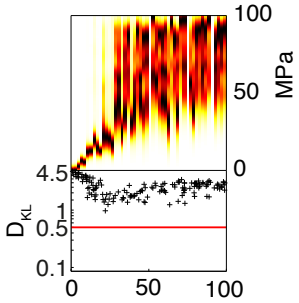


Figure 5.8: Yield stress inferred from the amplitude spectra of the temperature field after 4.5 Gyr of run time.

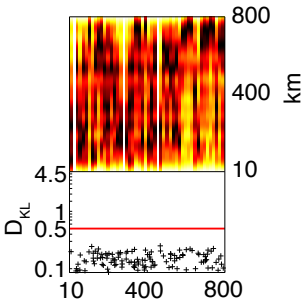


Figure 5.9: Initial thickness of primordial material inferred from the amplitude spectra of the temperature field after 4.5 Gyr of run time. The signal for this parameter is lost between 3 and 4.5 Gyr.

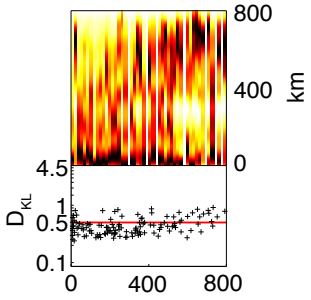


Figure 5.10: Mean resultant thickness of primordial material after 4.5 Gyr of convection inferred from the amplitude spectra of the temperature field after 4.5 Gyr of run time.

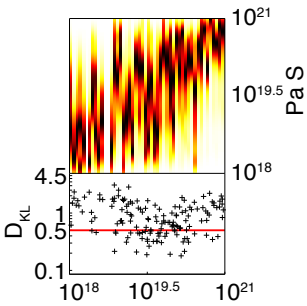


Figure 5.11: Reference viscosity inferred from the amplitude spectra of the temperature field after 4.5 Gyr of run time.

5.3 Discussion

Modelling Earth-like convection relies on poorly constrained estimates for many key input parameters, which include both initial conditions and constant physical parameters which appear in the equations of mantle convection. In this chapter I present a new method which allows us to invert the thermal structure of mantle simulations at some time steps for convection parameters. I find that I can invert for yield stress, reference viscosity (both constant physical parameters) and initial thickness of primordial material. Whilst there are currently other methods to estimate these values for the Earth, my method is novel in both its self consistency and its use of a static observation of convection. Whilst I cannot adequately recover other parameters in this study, they may be recoverable by using different observations or with larger training sets. This method is not perfect, but it shows a significant improvement compared to the methods presented in chapter 4, which leads me to conclude that it is a success as a proof of concept. The limitations and results seen here are discussed in more detail below.

Yield stress is the best constrained parameter when inverting the amplitude spectrum of the temperature field. It is particularly well inferred at low yield stresses, where the prediction is accurate with low uncertainty. The yield stress parameter determines how much stress the material can withstand before it begins to undergo plastic or brittle deformation. If the lithosphere is weak enough, relative to convective stresses, it will yield forming a mobile-lid regime. The yield stress has been observed in many previous studies to be the major factor in determining whether a planet has a stagnant lid or evolves a mobile lid (e.g. Moresi and Solomatov, 1998; Valencia et al., 2007; van Heck and Tackley, 2011; Lenardic and Crowley, 2012), and when continents are present, the strength is a factor in determining the wave-length of convective flow (e.g. Zhong et al., 2007; Rolf et al., 2014)

If I define a mobile lid to have a mean surface velocity of > 1cm/yr, as in Lourenço et al. (2016), approximately 50% of my simulations are in a mobile lid regime at each time step. However, I do not explicitly provide the networks with any information about plate velocity, therefore they can only identify that the simulations are in a stagnant or mobile regime by finding the relevant patterns in the temperature spectra. Similarly, the history of the crust, whether it is stable or has changed regimes during its evolution, is dependent on the input parameters. The temperature structure is dependent on this history. I provide the networks with no explicit historic data, so if they require any information

on the history in order to make inferences about the rheology, they must be finding it from within a snapshot of a single time step.

Figure 5.12 shows thirty randomly selected simulations which have run for 3 Gyr, grouped according to the yield stress, but with all other parameters varying randomly. Whilst the sample is quite small, there is a pattern from low to high yield stress (left to right). Almost all of my simulations with very low yield stress (0-20 MPa) have lower than average mid-mantle temperature, and the reverse is true for very high yield stress simulations (79-99 MPa). The low yield stress simulations also have larger lateral temperature variations, with heterogeneity patterns which saturate the colour map in figure 5.12, and narrower, more distinct upwellings extracting heat more efficiently, leading to the observed cooler mid-mantle. The network is probably using these differences to classify the simulations into low or high yield stress, and the large lateral variations in the low yield stress simulations are why the networks infer the yield stress with such low uncertainty at low values. How they are separating the mid-range simulations is less clear, but demonstrates how neural networks can pick out subtle relationships.

The difference in inference quality between high and low yield stress simulations is a very similar result to that seen when using just the mean mantle temperature in figure 4.1, suggesting that the mean mantle temperature is the part of the observation that the networks are relying on the most to make their inferences. The networks do however still produce better results than the simple linear analysis presented in figure 4.1.

The ability of the networks to find yield stress is probably also enhanced by the low temperature dependence of viscosity used in my simulations. This reduces the variations in viscous stress that would be caused by the near-surface temperature variations resulting from blanketing by the crust (e.g Rolf et al., 2012; Heron and Lowman, 2014). The lithosphere has the same strength throughout my simulations. The evolution of the atmosphere, and therefore the addition of water to the crust may reduce yield stress with time (Valencia et al., 2007). The way in which a planet's lithosphere evolves to reach a particular strength may also determine its tectonic state, as much as the final strength (Lenardic and Crowley, 2012; Weller et al., 2015), introducing further complications and trade-offs. I do not attempt to replicate any of these processes here, but could do with more complex simulations. However, if I were to use more complex simulations in my training set, the trade-offs would simply be represented in the width of the marginal PDFs.

Viscosity is another important factor controlling the patterns of mantle con-

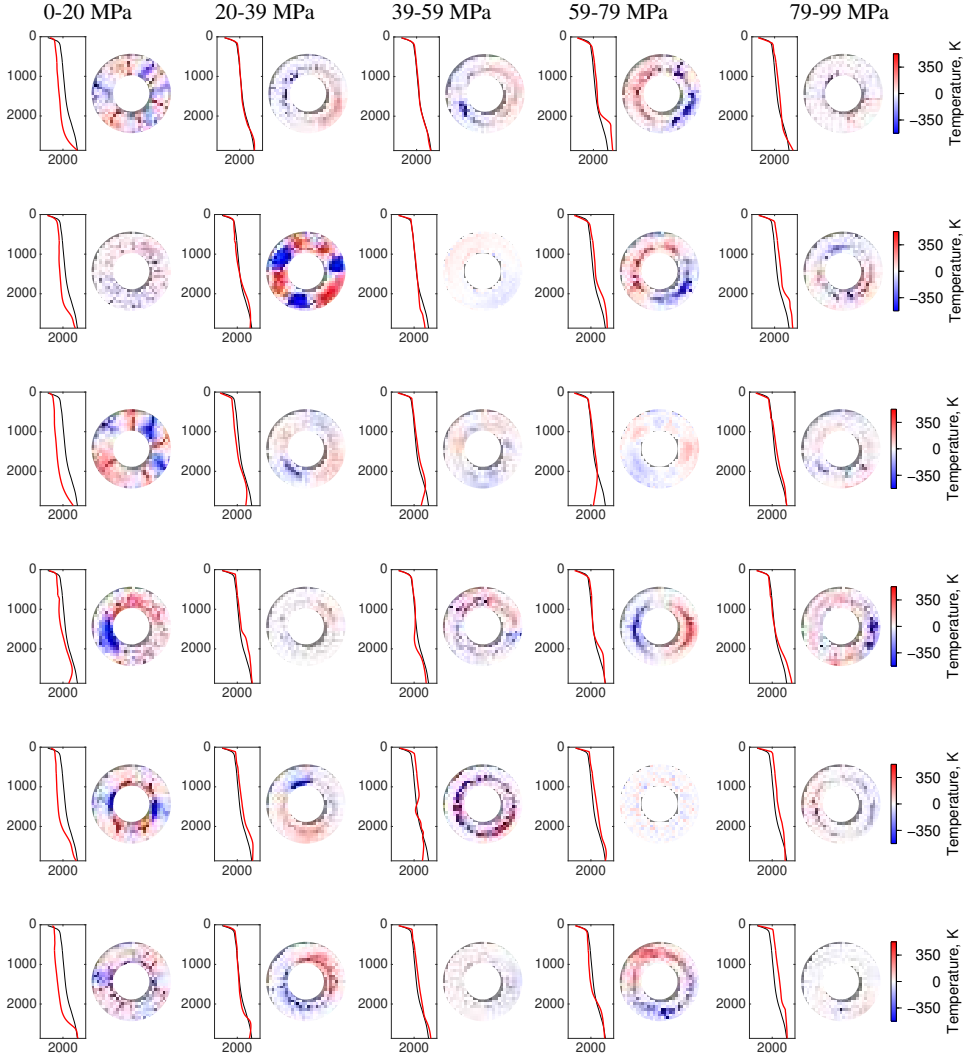


Figure 5.12: Temperature structure for 30 simulations after 3 Gyr, grouped into columns according to yield stress used. On the left for each simulation is the temperature profile for the simulation in red. The mean profile for the whole group of simulations is plotted in black to aid comparison, and is identical in each case. The lateral variation from the 2-D mean is plotted on the right to highlight the convection patterns. The same colour scale is used for all simulations.

vection. Viscosity is exponentially dependent on temperature, so small lateral temperature variations can have large effects on viscosity. The original estimate of 10^{21} Pa s by Haskell (1935) is still considered valid as an approximate average mantle value, although newer studies, (e.g. Whitehouse et al., 2012; Argus et al., 2014) include much more complex lateral and radial variations. There is expected to be a viscosity jump of at least an order of magnitude around the transition zone, although the size and location of the jump varies between studies. However, these viscosity models inherit uncertainties from climate history and sea level models, and large uncertainties when converting seismic velocities to temperature or density. In my convection simulations, viscosity has a clearly identifiable and quantifiable effect on the temperature variations within the mantle. Using my approach, I can estimate the order of magnitude of the reference viscosity directly from a single observation with a reasonable degree of certainty. This method may therefore provide a more direct method for inferring mantle viscosity in the future.

In this study, I only varied the viscosity prefactor and pressure dependence (η_0 and V_{η_0} in equations 3.7 and 3.8). However, Yoshida (2008) found that the temperature dependence of viscosity can determine the wavelength of convection, although lithospheric yield strength was found to be the dominant factor. High temperature dependence increases the chance of a stagnant lid regime because a higher viscosity contrast promotes decoupling in upper mantle, while increasing pressure dependent viscosity promotes mobile lids, because it increases the convective stresses exerted by the mantle (Stein et al., 2013). The magnitude of a mid-mantle viscosity jump also affects the convection pattern (e.g. Davaille, 1999; Lowman et al., 2011), which I neglect here. If I were to vary more viscosity parameters in my training simulations, such as the temperature dependence (E_η), it is therefore possible given the presented results that I may be able to invert for them using the patterns produced by convection. However, more complex viscosity dependence may equally well just introduce more trade-offs, increasing the width of the posterior PDFs.

The presence of primordial material at the base of the mantle has also been observed to affect convection patterns and even to lead to stagnation (e.g. McNamara and Zhong, 2004; Nakagawa and Tackley, 2008; Deschamps et al., 2011; Stamenković et al., 2012; Trim et al., 2014). The community is currently divided about the existence of dense material at the base of the mantle and estimates of the lifespan, stability and origin of such material vary wildly. My networks only give an approximate estimate with large uncertainties (figure 5.7l) for the initial thickness of primordial material when the networks are trained on the temper-

ature patterns taken from the mantle convection simulations after 3 Gyr of run time. Even an estimate for an initially thin, medium or thick layer is a significant improvement on current knowledge, especially if my method also works for three dimensional cases over 4.5 Gyr. It is also surprising that I can identify a primordial layer using only the temperature field, since the concentration of heat producing elements in the primordial material is not successfully found by the network. The dense material must therefore affect the temperature distribution throughout the mantle since the networks are not simply identifying a hot, highly radioactive layer at the base of the mantle.

The advantage of investigating primordial material properties in this way is that no extra data are required because the simple patterns contain the information. Whilst I invert temperature structure here, which is imperfectly known for the Earth, I could use other more direct mantle observations such as seismic tomography to train my networks to identify signs of primordial material. Current methods to investigate anomalous material at the base of the mantle require either imperfect relationships between seismic velocity and chemical properties, or time dependent data such as the location of subduction zones which push dense material around into the desired locations (e.g. McNamara and Zhong, 2005; Bull et al., 2009; Steinberger and Torsvik, 2012). Using the spectra of thermal heterogeneities, as demonstrated here, or seismic heterogeneities therefore simplifies the inversion and removes some sources of uncertainty which are present in existing models.

I mentioned the existence of possible trade-offs between parameters. Whilst this is a problem in more classical approaches where only vary a few parameters are varied at a time, my results implicitly contain all information on the trade-offs within my chosen range for the input parameters. My networks return marginal probability density functions for a given parameter for a training set where all other parameters have changed as well. The width of the marginals therefore contain all the possible trade-offs. The trade-offs can also mean that the inferences are a long way from the true values. A particular example is in figure 5.7i, where one simulation which was initialised with around 400 km of primordial material is inferred to be most likely to have a very small amount of primordial material after 0.4 Gyr. The recovered PDF still encompasses the true value, although it is given a low probability. Within a probabilistic approach, this need not necessarily be regarded as a failure: the true value is explicitly included in the range of possibilities compatible with observations. However, the network regards other explanations for the observation as 'more likely', given the training information it has received. The 1-D marginal alone

does not inform us on the nature of the trade-offs, but this could easily be investigated by using higher dimensional marginals as for instance in de Wit et al. (2013).

There are several reasons why my networks may not be able to constrain the other model parameters varying in table 3.1. I are only using the encoded amplitude spectra for degrees 0 to 10 to train the networks. This removes much of the fine-scale variation in the temperature field, and means that I discard all the phase information, therefore losing all the details about how variations are spaced relative to one another. Some parameters may have more pronounced effects in the small wavelength variations in the spectra. These unresolved parameters may also only have very small effects which I could observe if I were to use much larger networks and with many more training sets. The networks may then be able to recognise the very small changes caused by these parameters, which are currently below the noise level. However, larger networks with more input dimensions are harder to train and are less stable, given that I only have small training sets.

I have also tried to train networks using only the radial mean temperature structure (degree 0 of the amplitude spectrum). This was significantly less successful than using degrees 0 to 10, implying that most of the signal of the parameters is contained in finer details of the patterns of convection, rather than mean temperature profile.

I tried to train networks to identify the molar percentage of iron oxide used in each constituent rock type but without success. The oxide composition of MORB has previously been found to affect compositional stratification in the transition zone and segregation at the CMB (Nakagawa et al., 2010), and the iron oxide concentration in primordial material affects density and therefore the shape and stability of primordial layers (e.g. Deschamps et al., 2012). My lack of success is probably because the temperature is unlikely to be the best observation from which to identify mantle chemical properties. My investigations were also very preliminary and I may have more success by using more subtle targets, such as oxide ratios or the presence of particular mineral phases, rather than bulk composition.

Here, I consider purely synthetic data sets and can therefore use the temperature structure of the mantle. If I were to apply this method to real data, I would have to rely on conversions from seismic velocity anomalies to temperature. For real data, it would be better to train my networks using the patterns of seismic heterogeneities. I can easily calculate P- and S-wave velocities for my convection simulations, because the mineral physics calculations include

the elastic parameters, meaning that no approximation is necessary to go from temperature to velocity. This would add additional uncertainties from mineral physics into my inversions, but these can be accounted for during network training. However, this study is a proof of concept, and other simplifications remain, including that my synthetic data are two-dimensional approximations to a three-dimensional Earth. I therefore currently use temperature observations as a first step to show that the simple patterns produced by convection do indeed contain information on these parameters.

For many parameters, the temperature structure is unlikely to be the best mantle observation from which to make inferences, even when inverting synthetic cases, because the temperature structure is not directly dependent on composition. It is already surprising that I can find the initial primordial layer thickness, which is a purely compositional parameter, from the temperature field. If I were to use an observation which is dependent on composition, such as density, seismic velocity, gravity, erupted basalt composition observed at the crust, or even mantle composition directly, I expect to be able to resolve the compositional parameters such as primordial thickness and basalt fraction much better.

One of the other parameters for which I expect to be able to make inferences is the viscosity contrast of primordial material with respect to the overlying pyrolitic mantle. In previous studies, (e.g. Davaille, 1999; McNamara and Zhong, 2004; Deschamps et al., 2011) the viscosity contrast was seen to determine the shape of any piles or ridges formed at the base of the mantle. I therefore expect to be able to find the viscosity contrast if I use a compositionally-dependent pattern to train my network. Using composition and temperature together may allow us to determine the relative variation of radiogenic element composition between different materials. This is one advantage of my sampling approach: in the future, I can use the same suite of forward simulations to investigate whether these parameters leave signals in other observables, without needing to run more forward simulations.

I also experimented with various neural network architectures and configurations, changing the number of Gaussian kernels, the number of hidden layers and the number of networks in the committee. Changing these made very little difference to the inferences, although in generally larger networks tended to perform less well. Since larger networks contain more free parameters that must be determined during learning, this is unsurprising given the limited amount of training data available to us.

5.4 Conclusion

As a proof of concept, this investigation demonstrates that the thermal structure of the mantle does contain information on the simulation input parameters, and that neural networks can be used to access this information. There are many limitations to this method, not least that the thermal structure of the mantle is poorly constrained for the Earth. However, by showing that it is possible with some simple cases, I demonstrate that further study may lead to a more refined method which can be applied to the Earth. Some refinements would require more complex convection simulations, but others simply require different observations or ways of interpreting the data, which shall be explored in the following chapters.

6

Developing neural network inversions and exploring different observations

In the previous chapter, I demonstrate that the thermal structure of the mantle contains enough information on convection simulation input parameters that they can be inferred using neural networks. The thermal structure of the mantle is a first order representation of the convective system: hot regions are upwelling, cold regions sinking. However, the thermal structure of the Earth cannot be observed directly and is very poorly constrained. The processes through which an approximation of the temperature structure of the mantle could be derived introduce a great many uncertainties. If I can find other observations of the Earth which contain information on these simulation input parameters, I may be able to reduce the uncertainties introduced by using the temperature structure. The need to use derived observations such as temperature of seismic

tomographic images may even be entirely removed.

There are many other observations of the Earth which I could potentially use to investigate the convective processes happening in the mantle. Ideally, I would use direct observations, such as seismograms and GPS recorded plate velocities. These have been through fewer preprocessing stages than derived observations such as seismic tomographic images so contain fewer uncertainties. There are also other derived observations, besides the temperature structure of the mantle, such as seismic tomography or density. Many of these are no better constrained that the thermal structure, but by using a combination of derived and direct observations they can support each other and possibly provide more insight into parameters.

In this chapter, I test how well my neural networks can make inferences using density, seismic velocity and surface plate motion as additional observations. I also investigate whether the neural networks can derive their own observations, for example trying to find temperature from seismic velocity and then viscosity from temperature, even when other parameters such as composition remain unknown.

6.1 Inversions using density

The first observation I use in place of the thermal structure of the mantle is density. This requires a slightly different method for preprocessing. For the density structure, I again consider the amplitude spectra, up to degree 10. For the thermal spectra, I used the logarithm spectra because the variations between different degrees in the temperature spectra were very large. The smaller scale variations then did not get swamped by much larger signals when I showed it to my networks. This is not necessary for the density spectra, as the amplitude varies much less between degrees and at different depths.

I use an auto-encoding neural network to reduce the dimensionality of the observation, although I must train a new auto-encoding network for density because the features are somewhat different. The density spectra has noise added with magnitude 12.5 kg m^{-3}. The noise is fixed for all depths, but since density increases with depth, the ratio of noise to the mean decreases with depth. However, the magnitude of the lateral variations in density is similar at all depths, with mean maximum variations of around 125 kg m^{-3}. The noise is added in the frequency domain to ensure that it does not get removed by the Fourier transformation.

Using observations of the density structure of my simulations, I can repeat

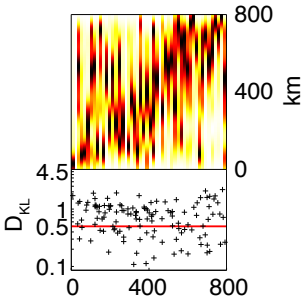


Figure 6.1: Initial thickness of primordial material inferred fromthe amplitude spectra of the density field after 4.5 Gyr of run time.

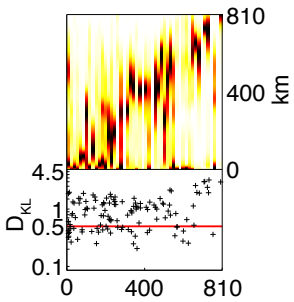


Figure 6.2: Thickness of primordial material after 4.5 Gyr of convection inferred from the amplitude spectra of the density field after 4.5 Gyr of run time.

the experiments to see if I can find the parameters better using a different observation. For the rheological parameters, viscosity and yield stress, the inferences made using density as an observation are very similar to those made using temperature. The main signal for these parameters is in the length scale of the convection cells and mean mantle temperature. The convection cells are thermal and since temperature is a contributing factor to lateral density anomalies, the networks will be using density as a proxy for temperature.

Using density as an observation does significantly improve the resolution of the primordial material, as seen in figures 6.1 and 6.2. All possible compositions for the primordial material are more iron-rich than the overlying pyrolitic mantle and therefore denser. This could make the change in rock-type obvious when studying the density field. However, the higher iron content can also mean that the increase in density reduces entrainment, keeping the material at the base of the mantle until it heats up and is hotter than the overlying mantle, making it nearly neutrally buoyant without a clear density contrast. The density jump is not significant enough to significantly alter the patterns of convection, which is why I cannot find the primordial material so well using temperature observations.

6.2 Inversions using density and temperature together

Lateral variation in density is a function of both temperature and composition. If I train the networks by giving them both the density and the temperature structure at the same time, I would expect that the networks can separate the

thermal and chemical contribution to density. I would therefore expect that they could better find the input parameters to the convection simulations.

To do this I use the density and temperature structures which have been pre-processed separately, using two separate sets of auto-encoders. I then present the auto-encoded patterns for both temperature and density to the networks together. The input layer therefore has 56 nodes, doubling the size of the network. I keep the range of the number of hidden nodes the same.

However, for the parameters considered in the previous section, there is no improvement in resolution of the inferences when using temperature and density together. For the thickness of primordial material, both the initial thickness and the residual thickness after 4.5 Gyr of convection, the temperature structure appears to contain little or no information, so adding it to the density simply increases the size of networks, making them less stable without contributing any extra information. In the previous chapter, I could find the initial thickness of the primordial material up to 3 Gyr using the temperature structure, but not beyond.

Yield stress and viscosity are not compositionally dependent in my simulations, therefore adding compositionally dependence does not help constrain these parameters. The signal for these parameters is in the mean mantle temperature and the shape of the convection. The shape of the convection cells can be seen equally well in the density or temperature spectra, therefore using both together does not provide any extra information.

By considering both density and temperature together, I am doubling the number of input nodes of my networks without increasing the number of training simulations. The potential improvement by using extra information is therefore be outweighed by the increased network size which is not able to train well enough to make use of this information.

6.3 Including surface velocity as an extra observation

As well as recording the density and temperature structure of the mantle, StagYY records the mean particle velocity in the cells. I can therefore extract the velocity with which the crust moves at each time step. This is analogous to plate tectonic velocity. The geological record shows changes in plate velocity, which may be due to plate boundary forces, but could also be due to instabilities in the mantle (e.g. King et al., 2002; Zhang et al., 2010; Lenardic and Crowley, 2012). The plate motion is closely linked to mantle convection (e.g. Hager and O'Connell, 1981), and therefore Rayleigh number, which is dependent on the

convection simulation input parameters.

I therefore investigate whether using mean plate velocity as an extra input to my networks improves the inferences when compared with the inferences made using temperature and density, which include no information about the rate of movement in the simulations. However, the results are no different to when I use just the temperature or density structure. This may be because the networks get all the information they need about the mobility of the simulations from the lateral heterogeneities in the temperature or density structure. The wavelength and magnitude of these heterogeneities are proxies for the convective vigour, which partially determines the velocity of the surface and therefore providing the actual velocity is unnecessary. If I use simulations with more complex crustal representations in the future, for example including continents, the plate velocity may prove to be of more importance.

6.4 Removing the time dependence of the inversion

In this and the next section, I use a slightly different type of observation for my inversions. The observations are different in two ways. Firstly, I mix together observations taken from different time steps in the simulations in the training set, rather than using only observations from a single time step. The training set can therefore include mantle structures from simulations that are 1 Gyr and 4.5 Gyr old at the same time, rather than all the observations being from 4.5 Gyr old simulations. Secondly, each network is trained on observations taken from a single depth slice in the simulations (i.e. one row in figure 3.1), rather than the whole mantle structure.

I investigate the effects of using observations of different ages together for two reasons. Firstly, I want to investigate how much the networks rely on knowing the age of the simulations. If all of the training observations are the same age, any difference between simulations is due to StagYY input parameters, rather than run length. This is especially true for observations which are strongly affected by temperature (including the temperature structure itself), which is inherently time dependent through conduction and radioactive decay. However, whilst simplifying the inversion for synthetic cases, relying on the fact that all of the training simulations being exactly the same age actually adds an extra source of uncertainty if I were to use real observations. This is because whilst the simulations use fully dimensional time in their calculations, a run length of 4.5 Gyr does not necessarily capture what would happen in 4.5 Gyr of Earth history. For example, my simulations begin with a solid mantle and no magma

ocean, therefore they skip the first few thousands to million years of Earth history. The way I initiate convectional instabilities, using small thermal perturbations, is also very slow. Generally it takes 0.5-1.5 Gyr for subduction to begin because these perturbations take time to grow into large enough anomalies to significantly influence the buoyancy. Because of this, I use observations from every time step after 1 Gyr, by which point the simulations are generally convecting. By using observations from different stages in the simulations, I therefore take into account the uncertainty in the difference between dimensional run time and Earth's evolutionary time.

The other reason for mixing observations of different ages is because it instantly increases the size of my data set. Each simulation goes through many time steps, all of which I can use. By doing this, I do not cover any more of the model space, and actually add another dimension (age), thereby increasing the dimensionality of the model space, but I do increase the number of samples in the data space significantly. If age turns out to be a not particularly important factor in the development of mantle structures, which could well be the case after the simulations have reached a stable convective state, I may be able to overcome some of the problems introduced by having a very limited data set.

Instead of using observations of the entire mantle, I use observations taken from a single depth slice. This means that the results are not directly comparable to those presented earlier. I do this because it removes the need for an auto-encoding neural network, thereby removing another source of uncertainty. Rather than encoding the amplitude spectra, I take degrees 0-10 for a single depth, giving me 11 observations for each simulation. This is a low-dimensional enough data space to no longer require encoding. By using a single depth, I keep the pressure constant for all observations.

I also change the noise model used to regularise these networks. In this case, I use Gaussian noise with standard deviation which is 5% of the mean value of whichever observation (temperature, density, seismic velocity) for whichever simulation I am considering.

Using this modified approach, I train the networks to find StagYY input parameters. I train networks at three different depths (564, 1468 and 2000 km) using observations of temperature, density and S-wave velocity (figure 6.3). The results from these networks are no better than those presented in the previous chapter.

Yet again, yield stress is the only accurately determined parameter. The yield stress is found best from the amplitude spectra of the temperature field (figure 6.3, top row), with only a very weak signal in the S-wave velocity or den-

sity observations. The accuracy of the inversion is comparable to those carried out at set time intervals using the whole amplitude spectra, as in chapter 5. I also train networks using just degrees 1–10 of the amplitude spectra of the temperature field. Degrees 1–10 capture the magnitude of lateral variations in the temperature, but do not include the mean temperature from which these variations deviate. The PDFs for the test sets using these networks are shown in figure 6.4. The inferences which include the mean temperature (figure 6.3) are significantly better than those using just the heterogeneity magnitude (figure 6.4), although there is a small signal for yield stress shown in the test sets for the networks trained at 564 km, but again only for simulations with low yield stress. The networks do not find much signal for yield stress when trained on observations of density or S-wave velocity, despite including the mean for both of these observations. This is probably because whilst these two observations are strongly influenced by temperature, they are also affected by compositional parameters, which obscure the temperature difference between simulations, which is where the yield stress has most effect.

By considering the results in this section and chapters 4 and 5, I can draw several conclusions about the effects of yield stress on the temperature structure of the mantle. Firstly, as seen from the linear correlation in figure 4.1, yield stress is directly correlated to mean mantle temperature, but only for low yield stress simulations. Secondly, the networks do not need to know the whole mantle temperature or the variation in temperature with depth, the relative difference in temperature at any depth between the sample is enough. This is shown by the similarity between results using the entire mantle versus results using a single depth slice.

The cooling caused by low yield stress probably happens very early in the simulations and then the temperature remains more or less constant, given that the same patterns are seen in the test set PDFs for mixed age simulations as for 4.5 Gyr simulations. The mixed age simulations are all more than 1 Gyr old, so the cooling must happen before this. This is in agreement with the results presented in chapter 5 (figure 5.7), where the low yield stress simulations always have their yield stress well inferred, regardless of the age of the observation. It is somewhat disappointing that using more samples did not improve the quality of the inference for yield stress, with respect to the results presented in chapter 5. The extra samples are in new regions in the data space, but occupy identical regions in the model space, because all the samples are simply repeats from the same set of simulations. This suggests, that for yield stress at least, I require greater sample coverage in the model (parameter) space to improve the

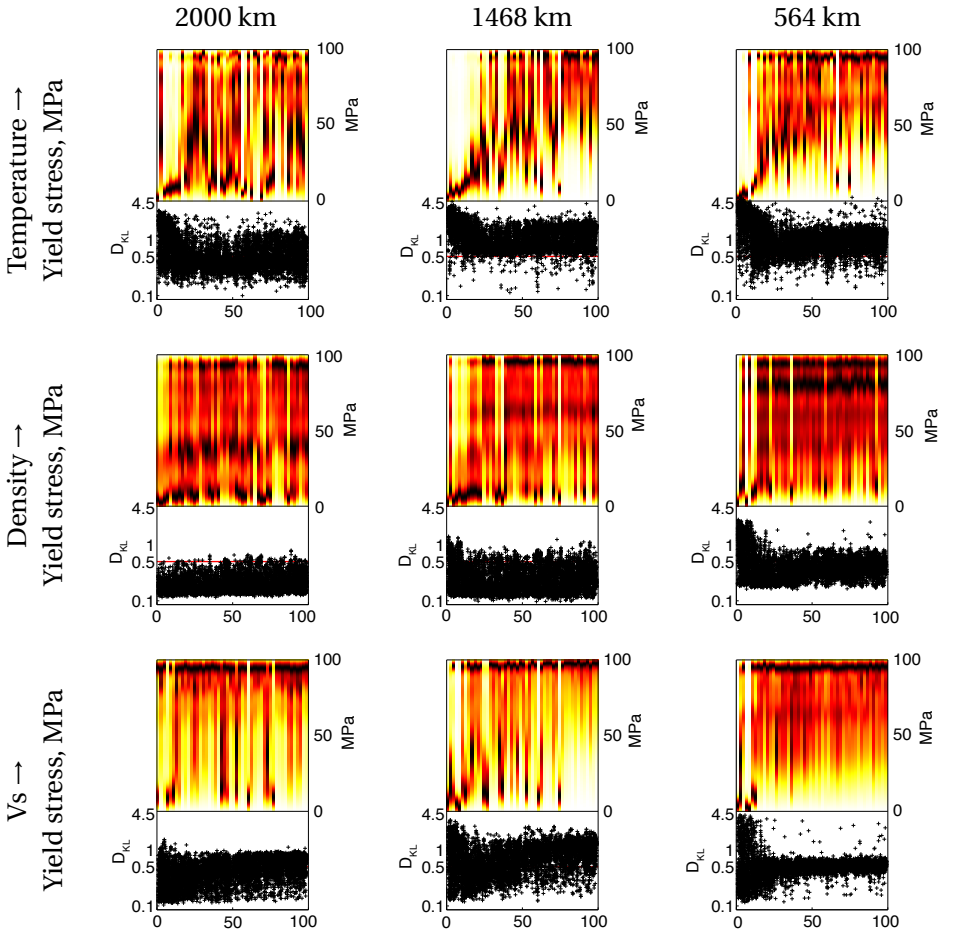


Figure 6.3: Inversions for yield stress from observations of the amplitude spectra of density, temperature and Vs at three different depths. Observations come from simulations of all different ages.

inferences.

6.5 Instantaneous local parameters

As well as attempting to invert for StagYY input parameters, I can use observations to make inferences about parameters which vary spatially or temporally, such as local viscosity or temperature. Figure 6.6 shows the PDFs when the

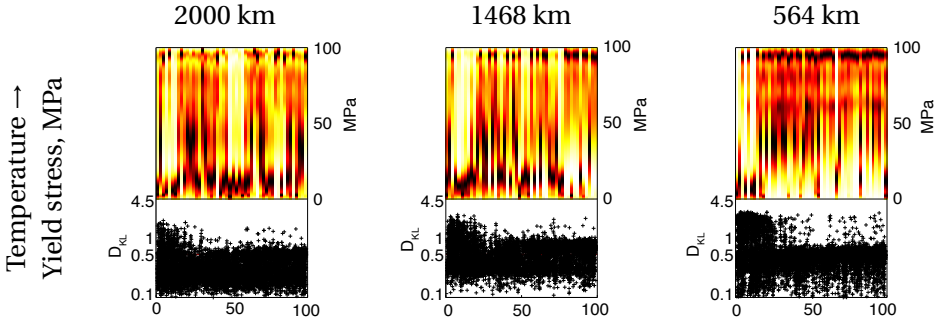


Figure 6.4: Inversions for yield stress from observations of the amplitude spectra using degrees 1-10 of temperature at three different depths. Observations come from simulations of all different ages. By removing the mean (degree 0) I can investigate the degree to which the networks require the mean temperature.

temperature spectra at 1468 km are used to make inferences about the mean viscosity at 1468 km. Viscosity is dependent on temperature according to:

$$\eta(T,p) = \eta_0 \exp\left(\frac{E_\eta + pV_\eta}{RT} - \frac{E_\eta}{RT_0}\right) \tag{6.1}$$

At a particular depth, p, R, T_0 and E_η are constant as discussed in chapter 3. The networks are not given any information on η_0 or V_η, which vary between simulations, so must decide how much of the viscosity is due to these two unknown parameters and how much is due to the temperature, which is given to the network. The parameter with the greatest influence on viscosity is η_0, which varies between 10^{18} and 10^{21} Pa s between simulations. However, temperature has a much greater effect on viscosity, as shown in figure 6.5. The temperature effects would therefore be expected to swamp the variations in η_0.

Despite the difference in the magnitude of contributions to viscosity from the temperature and η_0, the networks find local viscosity from the temperature spectra. Most of the information seems to come from the lateral variations in temperature as can be seen by comparing the two panels in figure 6.6. The one on the left uses degrees 0-10 and on the right only degrees 1-10. Removing the degree 0 from the amplitude spectra before training the networks removes some of the accuracy, particularly for lower viscosity models, but the inferences are still good, compared with the results including the mean temperature (left, figure 6.6). This is in stark contrast to most of my inference results.

I cannot currently find local viscosity reliably when using density, P- or S-wave velocity as the observation, or by using any combination of those three

observations. This is unfortunate, since using seismic velocity to constrain mantle viscosity would be a very powerful tool. In my simulations, local variations in viscosity are due to changes in temperature. The seismic and density observations are affected by temperature, but also composition. The composition effects therefore add an extra layer of complexity for the networks to unravel. Given how successfully this works with just the temperature field, it seems likely that it would be possible using density or seismic velocity, but a different network setup may be needed, more simulations would definitely help and perhaps some other constraining observation, such as information about the assumed composition of the mantle. These are all things which could be explored further.

The amplitude spectra of seismic velocities do however work particularly well for finding mean temperature, giving uncertainties of around 100 K for each observation (figure 6.7). The mantle is expected to be adiabatic, but since the heat flow from the core is uncertain, this potentially provides a new method for studying the temperature structure of the Earth.

6.6 Conclusions

I investigate whether using observations of the amplitude spectra of the density structure improve the inferences of StagYY input parameters. The only improvement seen is for the thickness of primordial material, both when finding the initial thickness from observations at 4.5 Gyr and for finding the resulting thickness of primordial material after 4.5 Gyr of convection. The primordial material is a chemical heterogeneity and is generally more dense due to higher iron content, which is why the inference is better when using density observations. My inferences for viscosity parameters are no better when using observations of density and temperature together than when using just density.

Using surface mobility as an extra input parameter does not improve the inferences at all. My simulations do not have granitic continents, so the movement of the basaltic crust is driven in a relatively simple manner by the underlying mantle. The shape of the convection cells below the crust therefore contain as much information about surface movement as the networks need and adding the extra observation does not provide any more information.

By using observations of multiple ages together in the training set, I investigate whether the networks require age information to make their inferences. In the case of yield stress, this does not seem to be the case and the inferences are as good as when using observations which all have the same age. The

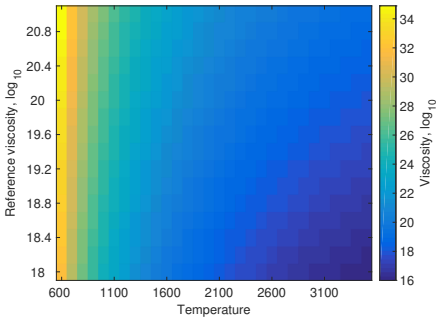


Figure 6.5: The relative effects of reference viscosity versus temperature at a pressure of 60GPa. Viscosity is calculated according to equation 6.1, where $V_\eta = 9.21 \times 10^{-7}$ (including decay with depth), $E_\eta = 1.62 \times 10^5$, and $T_0 = 1600$ K.

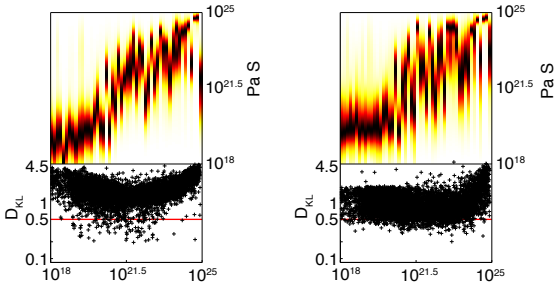


Figure 6.6: Using temperature spectra at 1468 km depth to find mean viscosity at that depth. The PDFs on the left use degrees 0-10 of the amplitude spectra, on the right degrees 1-10.

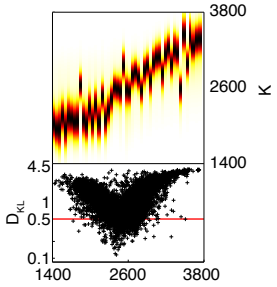


Figure 6.7: Using seismic velocity spectra at 1468 km depth to find mean temperature at that 1468 km.

mean temperature is necessary as an input to find yield stress, suggesting that the dominant signal is in the mean temperature rather than the degree of lateral heterogeneity. This also agrees with the results published in Atkins et al. (2016), and together they suggest that the distinctive cooling seen in low yield stress simulations occurs very early in the simulation and is then maintained throughout the run.

I can also find two different local parameters: mean viscosity and mean temperature at a given depth. Mean viscosity can only be found using temperature, but since this works so well it seems likely that, with more simulations and possibly a different network training approach, similar results could be achieved using seismic velocity. This would be an interesting way in which to constrain the viscosity of the mantle and would complement existing methods. Mean temperature can be found using the seismic structure of a particular depth interval, potentially leading to a new method to study the one-dimensional temperature structure of the mantle.

7

Inversions for Composition

Currently in preparation for publication

One of the many unknowns about the Earth is the bulk composition of the rocky mantle and crust. The bulk composition of this silicate based-part of the Earth is an important parameter when building models and theories about the current state and history of the mantle. For example, the interpretation of seismic tomography relies on assumptions about composition, as do models about the origins of erupted igneous rocks found at the surface, simulations modelling the convecting mantle, and theories about how the Earth accreted in the early stages of the solar system. In this chapter, I present a new method for inferring bulk silicate Earth (BSE) composition by using machine learning to identify the relationship between convection patterns and composition.

There currently exist several methods for estimating bulk silicate Earth composition, each with strengths and weaknesses. The first approach is to use material from elsewhere in the solar system. Chondritic meteorites are often as-

sumed to be the original solar building blocks of the Earth and are thus used as a starting point for bulk composition (e.g. Drake and Righter, 2002; Palme and O'Neill, 2003; Javoy et al., 2010). However there are several classes of chondrites which all have different chemistry, none of which perfectly match that of the Earth. It is likely that some of the Earth's early crust was lost by collisional erosion (O'Neill and Palme, 2008) and therefore the BSE composition may no longer be chondritic at all (Campbell and O'Neill, 2012). The uncertainties of these chondritic compositional models are little better than guess work (e.g. McDonough and Sun, 1995).

Instead of using meteorites, samples of Earth rocks can be used. The composition of rocks found at the surface can be used to infer the chemistry of the mantle (e.g. Lyubetskaya and Korenga, 2007). However, these rocks only sample the upper mantle so can only be used to estimate the primitive upper mantle composition. Geophysical observations allow us to make observations of the deep mantle, and provide a third way for investigating the composition of the mantle. To investigate lower mantle composition, I must use the relationship between seismic velocities and mineral physics (e.g. Deschamps and Trampert, 2004; Matas et al., 2007; Mosca et al., 2012; Cobden et al., 2012). The estimates produced from tomographic inversion are very uncertain, due to the inherently non-unique relationship between velocity and mineralogy, as well as the uncertainties in mineral physics properties. A fully probabilistic method which constrains the uncertainties associated with geochemical and cosmochemical models, whilst providing inferences with greater certainty than those made from the inversion of seismic tomography, may therefore prove to be a very useful tool.

I present a fully probabilistic inversion method that uses the dependence of convection on composition to make inferences about composition for simulated mantles. Composition determines the density, thermal expansivity, conductivity and phase of the minerals present in the mantle. These in turn determine the depth of phase changes and the buoyancy changes with temperature, both of which control the size and shape of convection cells in the mantle. I use machine learning to find the relationship between these simulated mantle structures and the composition. By taking this convectional dependence into account, I produce inferences with more certainty than by simply attempting to map density and seismic velocity to composition. All of my inferences are made as probability density functions, which take into account uncertainties from all stages of the inversion process, such as theoretical and observational uncertainties. When extended, this method will have the potential to signifi-

cantly improve the compositional models for the Earth.

7.1 Method particulars

I use observations of the temperature and density structure taken from simulations which have completed 4.5 Gyr of convection. As in the previous chapter, I add noise of with a standard deviation of 50 K and 12.5 kg m^{-3} at all depths. The ratio of noise to mean temperature and density therefore varies, since both temperature and density increase with depth, but the maximum magnitude of lateral variations in both are fairly constant with depth, at around ± 500 K and 125 kg m^{-3} respectively. The noise is added in the frequency domain to both the real and imaginary parts of the spectra. The amplitude in the frequency domain is scaled so that when an inverse Fourier transform is performed, the difference between the original and noisy temperature or density has a Gaussian distribution with magnitude 50 K or 12.5 kg m^{-3}. I do this because if I added the Gaussian noise added to the temperature or density field before taking the Fourier transform, most of it would be stripped out when I throw away the higher degrees because the variations in the noise have too short a wavelength. Adding the noise in the frequency domain is the easiest way to ensure that there is noise of the same magnitude at all wavelengths.

In this chapter, I vary the number of spectral degrees used as observations and network inputs. This is in order to investigate how composition affects the wavelength of heterogeneities, and where the dominant signal of the composition is in the mantle structure. The input observations I consider are any combination of: the degree zero variation in temperature and density, which is the same as the mean 1-D profile; the degree 1–10 amplitude spectra of temperature and density, which includes the amplitude of the lateral variations away from from the 1-D mean; and the mean temperature integrated over the whole mantle. The mean temperature is a single number for each simulation, but the degree zero and degree 1–10 parts of the spectra initially have dimensions 1×64 and 10×64 respectively. To ease network training, these are preprocessed using an autoencoding neural network to reduce their dimensionality. Four different autoencoding networks are used: one for each of degree 0 for temperature and density, and one each for degrees 1–10 for each observation. This reduces the dimensionality to 10 and 22 discreet numbers respectively, and generally results in smoothing of the amplitude spectra. I encode the degree 0 separately to degrees 1–10 because they have very different characteristics and amplitude. Degree 0 has much greater amplitude, with the depth dependent increase be-

ing significantly greater than the magnitude of lateral variations. It also makes investigating the information content of each slightly easier because I can separate the two without having to encode a new data set. I can consider the full degree 0–10 spectra if I train the networks by combining the two encodings into one vector which has 32 dimensions. However, the inputs to the network are therefore a combination of two highly non-linear functions. To test how much difference this makes to the results, I also train another neural network auto-encoder on degrees 0–10 at once. The difference between the encoded results is not significantly (as shown in figure 2.8, using an example of the temperature spectra). I also train an inverting neural network on the results from this second auto-encoder to compare the compositional inferences (figure 7.4).

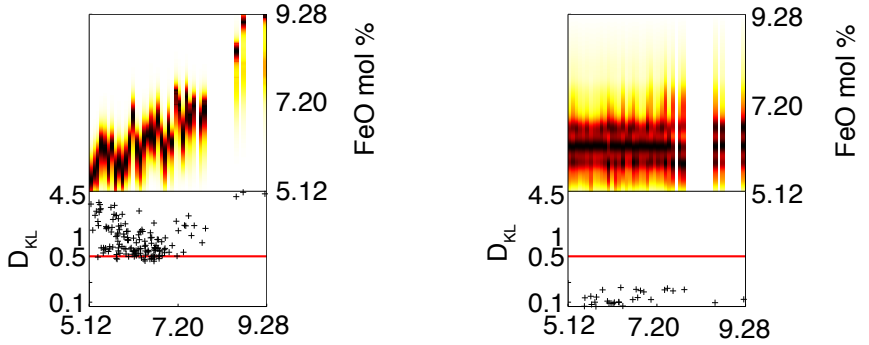
7.2 Demonstration

The results in this chapter investigate three main questions:

1. Does composition have a significant effect on lateral heterogeneities in temperature and density in the mantle?
2. How well do different observations of the mantle constrain composition?
3. Is the convection history of the mantle important when looking for information about the composition?

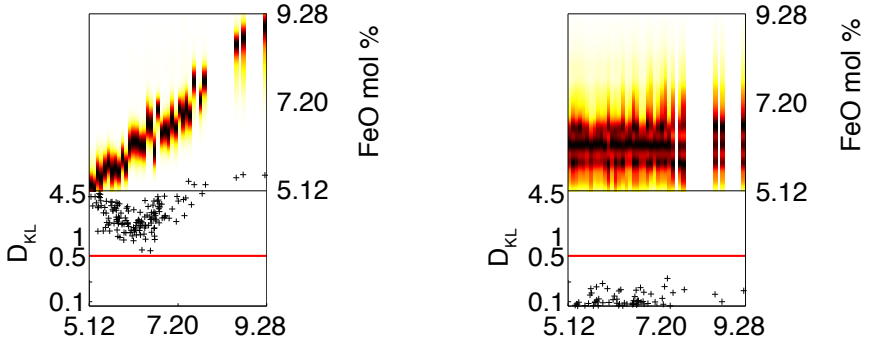
Inferences for FeO using different mantle observations

I first demonstrate how the signal of one compositional component varies when different observations are used. For this, I concentrate on the FeO mol % in the bulk mantle because the bulk iron content has a first order effect on mean mantle density. Using the degree 0 of the amplitude spectra of the density field, which gives the 1-D density profile as a function of depth, I can make a reasonable inference for FeO content of the mantle, as seen in figure 7.1a. The PDFs in figure 7.1a are all relatively wide, indicating the uncertainty on the inference. Much of this uncertainty arises because the mean mantle density is a function of temperature, which is a function of the convection parameters used and therefore varies between simulations. The marginal posterior PDFs also contain effects from the many other varying parameters which are integrated out. Using the observations of the density structure alone, it is impossible to distinguish how much of the variations in the spectra are due to temperature and compositional contributions respectively.



(a) *(b)*

Figure 7.1: Inferences for FeO mol % using various parts of the amplitude spectra of density field from simulations which have run for 4.5 Gyr. (a) shows the results using only degree 0, (b) shows the inferences made when the network is trained using degrees 1–10.



(a) *(b)*

Figure 7.2: Inferences for FeO mol % when networks are shown various parts of the amplitude spectra for both density and temperature at the same time. (a) shows posterior PDFs for networks trained using only degree 0 of each (the mean 1-D profile of temperature and density), and (b) shows inferences using degree 1–10 of temperature and density.

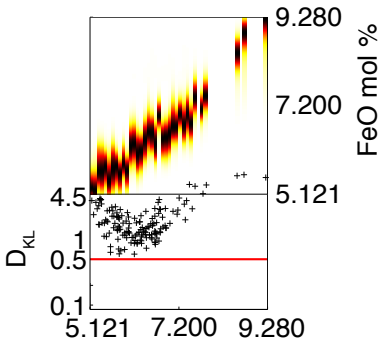


Figure 7.3: Inferences for FeO mol % using mean 1-D density profile (degree 0 of the amplitude spectra) as a function of depth with mean whole mantle temperature as a single extra input.

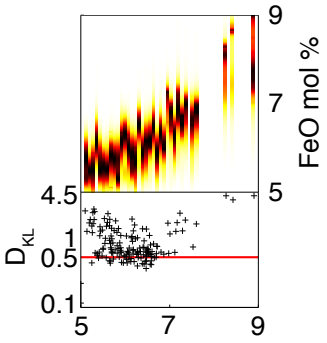


Figure 7.4: Inference for mantle molar percent iron oxide using degrees 0–10 of the density spectrum after 4.5 Gyr. This inversion uses a different dimension reduction method to figure 7.1. In this case, I train a single auto-encoding neural network on degrees 0–10. In figure 7.1, one auto-encoder deals with degree 0 and a second does degrees 1–10. There is not significant difference between the two, suggesting that the encoding strategy does not make much difference to the results, despite its inherent non-linearity.

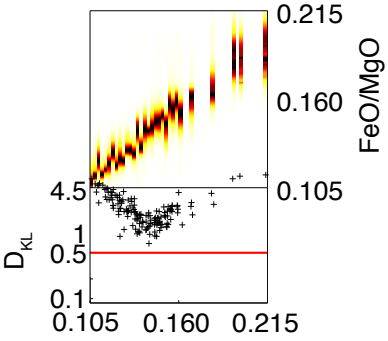


Figure 7.5: Inference for FeO/MgO. The network is trained on the amplitude spectra of both density and temperature, using degrees 0–10.

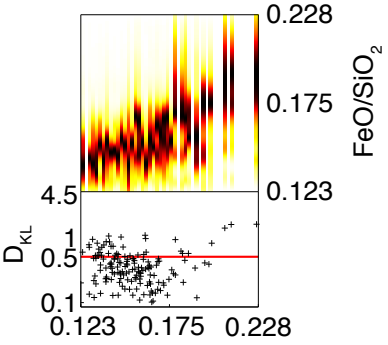


Figure 7.6: Inference for FeO/SiO_2. The network is trained on the amplitude spectra of both density and temperature, using degrees 0–10.

Using more details of the density structure of the mantle does not improve the inference of FeO in the mantle in my case. Figure 7.1b shows the inferences made by neural networks trained to find mappings using only the higher degrees of the amplitude spectra of the density field as input. The relative densities of model constituent end-members has been shown to significantly affect the convection patterns which develop (e.g. Nakagawa et al., 2009; Deschamps and Tackley, 2009; Nakagawa et al., 2010, 2012). Changing the FeO content changes the density of the density, therefore I would expect that using higher order component of the amplitude spectra would identify patterns arising from the interaction between different mantle components. However, when I test the signal of FeO in degrees 1–10 of the amplitude spectra, thereby removing the mean density variations with depth and considering only the relative amplitude of lateral variations at each depth slice, the signal of FeO fraction disappears completely, as seen in figure 7.2b. This is most likely to be because there is no signal in degrees 1–10, because iron content does not significantly affect the magnitude of lateral density heterogenities. However as in all the previous cases presented, there may be a signal, but my sample size is too small and the neural networks cannot interpolate between the samples to make inferences from them, or the signal is lost during preprocessing. I cannot know which of these possibilities is the case here.

The uncertainty in the results produced by inverting only the density structure arises in part from the combined contribution of chemistry and temperature to density. By including the temperature as an input, the networks can implicitly separate temperature and chemistry effects, because any part of the density structure that does not scale, linearly or not, with temperature must be due to chemistry (figure 7.2a). This significantly improves the quality of the inferences, reducing the uncertainty relative to the density only case (figure 7.1a). Adding the higher degrees of the temperature spectra to the higher degrees of the density spectra does not improve the inversion (figure 7.2b), suggesting that there is very little signal of bulk FeO content in the lateral variations in the mantle, although as previously said, this may be due to my network configuration, preprocessing of observations or small sample size.

As a test of the significance of the scaling between temperature and density, I also trained a network to find a mapping between the 1-D density profile with the addition of a single integrated mantle temperature value. This forces the network to find a scaling relationship to eliminate the temperature contribution to density before inferring composition. Surprisingly, I find that this performs as well as using the 1-D temperature profile (figure 7.3), although the

information gain (Kullback-Leibler distance, D_{KL}) is very slightly lower.

Figure 7.4 uses a slightly different dimensionality reduction approach. In this inversion for FeO mol % from density, a single auto-encoder was used for degrees 0–10. The results have slightly greater uncertainties that when two auto-encoders are used. As shown in figure 7.1, the signal for FeO content is predominantly in the 1-D mantle structure. By separating the two parts of the spectrum, the degree 0 may be better resolved, improving the inversion result. I tested this because the autoencoding is an inherently non-linear approach. Combining two separately encoded parts of the amplitude spectra into one matrix may therefore have produced some unexpected results. However, looking at this inversion test, and the example of the encoded spectra in figure 2.8, suggests that the encoding strategy (one or two encoders) does not make a significant difference to the results.

Other oxides and ratios

Whilst bulk FeO is interesting, it still leaves the remaining composition of the mantle unknown. I therefore go on to test other oxides, and their relative ratios, which can help bring clarity to this. Oxide ratios are useful because they give an indication of mineralogy and therefore properties, and also give an indication of melting history. I try to invert for several bulk oxide ratios in the mantle, using the density and temperature spectra. Figure 7.5 gives the ratio of bulk mantle FeO mol % to MgO mol %.

I would also like to know the SiO_2 concentration of the mantle, as this Mg/Si cannot be gained from meteorite observations due to fractionation effects in accreting disks (van Boekel et al., 2004). Again, I cannot directly invert for SiO_2 molar percentage. However, I can get an approximate bulk FeO/SiO_2 ratio, as shown in figure 7.6. This inversion is less successful than some of the others, but still provides a rough constraint on SiO_2 mol %, with uncertainties. Combining these inversions to make conditional PDFs would give us an estimate for total Mg, Fe and Si content of the mantle.

I tried several other oxide ratios, including MgO/SiO_2 and MgO/Al_2O_3, but could find no signal in my simulations. As with all of the inferences in this thesis, this could be due to any number of factors and these ratios may be worth exploring in the future with more simulations.

7.3 The effects of convection on my inferences

The results presented in the previous section are inferences made from convection simulations. In these simulations, in each grid cell, the density and thermal expansivity and therefore the buoyancy which drives convection are compositionally dependent. The patterns which evolve during the simulations are therefore to some extent compositionally dependent. I hypothesise that the networks recognise the compositional dependence of the shape of these convection patterns, giving the networks some extra information with which they can better determine the composition.

To test this hypothesis, I generate random temperature profiles, which have no convection history behind them, but fall within the ranges produced by the simulations. The density for these temperature profiles is then calculated given a mechanical mixture of harzburgite and basalt compositions, drawn randomly from the same prior distributions as the convection simulation compositions. The basalt-harzbugite fraction is are also allowed to vary. I encode the random profiles in the same way as the convectional ones. If I can find the composition from these observations as well as I manage in the previous section, the networks are simply finding a relationship between density, temperature and composition, which is in no way affected by the convective movements which generated those profiles. If the inference is less good, it suggest that the networks are using extra information from the compositional dependence of convection.

Figures 7.7 and 7.8 show the full set of PDFs for the randomly generated profiles. The Kullback-Leibler distances below them are between these PDFs and the prior of the random training set. The prior ranges of values (and therefore the axis length) is different to those in figure 7.3 making it difficult to compare inferences for the random profiles to those produced from observations of the convection simulations (7.3 and figures 7.5). This difference is because the random profiles have no iron-rich primordial material, giving a smaller range of possible FeO %. At first glance, it is apparent that the Kullback-Leibler distances for bulk FeO mol % are much higher than when the convective profiles are used, and the PDFs for FeO/MgO in the random profile case seem much broader with greater uncertainty. To aid comparison, the PDFs for both random and convective cases are replotted with the same prior range in figure 7.9. The PDFs are normalised so that the integrate to 1 over the range shown. The Kullback-Leibler distance is recalculated to be between the PDFs and a uniform distribution which is the same for both random and convective inputs.

The convective and random profiles can then be better compared.

The inferences for FeO mol % (top row, figure 7.9) are better using the randomly generated profiles, with narrower PDFs. There are several reasons for this. Firstly, if the ranges of the full training set are compared (using the axes limits in figures 7.3 and 7.7), the convective simulations have a much wider prior when the networks are being trained. This is because some simulations include primordial material which is iron-rich. The networks therefore have more work to do to narrow the prior in the convectional case because the samples are scattered over a much larger data space. This accounts for some of the difference in certainty. Secondly, the compositional profile of the convection simulations is significantly more complicated (even excluding the primordial material) because melting and subduction take place. The harzburgitic and basaltic end-members are therefore not evenly distributed throughout the profile, unlike in the random case. The greater uncertainty in the convective profiles reflects this complexity. Thirdly, I use 5422 random profiles to train the network, compared to 457 convective profiles. The difference in samples size alone may explain the certainty. If I use 500 random profile samples to train the network, the resolution decreases significantly and is comparable to the resolution from the convective profiles, with a slightly lower Kullback-Leibler distance, as seen in figure 7.10. This hints at the possibility that, with much more computer power, I may be able to significantly improve my inferences when using observations from convection simulations, given all the other complexities that the networks overcome in the convective cases.

Despite all of this convectional complexity and small sample size, the profiles produced by convection allow a much better inference for the ratio of MgO/FeO. The convective profiles give far more certain inferences than the random profiles in the lower panels of figure 7.9. There is still a clear trend in the random cases, but with much greater uncertainty. It therefore appears that the FeO/MgO ratio has a significant effect on the evolution of the temperature and density profiles in convection simulations.

7.4 Discussion

In the previous section, I showed that it is possible to use pattern recognition to make probabilistic inferences about the composition of the mantle using density observations taken from convection simulations. The inference improves if I invert temperature and density observations at the same time.

There are potentially several major advantages to my new method, besides

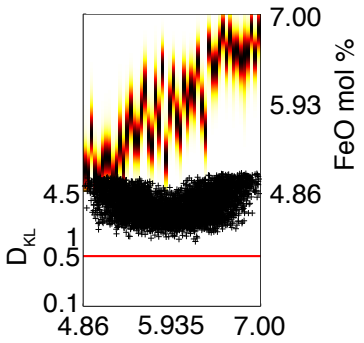


Figure 7.7: Inference for FeO from the temperature and density profiles, when a random temperature profile is used, which is not dependent on any convection history.

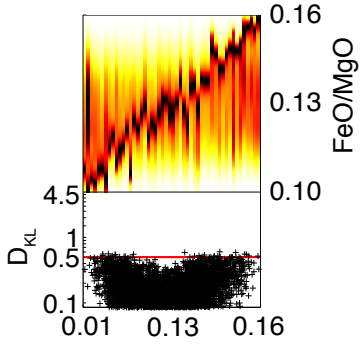
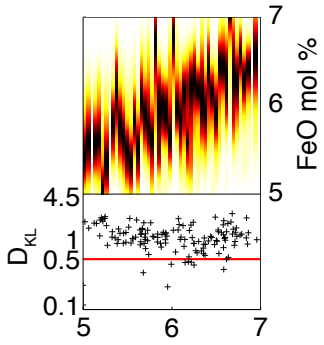
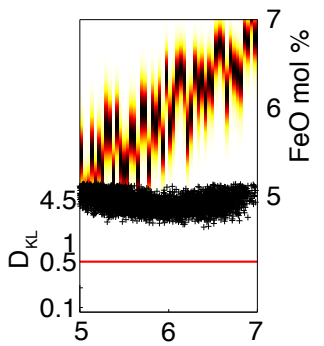


Figure 7.8: Inference for FeO/MgO from the temperature and density profiles, when a random temperature profile is used, which is not dependent on any convection history.

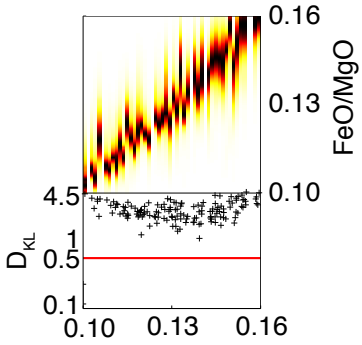
those presented in chapter 5. Firstly, I use *in situ* observations of the mantle. I therefore do not need to rely on samples collected at the surface or to make assumptions about how conclusions drawn from these can be extended down into the mantle. The observations are instantaneous, therefore I do not need to worry about how the composition of the Earth may have changed since the samples formed, which is not the case when making models based on meteorite composition. Secondly, I can use observations which are readily available. In this paper I present inferences made from the temperature and density structure of mantle simulations. The 1-D density structure of the mantle was established by PREM (Dziewonski and Anderson, 1981), although by using the Adams-Williamson method as the starting model, the density profile for PREM assumes that the mantle has an adiabatic temperature profile. Seismic tomography for mantle density variations is expected within the next few years (e.g. Płonka et al., 2016) which will improve my ability to make inferences. The 1-D mantle temperature structure is not well constrained, although the mantle potential temperature is relatively well constrained. This is all my inversions require, with a single integrated mantle temperature value used as a reference point to anchor the temperature profile which determines the density profile. Estimates of mantle potential temperature vary between studies, but are probably in the range 1300–1400 °C beneath mid-ocean ridges (e.g. Lee et al., 2009). I add Gaussian noise with a standard deviation of 50 K to my mean mantle temperature values, making my inferences robust with respect to these uncertainty



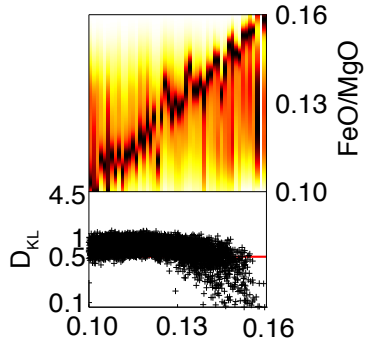
Inferences for FeO mol % using temperature and density profiles produced by StagYY, limited to the same range as the random cases.



Inferences for FeO mol % using temperature and density profiles, which are generated using random temperature profiles.



Inferences for Fe/Mg using temperature and density profiles produced by StagYY, limited to the same range as the random cases.



Inferences for Fe/Mg using temperature and density profiles, which are generated using random temperature profiles.

Figure 7.9: PDFs within the same ranges for convection profiles (left) and random profiles (left). The Kullback-Leibler distance for both data sets are calculated with respect to the same prior, rather than the prior of the training set, making them directly comparable.

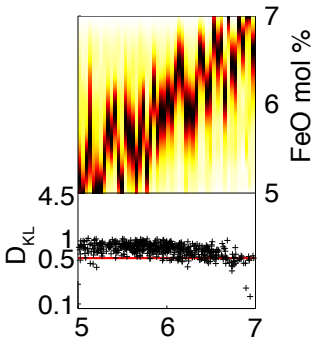


Figure 7.10: Inferences for FeO mol % using temperature and density profiles, which are generated using random temperature profiles, 500 of which are used to train the network.

levels.

In my simulations I know the temperature and composition, so I can easily calculate the seismic velocities. Since seismic velocity is dependent on composition, density and temperature, the results are very similar for the cases I tried. I do not present these results simply because this is currently a proof of concept with many limitations and I want to be clear that my method is not yet ready to apply to real data.

The third major advantage of my method is that it is fully probabilistic. I can account for every uncertainty in the process, including those in the theory, the assumptions made in the forward model and the uncertainties associated with the data, both through observation and in converting observations to a mantle image. Robust error estimates for the composition then mean that the uncertainties in subsequent modelling and theoretical developments can be more accurately assessed. This is in contrast to many petrological studies, where the measurements are very precise with well constrained errors, but the subsequent theoretical extension to bulk silicate Earth composition lack robust error analysis. This brings me to the fourth advantage of this method. Probabilistic inversion of seismic structure for composition has been undertaken before (e.g. Matas et al., 2007; Mosca et al., 2012), but only using a single snapshot of one possible convection end-state: that of Earth after 4.5 Gyr. By considering many possible convective states resulting from different compositions, I implicitly get an extra piece of information in my inversions, the compositional dependence of convection. This reduces the uncertainty of the inferences, despite adding in extra complications and chemical heterogeneities such as subducting slabs.

There are several limitations inherent in this method. Most importantly, the inferences inherit all the assumptions made in the forward modelling. My results can therefore only be as good as the choices I make in my simulation setup. For example, my simulations assume pyrolitic mantle with no compositional difference between upper and lower mantle. The networks therefore do not know that any other types of mantle could exist and will make all their inferences assuming that the mantle was initially well mixed pyrolite. Some modelling assumptions can be accounted for using the noise, such as the averaged decay rate of radioactive elements. Others, such as a homogeneous mantle, can be taken into account by varying more different parameters in the forward simulations (e.g. degree of mantle homogeneity). The physics behind some convection processes, such as the link between grain size and damage and rheology are not yet fully understood. In this case, I must be clear what assumptions were made in order to make my inferences and what the likely effects of these assumptions are. However, in general, convection simulations can currently capture the large-scale processes operating in the mantle, producing recognisably Earth-like simulations. For large-scale analysis of the composition of the mantle, this is probably enough.

The other main limitation is practical. The quality of the inversion is highly dependent on the number of samples, because the networks interpolate between samples. If the samples are too widely distributed in the model space and the observation does not vary smoothly between the samples, the networks will have limited success making inferences for a new sample that lies away from the training set samples. This problem increases with the number of dimensions in the model space, and is generally known as the curse of dimensionality. However, convection simulations are computationally very expensive, limiting the number of forward models I can run. The number of parameters I can investigate is therefore limited by the number of simulations I can run.

In the future, I may be able to circumvent some of the cost of the forward simulations by running regional models until the stabilise, rather than simulating the whole mantle for 4.5 Gyr. This may allow me to make inferences about local variations in composition. I could then map heterogeneities in the mantle, to study features such as subducting slabs, potential basalt ponding at the transition zone, heterogeneities at the base of the mantle, in a manner similar to existing studies (e.g. Nakagawa et al., 2010; Deschamps et al., 2012; Ballmer et al., 2015), but with enough samples to be able to apply statistical inversion methods. However, with smaller scale simulations it is tempting to attempt to

capture finer-scale processes, which would make them just as expensive.

7.5 Conclusion

In this chapter, I present a new probabilistic method for constraining mantle major element geochemistry. By considering the convection patterns from many different simulations, all with different chemistry, I implicitly take into account the dependence of mantle evolution on geochemistry. This gives us an extra piece of information, which improves the quality of our inferences, compared to simply trying to map from density to composition, which is a very non-unique problem and therefore produces large uncertainties. The major strengths of my method, besides the extra constraints from the convection history, are that it is fully probabilistic, allowing robust analysis of uncertainties, and that it uses *in situ* measurements of the deep mantle, rather than relying on surface samples such as igneous rocks. Whilst I must make assumptions in my forward modelling, which are constrained by both the existence of the appropriate theory and by the availability of computing power, I can fully describe these assumptions and make it clear that my results are subject to their validity.

At this stage, this study is a demonstration of a concept, but with more realistic convection simulations, I feel this method could be applied to make inferences about composition using seismic tomography. This may allow us to begin to unravel the controversy over the relative contribution of chemistry and temperature to seismic heterogeneities in the mantle. Using regional geodynamic models, I may also be able to study the local variation in chemistry by considering their effects on dynamics in a large-scale statistical manner.

8

Emulating thermodynamical equilibrium calculations

Here I consider using neural networks to predict thermodynamic equilibria for deep mantle mineral physics. Machine learning is already used to study chemical processes in computational biochemistry, e.g. by predicting the stability of proteins (Masso and Vaisman, 2008) or for compound screening for potential drugs (Byvatov et al., 2003). It therefore seems a logical application to apply machine learning techniques to mineral physics. In this example, I do not try to calculate any precise mineral physics properties *ab initio*, but instead look at the possibility of using a neural network to emulate existing mineral physics calculation tools such as Perple_X to make them more flexible.

Currently, to calculate the mineral physics properties for my convection simulations, I use Perple_X which finds the most stable phase assemblage at

each pressure and temperature, given a mineral physics database. This information goes into a look up table and the properties of each cell are found based on pressure, temperature and the relative proportions of the end-member rock types in a mechanical mixture. This is practical in that there is no need to calculate the equilibrium assemblage of minerals at every time step. It is also potentially a realistic model of the mantle, given the very low chemical diffusivity of mantle mineral (e.g. Hofmann and Hart, 1978) and long stirring time of the mantle and evidence of basaltic slabs penetrating the mantle without mixing and disintegrating. However, the choice to treat the mantle as a mechanical mixture or as an equilibrium assemblage has important mineral physical implications, particularly for the calculation of seismic velocity.

Figures 8.1 and 8.2 show the temperature structure and distribution of basaltic material in one of my convection simulations after 4.5 Gyr of convection. The simulation has a marble cake structure with lots of sinuous strands of basalt (Allègre and Turcotte, 1986). I calculate the density and seismic structure of this simulation twice, once using a mechanical mixture approach and once where the components are assumed to be fully equilibrated. For the mechanical mixture, the properties in each cell are the weighted mean of the properties of the two end members. For the equilibrium assemblage, I use Perple_X to find the equilibrium phase assemblage for the fully mixed chemical composition in each cell. Figures 8.3 to 8.5 show the difference in density and velocity between the mechanical mixture results and the equilibrium results. In the lower mantle, the equilibrium assemblage generally produces a slightly lower density and very slightly higher s-wave velocity than the mechanical mixture. The p-wave velocities are almost identical. The big differences between the two approaches are apparent around the transition zone (between around 400 and 1000 km, dividing the upper and lower mantle) and just above where we expect the D" layer to be at the base of the mantle. This simulation has no primordial material, so all jumps in physical properties are due to sharp phase changes in the basalt (garnet) and harzburgite (pyroxene) systems. The differences between the approaches show up at the major phase changes: olivine to wadsleyite at around 410 km, ringwood to bridgemanite at 660 km and bridgemanite to post-perovskite at the base of the mantle. These differences have been observed previously when trying to find a geochemical fit to seismic data, for example by Xu et al. (2008).

The difference between the mechanical mixture and equilibrium approaches has implications for studies such as mine in several ways. The difference between the density calculated in each case is small, but is potentially enough

to alter the convection patterns. The largest density discrepancies are around phase changes, which may change the depth and degree to which the upper and lower mantle interact. A change in the depth of phase boundaries may also have implications for slab subduction, changing the depth of slab stagnation by changing the depth of critical viscosity and buoyancy contrasts (Ballmer et al., 2015). It also determines how I interpret any observations taken from seismic tomography, and therefore assessments about how well a simulation produces Earth-like structures. For example, topography on the 660 km discontinuity is often used as a proxy for temperature and composition (e.g. Jenkins et al., 2016). The choice of a fully chemically mixed equilibrium assemblage versus a mechanical mixture will change the degree to which thermochemical variations deflect this boundary in simulations and therefore the similarity between simulations and real seismic tomographic data.

Calculating thermodynamic equilibria during a convection simulation

To assess the differences in convection patterns produced, I would have to be able to calculate the phase equilibrium assemblage for every grid point at every time step during a simulation, to see how it compares with using a mechanical mixture of the end-member rock types. It would be very expensive to run Perple_X at each time step and phase transitions can cause numerical instabilities (Connolly, 2009) making this an impractical option. There are codes which calculate the stepped composition between two end-members (e.g. Zunino et al., 2011), which produces a look-up table similar to the one I am currently using. However, this assumes that as a two end-member system melts, the oxide-composition of the melt varies linearly as a function of melt fraction. The calculations also become more complicated if a third component such as primordial material is added, requiring a either mixture-assemblage approach or a huge amount of memory to store a much enlarged look-up table. Alternatively, a large amount of data can be compressed, by methods such as tensor rank decomposition (Afonso et al., 2015). This still requires the mineral physics data to be initially calculated using a package such as Perple_X covering the likely range of compositions prior to compression.

I instead propose a novel solution to the problem of calculating mineral physics properties on the fly, using, unsurprisingly, neural networks. For this problem, neural networks have four obvious advantages: firstly, they are very fast to evaluate, potentially only requiring three matrix operations depending on the network architecture. This will not slow the simulations down signif-

icantly even when performed at every time step. Secondly, I can train a network using a very wide range of compositions, meaning that any possible melt-residual-primordial combination can be covered. The only requirement would be to track the movement of oxide components. Thirdly, once a reliable network has been trained, it will be highly portable and can be used in many different scenarios, removing the need to recalculate look-up tables before using a new composition in a simulation. Finally, it will make it very easy to initialise simulations with non-homogeneous composition. It would be straightforward to change composition through the mantle to represent crystallisation fronts and differentiation caused by a cooling magma ocean or to have a chemically distinct upper and lower mantle.

As an initial investigation into the feasibility of this proposal, I train networks at each depth slice to calculate the equilibrium properties of the mineral assemblage give the major element oxide composition and temperature. I give a few examples of the resulting PDFs below (figure 8.7). For this initial exploration, I train one network at each depth because pressure has the dominant effect on the properties. However, with different pre-processing, such as using the deviation of the property from a fixed depth-dependent mean, I may be able to use one network at all depths.

This method is not particularly successful for the convection simulation shown in figures 8.1 to 8.5. The density is generally underestimated (figure 8.6), which has a knock-on effect for velocity. Besides the usual possible reasons for poor network performance, such as too small a dataset or non-optimal network architecture, the results suffer from having too wide a prior. The priors include extremes of temperature, going from 100-8000 K at all depths. This produces instabilities in the Perple_X calculated results which propagate through into the network training. The range of possible compositions also vary massively and are not all evenly spaced. This is because I use my previously calculated look-up tables for basalt, harzburgite and primordial material, which are drawn from 4 different priors, which are themselves not even. Smoothing these priors and removing some of the extremes of temperature should instantly improve the network performance.

Theoretical problems with an equilibrium assemblage

Whilst this new method may give me a way to run simulations in chemical and thermodynamical equilibrium there are several geophysical constraints which would need to be investigated before I chose to run my simulations in this way. One big unanswered question in geophysics is: to what extent is the mantle ac-

tually homogeneous even on a small scale? Tirone et al. (2016) found that two lithologies can coexist in thermodynamic equilibrium without homogenising. This is in agreement with the very short length scales over which diffusion takes place in the mantle. It is therefore possibly quite unreasonable to assume that subducted basalt re-mixes into the mantle. This is supported by seismic observations of scatterers in the mantle, suggestive of chemical heterogeneities, and seismic velocities which are better modelled by a mixture assemblage than an equilibrium assemblage, particularly around the transition zone (Xu et al., 2008).

However, this still leaves applications for mid-run mineral physics calculations. If the mantle is assumed to have formed from a cooling magma ocean, it is reasonable to assume that it started in chemically well-mixed equilibrium over at least some length scales. Molten basalt at any point in time is also going to be in equilibrium with itself, but the physical properties of this equilibrium depend on the composition. An on-the-hoof calculation mechanism will therefore allow me to have basalts of different compositions which are in equilibrium. The same reasoning applies to residual harzburgite. I currently track basalt and harzburgite on tracers and it should be possible to track the oxide composition of each of these so that I could calculate the varying properties of varying compositions of each member, and then calculate the mechanically mixed properties of a mixture of two properly equilibrated end-members. In this case, I can also track any partial melting and how long it remains in the cell, because this would increase the likelihood that the two end-members remix.

The last problem is how to determine where the oxides go. To track the movement of one oxide, such as iron, I would need a simple enough petrological law to determine the partition coefficient of each oxide upon melting of any composition at any temperature and pressure. Current laws are also subject to a high degree of uncertainty. Whilst this uncertainty potentially removes the advantages of being able to calculate the mineral physics properties for any composition at any time step, this ability may give me a way to assess the results of different petrological laws. This could be of particular use at high pressures, where some iron rich melts may be negatively buoyant, with potentially very interesting geodynamical implications.

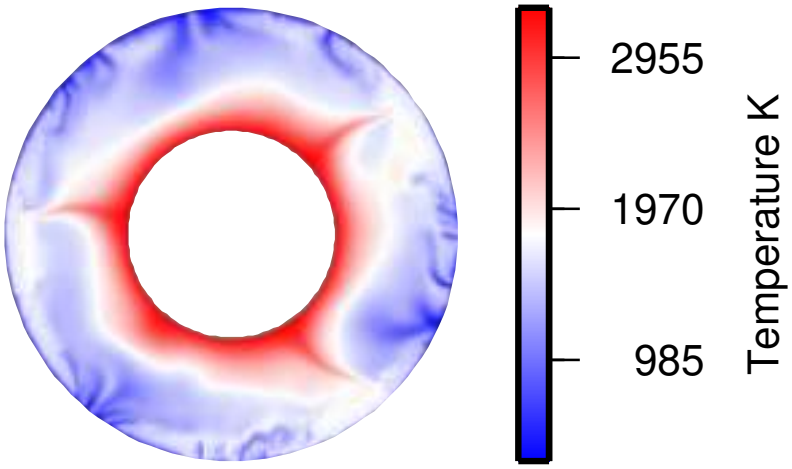


Figure 8.1: Temperature after 4.5 Gyr.

Figure 8.2: Basalt fraction after 4.5 Gyr.

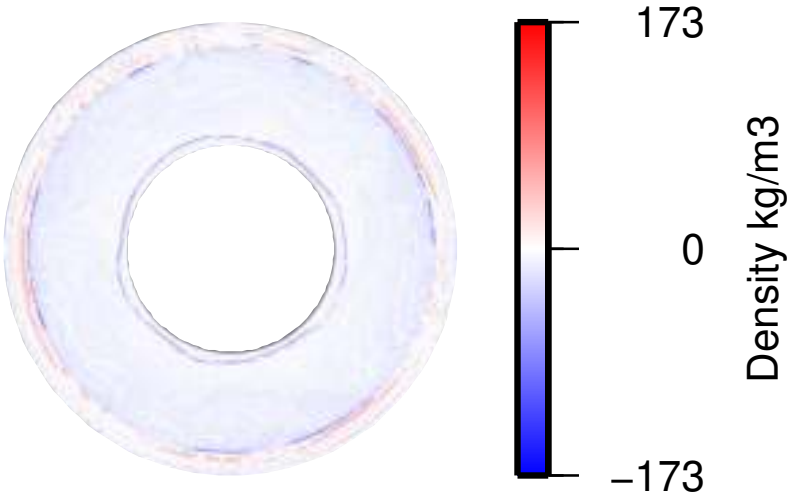


Figure 8.3: Difference between density calculated for a mechanical mixture and a fully chemically mixed assemblage.

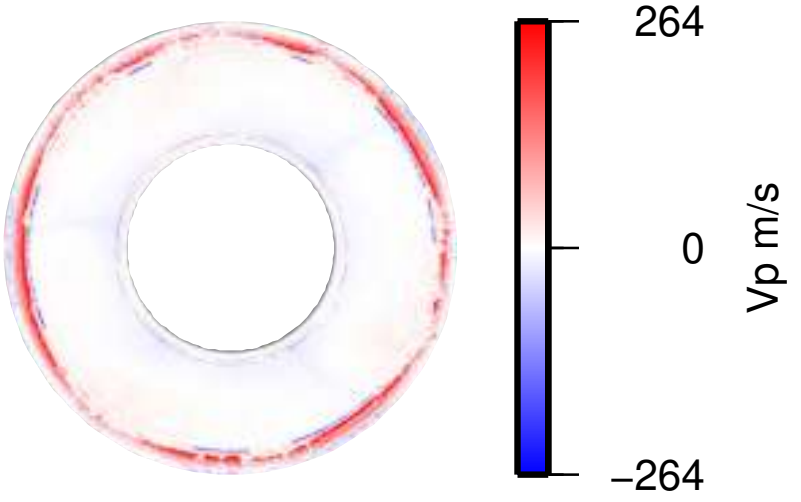


Figure 8.4: Difference between p-wave velocity calculated for a mechanical mixture and a fully chemically mixed assemblage.

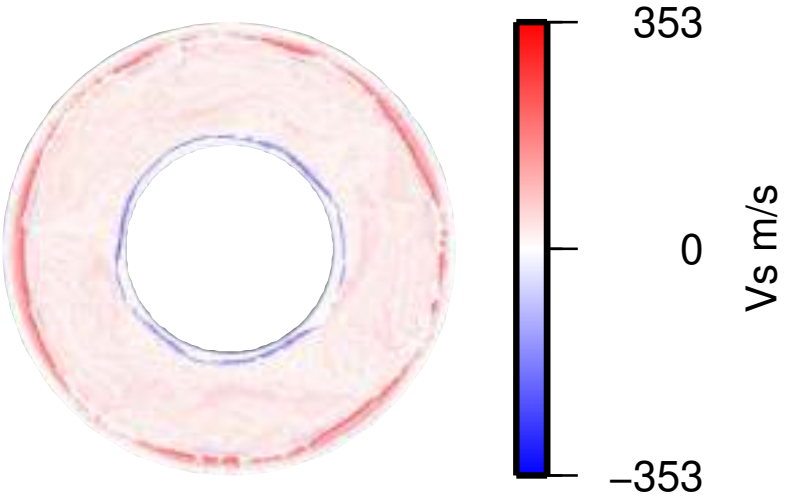


Figure 8.5: Difference between s-wave velocity calculated for a mechanical mixture and a fully chemically mixed assemblage.

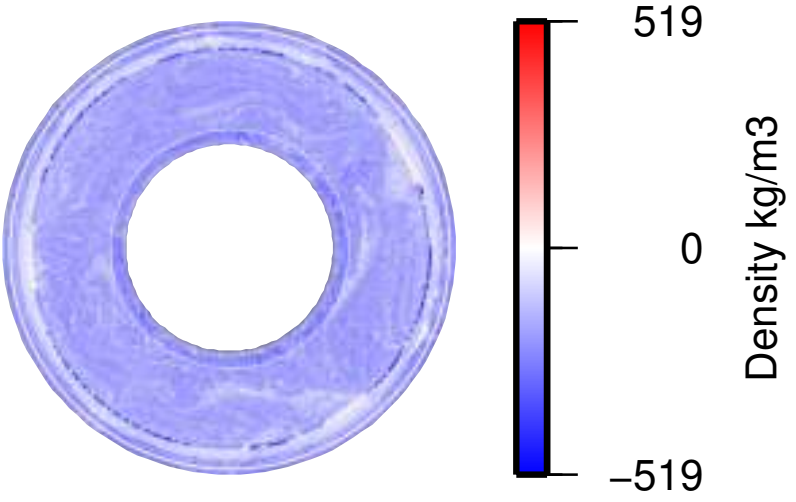


Figure 8.6: Difference between density calculated for a mechanical mixture and a fully chemically mixed assemblage, when approximated using a neural network.

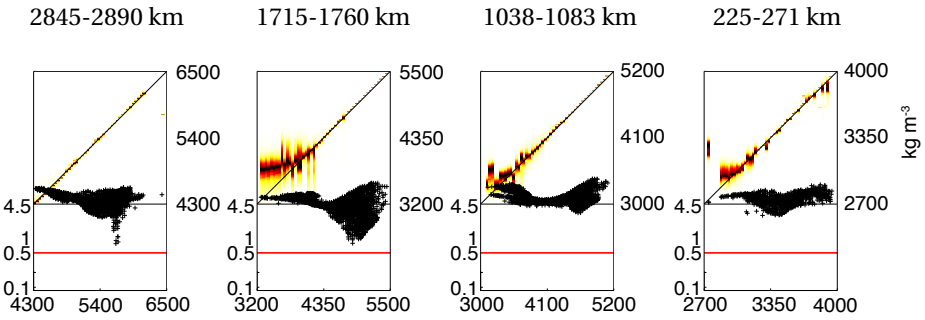


Figure 8.7: Network predictions for density at a variety of depths. The networks are given oxide composition and temperature. The low density predictions are generally at unrealistically high temperatures, up to 8000 K.

9

Conclusion

In this thesis, I have demonstrated a proof of concept for a new method for studying geodynamics. I show that I can use neural networks both to make inferences from observations from the end-state of convection simulations and to predict this end-state given all the simulation input parameters. This method has many advantages, most of which stem from the probabilistic treatment of geodynamic problems.

Firstly, it gives me a way to describe uncertainties for geodynamical problems in a computationally feasible manner. The prior sampling approach, combined with the use of neural networks to interpolate between these samples means that I can begin to make inferences with even a small data set. The data set is both reusable and flexible, so whilst it is still very computationally expensive to produce, it can at least be fully utilised for a variety of purposes. This

gives me an advantage over many other statistical methods used in geodynamics. Monte Carlo style approaches produce a single-use data set. It is therefore difficult to justify the massive computational expense necessary to do full posterior sampling for an inverse problem for a large-scale geodynamical study.

The second advantage to using a fully probabilistic approach is that I can include the many sources of uncertainty in geodynamics in my inference. These appear in the shapes of the probability density functions which describe any inference I make. Should inferences be made about the Earth in this way, they are therefore more useful because they acknowledge the uncertainty of the result and give an indication of their reliability before the inferences are used for other purposes.

The probabilistic framework within which my method operates also provides a novel way to access the history of the Earth. Using this method, I can attempt to make inferences about the Earth at time and length scales beyond which deterministic approaches are thwarted by the non-linearity and chaos of convection.

In this thesis, I have applied my method with several aims. In chapter 4, I show that it is possible to train a neural network to emulate a complex non-linear convection code such as StagYY. This allows me to rapidly predict the likely results of a simulation before running it, potentially saving much computational expense and frustration. For each set of parameters, the networks simply interpolate between the existing samples to find a different posterior probability density function. Whilst this approach is therefore not a replacement for a Monte Carlo style investigation as no new data is added with each sample, the interpolation of existing samples is what makes the parameter space exploration so rapid. With enough samples it could become a very powerful tool for model space searches for likely parameter combinations. The predictive networks can also be used to test the sensitivity of the simulations to each parameter, potentially providing an inexpensive method to estimate the uncertainty of results.

Neural network emulators may also have potential in the study and comparison of many different simulations at once. With some thought and preprocessing, simulations from different codes could be compared in this manner and used to boost sample numbers. Features which vary between codes and research groups would affect the marginal PDFs, but could be integrated out. The challenges in combining codes would be extensive, and would include preprocessing the results in a way to eliminate as many differences as possible and to ensure that parameters have the same meaning and scaling across all the

codes, as well as collecting together simulations into a usable database.

The predictive abilities of the neural networks may also find applications within the full convection simulation code, with one possibility demonstrated in chapter 8. This may allow us to include approximations to processes which would otherwise be far too computationally expensive to include in the forward simulations.

I then go on to show that besides predicting the expected outcome of a convection simulation, neural networks can be used to take the end-state of a convection simulation and make inferences about the parameters used to run it. This works well with rheological and compositional parameters. At present, most of my inferences are very uncertain. This is most likely to be because I have very few samples with which to train the networks. Hopefully, by expanding the data set, I will find that I can make inferences about many more parameters, with a much greater degree of certainty.

This introduces the biggest current limitation in my method, which is the size of my data set. With more samples I expect to be able to resolve many more parameters and characteristics from the convection simulation observations. This is simply because a few hundred simulations are not enough to adequately sample a 29 dimension model space, especially when the relationship between model space and observation is so non-linear. Generating more samples requires more computational expense, but I expect the result to improve rapidly. Figures 7.9 and 7.10 show the difference that a factor of 10 increase in samples has when inverting for composition. Running more simulations will also allow me to change some choices I made during model setup. For example, I will be able to use a more realistic viscosity profile with a viscosity jump at the transition zone.

If I can make better inferences from somewhat more realistic model setups, I can begin to consider the possibility of using my synthetically trained networks to make inferences about the condition of the Earth's mantle. There are many caveats to this, as discussed in chapters 3 and 5. I have to ensure that the amplitude spectra in spherical harmonics of 3-D simulations have a similar relationship to the convection simulation parameters as in the 2-D cases. Even with modern supercomputers it will currently be nearly impossible to run enough 3-D case of StagYY to train a neural network. I therefore have to hope that with appropriate preprocessing, a network trained on 2-D cases can be used to make inferences about the Earth's 3-D mantle. I also need to ensure that the convection simulations capture enough mantle processes to be suitable models when I make an inference. Any inference I make will only ever be

as good as the simulation code used and the choices made therein.

Whilst this study is still very much a proof of concept, the advantages of this method are very real and provide ways to overcome many challenges faced by geodynamicists, and may give the community access to parts of Earth history which were simply not possible before. The applications of this method also extend beyond the Earth. Planetary science is plagued by many of the same problems as deep-Earth geophysics, but often amplified. Very few observations exist for extraterrestrial bodies, compared to the number of observations of the Earth and obtaining more observations is incredibly expensive and limited by available technology. This means that the prior range of possible solutions is generally enormous and any solution will have huge uncertainties. A prior sampling approach may therefore have benefits, both for making fully probabilistic inferences from the limited data available and conducting rapid parameter space searches to focus on regions of potential interest, which could potentially be of use before and during missions.

Bibliography

Afonso, J. C., Zlotnik, S., and Díez, P. (2015). An efficient and general approach for implementing thermodynamic phase equilibria information on geophysical and geodynamic studies. *Geochemistry, Geophysics, Geosystems*, 16(10):3767–3777.

Allègre, C. J., Staudacher, T., Sarda, P., and Kurz, M. (1983). Constraints on evolution of Earth's mantle from rare gas systematics. *Nature*, 303:762–766.

Allègre, C. J. and Turcotte, D. L. (1986). Implications of a two-component marble-cake mantle. *Nature*, 323:123–127.

Andrault, D., Bolfan-Casanova, N., Lo Nigro, G., Bouhifd, M. A., Garbarino, G., and Mezouar, M. (2011). Solidus and liquidus profiles of chondritic mantle: Implications for melting of the Earth across its history. *Earth and Planetary Science Letters*, 304(1-2):251–259.

Argus, D. F., Peltier, W. R., Drummond, R., and Moore, A. W. (2014). The Antartica component of postglacial rebound model ICE-6G_C (VM5a) based on GPS positioning, exposure age dating of ice thicknesses, and relative sea level histories. *Geophysical Journal International*, 198:537–563.

Atkins, S., Valentine, A. P., Tackley, P. J., and Trampert, J. (2016). Using pattern recognition to infer parameters governing mantle convection. *Physics of the Earth and Planetary Interiors*, 257:171–186.

Austermann, J., Kaye, T. B., Mitrovica, J. X., and Huybers, P. (2014). A statistical analysis of the correlation between large igneous provinces and lower mantle seismic structure. *Geophysical Journal International*, 197:1–9.

Ballmer, M. D., Schmerr, N. C., Nakagawa, T., and Ritsema, J. (2015). Compositional mantle layering revealed by slab stagnation at 1000-km depth. *Science Advances*, 1(11):DOI: 10.1126/sciadv.1500815.

Baum, E. B. and Haussler, D. (1989). What size net gives valid generalization? *Neural Computation*, 1:151–160.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. by the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Cantone, M.A. and F.R.S. *Philosophical Transactions of The Royal Society*, 53:370–418.

Becker, T. W. and Boschi, L. (2002). A comparison of tomographic and geodynamic mantle models. *Geochemistry, Geophysics, Geosystems*, 3:DOI 10.1029/2001GC000168.

Becker, T. W., Kellogg, J. B., and O'Connell, R. J. (1999). Thermal constraints on the survival of primitive blobs in the lower mantle. *Earth and Planetary Science Letters*, 171:351–365.

Bello, L., Coltice, N., Rolf, T., and Tackley, P. J. (2014). On the predictability limit of convection models of the Earth's mantle. *Geochemistry, Geophysics, Geosystems*, 15:2319–2328.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, UK.

Blum, A. L. and Rivest, R. L. (1992). Training a 3-node neural network is NP-complete. *Neural Networks*, 5:117–127.

Bocher, M., Coltice, N., Fournier, A., and Tackley, P. J. (2016). A sequential data assimilation approach for the joint reconstruction of mantle convection and surface tectonics. *Geophysical Journal International*, 204:200–214.

Bower, D. J., Gurnis, M., and Flament, N. (2015). Assimilating lithosphere and slab history in 4-D Earth models. *Physics of the Earth and Planetary Interiors*, 238:8–22.

Bower, D. J., Gurnis, M., and Seton, M. (2013). Lower mantle structure from paleogeographically constrained dynamic Earth models. *Geochemistry, Geophysics, Geosystems*, 14(1):DOI 10.1029/2012GC004267.

Buffett, B. A., Huppert, H. E., Lister, J. R., and Woods, A. W. (1992). Analytical model for the solidification of the Earth's core. *Nature*, 356:329–331.

Bull, A. L., McNamara, A. K., and Ritsema, J. (2009). Synthetic tomography of plume clusters and thermochemical piles. *Earth and Planetary Science Letters*, 278:152–162.

Bunge, H.-P., Hagelberg, C. R., and Travis, B. J. (2003). Mantle circulation models with variational data assimilation: Inferring past mantle flow and structure from plate motion histories and seismic tomography. *Geophysical Journal International*, 152:280–301.

Byvatov, E., Fechner, U., Sadowski, J., and Schneider, G. (2003). Comparison of support vector machine and artificial neural network systems fro drug/nondrug classification. *Chemical Information and Modeling*, 43(6):1882–1889.

Campbell, I. H. and O'Neill, H. S. C. (2012). Evidence against a chondritic earth. *Nature*, 483:553–558.

Cobden, L., Mosca, I., Trampert, J., and Ritsema, J. (2012). On the likelihood of post-perovskite near the core-mantle boundary: A statistical interpretation of seismic observations. *Physics of the Earth and Planetary Interiors*, 210-211:21–35.

Coltice, N. and Ricard, Y. (1999). Geochemical observartions and one layer mantle convection. *Earth and Planetary Science Letters*, 174(1-2):125–137.

Coltice, N. and Ricard, Y. (2002). On the origin of noble gases in mantle plumes. *Philosophical Transactions of The Royal Society*, 360:2633–2648.

Connolly, J. A. D. (2009). The geodynamic equation of state: What and how. *Geochemistry, Geophysics, Geosystems*, 10:DOI:10.1029/2009GC002540.

Conrad, C. P. and Gurnis, M. (2003). Seismic tomography, surface uplift, and the breakup of Gondwanaland: Integrating mantle convection backwards in time. *Geochemistry, Geophysics, Geosystems*, 4(3):DOI 10.1029/2001GC000299.

Davaille, A. (1999). Simultaneous generation of hotspots and superswells by convection in a heterogeneous planetary mantle. *Nature*, 402:756–760.

Davies, G. F. (2010). Noble gases in the dynamic mantle. *Geochemistry, Geophysics, Geosystems*, 11(3):DOI: 10.1029/2009GC002801.

Davies, J. H. and Davies, D. R. (2010). Earth's surface heat flux. *Solid Earth*, 1:5–24.

de Wit, R. W. L., Valentine, A. P., and Trampert, J. (2013). Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks. *Geophysical Journal International*, (195):408–422.

Deschamps, F., Cobden, L., and Tackley, P. J. (2012). The primitive nature of large low shear-wave velocity provinces. *Earth and Planetary Science Letters*, 349-350:198–208.

Deschamps, F., Kaminski, E., and Tackley, P. J. (2011). A deep mantle origin for the primitive signature of ocean island basalt. *Nature Geoscience*, 4:879–882.

Deschamps, F. and Tackley, P. J. (2008). Searching for models of thermochemical convection that explain probabilistic tomography. I - Principles and influence of rheologial parameters. *Physics of the Earth and Planetary Interiors*, 171:357–373.

Deschamps, F. and Tackley, P. J. (2009). Searching for models of thermochemical convection that explain probabilistic tomography. II - Influence of physical and compositional parameters. *Physics of the Earth and Planetary Interiors*, 176:1–18.

Deschamps, F. and Trampert, J. (2004). Towards a lower mantle reference temperature and composition. *Earth and Planetary Science Letters*, 222(1):161–175.

Drake, M. J. and Righter, K. (2002). Determining the composition of the Earth. *Nature*, 416:39–44.

Dymkova, D. and Gerya, T. (2013). Porous fluid flow enables oceanic subduction initiation on Earth. *Geophysical Research Letters*, 40(21):5671–5676.

Dziewonski, A. M. and Anderson, D. L. (1981). Preliminary reference Earth model. *Physics of the Earth and Planetary Interiors*, 25:297–356.

Fiquet, G., Auzende, A. L., Siebert, J., Corgne, A., Bureau, H., Ozawa, H., and Garbarino, G. (2010). Melting of peridotite to 140 gigapascals. *Science*, 329(5998):1516–1518.

Forte, A. M. and Mitrovica, J. X. (2001). Deep-mantle high-viscosity flow and thermochemical structure inferred from seismic and geodynamic data. *Nature*, 410:1049–1056.

Garnero, E. J. (2000). Heterogeneity of the lowermost mantle. *Annual Review of Earth and Planetary Sciences*, 28:509–537.

Gerya, T. V. (2010). *Introduction to Numerical Geodynamic Modelling*. Cambridge University Press, Cambridge, UK.

Gomi, H., Ohta, K., Hirose, K., Labrosse, S., Caracas, M. J., Verstraete, M. J., and Hernlund, J. W. (2013). The high conductivity of iron and thermal evolution of the Earth's core. *Physics of the Earth and Planetary Interiors*, 224:88–103.

Hager, B. H. and O'Connell, R. J. (1981). A simple global model of plate dynamics and mantle convection. *Journal of Geophysical Research*, 86:4843–4867.

Haskell, N. A. (1935). The motion of a fluid under a surface load. *Physics*, 6:265–269.

Hedlin, M. A. H., Shearer, P. M., and Earle, P. S. (1997). Seismic evidence for small-scale heterogeneity throughout the Earth's mantle. *Nature*, 387:145–150.

Hernlund, J. W. and Tackley, P. J. (2008). Modelling mantle convection in the spherical annulus. *Physics of the Earth and Planetary Interiors*, 171:48–54.

Heron, P. J. and Lowman, L. P. (2014). The impact of Rayleigh number on assessing the significance of supercontinent insulation. *Journal of Geophysical Research*, 119:711–733.

Herzberg, C., Raterron, P., and Zhang, J. (2000). New experimental observations on the anhydrous solidus for peridotite KLB-1. *Geochemistry, Geophysics, Geosystems*, 1:DOI 10.1029/2000GC000089.

Hinton, G. and Salakhutdinov, R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313:504–507.

Hofmann, A. W. (1997). Mantle geochemistry: The message from oceanic volcanism. *Nature*, 385:219–229.

Hofmann, A. W. and Hart, S. R. (1978). An assessment of local and regional isotopic equilibrium in the mantle. *Earth and Planetary Science Letters*, 38:44–62.

Höink, T., Lenardic, A., and Richards, M. A. (2012). Depth-dependent viscosity and mantle stress amplification: Implications for the role of the asthenosphere in maintaining plate tectonics. *Geophysical Journal International*, 191:30–41.

Horbach, A., Bunge, H.-P., and Oeser, J. (2014). The adjoint method in geodynamics: derivation from a general operator formulation and application to the initial condition problem in a high resolution mantle circulation model. *International Journal on Geomathematics*, pages DOI 10.1007/s13137–014–0061–5.

Hornik, K., Stinchcombe, M., and White, H. (1989). Mulitlayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366.

Huang, Y., Chubakov, V., Mantovani, F., Rudnick, R. L., and McDonough, W. F. (2013). A reference Earth model for the heat-producing elements and associated geoneutrino flux. *Geochemistry, Geophysics, Geosystems*, 14(6):2003–2029.

Igel, C. and Hüsken, M. (2000). Improving the Rprop learning algorithm. In *Proceedings of the Second International Symposium on Neural Computation, NC'2000*, volume 2000, pages 115–121. ICSC Academic Press.

Ismail-Zadeh, A., Schubert, G., Tsepelev, I., and Korotkii, A. (2004). Inverse problem of thermal convection: Numerical approach and application to mantle plume restoration. *Physics of the Earth and Planetary Interiors*, 145:99–114.

Ismail-Zadeh, A. and Tackley, P. J. (2010). *Computational Methods for Geodynamics*. Cambridge University Press, Cambridge, UK.

Javoy, M., Kaminski, E., Guyot, F., Andrault, D., Sanloup, C., Moreira, M., Labrosse, S., Jambon, A., Agrinier, P., Davaille, A., and Jaupart, C. (2010). The chemical composition of the Earth: Enstatite chondrite models. *Earth and Planetary Science Letters*, 293:259–268.

Jenkins, J., Cottaar, S., White, R. S., and Deuss, A. (2016). Depressed mantle discontinuities beneath Iceland: Evidence of a garnet conrolled 660 km disconituity? *Earth and Planetary Science Letters*, 433:159–168.

Johnson, D. H. and Sinanović, S. (2000). Symmetrizing the Kullback-Leibler distance. Technical report, IEEE Transactions on Information Theory.

Johnson, K. T. M., Dick, H. J. B., and Shimizu, N. (1990). Melting in the oceanic upper mantle: An ion microprobe study of diopsides in abyssal peridotites. *Journal of Geophysical Research*, 95(B3):2661–2678.

Käufl, P., Valentine, A. P., de Wit, R. W. L., and Trampert, J. (2016). Solving probabilistic inverse problems rapidly with prior samples. *Geophysical Journal International*, 205:1710–1728.

Käufl, P., Valentine, A. P., O'Toole, T. B., and Trampert, J. (2014). A framework for fast probabilistic centroid-moment-tensor determination - Inversion of regional static displacement measurements. *Geophysical Journal International*, 196:1676–1693.

Kaus, B. J. P. and Podladchikov, Y. Y. (2001). Forward and reverse modeling of the three-dimensional viscous Rayleigh-Taylor instability. *Geophysical Research Letters*, 28(6):1095–1098.

King, S. D., Lowman, L. P., and Gable, C. W. (2002). Episodic tectonic plate reorganizations driven by mantle convection. *Earth and Planetary Science Letters*, 203(1):83–91.

Klein, E. M. and Langmuir, C. H. (1987). Global correlations of ocean ridge basalt chemistry with axial depth and crustal thickness. *Journal of Geophysical Research*, 92(B8):8089–8115.

Labrosse, S. (2015). Thermal evolution of the core with a high thermal conductivity. *Physics of the Earth and Planetary Interiors*, 247:36–55.

Labrosse, S., Hernlund, J. W., and Coltice, N. (2007). A crystallizing dense magma ocean at the base of the Earth's mantle. *Nature*, 450:866–869.

Le Pichon, K. and Huchon, P. (1984). Geoid, Pangea and convection. *Earth and Planetary Science Letters*, 67:123–135.

LeCun, Y., Bengio, Y., and Hinton, G. E. (2015). Deep learning. *Nature*, 521:436–444.

Lee, C.-T. A., Luffi, P., Hoink, T., Li, J., Dasgupta, R., and Hernlund, J. W. (2010). Upside-down differentiation and generation of a 'primordial' lower mantle. *Nature*, 463(7283):930–933.

Lee, C.-T. A., Luffi, P., Plank, T., Dalton, H., and Leeman, W. P. (2009). Constraints on the depths and temperatures of basaltic magma generation on earth and other terrestrial planets using new thermobarometers for mafic magmas. *Earth and Planetary Science Letters*, 279:20–33.

Lekić, V., Cottaar, S., Dziewonski, A. M., and Romanowicz, B. A. (2012). Cluster analysis of global lower mantle tomography: A new class of structure and implications for chemical heterogeneity. *Earth and Planetary Science Letters*, (357-358):68–69.

Lenardic, A. and Crowley, J. W. (2012). On the notion of well-defined tectonic regimes for terrestrial planets in the solar system and others. *The Astrophysical Journal*, 755(2):DOI 10.1088/0004–637X/755/2/132.

Liu, L. and Gurnis, M. (2008). Simultaneous inversion of mantle properties and initial conditions using an adjoint of mantle convection. *Journal of Geophysical Research*, 113(B08405):DOI 10.1029/2008JB005594.

Lorenz, E. N. (1963). Determinisitic nonperiodic flow. *Journal of atmospheric sciences*, 20:130–141.

Lourenço, D. L., Rozel, A., and Tackley, P. J. (2016). Melting-induced crustal production helps plate tectonics on Earth-like planets. *Earth and Planetary Science Letters*, 439:18–28.

Lowman, L. P., King, S. D., and Trim, S. J. (2011). The influence of plate boundary motion on planform in viscously stratified mantle convection models. *Journal of Geophysical Research*, 116:DOI 10.1029/2011JB008362.

Lyubetskaya, T. and Korenga, J. (2007). Chemical composition of Earth's primitive mantle and its variance: 1. Methods and results. *Journal of Geophysical Research*, 112(B03211):DOI 10.1029/2005JB004223.

MacKay, D. A. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, UK.

Masso, M. and Vaisman, I. I. (2008). Accurate prediciton of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. *Bioinformatics*, 24(18):2002–2009.

Matas, J., Bass, J., Ricard, Y., Mattern, E., and Bukowinski, M. S. T. (2007). On the bulk composition of the lower mantle: Predictions and limitations from

generalized inversion of radial seismic profiles. *Geophysical Journal International*, 170:764–780.

McDonough, W. F. (2016). *Deep Earth: Physics and Chemistry of the Lower Mantle and Core, Geophysical Monograph 217*, chapter The composition of the lower mantle and core. John Wiley and Sons, New York, USA.

McDonough, W. F. and Sun, S.-s. (1995). The composition of the Earth. *Chemical Geology*, 120:223–253.

McKenzie, D. and Bickle, M. J. (1988). The volume and composition of melt generated by extension of the lithosphere. *Journal of Petrology*, 29(3):625–679.

McLachlan, G. J. and Chang, S. U. (2004). Mixture modelling for cluster analysis. *Statistical Methods in Medical Research*, 13(5):347–361.

McNamara, A. K. and Zhong, S. (2004). Thermochemical structures within a spherical mantle: Superplumes or piles? *Journal of Geophysical Research*, 109(B07402):DOI 10.1029/2003JB002847.

McNamara, A. K. and Zhong, S. (2005). Thermochemical structures beneath Africa and the Pacific Ocean. *Nature*, 437:1136–1139.

Meier, U., Curtis, A., and Trampert, J. (2007). Global crustal thickness from neural network inversion of surface wave data. *Geophysical Journal International*, (169):706–722.

Moresi, L. and Solomatov, V. (1998). Mantle convection with a brittle lithosphere: Thoughts on the global tectonic styles of the Earth and Venus. *Geophysical Journal International*, 133:669–682.

Mosca, I., Cobden, L., Deuss, A., Ritsema, J., and Trampert, J. (2012). Seismic and mineralogical structures of the lower mantle from probabilistic tomography. *Journal of Geophysical Research*, 117(B06304):DOI 10.1029/2011JB008851.

Murthy, V. R., van Westrenen, W., and Fei, Y. (2003). Experimental evidence that potassium is a substantial radioactive heast source in planetary cores. *Nature*, 423:163–165.

Nakagawa, T. and Tackley, P. J. (2008). Lateral variations in CMB heat flux and deep mantle seismic velocity caused by a thermal–chemical-phase boundary layer in 3D spherical convection. *Earth and Planetary Science Letters*, 271:348–358.

Nakagawa, T. and Tackley, P. J. (2010). Influence of initial CMB temperature and other parameters on the thermal evolution of Earth's core resulting from thermochemical spherical mantle convection. *Geochemistry, Geophysics, Geosystems*, 11(6):DOI 10.1029/2010GC003031.

Nakagawa, T. and Tackley, P. J. (2014). Influence of combined primoridal layering and recycled morb on the coupled thermal evolution of earth's mantle and core. *Geochemistry, Geophysics, Geosystems*, 15:619–633.

Nakagawa, T., Tackley, P. J., Deschamps, F., and Connolly, J. A. D. (2009). Incorporating self-consitently calculated mineral physics into thermochemical mantle convection simulations in a 3-D spherical shell and its influence on seismic anomalies in Earth's mantle. *Geochemistry, Geophysics, Geosystems*, 10(3):DOI 10.1029/2008GC002280.

Nakagawa, T., Tackley, P. J., Deschamps, F., and Connolly, J. A. D. (2010). The influence of MORB and harzburgite composition on thermo-chemical mantle convection in a 3-D spherical shell with self-consistently calculated mineral physics. *Earth and Planetary Science Letters*, 296:403–412.

Nakagawa, T., Tackley, P. J., Deschamps, F., and Connolly, J. A. D. (2012). Radial 1-D seismic structures in the deep mantle in mantle convection simulations with self-consistently calculated mineralogy. *Geochemistry, Geophysics, Geosystems*, 13(11):doi:10.1029/2012GC004325.

Nomura, R., Hirose, K., Uesugi, K., Ohishi, Y., Tsuchiyama, A., Miyake, A., and Ueno, Y. (2014). Low core-mantle boundary temperature inferred from the solidus of pyrolite. *Science*, 343:522–525.

Olson, R., Sriver, R., Goes, M., Urban, N. M., Matthews, H. D., Haran, M., and Keller, K. (2012). A climate sensitivity estimate using Bayesian fusion of intrumental observations and an Earth System model. *Journal of Geophysical Research*, 117(D4):DOI 10.1029/2011JD016620.

O'Neill, H. S. C. and Palme, H. (2008). Collisional erosion and the nonchondritic composition of the terrestrial planets. *Philosophical Transactions of The Royal Society*, 366:4205–4238.

Palme, H. and O'Neill, H. S. C. (2003). *Treatise on Geochemistry*, volume 2, chapter Cosmochemcical estimates of mantle composition. Elsevier, Amsterdam, The Netherlands.

Pearson, D. G., Brenker, F. E., Nestola, F., McNeill, J., Nasdala, L., Hutchinson, M. T., Matveev, S., Mather, K., Silversmit, G., Schmitz, S., Vekemans, B., and L., V. (2014). Hydrous mantle transition zone indicated by ringwoodite included within diamond. *Nature*, 507:221–224.

Peltier, W. R. (1998). Postglacial variations in the level of the sea: Implications for climate dynamics and solid-Earth geophysics. *Reviews of Geophysics*, 36:603–689.

Peronne, M. P. and Cooper, L. N. (1993). *Artificial neural networks for speech and vision*, chapter When networks disagree: Ensemble methods for hybrid neural networks, pages 126–142. Chapman and Hall, London, UK.

Płonka, A., Blom, N., and Fichtner, A. (2016). The imprint of crustal density heterogeneities on regional seismic wave propagation. *Solid Earth*, 7:1591–1608.

Pozzo, M., Davies, C., Gubbins, D., and Alfè, D. (2014). Thermal and electrical conductivity of iron and iron-silicon mixtures at Earth's core conditions. *Earth and Planetary Science Letters*, 393:159–164.

Ratnaswamy, V., Stadler, G., and Gurnis, M. (2015). Adjoint-based estimation of plate coupling in a non-linear mantle flow model: Theory and examples. *Geophysical Journal International*, 202:768–786.

Ritsema, J., van Heijst, H. J., and Woodhouse, J. H. (1999). Complex shear wave velocity strucutre imaged beneath Africa and Iceland. *Science*, 286:1925–1928.

Rolf, T., Coltice, N., and Tackley, P. J. (2012). Linking continental drift, plate tectonics and the thermal state of the Earth's mantle. *Earth and Planetary Science Letters*, 351-352:134–146.

Rolf, T., Coltice, N., and Tackley, P. J. (2014). Statistical cyclicity of the supercontinent cycle. *Geophysical Research Letters*, 41:DOI 10.1002/2014GL059595.

Rudolph, M. L., Lekić, V., and Lithgow-Bertelloni, C. (2015). Viscosity jump in Earth's mid-mantle. *Science*, 350(6266):1349–1352.

Sambridge, M. (1999). Geophysical inversion with a Neighbourbood algorithm, I, Searching a parameter space. *Geophysical Journal International*, 138:479–494.

Sambridge, M. and Mosegaard, K. (2002). Monte Carlo methods in geophysical inverse problems. *Reviews of Geophysics*, 40(3):DOI 10.1029/2000RG000089.

Schaeffer, A. J. and Lebedev, S. (2015). *The Earth's Heterogeneous Mantle*, chapter Global heterogeneity of the lithosphere and underlying mantle: A seismological appraisal based on multimode surface-wave dispersion analysis, shear-velocity tomography, and tectonic regionalization, pages 3–46. Springer Geophysics. Berlin, Germany.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117.

Schubert, G., Turcotte, D. L., and Olson, P. L. (2001). *Mantle convection in the Earth and planets*. Cambridge University Press, Cambridge, UK.

Schuberth, B. S. A., Bunge, H.-P., and Ritsema, J. (2009). Tomographic filtering of high-resolution mantle circulation models: Can seismic heterogeneity be explained by temperature alone? *Geochemistry, Geophysics, Geosystems*, (Q05W03):DOI 10.1029/2009GC002401.

Seagle, C. T., Cottrell, E., Fei, Y., Hummer, D. R., and Prakapenka, V. B. (2013). Electrical and thermal transport proeprties of iron and iron-silicon alloy at high pressure. *Geophysical Research Letters*, 40:5377–5381.

Sharpe, H. N. and Peltier, W. R. (1978). Parameterized mantle convection and the Earth's thermal history. *Geophysical Research Letters*, 5(9):737–740.

Shephard, G. E., Flament, N., Williams, S., Gurnis, M., and Müller, R. D. (2014). Circum-Arctic mantle structure and long-wavelength topography since the Jurassic. *Journal of Geophysical Research*, 119:DOI 10.1002/2014JB011078.

Solomatov, V. S. (2001). Grain size-dependent viscosity convection and the thermal evolution of the Earth. *Earth and Planetary Science Letters*, 191:203–212.

Spiller, E. T., Bayarru, M. J., Berger, J. O., Calder, E. S., Patra, A. K., Pitman, E. B., and Wolpert, R. L. (2014). Automating emulator construction for geophysical hazard maps. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):126–152.

Stacey, F. D. and Anderson, O. L. (2001). Electrical and thermal conductivities of Fe-Ni-Si alloy under core conditions. *Physics of the Earth and Planetary Interiors*, 124(3-4):153–162.

Stamenković, V., Noack, L., Breuer, D., and Spohn, T. (2012). The influence of pressure-dependent viscosity on the thermal evolution of super-Earths. *The Astrophysical Journal*, 748(41):DOI 10.1088/0004–637X/748/1/41.

Stein, C., Lowman, L. P., and Hansen, U. (2013). The influence of mantle internal heating on lithospheric mobility: Implications for super-Earths. *Earth and Planetary Science Letters*, 361:448–459.

Steinberger, B. and Torsvik, T. H. (2012). A geodynamic model of plumes from the margins of Large Low Shear Velocity Provinces. *Geochemistry, Geophysics, Geosystems*, 13(1):DOI 10.1029/2011GC003808.

Stixrude, L. and Lithgow-Bertelloni, C. (2005). Thermodynamics of mantle minerals - I. Physical properties. *Geophysical Journal International*, 162:610–632.

Stixrude, L. and Lithgow-Bertelloni, C. (2011). Thermodynamics of mantle minerals - II. Phase equilibria. *Geophysical Journal International*, 184(3):1180–1213.

Tackley, P. J. (2008). Modelling compressible mantle convection with large viscosity contrasts in a three-dimensional spherical shell using the yin-yang grid. *Physics of the Earth and Planetary Interiors*, 171:7–19.

Tarantola, A. (2005). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society of Industrial and Applied Mathematics.

Thoraval, C. and Richards, M. A. (1997). The geoid constraint in global dynamics: Viscosity structure, mantle heterogeneity models and boundary conditions. *Geophysical Journal International*, 131:1–8.

Tirone, M., Buhre, S., Schmück, H., and Kaak, K. (2016). Chemical heterogeneities in the mantle: The equilibrium thermodynamic approach. *Lithos*, 244:140–150.

Tolstikhin, I. and Hofmann, A. W. (2005). Early crust on top of Earth's core. *Physics of the Earth and Planetary Interiors*, 148:109–130.

Torsvik, T. H., Steinberger, B., Gurnis, M., and Gaina, C. (2010). Plate tectonics and net lithosphere rotation over the past 150 My. *Earth and Planetary Science Letters*, 291:106–112.

Torsvik, T. H., van der Voo, R., Doubrovine, P. V., Burke, K., Steinberger, B., Ashwal, L. D., Trønnes, R. G., Webb, S. J., and Bull, A. L. (2014). Deep mantle structure as a reference frame for movements in and on the Earth. *Proceedings of the National Academy of Sciences*, 111(24):8735–8740.

Trampert, J., Deschamps, F., Resovsky, J., and Yuen, D. (2004). Probabilistic tomography maps chemical heterogeneities throughout the lower mantle. *Science*, 306:853–856.

Trim, S. J., Heron, P. J., Stein, C., and Lowman, L. P. (2014). The feedback between surface mobility and mantle compositional heterogeneity: Implications for the Earth and other terrestrial planets. *Earth and Planetary Science Letters*, 405:1–14.

Turcotte, D. L., Cooke, F. A., and Willeman, R. J. (1979). Parameterized convection within the moon and the terrestrial planets. *Proceedings of Lunar and Planetary Science Conference*, 3:2375–2392.

Valencia, D., O'Connell, R. J., and Sasselov, D. D. (2007). Inevitability of plate tectonics on super-Earths. *The Astrophysical Journal*, 670:L45–L48.

Valentine, A. P. and Trampert, J. (2012). Data space reduction, quality assessment and searcing of seismograms: Autoencoder networks for waveform data. *Geophysical Journal International*, (189):1183–1202.

van Boekel, R., Min, M., Leinert, C., Waters, L. B. F. M., Richichi, A., Chesneau, O., Dominik, C., Jaffe, W., Dutrey, A., Graser, A., Henning, T., de Jong, J., Köhler, R., de Koter, A., Lopez, B., Malbet, F., Morel, S., Paresce, F., Perrin, G., Preibisch, T., Przygodda, F., Schöller, M., and Wittkowski, M. (2004). The building blocks of planets within the 'terrestrial' region of protoplanetary disks. *Nature*, 432:479–482.

van der Hilst, R. D., Widiyantoro, S., and Engdahl, E. R. (1997). Evidence for deep mantle circulation from global tomography. *Nature*, 386:578–584.

van Heck, H. and Tackley, P. J. (2011). Plate tectonics on super-Earths: Equally or more likely than on Earth. *Earth and Planetary Science Letters*, 310:252–261.

Weller, M. B., Lenardic, A., and O'Neill, C. (2015). The effects of internal heating and large scale climate variations on tectonic bi-stability in terrestrial planets. *Earth and Planetary Science Letters*, 420:85–94.

Whitehouse, P. L., Bentley, M. J., Milne, G. A., King, H., and Thomas, I. D. (2012). A deglacial model for Antarctica: Geological constraints and glaciological modelling as a basis for a new model of Antartic glacial isostatic adjustment. *Quaternary Science Reviews*, 32:1–24.

Wookey, J., Kendall, J. M., and Barruol, G. (2002). Mid-mantle deformation inferred from seismic anisotropy. *Nature*, 415:777–780.

Workman, R. K. and Hart, S. R. (2005). Major and trace element composition of the depleted MORB mantle (DMM). *Earth and Planetary Science Letters*, (231):53–72.

Worthen, J., Stadler, G., Petra, N., Gurnis, M., and Ghattas, O. (2014). Towards adjoint-based inversion for rheological parameters in nonlinear viscous mantle flow. *Physics of the Earth and Planetary Interiors*, 234:23–34.

Xu, W., Lithgow-Bertelloni, C., Stixrude, L., and Ritsema, J. (2008). The effect of bulk composition and temperature on mantle seismic structure. *Earth and Planetary Science Letters*, 275:70–79.

Yoshida, M. (2008). Mantle convection with longest-wavelength thermal heterogeneity in a 3-D spherical model: Degree one or two? *Geophysical Research Letters*, 35:DOI 10.1029/2008GL036059.

Zerr, A., Diegeler, A., and Boehler, R. (1998). Solidus of Earth's deep mantle. *Science*, 281(5374):243–246.

Zhang, N., Zhong, S., Leng, W., and Li, Z.-X. (2010). A model for the evolution of the Earth's mantle structure since the Early Paleozoic. *Journal of Geophysical Research*, 115:1029/2009JB006896.

Zhong, S., Zhang, N., Li, Z.-X., and Roberts, J. H. (2007). Supercontinent cycles, true polar wander, and very long-wavelength mantle convection. *Earth and Planetary Science Letters*, 261:551–564.

Ziv, J. and Lempel, A. (1977). A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, 23:337–343.

Zunino, A., Connolly, J. A. D., and Khan, A. (2011). Precalculated phase equilibrium models for geophysical properties of the crust and mantle as a function of composition. *Geochemistry, Geophysics, Geosystems*, 12(4):DOI 10.1029/2010GC003304.

Samenvatting (Summary in Dutch)

Convectie in de aardmantel is een niet-lineair systeem. Also gevolg hiervan is het moeilijk om de geschiedenis van de aardmantel te reconstrueren. In dit proefschrift presenteer ik een nieuwe methode om convectie terug te spoelen. Ik gebruik patroonherkenningsalgoritmen die gebruik maken van neurale netwerken. Deze herkennen onderliggende statistische patronen die informatie over het convectiesysteem bevatten. Hierdoor kan ik parameters zoals mantelcompositie of de zogeheten 'yield stress' vinden en ook historische kenmerken zoals de geschiedenis van LLSVPs. Verder kan ik de neurale netwerken gebruiken om de convectiepatronen te voorspellen. Alles wordt gedaan in een probabilistische Bayesiaanse context met volledige onzekerheidsanalyse. Daarom is dit onderzoek helemaal nieuw in de geodynamica.

Acknowledgements

I'm really glad that I took this PhD project. It probably rates as one of the best decisions I ever made and gave me the chance to meet and work with a bunch of really great people, to learn a lot about almost every aspect of geophysics, and generally to broaden my horizons more than I ever expected. So first off, I'd like to thank Jeannot for hiring me on a such a cool project and making it happen. Most people seem to write about ups and downs with their supervisor, but it genuinely felt like a very smooth journey. Although I didn't always take your advice about not drinking with geodynamicists...

My reading thesis committee, Boris, Malcolm, Nicholas, Patrick and Paul, provided some really helpful comments on my thesis, so than you very much for taking the time to do so.

The Utrecht seismologists have always been fun and supportive. I worked most closely with Andrew. We've shared a lot of beers, gossip, moans, a few arguments and you've given me a lot of scientific support, not least the challenge of trying to turn my scientific English into something intelligible! I definitely owe you a beer. Denise has always been a wonderfully supportive office mate, surprisingly tolerant of my untidiness and heated political discussions with Maria. We've had three offices together, bonding over multiple asbestos crises, tea, skating and Christmas parties. Maria and I have have discussed/rowed our way through almost every political topic under the sun and are surprisingly still friends. Things seem to have been strategically organised so that Nienke and I never shared an office, which was probably better for everyone else's sanity, given the likely ensuing noise levels! As internationally (in)famous partners in conference crime, we've certainly had some good times. Sahar has been the bringer of laughter, skating, saffron and and dancing to the group. And to all the other seismologists, Ralph, Paul, Wen, Agnieszka, Laura, Nesli, Elmer, Su,

Simon, Hanneke, Arwen, Ivan, Kabir, Jacqueline, Areti, and last but in no ways least, Henk, Theo and Arie, a big thank you for all the support in so many ways.

I'd also like to thank everyone in the geophysical fluid dynamics group at ETH Zürich, both for all their help and support during my PhD and for making me so welcome as a new colleague. Particular thanks goes to Paul who lead the Zürich-half of the iGEO project and has been a great source of ideas and support, and included me in the broader research group throughout my PhD. The other Zürich-based members of the iGEO project, Antoine, Charitra and Ilya have also been so welcoming and inclusive. We all got to know one another, and many more people, when we were isolated in a Japanese hotel with not much besides sake and a view of Fuji, which has to be one of the best ways to start.

Whilst probably detrimental to my scientific output, all of the Usual Suspects (you know who you are) have made my time in Utrecht so much fun. Highlights include the floating hot tub in Rotterdam harbour, swimming in the Fort Hoofddijk lake and some brilliant nights out.

Finally, I'd like to thank my husband Chris without whom I probably wouldn't ever as got as far as needing to write the acknowledgements for a thesis. From baking me cake to bribe my colleagues to supporting me through all the ups and downs, I couldn't have done it without you!

Curriculum Vitae

I grew up in the UK. Having completed high school in Birmingham, I spent a year working in geotechnical engineering for both Atkins (no relation!) and AECOM under the Year in Industry scheme. After deciding that embankments, groundwater remediation and great crested newts really were not going feature in my future, I started a degree in Earth Sciences at St Anne's College, Oxford. There I survived weeks of fieldwork with my sanity in tact before escaping to the relative civilisation of the seismology lab for a masters project with John Woodhouse. I then moved to Utrecht to study for my PhD. I am now continuing my research in machine learning and geodynamics at ETH Zürich.