Physics of the Earth and Planetary Interiors 257 (2016) 171–186

Contents lists available at ScienceDirect

Physics of the Earth and Planetary Interiors

journal homepage: www.elsevier.com/locate/pepi

Using pattern recognition to infer parameters governing mantle convection

Suzanne Atkins [a,*], Andrew P. Valentine [a], Paul J. Tackley [b], Jeannot Trampert [a]

[a] Department of Earth Sciences, Utrecht University, P.O. Box 80.115, 3508 TC Utrecht, The Netherlands
[b] Institute of Geophysics, ETH Zurich, Sonneggstrasse 5, 8092 Zurich, Switzerland

ARTICLE INFO

Article history:
Received 5 July 2015
Received in revised form 19 May 2016
Accepted 27 May 2016
Available online 3 June 2016

Keywords:
Mantle convection
Mantle evolution
Probabilistic inversion
Neural networks

ABSTRACT

The results of mantle convection simulations are fully determined by the input parameters and boundary conditions used. These input parameters can be for initialisation, such as initial mantle temperature, or can be constant values, such as viscosity exponents. However, knowledge of Earth-like values for many input parameters are very poorly constrained, introducing large uncertainties into the simulation of mantle flow. Convection is highly non-linear, therefore linearised inversion methods cannot be used to recover past configurations over more than very short periods of time, which makes finding both initial and constant simulation input parameters very difficult. In this paper, we demonstrate a new method for making inferences about simulation input parameters from observations of the mantle temperature field after billions of years of convection. The method is fully probabilistic. We use prior sampling to construct probability density functions for convection simulation input parameters, which are represented using neural networks. Assuming smoothness, we need relatively few samples to make inferences, making this approach much more computationally tractable than other probabilistic inversion methods. As a proof of concept, we show that our method can invert the amplitude spectra of temperature fields from 2D convection simulations, to constrain yield stress, surface reference viscosity and the initial thickness of primordial material at the CMB, for our synthetic test cases. The best constrained parameter is yield stress. The reference viscosity and initial thickness of primordial material can also be inferred reasonably well after several billion years of convection.

1. Introduction

Convection in the mantle is a major Earth process and plays a crucial role in driving plate tectonics and the vertical motion of the Earth's surface, changing sea-level, erosion and sedimentation rates, and therefore climate (Spasojevic and Gurnis, 2012), whilst changes in insulation at the core mantle boundary affect the heat flux (e.g. Nakagawa and Tackley, 2010), which in turn affects the intensity, dipolarity and stability of the magnetic field (Aubert et al., 2009). To understand these and many more Earth processes, we must understand mantle convection throughout Earth's history.

In this paper we present a new probabilistic method for gaining information on the governing parameters for mantle convection. Mantle flow can be simulated by solving a coupled system of conservation, momentum and energy equations, and many codes exist with which to model mantle convection. These codes tend to be very computationally expensive. The success with which Earth-like convection is simulated depends on approximations in the equations and the choice of values for parameters such as viscosity and density structure, phase changes, initial temperature and the distribution of continents. However, many of these parameters are very poorly constrained by geological, geochemical or geophysical observations. Seismic tomography, for instance, provides maps of heterogeneities in wave speeds in the mantle, but these have large uncertainties, and separating the contributions of chemistry and temperature to these heterogeneities is challenging (e.g. Trampert et al., 2004; Schuberth et al., 2009). The locations and amplitude of seismic discontinuities in the mantle give an indication of phase changes and therefore mineralogy (e.g. Deuss, 2009), but are hampered by large uncertainties (e.g. Koroni and Trampert, 2016). Viscosity structure can be found from glacial rebound and geoid studies (e.g. Mitrovica and Forte, 2004; Rudolph et al., 2015) and geodynamic modelling (e.g. Forte and Mitrovica, 2001). These viscosity inferences rely on many

* Corresponding author.
E-mail addresses: s.e.atkins@uu.nl (S. Atkins), a.p.valentine@uu.nl (A.P. Valentine), paul.tackley@erdw.ethz.ch (P.J. Tackley), j.a.trampert@uu.nl (J. Trampert).

http://dx.doi.org/10.1016/j.pepi.2016.05.016
0031-9201/

assumptions and the resulting uncertainties are poorly understood. Plate reconstructions provide a history of the surface motion of the mantle, constraining flow evolution and viscosity (e.g. Bower et al., 2013; Worthen et al., 2014). Detailed reconstructions are available for about 150 Ma (Torsvik et al., 2010), before which reconstructions must rely on inferences which contain much greater uncertainty (e.g. Stampfli and Borel, 2002). Given the computational expense of convection simulations, these uncertainties present a great challenge to geodynamists. Finding new ways to constrain parameters such as these is therefore valuable.

Because mantle convection can be simulated, we could in principle apply inverse theory to try to find the input parameters necessary for a mantle convection simulation to be Earth-like. However, no general solution to the inverse flow equations exists and the physics of mantle convection is not completely understood meaning that many approximations are necessary. Furthermore, mantle convection is a non-linear and chaotic process, therefore even small changes in the initial distribution of thermochemical anomalies between two simulations can lead to large differences in the end-state of the simulation (Bello et al., 2014), making inversion for simulation input parameters challenging. Nevertheless, several attempts have been made to infer initial conditions and physical parameters from geodynamic data.

One possible approach is to solve the time dependent flow equations in reverse. This means that, starting with the present-day state of the mantle, the convection can be forced to flow backwards. Conrad and Gurnis (2003) found this method to be successful for short timescales, but, because thermal diffusion is time-irreversible and must thus be neglected and because of the chaotic nature of convection, backwards simulations cannot be used for more than about 75 Myr. Eventually backwards convection produces a stable stratified configuration (Kaus and Podladchikov, 2001). Even without the limitations imposed by thermal diffusion, the uncertainties and errors in the initial mantle state grow uncontrollably going back in time. Bello et al. (2014) showed that small perturbations in the temperature structure of two otherwise identical mantle convection simulations caused them to diverge unrecognisably after around 95 Myr of run time, which applies equally to simulations run forward or backward in time. 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. 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). However, this also suffers from the chaotic nature of convection, with small differences or errors in the starting conditions in the forward or the adjoint simulations potentially causing large differences in the inversion result. By assimilating geological observations, such as plate reconstructions (e.g. Bower et al., 2013; Shephard et al., 2014), this time limit may be extended. 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 assimilation of data adds further uncertainty into the process and these uncertainties may not be well quantified in, for instance, geological observations or seismic tomography.

The effects of varying convection simulation parameters are generally investigated a few at a time by running multiple simulations (e.g. Deschamps and Tackley, 2008; Lenardic and Crowley, 2012; Rolf et al., 2014). Conducted on a larger scale, this sampling can be used for sampling-based inversion. Sampling based inversions can be put into a probabilistic framework and include all of the non-linearities in the simulation, allowing us to tackle inversion of convection, potentially over any timescale. In this paper we demonstrate that a sampling based inversion method using pattern recognition, as described by Käufl et al. (2016), can be applied successfully to mantle convection simulations.

Sampling based approaches require large numbers of forward simulations to explore the relationship between the inputs for the convection simulations and the end-state of the simulated mantle. The inputs can be varied to cover a wide range of values (which we refer to as prior sampling) or to preferentially produce simulation results which resemble the observations (known as posterior sampling). Monte Carlo methods use the latter approach (e.g. Sambridge and Mosegaard, 2002). Before sampling begins, a misfit function is defined between the end-state of the simulation and an observation. The simulation inputs may then be varied to preferentially find end-states with low misfit, by using algorithms such as the neighbourhood algorithm of Sambridge (1999). The resulting distribution of models approximates the probability distribution of simulation parameters being responsible for the observation. Monte Carlo methods are widely used in geophysics (e.g. Forte et al., 2002; Cobden et al., 2012; Höink et al., 2013; Austermann et al., 2014; Baumann et al., 2014).

While a posterior sampling Monte Carlo technique could in principle be applied to our problem, the computational costs involved make this impractical. Instead, we use a prior sampling approach which has several benefits when compared with a traditional posterior approach. We assume that the probability distributions are smooth between sample simulations, allowing us to interpolate between them. This means that we can work with far fewer convection simulations, making the inverse approximation tractable despite the computational expense of running convection simulations. Secondly, our method is very flexible. The inputs to the convection simulations are selected without aiming to replicate any observation of Earth, and each simulation produces many outputs, including temperature structure, seismic velocity, and gravity, meaning they can be reused to investigate many different observations without the need for further sampling.

Despite the rapid evolution of differences between models seen in the experiments of Bello et al. (2014), certain statistics of convection remain stable with respect to the input parameters, i.e. although the exact positions of upwellings and downwellings may rapidly diverge between simulations, statistical measures of the convective pattern do not. In this paper, we show that the amplitude spectra of the temperature fields resulting from two-dimensional convection simulations contain sufficient information on certain input parameters of the simulations that we can use them to perform inversions for these parameters after several billion years of convection.

2. Method

Inversion methods formally consist of finding a mapping between a region in some data space and a region in a model space. The model space contains each parameter which is necessary to simulate the observations (which belong to the data space) given some theoretical relationship described by a mathematical function. In our study, the model parameters are the input parameters to the convection simulation code StagYY, both initial conditions such as temperature profile and the distribution of chemical heterogeneities, and constant convection parameters such as the reference viscosity. The convection code implements the theoretical relationship between the model parameters and the observations in the data space, in this case the amplitude spectrum of the temperature field of the mantle, according to a set of assumptions.

Probabilistic inversion techniques use a posterior probability density function (PDF) to capture the extent to which a particular

value of a model parameter is likely to be responsible for an observation. The PDF describes the probability that the model parameter lies within any range of values. The conditional posterior PDF, the most general way to express the solution to an inverse problem, can be calculated using Bayes' theorem:

$$P(\mathbf{m}|\mathbf{d}) = \frac{P(\mathbf{m})P(\mathbf{d}|\mathbf{m})}{P(\mathbf{d})} \quad (1)$$

where $\mathbf{d}$ is an observation; in our case, the amplitude spectrum of the mantle temperature field. We use the vector $\mathbf{m}$ to describe the model parameter values which determine $\mathbf{d}$. $P(\mathbf{m})$ is the prior probability density of each element of $\mathbf{m}$, which contains all the information that we know about each parameter before we consider any observation. For our purposes, $P(\mathbf{d})$ is a normalisation constant. $P(\mathbf{d}|\mathbf{m})$ describes the extent to which observations are compatible with a given set of model parameters, and is called the likelihood of $\mathbf{m}$. Technically this is not a PDF but a description of how well a particular parameter explains the data. Most Monte Carlo techniques directly evaluate the right-hand-side of Eq. (1) (e.g. Mosegaard and Tarantola, 1995; Sambridge, 1999). This is referred to as posterior sampling, because for each sample the likelihood is explicitly evaluated.

We use neural networks to directly represent the marginal PDF of each input parameter to the convection simulations without ever evaluating the right-hand-side of Eq. (1). Given a set of samples drawn from the prior model space, we calculate the corresponding observations of the convective state at time t using a convection simulation.

The neural network represents the marginal PDF for all possible data corresponding to the prior model space, interpolating between samples, and can evaluate the marginal PDF very quickly for any new observation shown to the network. If we assume that the variations in the joint data-model space are smooth with respect to variations in the data, we need many fewer samples than in more traditional Monte Carlo techniques (Käufl et al., 2016).

The major advantage of using samples which are distributed according to prior distributions is that they are not tuned to any particular observation. They can therefore be reused repeatedly to find an inference of the model parameters associated with any observation. This is a particular strength for geophysical applications. For example, there are many seismic tomographic models of the mantle, which vary in their details and amplitudes. We can therefore reuse the same set of samples to invert different tomographic models, or the temperature structures derived from them, to compare how robust our inferences are with respect to different tomographic models. We can also compare *a priori* the expected resolution of each model parameter by comparing the information contained in the posterior PDFs for the synthetic samples, given, for example, gravity maps versus seismic tomographic images, allowing us to focus on the best Earth observation to use to infer a particular parameter if we want to consider real data. This can all be done with a single suite of convection simulations, provided that the prior ranges of the simulation input parameters are appropriate.

In principle, prior sampling and posterior Monte Carlo sampling methods should give identical results if an infinite number of samples are available. In practice, with a finite number of samples, there is no guarantee that a specific implementation of a neural network is performing optimally or returning the same results as a traditional Monte Carlo method would. However, in our experience, we find that interpolation between samples from the prior space produces a conservative estimate for the posterior PDF compared to a Metropolis–Hastings inference (Käufl et al., 2016). By this, we mean that the PDF obtained by a Metropolis–Hastings inference will typically have a lower variance than that inferred from prior samples and will thus provide a stronger constraint on the solution. This is unsurprising, given the more targeted nature of Metropolis–Hastings sampling. However, the parameter range indicated by Metropolis–Hastings sampling is usually a subset of the range indicated by prior sampling, so the results from the two approaches may be regarded as compatible.

The uncertainty indicated by the network decreases with higher sampling density. If samples are too widely spaced, the networks simply make an inference with very wide standard deviation, and therefore great uncertainty. The sampling density and the smoothness between samples varies in different regions of the model space. We may therefore be able to make inferences in some regions and not in others.

We initialise our networks to return the prior model parameter distribution before training begins, and the networks should only move away from the prior if there is positive information about this parameter in the training set. If the network can find no relationship between the temperature structure and the model parameter, it will simply return the prior distribution for that model parameter.

To describe our procedure in more detail, we follow the flowchart in Fig. 1. The model parameter vector $\mathbf{m}$ in our study has 29 dimensions, which are the inputs to the 2D mantle convection

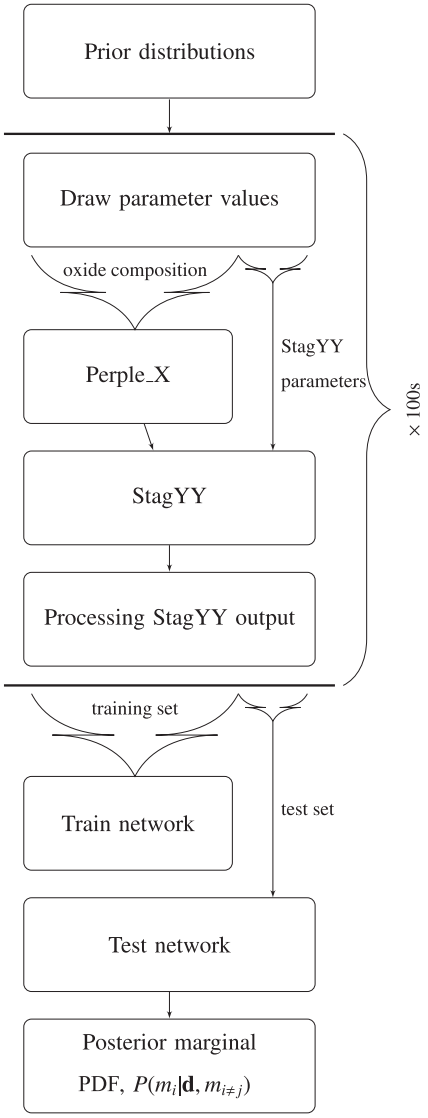

Fig. 1. Workflow to train a network to infer a convection model parameter.

simulation code StagYY (Tackley, 2008; Hernlund and Tackley, 2008). Details of the simulation setups can be found in Appendix A. Every dimension has an independent prior PDF, and the parameters are sampled randomly from these PDFs. The prior PDFs are chosen to cover a wide region of the model space and to include values very different to those expected for Earth. Twelve of these dimensions are direct input parameters, and include both initial conditions and constant physical values, such as initial core temperature, initial mantle potential temperature, yield stress and reference viscosity. The ranges of the PDFs from which each are selected can be found in Table A.1 in Appendix A. The other 17 dimensions are the major element components of the three rock types used in the model, which are given in Tables A.2 and A.3. The resulting mineral physics properties are calculated using the Perple_X package (Connolly, 2009), using the database of Stixrude and Lithgow-Bertelloni (2011). The runtime of the simulations varies massively with the input values, and some input values give rise to computational difficulties, particularly for high initial core temperatures.

Our simulation setup produces observations at set time intervals. We consider separately all of the simulations which have run for 0.4, 1, 2 or 3 Gyr (see (Table B.4) for the exact number), and train networks at each of these stages to retrieve the input parameters. We currently only have enough completed simulation runs to consider up to 3 Gyr of convection, due to limited computational resources, but it is straightforward to extend the study to 4.5 Gyr.

The outputs produced by StagYY include temperature, density and viscosity, each stored as a grid with resolution 64×512 points radially and horizontally, respectively. In this study, we consider the temperature field only. We first post-process the temperature field to condition it to be suitable for efficient network training. We use the amplitude spectrum of the temperature field rather than the original field, since removing the phase renders the observables more stable with respect to the small random perturbations used to initiate convection than observations in the spatial domain. Fig. 2 demonstrates how models which differ in the spatial domain are similar in the spectral domain. We take 64 1D Fourier transforms of the temperature field, one at each depth slice. The full set of amplitude spectra also has dimensions 64×512. To ease network training, we reduce the dimensionality of this input vector. First we only consider the longest wavelength features and retain degrees 0−10, reducing the dimensionality of the amplitude spectrum to 64×11. The dimensionality is then further reduced by using an auto-encoding neural network, developed by Valentine and Trampert (2012), based on the work of Hinton and Salakhutdinov (2006). The auto-encoder produces a lower dimensional representation of the amplitude spectrum. The encoding process reduces the 64×11 elements of the amplitude spectrum to 28 discrete numbers in the encoded version. By trial and error we 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 loss-less, with the loss being in the fine details of the amplitude spectrum, as can be seen in the example in Fig. 3. The spectrum is smoothed, but the broad patterns are retained.

The encoded amplitude spectrum can then be used to train the mixture density network (MDN) (Bishop, 1995; MacKay, 2003). Neural networks have been used for a variety of applications in Earth sciences, including inversion. Examples from seismology including finding the depth of the Moho discontinuity using surface waves (Meier et al., 2007), determining radial structure of the Earth from body-wave traveltimes (de Wit et al., 2013), earthquake source centroid moment tensor determination (Käufl et al., 2014) and automated identification of seamounts from bathymetric data (Valentine et al., 2013).

Visualising and interpreting a high dimensional PDF such as the left-hand-side of Eq. (1) can be challenging. We therefore choose to work with its marginalised form, so that only one parameter is considered at a time. The marginal PDF is

$$P(m_i|\mathbf{d}) = \int P(\mathbf{m}|\mathbf{d}) \prod_{j \neq i} \mathrm{d}m_j \tag{2}$$

This PDF for parameter i depends on the variations of every other parameter ($j \neq i$) and takes into account the covariance between the included and excluded model parameters.

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. Details of the network architecture are included in Appendix B. For each training observation the true value of the model parameter of interest is known.

The posterior PDF is parameterised using a mixture of Gaussians. The trained MDN outputs the mean, standard deviation and a weighting factor for three to five of these Gaussian kernels, which together give a PDF representing the marginal posterior distribution. This PDF encapsulates our knowledge on a model parameter given a particular observation and the choices made during network training.

3. Proof of concept

We assess the performance of each committee of networks (see Appendix B for details) after training by carrying out a series of synthetic tests. We 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, we can be confident that any positive results we 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 the training and test sets for each age group are given in Table B.4 in Appendix B.

We use the Kullback–Leibler distance

$$D_{KL} = \int P(m_i) \log_2 \frac{P(m_i)}{P(m_i|\mathbf{d}, m_{j \neq i})} dm_i \tag{3}$$

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 (Tarantola and Valette, 1982). 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}. Fig. 4 shows the Kullback–Leibler distance between a Gaussian mixture approximation to a uniform distribution ($P(m_i)$ in Eq. (3)) and Gaussian distributions with decreasing standard deviation, $P(m_i|\mathbf{d}, m_{j \neq i})$. 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.

The posterior PDF, and therefore the D_{KL}, include all the uncertainties from imperfect sampling, the assumption of smoothness between samples, the information content of the observation,

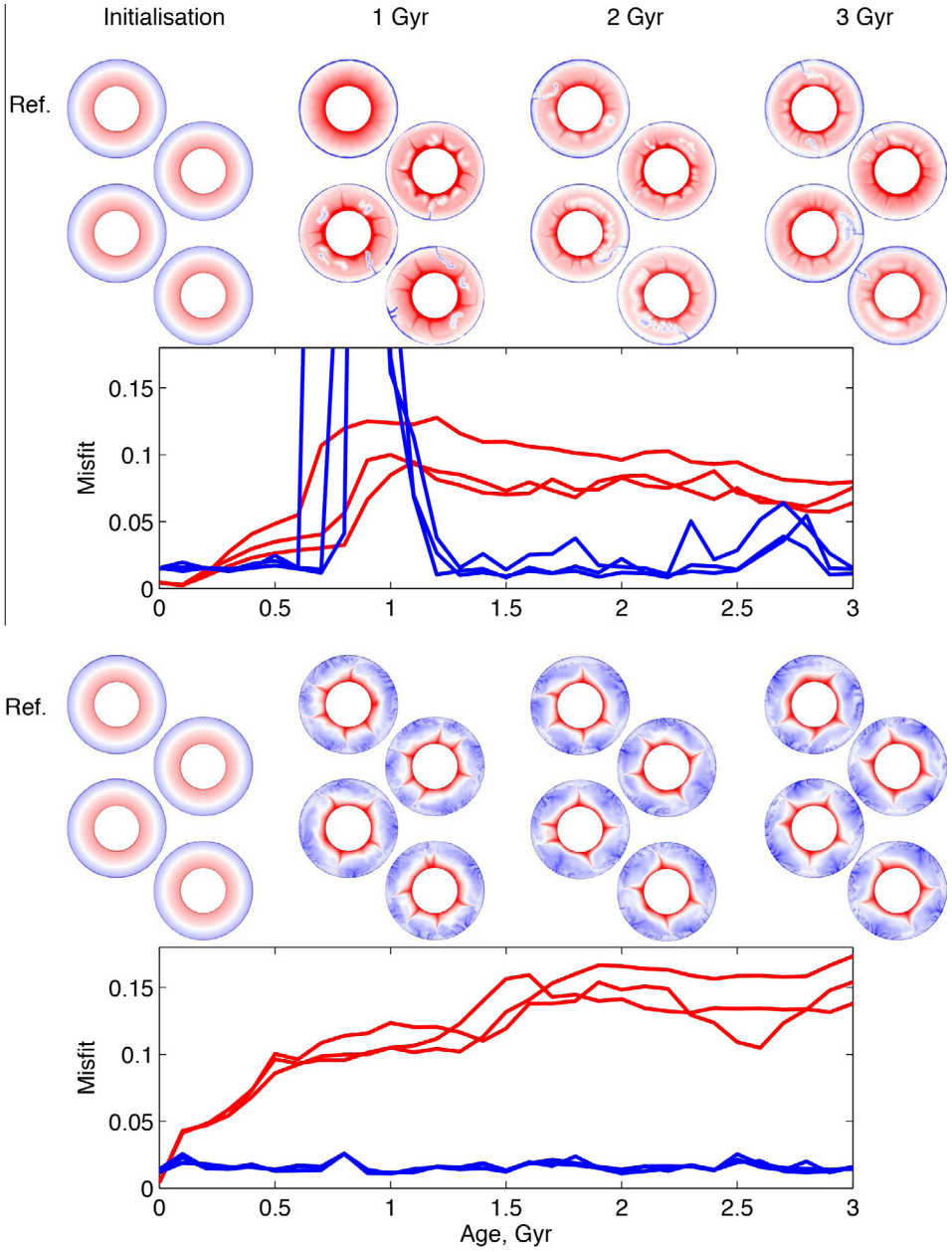

Fig. 2. The misfit between the full temperature field (red) and amplitude spectra (blue) when four simulations are started with identical model parameters, but initial 20 K perturbations located in different places. Two different comparison sets are run with different model parameters. The misfit is calculated according to the method of Bello et al. (2014), where the misfit $= 1/N \sum |t_{ij} - t_{ij}^{\text{ref}}|$, 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. The large peak in the misfit for the upper set of simulations 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. The relative error in the amplitude spectra peaks at 1.4, which is a very significant difference.

and the network training procedure, for every inversion performed. The posterior PDF is dependent on the observation and all the other varying model parameters, therefore the PDF and D_{KL} are different for every simulation in the test set.

We 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. Fig. 5 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.

The networks may find less information after longer simulation run times or not gain as much information on the other parameters for a variety of reasons. It may be that there is no information to be learnt on that input parameter from the observations shown to the network. Alternatively, a signal may be present in the temperature

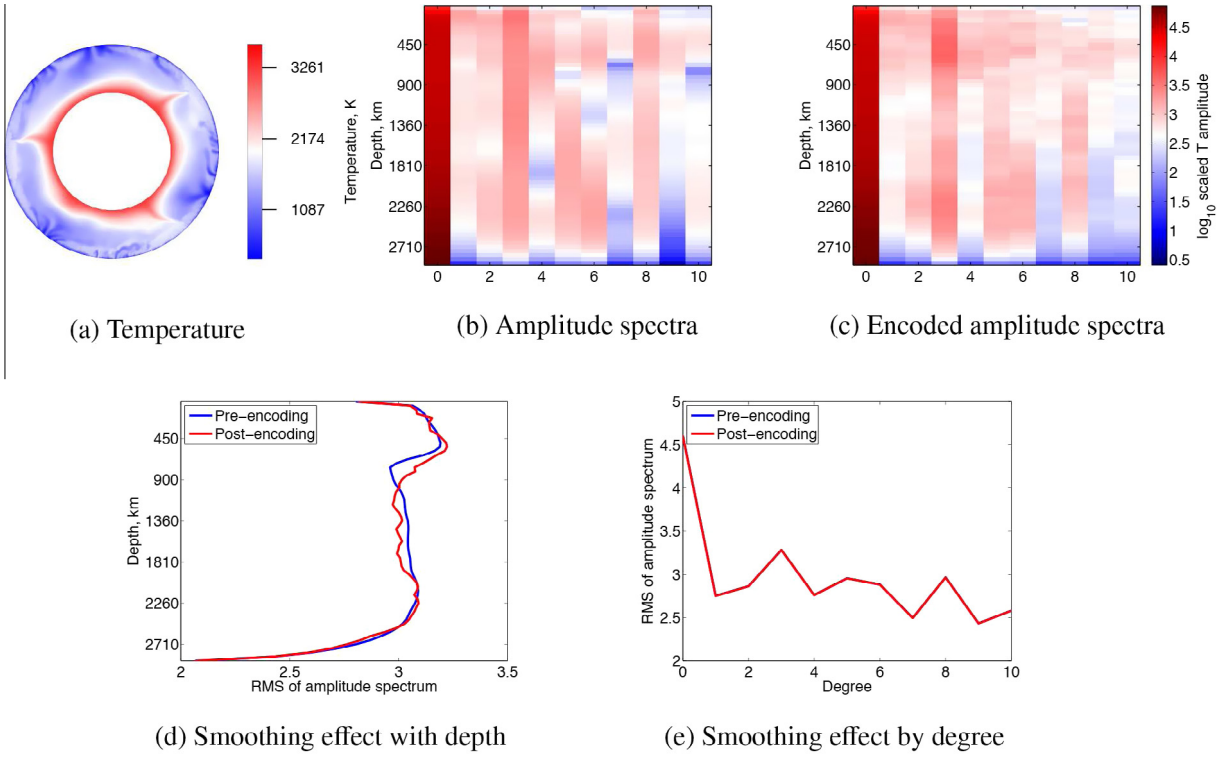

Fig. 3. Example of the effects of using an auto-encoding neural network to reduce dimensionality. (a) The original temperature field; (b) the original amplitude spectrum for the temperature field for degree 0–10; (c) amplitude spectrum after encoding and decoding. Both spectra are on the same scale. 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.

field, but the training set may not contain enough simulation samples to allow the network to find a mapping between observation and input parameter. It is impossible to distinguish these causes and we have no way to estimate how many samples we will need before training the network. We stress that although we do not find these parameters in this study they are not necessarily unknowable, 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.

However this is not going to mislead our inferences as our 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 we are 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 mixture density neural networks and Monte Carlo techniques can be found in Käufl et al. (2016). We have fewer samples at greater ages, therefore we would expect the D_{KL} to decrease with age, unless this is compensated by an increase of information in the data.

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, we 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 we 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, we train different networks with standard deviations of noise levels of 10, 50 and 100 K. The noise level that produces the highest mean information gain for the test set depends on the simulation input parameter of interest. The noise is added to the mantle temperature field before any dimensionality reduction takes place. For a given network, the noise has the same standard deviation at all depths throughout the mantle. If we 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. Fig. 6 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

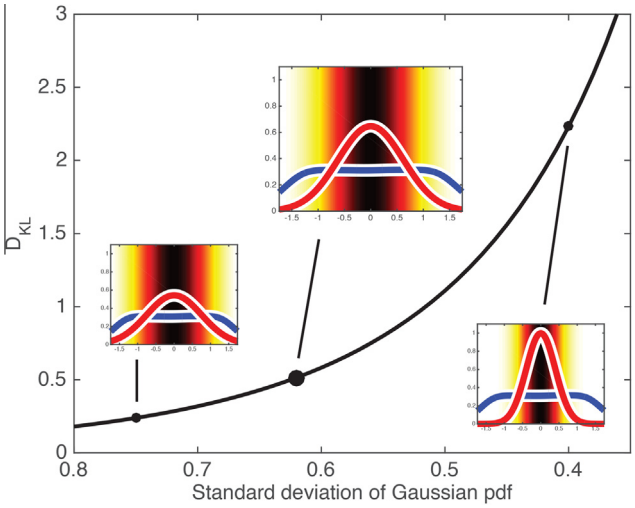

Fig. 4. D_{KL} between a Gaussian mixture approximation to a uniform distribution and Gaussian distributions with varying standard deviation. The inset distributions show Gaussians with standard deviation of 0.75, 0.62 and 0.4 respectively (in red), with the Gaussian mixture distribution plotted in blue behind each. The D_{KL} is 0.24, 0.5 and 2.23 respectively. The colour in the background corresponds to those used in Fig. 8. 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

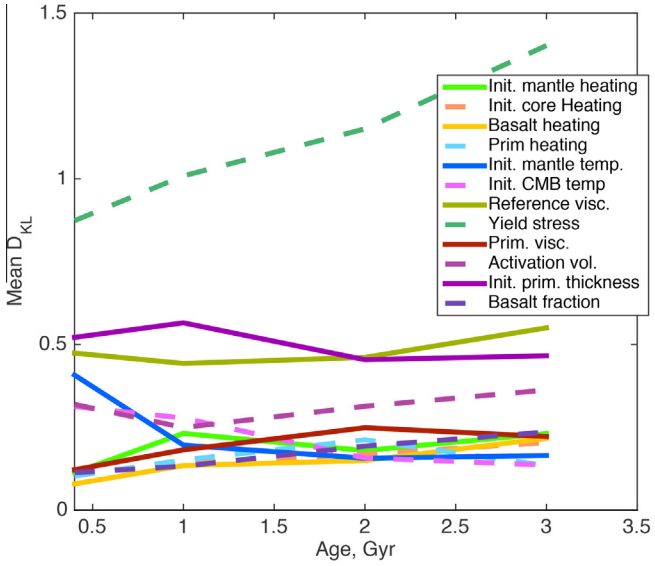

Fig. 5. Mean D_{KL} between the prior distribution of parameter values in the training set and the network calculated posterior distribution for the test set simulations. Each point represents the outputs for one committee of networks, trained to find the particular model parameter.

and the true model parameter value of each simulation in the test set in terms of the variance of the 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 Fig. 5 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 us to get more insight into the inversion, for example to look at how performance varies in different regions of the model space. Fig. 8 shows 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 B.4. 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 Fig. 7. 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. Fig. 7 demonstrates how the PDFs for six test simulations are placed into the grids in Fig. 8. If the network effectively infers 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 Fig. 8(c). We also need to know how certain the networks are, therefore underneath each grid we 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 Fig. 4. 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 network, shown in Fig. 8(a)–(d), 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 over-estimates 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 we can make inferences about these parameters from the temperature field. We discuss this further in Section 4.

The thickness of primordial material can be determined from the temperature field after 0.4 Gyr (Fig. 8(i)). 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 (Fig. 8(l)).

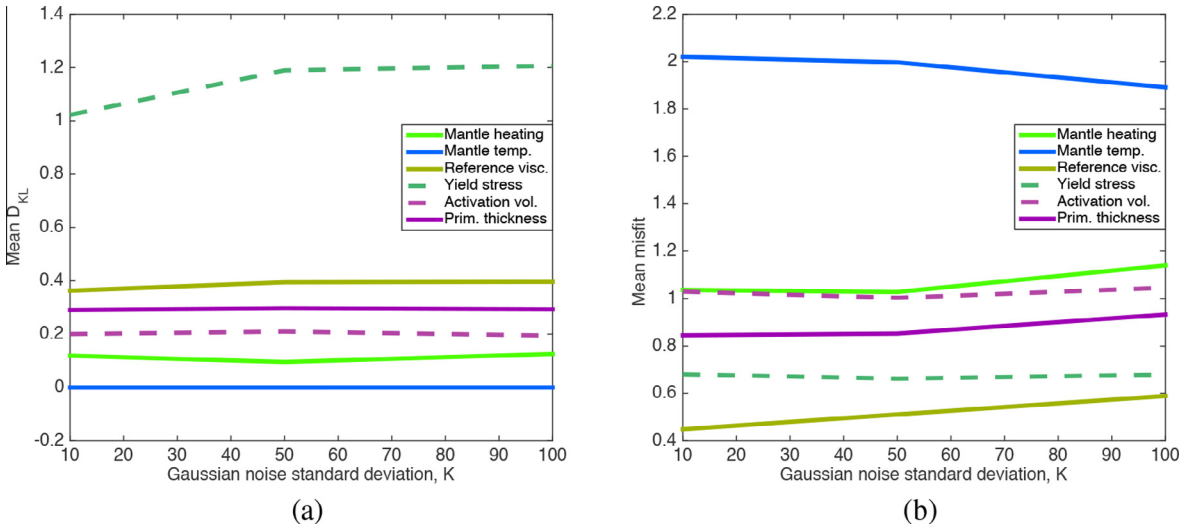

Fig. 6. (a) Mean D_{KL} with different levels of noise. (b) Mean difference between PDF maximum and true simulation parameter for each simulation with different noise levels after 3 Gyr. The unit is the variance of the prior distribution of each parameter.

The inversions for initial mantle temperature are not particularly successful for any time step after 0.4 Gyr, but they demonstrate what happens when the network learns nothing about a parameter. For Fig. 8(p), 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 is very small.

4. 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 work we present a new method which allows us to invert the thermal structure of mantle simulations at some time steps for convection parameters. We find that we 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, our method is novel in both its self consistency and its use of a static observation of convection. Whilst we cannot adequately recover other parameters in this study, they may be recoverable by using different observations or larger training sets.

Yield stress is the best constrained parameter when inverting the amplitude spectrum of the temperature field, and is particularly successful 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. 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 we define a mobile lid to have a mean surface velocity of > 1 cm/yr, as in Lourenço et al. (2016), approximately 50% of our simulations are in a mobile lid regime at each time step. However, we do not explicitly provide the networks with any information about plate velocity, therefore they can only identify that the

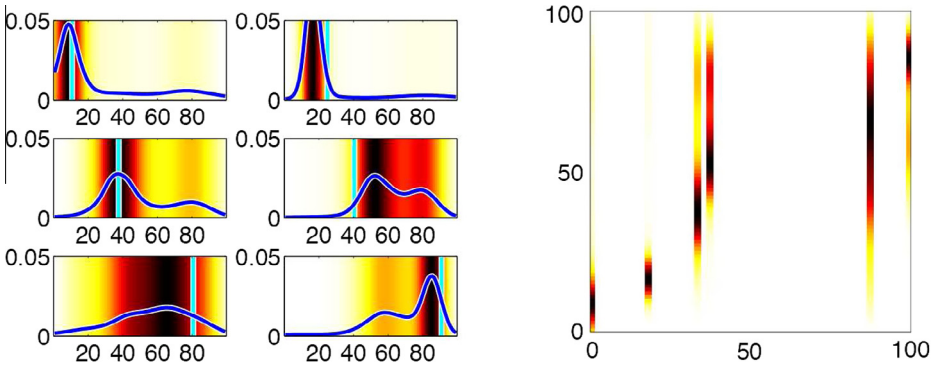

Fig. 7. Some randomly selected examples for posterior PDFs inferring yield stress after 3 Gyr, taken from the test set of simulations. The six PDFs to the left are the committee output, coloured with the same colour scale as in the right-hand panel and Fig. 8. The colour scale is black at the maximum 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

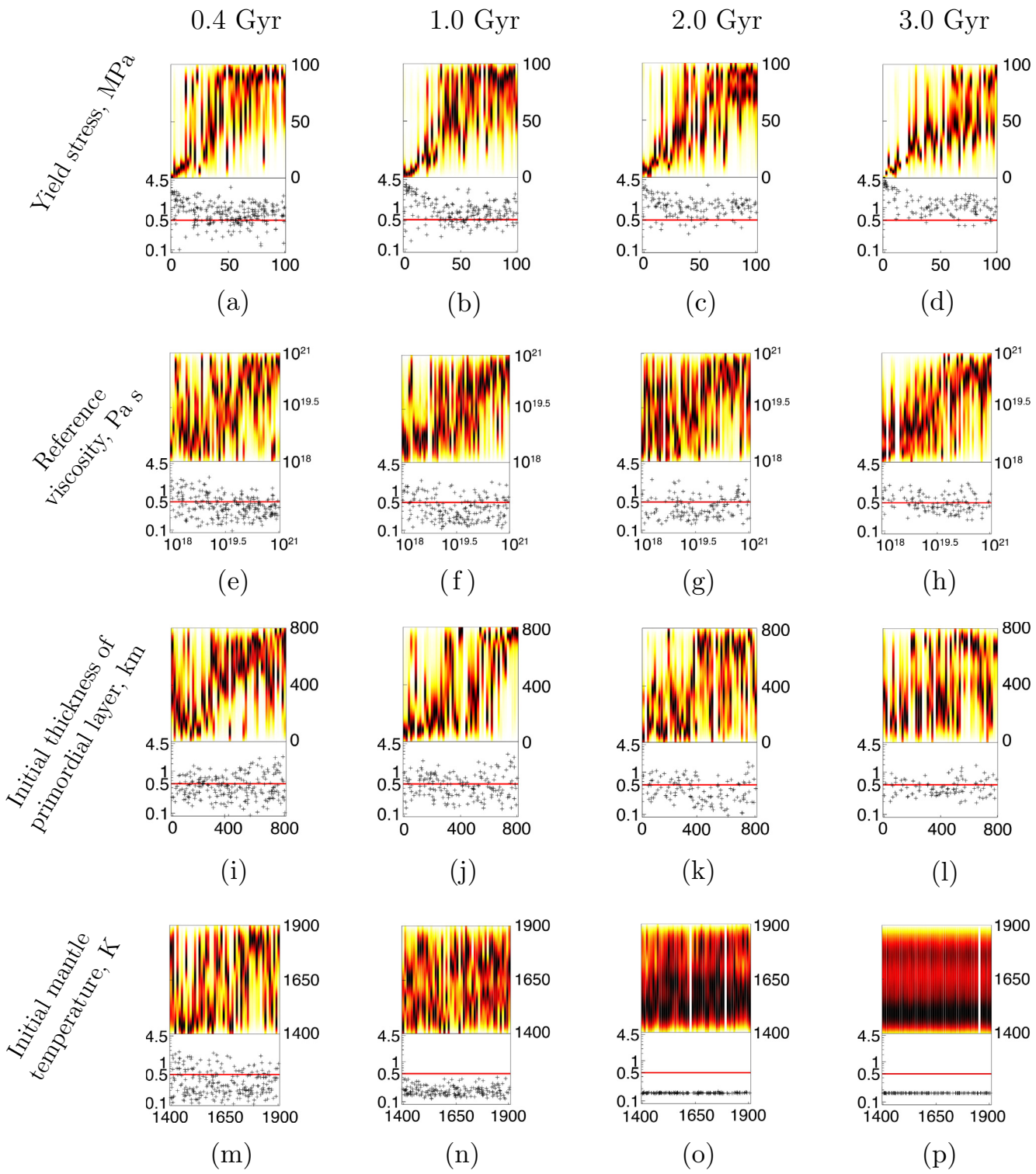

Fig. 8. PDFs for the test set of simulations at each age which provide an independent demonstration of network performance. In the coloured grids, each vertical column is one posterior PDF for the relationship between the temperature structure of a single simulation and the model parameter given on the 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 Fig. 7 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 Fig. 4. 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. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

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 but since we provide the networks with no explicit historic data, they must infer the history

from information contained within a snap shot of a single time step.

Fig. 9 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 our 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 Fig. 9, and narrower, more distinct upwellings extracting heat more efficiently leading to the observed cooler mid-mantle. The network is probably using these observations 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.

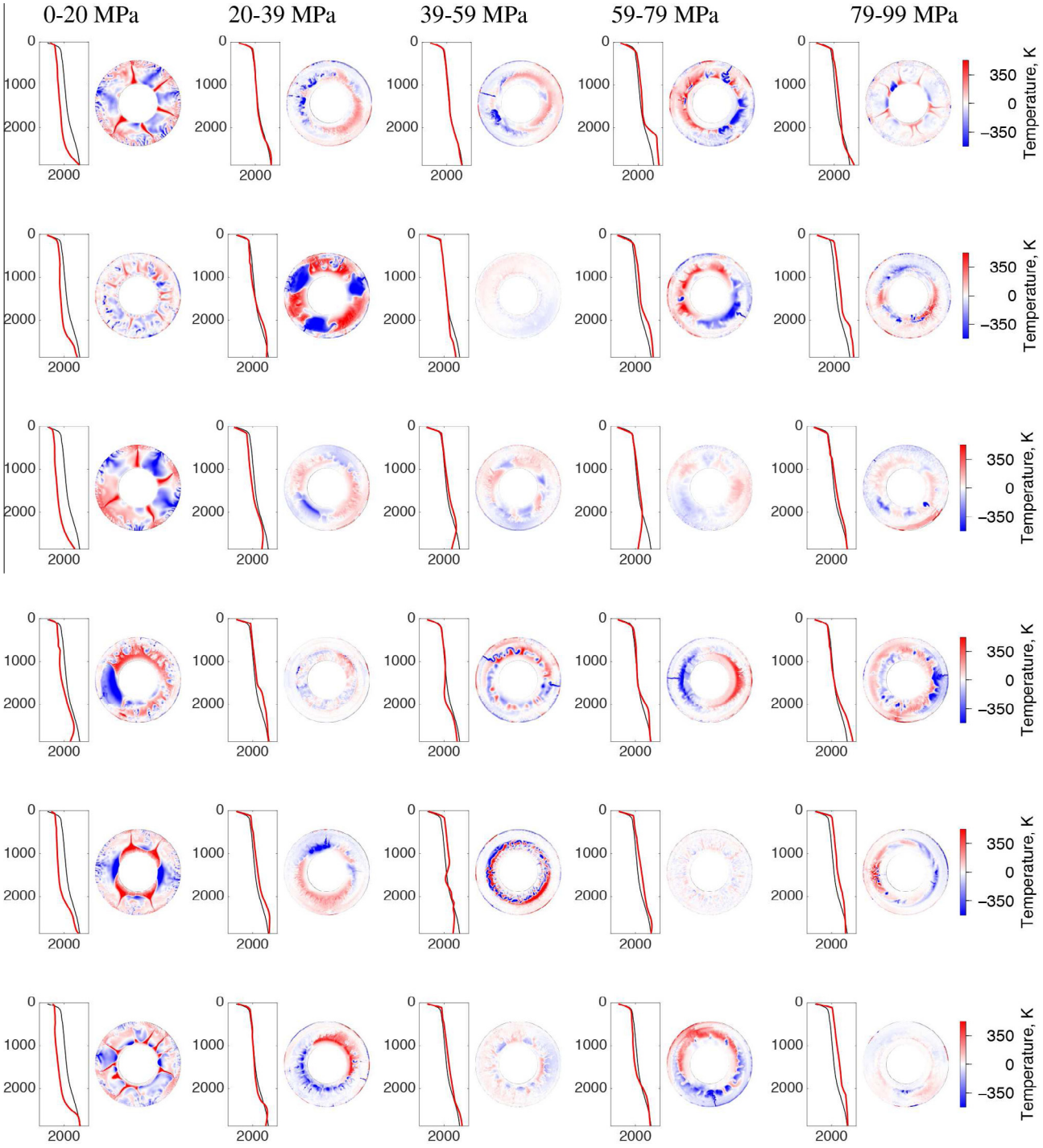

Fig. 9. Temperature structure for 30 simulations after 3 Gyr, grouped into columns according to yield stress used. On the left for each simulation is the temperature profile for the simulation in red. The mean profile for the whole group of simulations is plotted in black to aid comparison, and is identical in each case. The lateral variation from the 1D mean is plotted on the right to highlight the convection patterns. The same colour scale is used for all simulations. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The ability of the networks to find yield stress is probably also enhanced by the low temperature dependence of viscosity used in our simulations. This reduces the variations in viscous stress that would be caused by the 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 our simulations. The evolution of the atmosphere, and therefore the addition of water to the crust may reduce yield stress (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. However, if we were to use more complex simulations in our 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 convection. 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 the uncertainties from climate history and sea level models, and large uncertainties when converting seismic velocities to temperature or density. In our convection simulations, viscosity has a clearly identifiable and quantifiable effect on the temperature variations within the mantle. Using our approach, we 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, we only varied the viscosity prefactor and pressure dependence. 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 the 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 we neglect here. If we were to vary more viscosity parameters in our training simulations, such as temperature dependence, it is therefore possible given the presented results that we 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. Our networks only give an approximate estimate with large uncertainties (Fig. 8(l)) for the initial thickness of primordial material when the networks are trained on temperature 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 our method also works for three dimensional cases over 4.5 Gyr. We are currently testing this. It is also surprising that we 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 we invert temperature structure here, which is imperfectly known for Earth, we could use other more direct mantle observations such as seismic tomography to train our networks to identify signs of primordial material. 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.

We mentioned the existence of possible trade-offs between parameters. Whilst this is a problem in more classical approaches where authors only vary a few parameters at a time, our results implicitly contain all information on the trade-offs within our chosen range for the input parameters. Our 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 Fig. 8(i), 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 our networks may not be able to constrain the other model parameters varying in Table A.1. We are only using the encoded amplitude spectra for degrees 0–10 to train the networks. This removes much of the fine scale variation in the temperature field, and means that we 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 these small wavelength variations in the spectra. These unresolved parameters may also only have very small effects which we could observe if we 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 we only have small training sets.

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

We 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). Our lack of success is probably because the temperature is unlikely to be the best observation from which to identify mantle chemical properties. Our investigations were also very preliminary and we 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, we consider purely synthetic data sets and can therefore use the temperature structure of the mantle. If we were to apply this method to real data, we would have to rely on conversions from seismic velocity anomalies to temperature. For real data, it would be better to train our networks using the patterns of seismic heterogeneities. We can easily calculate P- and S-wave velocities for our 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 our inversions, but these can be accounted for during network training. However, this study is a proof of concept, and other simplifications remain, including that our synthetic data are two-dimensional approximations to a three-dimensional Earth. We 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 we can find the initial primordial layer thickness, which is a purely compositional parameter, from the temperature field. If we 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, we expect to be able to resolve the compositional parameters such as primordial thickness and basalt fraction much better.

In previous studies, (e.g. Davaille, 1999; McNamara and Zhong, 2004; Deschamps et al., 2011) the viscosity contrast of primordial material relative to the over-lying mantle was seen to determine the shape of any piles or ridges formed at the base of the mantle. We therefore expect to be able to find the viscosity contrast if we use a compositionally-dependent pattern to train our 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 our sampling approach: in the future, we 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.

We 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. Concluding remarks

We propose a new method to analyse a suite of convection simulations using pattern recognition techniques. We show that we can make inferences about simulations input parameters from simulation outputs over several billion years of convection. We make several choices, such as neural network architecture and the use of an auto-encoding neural network to reduce dimensionality which are guided by our previous experience rather than the necessity of the method. Other pattern recognition techniques could possibly perform equally well given an appropriate set of inputs.

Our method shows that some convection model parameters determine the convection pattern so significantly that the amplitude spectrum of the temperature alone can be used to find the values of those parameters, even after several billion years of convection time. Whilst the convection models we use are only simple two-dimensional cases with priors which are not realistic for the Earth, we expect this still to be the case for more complex three-dimensional models, although the relative importance of the model parameters may change. Moving into 3-D will present new challenges, both through the additional computational expense of running 3-D training sets, and through the significantly higher dimensionality of the observations. We hope that the signatures of some of these parameters will be similar in both two and three dimensions, allowing us to use 2-D approximations to invert the 3-D Earth.

In either 2-D or 3-D, we hope that our approach will prove to be a powerful way to constrain many unknown parameters necessary for better understanding the Earth. This approach may also prove to be particularly profitable for planetary science applications, where real observations are even more sparse than for Earth and vastly more expensive to obtain, particularly for exoplanets (Dorn et al., 2015). Rapid inversion of different data sets can therefore be used to guide future studies to fill the gaps in our knowledge, by providing rapid constraints on unknown characteristics, and the best ways in which to look for them.

Acknowledgments

We would like to thank the editor Mark Jellinek, reviewers Huw Davies, Mike Gurnis, Matt Weller and an anonymous colleague for their constructive comments. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement n. 320639, iGEO, and from the Netherlands Research Center for Integrated Solid Earth Science (ISES 2012-81). We would also like to thank Laura Cobden, Paul Käufl, Antoine Rozel and Ralph de Wit for constructive discussions and Theo van Zessen for managing the computers and making everything work.

Appendix A. StagYY

The forward models are run using the compressible mantle convection code StagYY (Tackley, 2008). A two-dimensional spherical annulus geometry is assumed, which does not require rescaling of the core-mantle boundary and surface radii in order to balance the ratio of bottom and internal heating, unlike 2D-cylindrical geometry (Hernlund and Tackley, 2008). The physical model is similar to that in Nakagawa et al. (2010), Nakagawa et al. (2009, 2012) with the addition of primordial material. Composition is expressed as a mixture of three end-members: basalt, harzburgite and primordial material; the relative proportions of which can vary within the mantle. Initially the mantle is composed of pyrolite (20–30% basalt and 70–80% harzburgite) except for a layer of primordial material above the CMB. The relative proportions of basalt and harzburgite vary between runs. Density, thermal expansivity, and thermal diffusivity are calculated for each end-member

composition (basalt, harzburgite and primordial) as a function of temperature and pressure using the Perple_X program (Connolly, 2009) according to the database of Stixrude and Lithgow-Bertelloni (2011); the properties for each cell depend on those of the end members, weighted according to the proportion of each. Density changes due to solid state phase changes are thus supplied by Perple_X and do not need to be imposed separately. The relative proportions of the bulk oxide components (Na_2O–CaO–FeO–MgO–Al_2O_3–SiO_2) vary between models. Harzburgite contains no Na_2O in our models. Compositional ranges are given in Table A.2. In 80% of the models, there is also an initial layer of primordial material at the base of the mantle. The composition of the primordial material is allowed to vary, and properties are also calculated using Perple_X. When present, the initial thickness of the primordial layer varies between models. The composition of the primordial material is chosen randomly according to one of three models (Table A.3). Twenty percent of models have no initial primordial layer. The primordial material can mechanically mix with pyrolitic material and is not confined to the base of the mantle.

Partial melting of the basaltic component occurs whenever the local temperature exceeds the solidus at which point enough melt is generated to bring the temperature back to the solidus, the function for which is based on results from Herzberg et al. (2000). Melt that is shallower than the depth of neutral buoyancy (set to 300 km) is instantly removed to form crust which is 100% basalt, leaving the cell more harzburgitic in composition.

Viscosity increases smoothly with depth to avoid imposing jumps at depths that are inconsistent with phase change locations calculated in Perple_X. A simple Arrhenius viscosity law is used, which is independent of composition and is continuous with depth:

$$\eta = \eta_0 \exp\left\{\frac{E_\eta + (1-z)V_\eta}{RT}\right\} \quad (A.1)$$

where T is temperature, z is depth, E_η (=162 kJ mol^{-1}) is activation energy and V_η activation volume, varying between 10^{-6} and 3×10^{-6} m^3 mol^{-1}, and η_0 is the reference viscosity, which is varied from $10^{18} - 10^{21}$ Pa s between runs. Viscosity values lower than 10^{18} Pa s are set to 10^{18} Pa s and values greater than 10^{25} Pa s are set to 10^{25} Pa s The ductile yield stress is given by:

$$\tau = \tau_0 + \tau_{dz} z \quad (A.2)$$

We vary the reference yield stress, τ_0. The depth-dependent yield stress, τ_{dz} is set to 0.005 Pa m^{-1}.

The Rayleigh number is not a separate input parameter but its value can be calculated from the various dimensional physical properties. Using the reference viscosity, which represents an asthenospheric value (i.e. at T = 1600 K and zero pressure) and surface values of other physical properties, we obtain values distributed uniformly on a logarithmic scale between 1.1×10^8 and 1.2×10^{11}. The volume-averaged Ra (using volume-averaged viscosity and other physical properties) is approximately 2–3 orders of magnitude lower, in line with common estimates.

The surface and core-mantle boundaries are free-slip and isothermal, with the surface temperature set to 300 K, and CMB temperature decreasing with time according to a parameterised core heat balance based on Buffett et al. (1992) and taking into account radioactive heating by ^{40}K in the core, which has a half-life of 1.25×10^9 years. The initial CMB temperature and initial concentration of ^{40}K in the core varies between runs. The mantle is also heated from within by radioactive elements, which are averaged into a single decay curve with a half-life of 2.43×10^9 years and an initial heating rate that varies between runs. Basaltic material is enriched in heat-producing elements by partitioning which occurs when basaltic partial melt is generated. The partition coefficient is varied between runs. All input parameter ranges used in this work are shown in Table (A.1).

Table A.1
Input parameter ranges to StagYY. All input parameters are drawn independently from uniform distributions between these ranges.

Parameter	
Initial mantle potential temperature at surface	1400–1900 K
Initial mantle heating	4.5–27 pW kg^{-1}
Basalt heating with HPE partition coefficient	$10^{-5} - 1$
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. (A.1))	$10^{18} - 10^{21}$ Pa s
Primordial viscosity contrast	Factor $10^{-2} - 10^2$
Viscosity activation volume (V_eta in Eq. (A.1))	$10^{-6} - 3 \times 10^{-6}$ m^3 mol^{-1}
Yield stress (τ_0 in Eq. (A.2))	1–100 MPa
Basalt fraction	0.2–0.3
Initial primordial layer thickness	0–800 km

Table A.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.

	Molar %
Basalt	
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 A.3
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.

Basalt + Chondritic 10% of models	(e.g. Tolstikhin and Hofmann, 2005)
	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 35% of models	(e.g. Lee et al., 2010)
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 35% of models	as in Table A.2
No primordial 20% of models	

The initial condition for the temperature field is adiabatic with the specified potential temperature, plus 30 km thick thermal boundary layers at the top and bottom and small random perturbations everywhere.

Velocity and pressure are solved for on an Eulerian staggered grid (finite volume discretisation) with 512 cells in azimuth by 64 cells in radius, while composition (bulk composition of solid and melt, and the concentration of heat-producing trace elements in them) and melt fraction are tracked by Lagrangian tracer particles (also called markers; a standard approach as detailed in Gerya (2010)), with approximately 15 tracer particles per cell, as is necessary according to benchmarks (Tackley and King, 2003). For more details of the solution method and equations see Tackley (2008) and Hernlund and Tackley (2008).

Appendix B. Neural networks

We use a class of networks called mixture density networks (MDN). For more information, see Bishop (1995) or MacKay (2003). A neural network consists of layers of nodes connected by weights, as illustrated in Fig. B.10. The first layer, d_i inputs the observation to the network and has one input node for each element of the encoded temperature amplitude spectrum. These are connected to a layer of hidden nodes, n_j^{hid} by a matrix of weights, w_{ij}^1, represented by lines in Fig. B.10. The hidden nodes are hyperbolic tangent functions activated by the product of the input nodes and the weight matrix, the output from each hidden node being $n_j^{\text{hid}} = \tanh(w_{ij}^1 d_i)$. The hidden layer then activates the output nodes n_k^{out}, according to a second layer of weights, w_{jk}^2 which give the mean, standard deviation and a weighting factor for each of three to five Gaussian kernels. The kernels combine to form the marginal posterior probability density function:

$$P(m_i|\mathbf{d}, m_{j\neq i}) = \sum_{g=1}^{G} \frac{\alpha_g(\mathbf{d})}{\sqrt{2\pi}\sigma_g(\mathbf{d})} \exp\left\{-\frac{||m_i - \mu_g(\mathbf{d})||^2}{2\sigma_g^2(\mathbf{d})}\right\} \quad \text{(B.1)}$$

where, for each input data vector $\mathbf{d}$, m_i is the target value of the model parameter. The conditional PDF is parameterised using a mixture of three to five Gaussian kernels, G, each with mean μ_g and standard deviation σ_g. The number of kernels is selected randomly. The contribution of each kernel to the conditional PDF is determined by a weighting factor, α_g. μ_g, σ_g, α_g are functions of the network weights $\mathbf{w}$. The bias weights are initialised so that before training the network output an approximation to the prior distribution of the model parameter of interest using a k-means clustering algorithm (McLachlan and Chang, 2004; de Wit et al., 2013; Käufl et al., 2014).

At the start of network training, the weights between each layer of nodes are initialised with values drawn randomly from a uniform distribution with range inversely proportional to the number of nodes in the lower of the two layers, that the weights are connecting, where the bottom layer is the observational input and the top layer are the Gaussian coefficients, as in Fig. B.10. At each iteration, we calculate the error function

$$E = -\ln(P(t|\mathbf{d})) \quad \text{(B.2)}$$

where t is the true value of the model parameter associated with the observation. The Rprop gradient-descent back-propagation algorithm of Igel and Hüsken (2000) is used to calculate the contribution of each weight in the network to the error function. The

Table B.4
Number of convection simulations in each age suite. 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 here are for inferences made given observations of the test set.

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

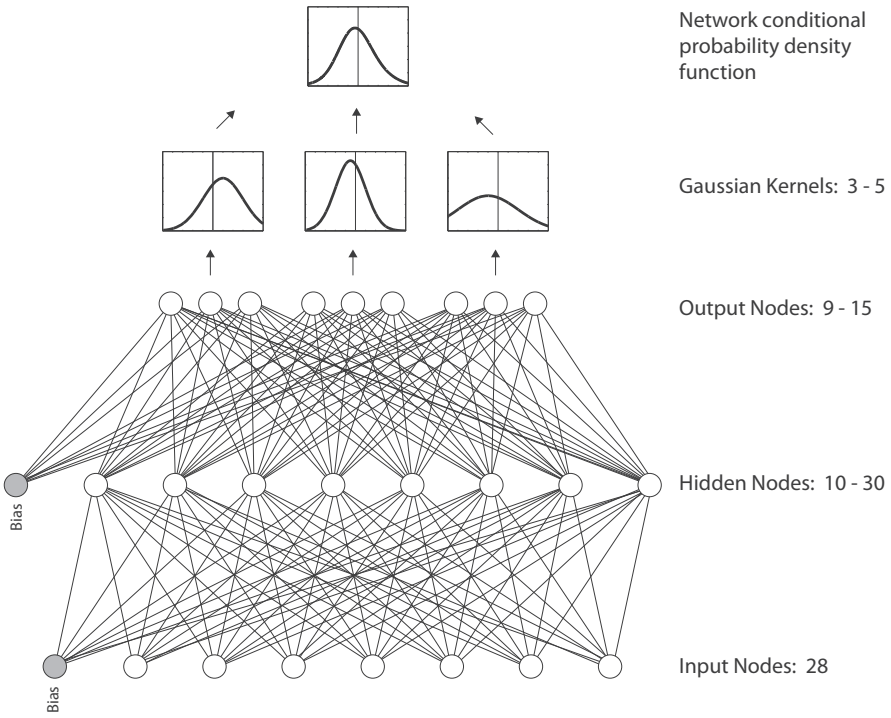

Fig. B.10. Mixture density network. The number of hidden units and Gaussian kernels are selected at random prior to training.

weights are then updated accordingly until the error function for a monitoring set of simulations reaches a minimum. After this point, the network overtrains, and the error for the monitoring set increases.

The weights in each network are initialised randomly before training and therefore every network performs slightly differently. A committee of networks generally performs better than a single network because the random initialisation can mean that networks are more effective at producing a reliable posterior PDF for a subset of the observations. Considering the PDFs from multiple networks therefore capitalises on the specialisms of each network. The more networks there are in the committee, the better the result is. However, the rate of improvement decreases rapidly, and we find 20 networks is a good balance between performance and computation time. To assemble the committee of networks we use a set of simulations that have not been used to update the weights during training. The committee assembly set is used to quantifying the difference in the value of the error function between the individual member networks. We then weight each member of the committee according to its performance using the method of Käufl et al. (2014).

References

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

Aubert, J., Labrosse, S., Poitou, C., 2009. Modelling the palaeo-evolution of the geodynamo. Geophys. J. Int. 179, 1414–1428.

Austermann, J., Kaye, T.B., Mitrovica, J.X., Huybers, P., 2014. A statistical analysis of the correlation between large igneous provinces and lower mantle seismic structure. Geophys. J. Int. 197, 1–9.

Baumann, T.S., Kaus, B.J.P., Popov, A.A., 2014. Constraining effective rheology through parallel joint geodynamic inversion. Tectonophysics 631, 197–211.

Bello, L., Coltice, N., Rolf, T., Tackley, P.J., 2014. On the predictability limit of convection models of the Earth's mantle. Geochem. Geophys. Geosyst. http://dx.doi.org/10.1002/2014GC005254.

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

Bocher, M., Coltice, N., Fournier, A., Tackley, P.J., 2016. A sequential data assimilation approach for the joint reconstruction of mantle convection and surface tectonics. Geophys. J. Int. 201, 200–214.

Bower, D.J., Gurnis, M., Seton, M., 2013. Lower mantle structure from paleogeographically constrained dynamic Earth models. Geochem. Geophys. Geosyst. 14 (1). http://dx.doi.org/10.1029/2012GC004267.

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

Bull, A.L., McNamara, A.K., Ritsema, J., 2009. Synthetic tomography of plume clusters and thermochemical piles. Earth Planet. Sci. Lett. 278, 152–162.

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

Cobden, L., Mosca, I., Trampert, J., Ritsema, J., 2012. On the likelihood of post-perovskite near the core-mantle boundary: a statistical interpretation of seismic observations. Phys. Earth Planet. Inter. 210–211, 21–35.

Connolly, J.A.D., 2009. The geodynamic equation of state: what and how. Geochem. Geophys. Geosyst. 10. http://dx.doi.org/10.1029/2009GC002540.

Conrad, C.P., Gurnis, M., 2003. Seismic tomography, surface uplift, and the breakup of Gondwanaland: integrating mantle convection backwards in time. Geochem. Geophys. Geosyst. 4 (3). http://dx.doi.org/10.1029/2001GC000299.

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

Deschamps, F., Tackley, P.J., 2008. Searching for models of thermo-chemical convection that explain probabilistic tomography. I – Principles and influence of rheologial parameters. Phys. Earth Planet. Inter. 171, 357–373.

Deschamps, F., Kaminski, E., Tackley, P.J., 2011. A deep mantle origin for the primitive signature of ocean island basalt. Nat. Geosci. 4, 879–882.

Deschamps, F., Cobden, L., Tackley, P.J., 2012. The primitive nature of large low shear-wave velocity provinces. Earth Planet. Sci. Lett. 349–350, 198–208.

Deuss, A., 2009. Global observations of mantle discontinuities using SS and PP precursors. Surv. Geophys. 30 (4), 301–326.

de Wit, R.W.L., Valentine, A.P., Trampert, J., 2013. Bayesian inference of Earth's radial seismic structure from body-wave traveltimes using neural networks. Geophys. J. Int. 195, 408–422.

Dorn, C.A.K., Heng, K., Alibert, Y., Connolly, J.A.D., Benz, W., Tackley, P.J., 2015. Can we constrain interior structure of rocky exoplanets from mass and radius measurements? Astron. Astrophys. 577 (A83).

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

Forte, A.M., Mitrovica, J.X., Espesset, A., 2002. Geodynamic and seismic constraints on the thermochemical structure and dynamics of convection in the deep mantle. Philos. Trans. R. Soc. 360, 2521–2543.

Gerya, T.V., 2010. Introduction to Numerical Geodynamic Modelling. Cambridge University Press, Cambridge.

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

Hernlund, J.W., Tackley, P.J., 2008. Modelling mantle convection in the spherical annulus. Phys. Earth Planet. Inter. 171, 48–54.

Heron, P.J., Lowman, L.P., 2014. The impact of Rayleigh number on assessing the significance of supercontinent insulation. J. Geophys. Res. 119, 711–733.

Herzberg, C., Raterron, P., Zhang, J., 2000. New experimental observations on the anhydrous solidus for peridotite KLB-1. Geochem. Geophys. Geosyst. 1, 2000GC000089.

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

Höink, T., Lenardic, A., Jellinek, A.M., 2013. Earth's thermal evolution with multiple convection modes: a Monte-Carlo approach. Phys. Earth Planet. Inter. 221, 22–26.

Igel, C., Hüsken, M., 2000. Improving the Rprop learning algorithm. In: Proceedings of the Second International Symposium on Neural Computation. ICSC Academic Press, pp. 115–121.

Ismail-Zadeh, A., Schubert, G., Tsepelev, I., Korotkii, A., 2004. Inverse problem of thermal convection: numerical approach and application to mantle plume restoration. Phys. Earth Planet. Inter. 145, 99–114.

Käufl, P., Valentine, A.P., O'Toole, T.B., Trampert, J., 2014. A framework for fast probabilistic centroid-moment-tensor determination – Inversion of regional static displacement measurements. Geophys. J. Int. 196, 1676–1693.

Käufl, P., Valentine, A.P., de Wit, R.W.L., Trampert, J., 2016. Solving probabilistic inverse problems rapidly with prior samples. Geophys. J. Int. 205, 1710–1728.

Kaus, B.J.P., Podladchikov, Y.Y., 2001. Forward and reverse modeling of the three-dimensional viscous Rayleigh–Taylor instability. Geophys. Res. Lett. 28 (6), 1095–1098.

Koroni, M., Trampert, J., 2016. The effect of topography of upper-mantle discontinuities on SS precursors. Geophys. J. Int. 204, 667–681.

Lee, C.-T.A., Luffi, P., Hoink, T., Li, J., Dasgupta, R., Hernlund, J.W., 2010. Upside-down differentiation and generation of a 'primordial' lower mantle. Nature 463.

Lenardic, A., Crowley, J.W., 2012. On the notion of well-defined tectonic regimes for terrestrial planets in the solar system and others. Astrophys. J. 755, 132.

Liu, L., Gurnis, M., 2008. Simultaneous inversion of mantle properties and initial conditions using an adjoint of mantle convection. J. Geophys. Res. 113 (B08405). http://dx.doi.org/10.1029/2008JB005594.

Lourenço, D.L., Rozel, A., Tackley, P.J., 2016. Melt-induced crustal production helps plate tectonics on Earth-like planets. Earth Planet. Sci. Lett. 439 (18–28).

Lowman, L.P., King, S.D., Trim, S.J., 2011. The influence of plate boundary motion on planform in viscously stratified mantle convection models. J. Geophys. Res. 116. http://dx.doi.org/10.1029/2011JB008362.

MacKay, D.A., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

McLachlan, G.J., Chang, S.U., 2004. Mixture modelling for cluster analysis. Stat. Methods Med. Res. 13 (5), 347–361.

McNamara, A.K., Zhong, S., 2004. Thermochemical structures within a spherical mantle: superplumes or piles? J. Geophys. Res. 109 (B07402). http://dx.doi.org/10.1029/2003JB002847.

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

Meier, U., Curtis, A., Trampert, J., 2007. Global crustal thickness from neural network inversion of surface wave data. Geophys. J. Int. 169, 706–722.

Mitrovica, J.X., Forte, A.M., 2004. A new inference of mantle viscosity based upon joint inversion of convection and glacial isostatic adjustment data. Earth Planet. Sci. Lett. 225, 177–189.

Moresi, L., Solomatov, V., 1998. Mantle convection with a brittle lithosphere: thoughts on the global tectonic styles of the Earth and Venus. Geophys. J. Int. 133, 669–682.

Mosegaard, K., Tarantola, A., 1995. Monte Carlo sampling of solutions to inverse problems. J. Geophys. Res. 100 (B7), 12431–12447.

Nakagawa, T., 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 Planet. Sci. Lett. 271, 348–358.

Nakagawa, T., 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. Geochem. Geophys. Geosyst. 11 (6). http://dx.doi.org/10.1029/2010GC003031.

Nakagawa, T., Tackley, P.J., Deschamps, F., 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. Geochem. Geophys. Geosyst. 10 (3). http://dx.doi.org/10.1029/2008GC002280.

Nakagawa, T., Tackley, P.J., Deschamps, F., 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 Planet. Sci. Lett. 296 (403–412).

Nakagawa, T., Tackley, P.J., Deschamps, F., Connolly, J.A.D., 2012. Radial 1-D seismic structures in the deep mantle in mantle convection simulations with self-consistently calculated mineralogy. Geochem. Geophys. Geosyst. 13 (11). http://dx.doi.org/10.1029/2012GC004325.

Ratnaswamy, V., Stadler, G., Gurnis, M., 2015. Adjoint-based estimation of plate coupling in a non-linear mantle flow model: theory and examples. Geophys. J. Int. 202, 768–786.

Rolf, T., Coltice, N., Tackley, P.J., 2012. Linking continental drift, plate tectonics and the thermal state of the Earth's mantle. Earth Planet. Sci. Lett. 351–352, 134–146.

Rolf, T., Coltice, N., Tackley, P.J., 2014. Statistical cyclicity of the supercontinent cycle. Geophys. Res. Lett. 41. http://dx.doi.org/10.1002/2014GL059595.

Rudolph, M.L., Lekić, V., 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. Geophys. J. Int. 138, 479–494.

Sambridge, M., Mosegaard, K., 2002. Monte Carlo methods in geophysical inverse problems. Rev. Geophys. 40 (3). http://dx.doi.org/10.1029/2000RG000089.

Schuberth, B.S.A., Bunge, H.-P., Ritsema, J., 2009. Tomographic filtering of high-resolution mantle circulation models: can seismic heterogeneity be explained by temperature alone? Geochem. Geophys. Geosyst. (Q05W03).

Shephard, G.E., Flament, N., Williams, S., Gurnis, M., Müller, R.D., 2014. Circum-Arctic mantle structure and long-wavelength topography since the Jurassic. J. Geophys. Res. 119. http://dx.doi.org/10.1002/2014JB011078.

Spasojevic, S., Gurnis, M., 2012. Sea level and vertical motion of continents from dynamic Earth models since the Late Cretaceous. AAPG Bull. 96 (11), 2037–2064.

Stamenković, V., Noack, L., Breuer, D., Spohn, T., 2012. The influence of pressure-dependent viscosity on the thermal evolution of super-Earths. Astrophys. J. 748 (41). http://dx.doi.org/10.1088/0004-637X/748/1/41.

Stampfli, G.M., Borel, G.D., 2002. A plate tectonic model for the paleozoic and mesozoic constrained by dynamic plate boundaries and restored synthetic oceanic isochrons. Earth Planet. Sci. Lett. 196, 17–33.

Steinberger, B., Torsvik, T.H., 2012. A geodynamic model of plumes from the margins of large low shear velocity provinces. Geochem. Geophys. Geosyst. 13 (1). http://dx.doi.org/10.1029/2011GC003808.

Stein, C., Lowman, L.P., Hansen, U., 2013. The influence of mantle internal heating on lithospheric mobility: Implications for super-Earths. Earth Planet. Sci. Lett. 361, 448–459.

Stixrude, L., Lithgow-Bertelloni, C., 2011. Thermodynamics of mantle minerals – II. Phase equilibria. Geophys. J. Int. 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. Phys. Earth Planet. Inter. 171, 7–19.

Tackley, P.J., King, S.D., 2003. Testing the tracer ratio method for modelling active compositional fields in mantle convection simulations. Geochem. Geophys. Geosyst. 4 (4). http://dx.doi.org/10.1029/2001GC000214.

Tarantola, A., Valette, B., 1982. Generalized nonlinear inverse problems solved using the least squares criterion. Rev. Geophys. 20, 219–232.

Tolstikhin, I., Hofmann, A.W., 2005. Early crust on top of Earth's core. Phys. Earth Planet. Inter. 148, 109–130.

Torsvik, T.H., Steinberger, B., Gurnis, M., Gaina, C., 2010. Plate tectonics and net lithosphere rotation over the past 150 My. Earth Planet. Sci. Lett. 291, 106–112.

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

Trim, S.J., Heron, P.J., Stein, C., Lowman, L.P., 2014. The feedback between surface mobility and mantle compositional heterogeneity: implications for the Earth and other terrestrial planets. Earth Planet. Sci. Lett. 405, 1–14.

Valencia, D., O'Connell, R.J., Sasselov, D.D., 2007. Inevitability of plate tectonics on super-Earths. Astrophys. J. 670, L45–L48.

Valentine, A.P., Trampert, J., 2012. Data space reduction, quality assessment and searching of seismograms: autoencoder networks for waveform data. Geophys. J. Int. 189, 1183–1202.

Valentine, A.P., Kalnins, L.M., Trampert, J., 2013. Discovery and analysis of topographic features using learning algorithms: a seamount case study. Geophys. Res. Lett. 40, 3048–3054.

van Heck, H., Tackley, P.J., 2011. Plate tectonics on super-Earths: equally or more likely than on Earth. Earth Planet. Sci. Lett. 310 (252–261).

Weller, M.B., Lenardic, A., O'Neill, C., 2015. The effects of internal heating and large scale climate variations on tectonic bi-stability in terrestrial planets. Earth Planet. Sci. Lett. 420, 85–94.

Whitehouse, P.L., Bentley, M.J., Milne, G.A., King, H., 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. Quatern. Sci. Rev. 32, 1–24.

Worthen, J., Stadler, G., Petra, N., Gurnis, M., Ghattas, O., 2014. Towards adjoint-based inversion for rheological parameters in nonlinear viscous mantle flow. Phys. Earth Planet. Inter. 234, 23–34.

Yoshida, M., 2008. Mantle convection with longest-wavelength thermal heterogeneity in a 3-D spherical model: degree one or two? Geophys. Res. Lett. 35. http://dx.doi.org/10.1029/2008GL036059.

Zhong, S., Zhang, N., Li, Z.-X., Roberts, J.H., 2007. Supercontinent cycles, true polar wander, and very long-wavelength mantle convection. Earth Planet. Sci. Lett. 261, 551–564.