Earth and Space Science

RESEARCH ARTICLE
10.1029/2020EA001484

Special Section:
Machine learning for Solid Earth observation, modeling and understanding

Key Points:

- Mixture Density Networks provide a probabilistic framework for inverting observables to infer parameters of Mars' interior evolution
- Reference viscosity, crustal enrichment in heat-producing elements and initial mantle temperature can be well constrained
- Activation energy of diffusion creep can be weakly constrained; constraining activation volume requires new observational signatures

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:
S. Agarwal,
agsiddhant@gmail.com

Citation:
Agarwal, S., Tosi, N., Kessel, P., Padovan, S., Breuer, D., & Montavon, G. (2021). Toward constraining Mars' thermal evolution using machine learning. *Earth and Space Science*, *8*, e2020EA001484. https://doi.org/10.1029/2020EA001484

Received 24 SEP 2020
Accepted 19 FEB 2021

Author Contributions:
Conceptualization: N. Tosi, S. Padovan, D. Breuer, G. Montavon
Data curation: S. Agarwal
Funding acquisition: N. Tosi, S. Padovan, D. Breuer, G. Montavon

Toward Constraining Mars' Thermal Evolution Using Machine Learning

S. Agarwal[1,2], N. Tosi[1], P. Kessel[2], S. Padovan[1], D. Breuer[1], and G. Montavon[2]

[1]Planetary Physics, Institute of Planetary Research, German Aerospace Center (DLR), Berlin, Germany, [2]Electrical Engineering and Computer Science, Berlin Institute of Technology, Berlin, Germany

Abstract The thermal and convective evolution of terrestrial planets like Mars is governed by a number of initial conditions and parameters, which are poorly constrained. We use Mixture Density Networks (MDN) to invert various sets of synthetic present-day observables and infer five parameters: reference viscosity, activation energy and activation volume of the diffusion creep rheology, an enrichment factor for radiogenic elements in the crust, and initial mantle temperature. The data set comes from 6,130 two-dimensional simulations of the thermal evolution of Mars' interior. We quantify the possibility of constraining a parameter using the log-likelihood value from the MDN. Reference viscosity can be constrained to within 32% of its entire range ($10^{19} - 10^{22}$ Pa s), when all the observables are available: core-mantle-boundary heat flux, surface heat flux, radial contraction, melt produced, and duration of volcanism. Furthermore, crustal enrichment factor (1–50) can be constrained, at best, to within 15%, and the activation energy ($10^5 - 5 \times 10^5$ J mol^{-1}) to within 80%. Initial mantle temperature can be constrained to within 39% of its range (1,600–1,800 K). Using the full present-day temperature profile or parts of it as an observable tightens the constraints further. The activation volume ($4 \times 10^{-6} - 10 \times 10^{-6}$ m^3 mol^{-1}) cannot be constrained. We also tested different levels of uncertainty in the observables and found that constraints on different parameters loosen differently, with initial temperature being the most sensitive. Finally, we present how a joint probability model for all parameters can be obtained from the MDN.

Plain Language Summary The mantle of rocky planets like Mars behaves like a highly viscous fluid over geological time scales. Key parameters and initial conditions for the non-linear partial differential equations governing mantle flow are poorly known. Machine Learning (ML) can help us avoid running several thousand computationally expensive fluid dynamic simulations each time to determine if an observable can constrain a parameter. Using an ML approach, we invert a set of synthetic observables such as present-day surface heat flux, duration of volcanism and radial contraction to constrain important parameters controlling the long-term evolution of the planet's interior, such as the reference mantle viscosity or the partitioning of radiogenic heat sources between mantle and crust. We demonstrate that by training a probabilistic ML algorithm on the data and applying it, we can quantify the constraints on parameters. This provides a high-dimensional framework for analyzing inverse problems in geodynamics.

1. Introduction

Detailed modeling of subsolidus mantle convection is important for studying how terrestrial planets like Mars evolve over billions of years. Subsolidus mantle convection is governed by the conservation equations of mass, momentum, and energy appropriate for a highly viscous fluid (the mantle viscosity is of the order of $\sim 10^{20}$ Pa s) with negligible inertia (e.g., Breuer & Moore, 2015, and references therein). These non-linear partial differential equations are typically solved numerically using fluid dynamics codes (e.g., Hüttig et al., 2013; Kronbichler et al., 2012; Tackley, 2008; Zhong et al., 2008).

Model parameters and initial conditions, which are inputs to these simulations, are often poorly known and unconstrained. However, certain outputs of the simulations can be processed to arrive at “observables”. Indeed, several quantities, which can be inferred from remote or in-situ observations, are tightly related to the thermal evolution of the interior. Measurements of topography, gravity, magnetic and seismic fields, surface spectra, and surface images, as well as meteoritic data can all be employed to infer a series of fundamental constraints for the evolution of the interior (see Tosi & Padovan, 2020, for a review). These include crustal and elastic lithosphere thickness, duration and timing of volcanism, surface heat flux, amount of

Investigation: S. Agarwal
Methodology: S. Agarwal, N. Tosi, P. Kessel, S. Padovan
Software: S. Agarwal, S. Padovan
Supervision: N. Tosi, P. Kessel, S. Padovan, D. Breuer, G. Montavon
Validation: S. Agarwal
Visualization: S. Agarwal
Writing – original draft: S. Agarwal
Writing – review & editing: S. Agarwal, N. Tosi, P. Kessel, S. Padovan, D. Breuer, G. Montavon

accumulated radial contraction, surface concentration of radioactive elements, or evolution of the mantle potential temperature. To some degree, all of these quantities are currently available for Mars (see Section 5 for more details). However, in this work, due to limitations in the physics accounted for by our forward models, we will only consider synthetic observables, that is, quantities that are obtained from our forward models.

The inverse problem of constraining mantle convection parameters from observations is ill-conditioned due to missing or poor-quality observables as well as due to non-linearity. Hence, such studies in geodynamics are typically formulated in a probabilistic framework. In principle, standard Markov Chain Monte Carlo (MCMC) methods could be used to perform the inversion (see Sambridge & Mosegaard, 2002 for an overview). However, these require sampling several forward models, a task that can become impractical or even unfeasible when the forward evaluations are computationally expensive.

Attempts to address this limitation have ranged from modifying the method (e.g., Sambridge, 1999a, 1999b) to completely bypassing it. An example of the latter was Käufl et al. (2016) proposing Mixture Density Networks (MDNs, Bishop, 1994) as an alternative. They noted that while MDNs provide a more conservative estimate of the posterior probability than MCMC methods, they are computationally more efficient since the inversion is carried out solely using a set of forward models calculated in advance. Meier et al. (2007) used MDNs to invert surface waves data for obtaining constraints on crustal thickness. Subsequently, MDNs have been employed in several geodynamic studies. For example, de Wit et al. (2013) inverted body-wave travel-times to constrain the Earth's radial seismic structure using MDNs. Atkins et al. (2016) showed that mantle convection can also be investigated as a pattern recognition problem. They inverted reduced representations of 2D temperature fields of simulations of the Earth's mantle to constrain convection parameters such as reference mantle viscosity, lithospheric yield stress and initial temperature. They argued that under the assumption of smooth probability distributions between training samples, MDNs need significantly fewer simulations than MCMC methods for constraining mantle convection parameters. Baumeister et al. (2020) used MDNs to predict the distribution of the possible interior structures of extrasolar planets for given mass and radius.

Some other examples of Machine learning (ML) studies include Baumann (2016), who approached the problem of constraining the dynamics and rheology of the lithosphere in collision zones using an unsupervised classification algorithm called self-organizing map (Vesanto & Alhoniemi, 2000) and Shahnas et al. (2018), who used support-vector machines to estimate the magnitude of density anomalies from mantle temperature fields (see Morra et al., 2020, for a recent review on the application of data science techniques in geodynamics).

The ability to learn complex non-linear mappings using ML methods can also be used to make MCMC inversions computationally feasible. Recently, Magali et al. (2020) retrieved temperature and viscosity fields generated with a numerical mantle convection model by inverting surface wave dispersion curves. The inversion was carried out using MCMC random walks, which sample from a forward surrogate model based on Neural Networks (NN).

NN-based forward surrogates have also been used in other mantle convection studies. Shahnas and Pysklywec (2020) showed that NNs can be used for predicting the heat flux of steady-state simulations given parameters such as Rayleigh number and variable physical processes such as mode of convection. Agarwal et al. (2020) used an NN as a forward surrogate model to study Mars' thermal evolution. They showed that the NN is able to predict the entire 1D temperature profile of the mantle at any time in the evolution, from 0 to 4.5 Gyr, with an average accuracy of 99.7%, given five key parameters: reference viscosity (η_{ref}), activation energy (E), activation volume (V) of a temperature- and pressure-dependent diffusion creep rheology, crustal enrichment factor of radiogenic elements (Λ) with respect to a given bulk silicate abundance, and initial mantle temperature (T_{ini}).

We now revisit the data set and the same five parameters used in Agarwal et al. (2020) and formulate this study as an inverse problem. Using MDNs, we study the extent to which the thermal evolution of a Mars-like planet can be constrained. In particular, we systematically isolate the degree to which a parameter can be constrained using different synthetic observables (i.e., observables generated by the forward model). We also quantify the effect of uncertainty in these observables on the ability to constrain different parameters.

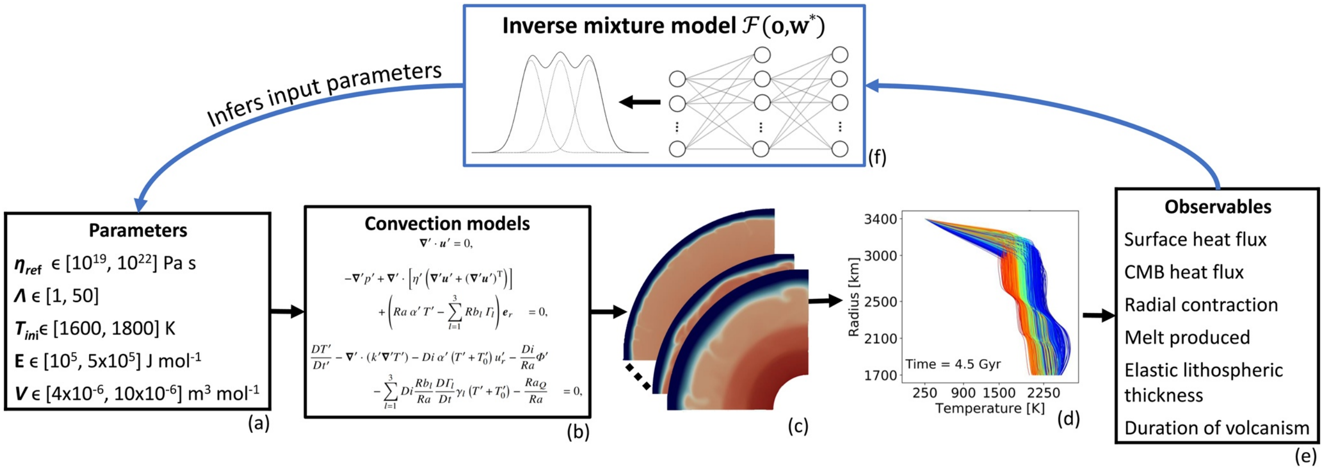

Inverse mixture model $\mathcal{F}(\mathrm{o},\mathrm{w}^*)$
(f)
Infers input parameters
Parameters
η_{ref} ∈ [10^{19}, 10^{22}] Pa s
Λ ∈ [1, 50]
T_{ini} ∈ [1600, 1800] K
E ∈ [10^5, 5×10^5] J mol^{-1}
V ∈ [4×10^{-6}, 10×10^{-6}] m^3 mol^{-1}
(a)
Convection models
$\nabla' \cdot \boldsymbol{u}' = 0,$
$-\nabla' p' + \nabla' \cdot \left[\eta' \left(\nabla' \boldsymbol{u}' + (\nabla' \boldsymbol{u}')^{\mathrm{T}}\right)\right] + \left(Ra\, \alpha' T' - \sum_{l=1}^{3} Rb_l\, \Gamma_l\right) \boldsymbol{e}_r = 0,$
$\frac{DT'}{Dt'} - \nabla' \cdot (k' \nabla' T') - Di\, \alpha' (T' + T'_0)\, u'_r - \frac{Di}{Ra}\Phi' - \sum_{l=1}^{3} Di \frac{Rb_l}{Ra} \frac{D\Gamma_l}{Dt} \gamma_l (T' + T'_0) - \frac{Ra_Q}{Ra} = 0,$
(b)
(c)
3400
3000
2500
2100
1700
Radius [km]
Time = 4.5 Gyr
250 900 1500 2250
Temperature [K]
(d)
Observables
Surface heat flux
CMB heat flux
Radial contraction
Melt produced
Elastic lithospheric thickness
Duration of volcanism
(e)

Figure 1. Illustration of our strategy for using MDNs to study constraints on Mars' thermal evolution. (a) We randomly generate several values of the following parameters: reference viscosity (η_{ref}), activation energy (E) and activation volume (V) of diffusion creep, crustal enrichment factor (Λ) with respect to a given bulk composition of radiogenic elements, and initial temperature (T_{ini}). (b) These are used as inputs to forward simulations run with the mantle convection code GAIA (Hüttig et al., 2013). (c) We then process the 2D temperature fields output by GAIA and (d) laterally average them to obtain 1D temperature profiles. (e) These temperature profiles, which are a reduced representation of the full 2D field, can be processed even further to arrive at observables such as heat flux at the surface, core-mantle boundary temperature, radial contraction, melt produced, elastic lithospheric thickness, and duration of volcanism. (f) We then train an MDN to build an inverse mixture model, which learns to constrain the parameters directly from the observables. 1D, one dimensional; 2D, two dimensional; MDNs, Mixture Density Networks.

The approach is shown in Figure 1. We start the study by randomly generating several values of five parameters and feeding these to two dimensional (2D) forward simulations performed with our finite-volume mantle convection code GAIA (Hüttig et al., 2013) on a quarter-cylindrical grid. We then process the 2D temperature fields output by GAIA and laterally average them to obtain one dimensional (1D) temperature profiles. These temperature profiles, which are reduced representations of the full 2D fields, can be processed even further to arrive at observables such as heat flux at the surface, core-mantle boundary (CMB) temperature, and elastic lithospheric thickness. We then employ MDNs to build an inverse mixture model, which learns to constrain the parameters directly from the processed, low-dimensional observables.

An obvious advantage of using the inverse approach is that we can remove any prediction inaccuracies in the observables (such as surface heat flux derived from the predicted temperature profile) from the ML surrogate and instead, directly use the data from the numerical simulations. The biggest advantage, though, is that we can directly test different numbers and combinations of observables as well as different sections of the temperature profiles to search for observational constraints on parameters governing mantle convection. With respect to Agarwal et al. (2020), to obtain a more balanced data set across the entire range of the various parameters, we re-ran the simulations with a higher grid resolution (300 × 392 nodes on a quarter cylindrical grid instead of 200 × 263). The higher resolution allowed us to overcome some numerical instabilities arising in cases where the parameters led to extremely vigorous convection.

The present work builds upon the study of Atkins et al. (2016), extending it in several ways. We study if and how well the thermal evolution of a stagnant-lid planet like Mars (rather than a mobile-lid planet like the Earth) can be constrained using synthetic observables at present-day, after 4.5 Gyr of evolution (rather than after up to 3 Gyr). Also, we deliberately consider observables that could potentially be measured for a Mars-like planet, instead of a reduced representation of the entire temperature field. Depending on the size of such reduced representation, it can still contain rich information about the convection structures, while not being realistically observable. This can potentially lead to overly optimistic constraints on parameters, especially for planets with sparser observational constraints than Earth. Furthermore, we demonstrate how to quantify constraints on parameters, instead of visually inspecting the over 1,000 cases tested in this paper. Finally, we show how an MDN can be modified to obtain the joint probability distribution of all parameters instead of the marginal probability.

Table 1
Fixed Parameter Values Shared by all Simulations as set by Plesa et al. (2015)

Parameter	Value	Unit
Initial temperature difference between core and surface	2,000	K
Surface temperature	250	K
Core density	7,000	kg m^{-3}
Mantle density	3,500	kg m^{-3}
Core specific heat capacity	850	J kg^{-1} K^{-1}
Mantle specific heat capacity	1,200	J kg^{-1} K^{-1}
Reference thermal conductivity	4	W m^{-1} K^{-1}
Reference thermal expansivity	2.50×10^{-5}	K^{-1}
Outer radius of the core	1,700	km
Planetary radius	3,400	km
Thickness of the crust	64	km
Reference pressure for viscosity	3	GPa
Reference temperature for viscosity	1,600	K

The outline of the paper is as follows. In Section 2, we explain the setup of the simulations used to generate the data. In the same section, we also visualize the data set and the processed observables. Then, in Section 3, we briefly present the basics of MDNs. In Section 4, we explain how we train our MDNs and define the degree to which parameters can be constrained. In the following subsections, we use this metric to explore which observables need to be measured and the required precision to infer different parameters. We also demonstrate that we have enough data for this study. In addition, we show, as a proof of concept, the potential of this procedure to search for possible observational signatures of parameters by looking at different parts of the temperature profiles. We also show how MDNs can be trained to obtain the joint probability distribution on all five parameters. Finally, in Section 5, we discuss in detail the next steps for inverting "real" data for Mars and then conclude by outlining the main findings and some interesting follow-ups to this work.

2. Data Set

2.1. Mantle Convection Simulations for a Mars-Like Planet

The data required for training MDNs comes from a subset of 10,080 evolution simulations for a Mars-like planet. A detailed description of the setup of the simulations is presented in Agarwal et al. (2020), where we applied ML to the same data set, albeit for forward 1D surrogate modeling. Here, we list the main features of the mantle convection model as well as some important quantities in Table 1.

We treat the mantle as a viscous fluid with infinite Prandtl number and Newtonian rheology. Given the low dissipation number for Mars (~ 0.13), we assume the extended Boussinesq approximation (e.g., King et al., 2010). The thermal expansivity and thermal conductivity are pressure- and temperature-dependent (Tosi et al., 2013a). This is also the case for the viscosity, which is calculated using the Arrhenius law for diffusion creep (Hirth & Kohlstedt, 2003) as follows

$$\eta(T,P) = \eta_{\text{ref}} \exp\left(\frac{E+PV}{T} - \frac{E+P_{\text{ref}}V}{T_{\text{ref}}}\right), \tag{1}$$

where T and P are temperature and pressure, η_{ref} is the reference viscosity attained at reference temperature and pressure $T_{\text{ref}} = 1600$ K and $P_{\text{ref}} = 3$ GPa, respectively, E is the activation energy, and V the activation volume. All simulations are initialized with a profile consisting of a constant mantle temperature T_{ini} supplemented by an upper and lower 300-km-thick thermal boundary layer and a small random perturbation to initiate convection. The mantle evolves in time due to the cooling core and decay of radiogenic elements. The bulk abundances of radiogenic elements, based on Wänke and Dreibus (1994), are modified via a crustal enrichment factor Λ, under the assumption that a crust of a fixed thickness ($d_{\mathrm{cr}} = 64$ km) formed very early in the evolution (Nimmo & Tanaka, 2005). Partial melting further depletes heat-producing elements in the mantle and affects the energy balance as detailed in Padovan et al. (2017). Finally, we complete the model by adding two phase-transitions in the olivine system using the standard phase-function approach of Christensen and Yuen (1985).

The simulations were run using the finite-volume code GAIA (Hüttig et al., 2013) on a 2D quarter-cylindrical domain with a grid resolution of 300 radial layers and 392 cells per layer. We considered all boundaries to be impermeable and free-slip. We assumed lateral walls to be insulating (i.e., with a zero heat flux across). We kept the surface temperature fixed at 250 K throughout the evolution and let the core-mantle boundary temperature evolve according to a standard core-cooling condition (e.g., Stevenson et al., 1983). Running several single-core simulations in parallel, we ended up generating over 10 TB of data after 1.7 million CPU hours.

2.2. Calculation of Observables

The output of the numerical simulations (velocity, temperature, viscosity, etc.) at any given time t can be post-processed to arrive at certain quantities of interest, some of which might be directly or indirectly observable. We consider here the laterally averaged temperature field, resulting in a 1D profile $T(r, t)$. Several quantities can be derived from $T(r, t)$ such as surface heat flux Q_{s} and CMB heat flux Q_{c}:

$$Q_{\mathrm{c}}(t) = k_{\mathrm{c}} \frac{T(R_{\mathrm{c}},t) - T(R_{\mathrm{c}} + \Delta R,t)}{\Delta R}, \tag{2}$$

$$Q_{\mathrm{s}}(t) = k_{\mathrm{s}} \frac{T(R_{\mathrm{p}} - \Delta R,t) - T(R_{\mathrm{p}},t)}{\Delta R}, \tag{3}$$

where, k_{c} and k_{s} are the thermal conductivity at CMB and at surface, respectively, R_{c} and R_{p} are radii of the core and planet, respectively, and ΔR is the radial distance between two neighboring shells (uniform throughout the computational domain). One can also use $T(r, t)$ to calculate the radial contraction of a planet (e.g., Grott et al., 2011; Tosi et al., 2013b). We consider the thermally induced radial contraction (and expansion) of the core and mantle as a post-processing step:

$$R_{\mathrm{th}}(t) = \alpha_{\mathrm{c}}(T(R_{\mathrm{c}},t) - T(R_{\mathrm{c}},0))\frac{R_{\mathrm{c}}^3}{3R_{\mathrm{p}}^2} + \frac{1}{R_{\mathrm{p}}^2}\int_{R_{\mathrm{c}}}^{R_{\mathrm{p}}} \alpha_{\mathrm{m}}(r,t)(T(r,t) - T(r,0))r^2\,dr, \tag{4}$$

where $\alpha_{\mathrm{m}}(r, t)$ and α_{c} are the coefficients of thermal expansion for mantle and core, respectively.

We calculate the elastic lithospheric thickness T_{e} from $T(r, t)$ based on the strength-envelope formalism (McNutt, 1984) and the same parameters and equations as used by Grott and Breuer (2010). This is given by the depth corresponding to the temperature T_{e} at which the lithosphere loses its mechanical strength, that is:

$$T_{\mathrm{e}} = \frac{Q_i}{R_{\mathrm{gas}}}\left(\log\left(\frac{\sigma_B^{n_i} A_i}{\dot{\epsilon}}\right)\right)^{-1}, \tag{5}$$

where Q_i, A_i and n_i are rheological parameters specific for the crust (assumed to have a diabase rheology) and mantle, R_{gas} is the gas constant, σ_B is a bounding stress, and $\dot{\epsilon}$ is the strain rate. From Equation 5, we compute the mechanical thickness of the crust and mantle $D_{\mathrm{e,cr}}$ and $D_{\mathrm{e,m}}$. The effective elastic thickness is then calculated as:

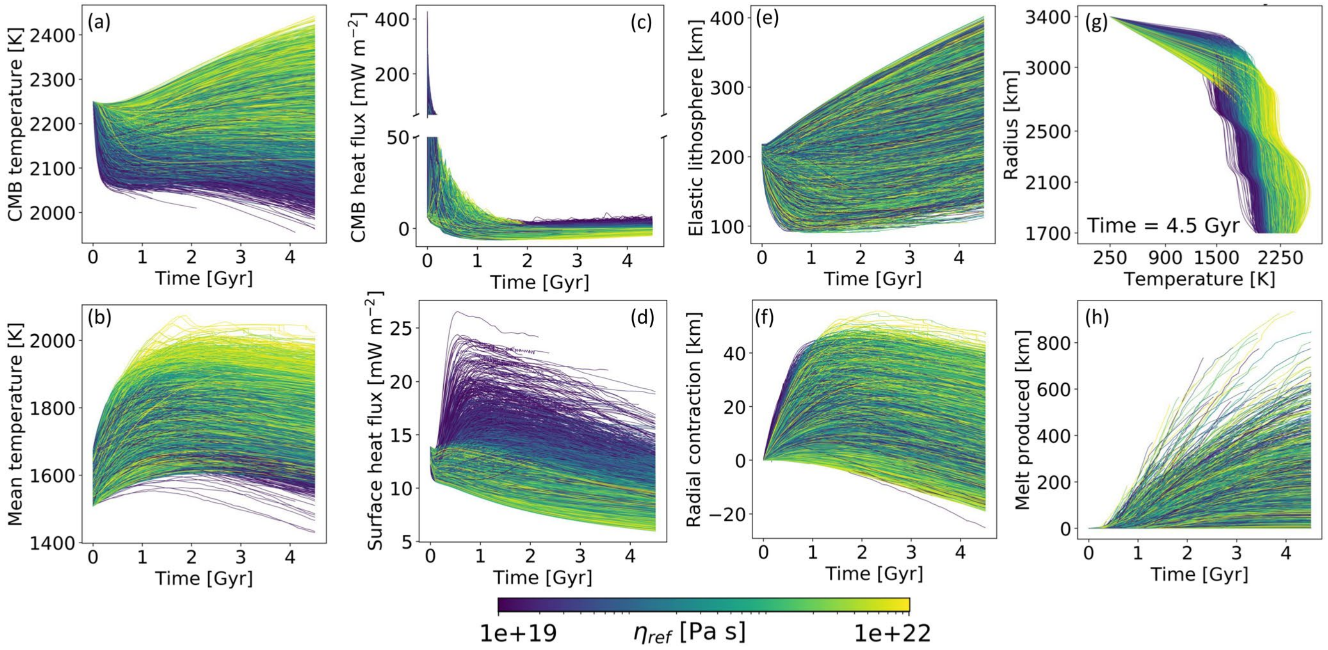


Figure 2. The data set consisting of 10,040 thermal evolution simulations with Mars-like parameters. For all simulations, the panels show the evolution of (a) CMB temperature, (b) mean mantle temperature, (c) CMB heat flux and (d) surface heat flux, (e) elastic lithospheric thickness, (f) thermally induced radial contraction, and (h) equivalent melt thickness produced. Panel (g) shows the present-day 1D temperature profiles plotted for the 6,130 simulations that could be run over the entire 4.5 Gyr of evolution. All the curves in panels a–h are colored by the value of the reference viscosity η_{ref} going from blue (low) to yellow (high). 1D, one dimensional; CMB, core-mantle boundary.

$$D_{\mathrm{e}} = \begin{cases} D_{\mathrm{e,m}} & \text{if } D_{\mathrm{e,cr}} > d_{\mathrm{cr}} \\ \left(\left(D_{\mathrm{e,m}} - d_{\mathrm{cr}}\right)^3 + D_{\mathrm{e,cr}}^3\right)^{\frac{1}{3}} & \text{otherwise} \end{cases} \tag{6}$$

where d_{cr} is the assumed crustal thickness of 64 km.

We also consider the cumulative amount of the melt produced during the evolution and calculate an equivalent thickness (D_{melt}) as well as the total duration of volcanism as two possible observables. Details on how the melt volume $V_{\mathrm{melt}}(t)$ is calculated can be found in Padovan et al. (2017) and Agarwal et al. (2020). The equivalent thickness is $D_{\mathrm{melt}}(t) = R_{\mathrm{p}} - R_{\mathrm{melt}}$, where R_{melt} can be obtained by solving the following equation:

$$V_{\mathrm{melt}}(t) = \frac{\pi}{4}\left(R_{\mathrm{p}}^2 - R_{\mathrm{melt}}^2\right). \tag{7}$$

It is worth noting that this model is simplified because the heat sources depleted at each time-step due to melt extraction are just lost in the current setup, instead of being redistributed into "new" crust. Furthermore, the duration of volcanism t_{volc} is the last time step at which melt was produced.

For a small number of specific parameter combinations (e.g., a very low reference viscosity and high activation energy) the iterative solver used by GAIA failed to converge. We filtered out such simulations by setting an upper bound for the root mean square of the magnitude of the non-dimensional velocity in the mantle u_{rms}. We empirically set the filtering criterion at $u_{\mathrm{rms}} \leq 20{,}000$ to ensure a balance between sufficient accuracy and number of lost simulations. 10,040 out of 10,080 simulations satisfied the criterion and are visualized in Figure 2 (even though, as we later explain that in this study we only consider the simulations that reached the end-time of 4.5 Gyr). The figure shows the evolution of CMB temperature (Figure 2a), mean mantle temperature (Figure 2b), CMB heat flux (Figure 2c), surface heat flux (Figure 2d), elastic lithospheric thickness (Figure 2e), thermally induced radial contraction (Figure 2f), and cumulative melt produced from all simulations (Figure 2h). The "present-day" 1D temperature profiles from simulations that finished are also plotted in Figure 2g. All the curves are colored by just one parameter in this figure. As expected, for higher reference viscosities, convection is more sluggish resulting in a less efficient heat transfer. Hence, the mantle cools more slowly, or even heats up. A lower viscosity, on the other hand, corresponds to more vigorous convection as exhibited by cooler temperature profiles with steeper surface heat flux.

It is also evident that while mean mantle temperature, for example, exhibits a color-pattern, elastic lithospheric thickness, radial contraction and melt produced show almost no correlation to reference viscosity. The colors seem scattered instead of showing a pattern that transitions from low to high values of the parameter.

In order to keep the number of computations manageable, we limit our focus to observables that can be related to the present-day temperature structure of the mantle. We split the subset of the 6,130 simulations that reached the end time of 4.5 Gyr into three parts: training (80%), test (10%) and validation (10%). The data set is non-dimensionalized to be between 0 and 1 using the maximum and minimum of each parameter and observable. It is common practice in machine learning to ensure that all the parameters are of the same order of magnitude for numerical stability of the optimizer. It also makes the error function (Equation 20) comparable across different parameters. The non-dimensionalized data set is shown in Figure 3. For each parameter in different rows, these synthetic present-day observables are plotted: CMB heat flux (Q_c), surface heat flux (Q_s), radial contraction (R_{th}), elastic lithospheric thickness (D_e), equivalent thickness of the cumulative melt produced (D_{melt}), and duration of volcanism (t_{volc}). The degeneracy of the problem is clear from the plot. Several ranges of parameters can lead to the same observation. Furthermore, for each parameter, in the very last row, 1D temperature profiles at 4.5 Gyr are also plotted for reference. They are colored from blue to yellow: from the minimum value of that parameter to its maximum. Parameters like η_{ref} and Λ seem to show a stronger pattern, while parameters like T_{ini} seem to show highly degenerate patters.

3. Mixture Density Networks

In this section, we outline the basics of Mixture Density Networks (MDN). For a more detailed explanation, we refer to the seminal paper by Bishop (1994). Consider a simple MDN, like the one illustrated in Figure 4. Here, only one hidden layer is shown. However, MDNs can have an arbitrary number of hidden layers. The MDN connects inputs **o** (observables) to outputs **p** (parameters) via a composition of neuron functions organized into multiple layers. Let *i* and *j* be indices for neurons of two consecutive layers of the MDN. The activation for neuron *j* in layer *h* depends on input *i* from the previous layer:

$$a_j = g\left(\sum_i a_i w_{ij} + b_j\right), \tag{8}$$

where w_{ij} are "weights" that can be learned from the data and g() is the activation function, which allows modeling of non-linearities. In this study, we use tanh() as the activation function. Furthermore, b_j are "biases" added to each neuron *j* in the given layer. They are also learned from the data and give the ability to translate the activation function to the left or to the right so that the origin of the activation function is no longer fixed at zero.

The problem of finding which parameters satisfy a given observable can be mathematically formulated as the conditional probability *p* (parameters|observables) or $p(\mathbf{p}|\mathbf{o})$:

$$p(\mathbf{o},\mathbf{p}) = p(\mathbf{p} \mid \mathbf{o})p(\mathbf{o}), \tag{9}$$

where, $p(\mathbf{o},\mathbf{p})$ is the joint probability density and $p(\mathbf{o})$ is the marginal probability density of observables. We can arrive at the conventional least-square formulation by assuming that the target data has the following distribution with a standard deviation σ:

$$p(\mathbf{p} \mid \mathbf{o}) = \frac{1}{(2\pi)^{c/2}\sigma^c}\exp\left\{-\frac{1}{2\sigma^2}\sum_{k=1}^{c}\left[F_k(\mathbf{o}) - \mathrm{p}_k\right]^2\right\}, \tag{10}$$

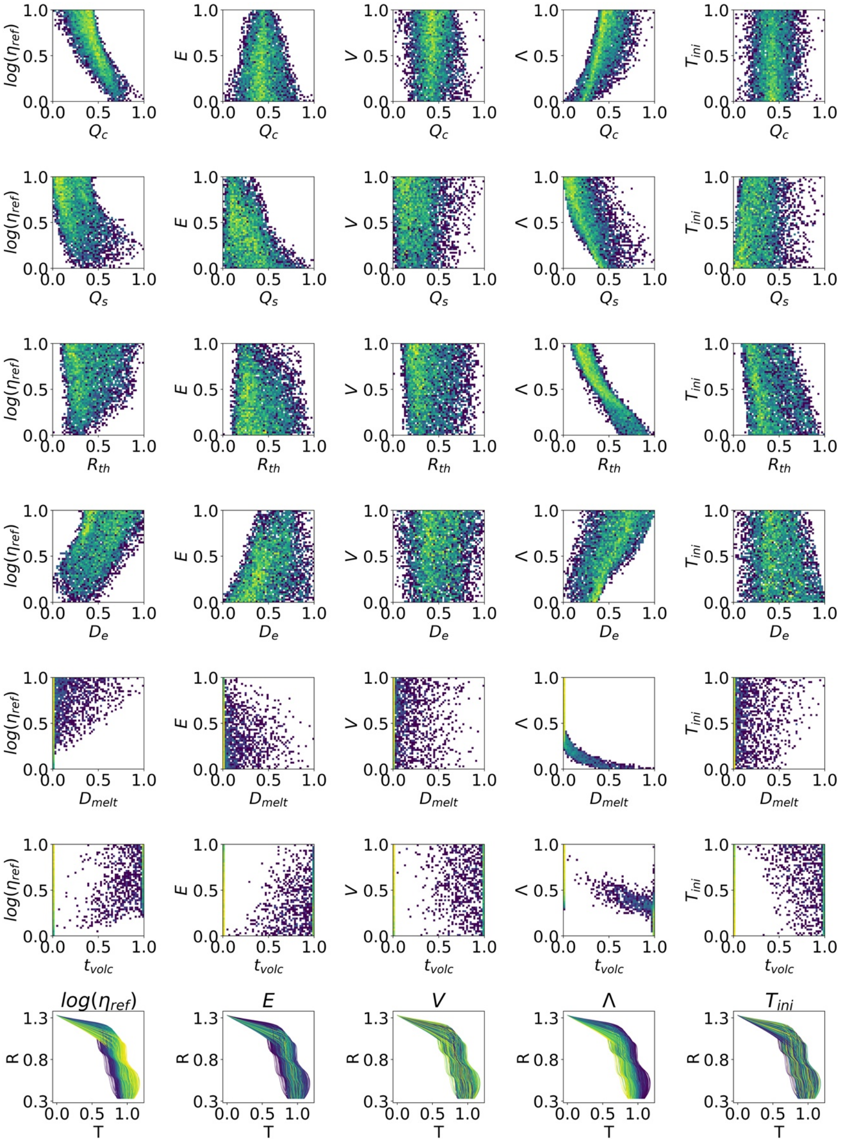


Figure 3. Two-dimensional histograms of the non-dimensionalized present-day observables and the five parameters: reference viscosity (η_{ref}), activation energy (E), activation volume (V), crustal enrichment factor (Λ) with respect to a given bulk composition of radiogenic elements, and initial temperature (T_{ini}). The corresponding dimensional values are: $\eta_{\text{ref}} \in [10^{19}, 10^{22}]$ Pa s, $E \in [10^5, 5 \times 10^5]$ J mol^{-1}, $V \in [4 \times 10^{-6}, 10 \times 10^{-6}]$ m^3 mol^{-1}, $\Lambda \in [1, 50]$, $T_{\text{ini}} \in [1{,}600, 1{,}800]$ K. The observables are CMB heat flux (Q_{c}), surface heat flux (Q_{s}), thermally induced radial contraction (R_{th}), elastic lithospheric thickness (D_{e}), cumulative melt produced (D_{melt}) and duration of volcanism (t_{volc}). For better visualization, the colorscale representing the number of simulations is normalized to each panel and is on a log scale. In the last row, 1D temperature profiles are also plotted. Unlike the previous panels, these are colored by the value of the parameter in the column: going from blue to yellow, that is, from low to high values of the parameter. 1D, one dmensional; CMB, core-mantle boundary.

where c is the total number of components of p. The function F_k is the mean of the target function and since it is unknown, it can be modeled by an NN $\mathcal{F}_k(\mathbf{o},\mathbf{w})$ from the training set $\left\{\mathbf{o}^q,\mathbf{p}^q\right\}_{q=1}^{n}$, with n examples. One can find the optimal values $\mathbf{w}*$ for $\mathcal{F}_k(\mathbf{o},\mathbf{w})$, by minimizing the negative log-likelihood:

$$\mathcal{E} = -\ln \mathcal{L} = -\ln \prod_{q=1}^{n} p(\mathbf{p}^q \mid \mathbf{o^q}) p(\mathbf{o}^q). \tag{11}$$

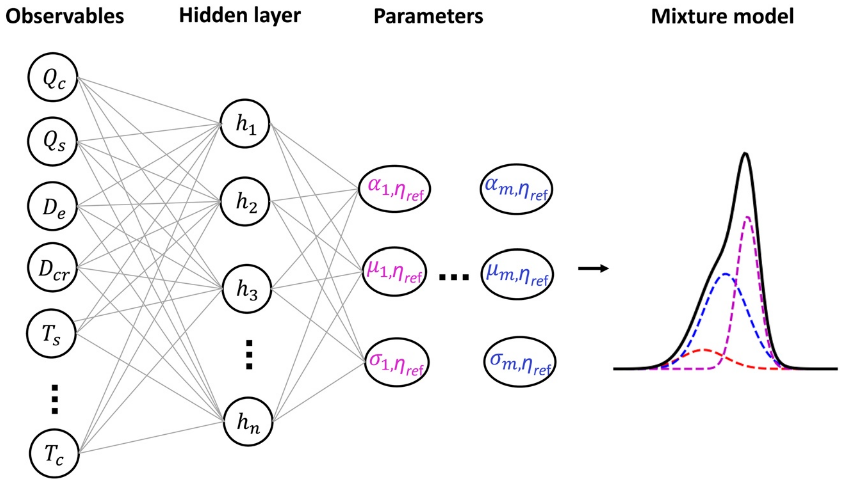


Figure 4. Using a marginal Mixture Density Network (MDNs) (Figure 1f) to constrain different parameters governing mantle convection. The input nodes **o** (observables) are connected to the output nodes **p** (parameters) via neurons in hidden layers. Each connection is quantified by an adjustable weight (***w***), which is optimized over several iterations by back-propagating the error in the prediction of the MDN. In this study, we consider several different "observables" as inputs: for example, surface heat flux Q_s, elastic lithospheric thickness D_e, as well as different parts of or the temperature profile T. For each parameter in p, say η_{ref}, the network predicts three components per Gaussian: mean μ, variance σ and weight α. We maintain a fixed architecture with two hidden layers containing 12 and 6 neurons, and 3 mixtures. These three mixtures (colored magenta, red, blue) are added to arrive at the combined probability distribution (black). We use the marginal MDN to train on only 1 parameter at a time, except in Section 4.5., where we use the joint MDN to train on all parameters at once.

Substituting, Equation 10 into 11 leads to an optimization problem, which can be further reduced to

$$\mathcal{E} = \frac{1}{2}\sum_{q=1}^{n}\sum_{k=1}^{c}\left[\mathcal{F}_k(\mathbf{o}^q,\mathbf{w}) - \mathrm{p}_k^q\right]^2, \tag{12}$$

Since this is the only term that depends on weights **w**, thereby yielding the standard least squares problem. $\mathcal{F}(\mathbf{o},\mathbf{w}^*)$, however, approximates the conditional average of parameters given the observables and thus, is not suitable for capturing all the possible solutions in an ill-conditioned problem. Therefore, Bishop (1994) proposes modeling $p\,(\mathbf{p}|\mathbf{o})$ by a mixture of distributions:

$$p(\mathbf{p} \mid \mathbf{o}) = \sum_{i=1}^{m}\alpha_i(\mathbf{o})\,\phi_i(\mathbf{p} \mid \mathbf{o}). \tag{13}$$

Here, m is the number of mixtures, $\phi_i(\mathbf{p}|\mathbf{o})$ the kernel function representing the conditional density and $\alpha_i(\mathbf{o})$ the mixing coefficient, such that:

$$\sum_{i=1}^{m}\alpha_i(\mathbf{o}) = 1, \tag{14}$$

which is achieved by using a softmax function on the network outputs p_i^{α}:

$$\alpha_i = \mathrm{softmax}(p_i^{\alpha}) = \frac{\exp(p_i^{\alpha})}{\sum_{j=1}^{m}\exp(p_j^{\alpha})}. \tag{15}$$

In principle, many choices for the kernel function are possible. However, since theoretically a mixture of Gaussian distributions can approximate any given density distribution (Mclachlan & Basford, 1988), we consider only Gaussian kernels in this study:

$$\phi_i(\mathbf{p} \mid \mathbf{o}) = \frac{1}{(2\pi)^{c/2}\sigma^c}\exp\left\{-\frac{\parallel \mathbf{p} - \mu_i(\mathbf{o}) \parallel^2}{2\sigma_i(\mathbf{o})^2}\right\}. \tag{16}$$

3.1. The Univariate Case

In the case where we are constraining one parameter at a time (Figure 4), for each mixture i, the marginal MDN outputs three quantities: p_i^u, p_i^σ and p_i^α. The mean μ_i, is taken directly as one of the network outputs:

$$\mu_i = p_i^\mu \tag{17}$$

and variance is

$$\sigma_i = \exp(p_i^\sigma). \tag{18}$$

Now, the log-likelihood can be calculated as

$$\ln \mathcal{L} = \ln\left\{\sum_{i=1}^{m} \alpha_i(\mathbf{o}^{\mathbf{q}})\,\phi_i\left(\mathbf{p}^q \mid \mathbf{o}^q\right)\right\}, \tag{19}$$

and its negative can be minimized, yielding the error function:

$$\mathcal{E} = \sum_q -\ln\left\{\sum_{i=1}^{m} \alpha_i(\mathbf{o}^{\mathbf{q}})\,\phi_i\left(\mathbf{p}^q \mid \mathbf{o}^q\right)\right\}. \tag{20}$$

Calculation of derivatives of the error with respect to output parameters (such as $\frac{\partial \mathcal{E}^q}{\partial p_i^\alpha}$, $\frac{\partial \mathcal{E}^q}{\partial p_i^\sigma}$ and $\frac{\partial \mathcal{E}^q}{\partial p_i^u}$) can be done analytically as shown in Bishop (1994). The error gradient can then be further propagated from mixture parameters to the neural network layers using back-propagation (e.g., Rumelhart et al., 1986; Werbos, 1982). For convenience, instead of using analytical expressions, we calculate the necessary derivatives using Automatic Differentiation (AD), now offered by several ML libraries. Specifically, we use TensorFlow (Abadi et al., 2015), where one only needs to set up the computational graph by defining the MDN architecture and specifying the cost function. TensorFlow then, takes care of the rest by using AD and one of the several already programmed optimizers to minimize the cost function. We use simple gradient descent with the Adam optimizer in this study due to the small size of the data set, with a pre-fixed learning rate, which controls how large the optimization strides are.

To train and evaluate the performance of the MDNs, we split the data into three parts. Training set: 80% of the data is used to train the network, validation set: 10% is used to test different hyperparameters of the MDN such as number and size of hidden layers and further make sure the network is not overfitting and test set: the remaining 10% is used to evaluate the performance of the trained network and present the results. This last subset of the data is helpful in assessing how well the MDN performs, because it is not seen by the network at any point in training or development. We use two techniques to prevent over-fitting. First, we employ weight-decay by modifying the cost function in Equation 20 as done by Hjorth and Nabney (1999):

$$\mathcal{E} = \left\{\sum_q -\ln\left\{\sum_{i=1}^{m} \alpha_i(\mathbf{o}^q)\,\phi_i\left(\mathbf{p}^q \mid \mathbf{o}^q\right)\right\}\right\} + \gamma \sum_{j=1}^{n}\sum_{k=1}^{r} \frac{w_{k,j}^2}{2}, \tag{21}$$

where the second term acts as a Gaussian regularizer and is summed over n training examples and r layers. The regularization parameter γ can be thought of as another hyperparameter of the MDN (like the number of hidden layers). Second, we only let the network train until the cost function on the test set starts increasing beyond a certain threshold known as early stopping.

3.2. The Multivariate Case

The marginal MDN illustrated in Figure 4 can be modified to predict the joint probability distributions for multiple parameters. Figure 5 shows the construction of a mixture of high-dimensional probability distributions. For m mixtures and c parameters, the network predicts mixture weights $\alpha \in \mathbb{R}^m$, means $\boldsymbol{\mu} \in \mathbb{R}^{m\times c}$ and l non-zeros components of a lower-diagonal matrix $\boldsymbol{\sigma} \in \mathbb{R}^{m\times l}$, where l is the sum of c diagonal components

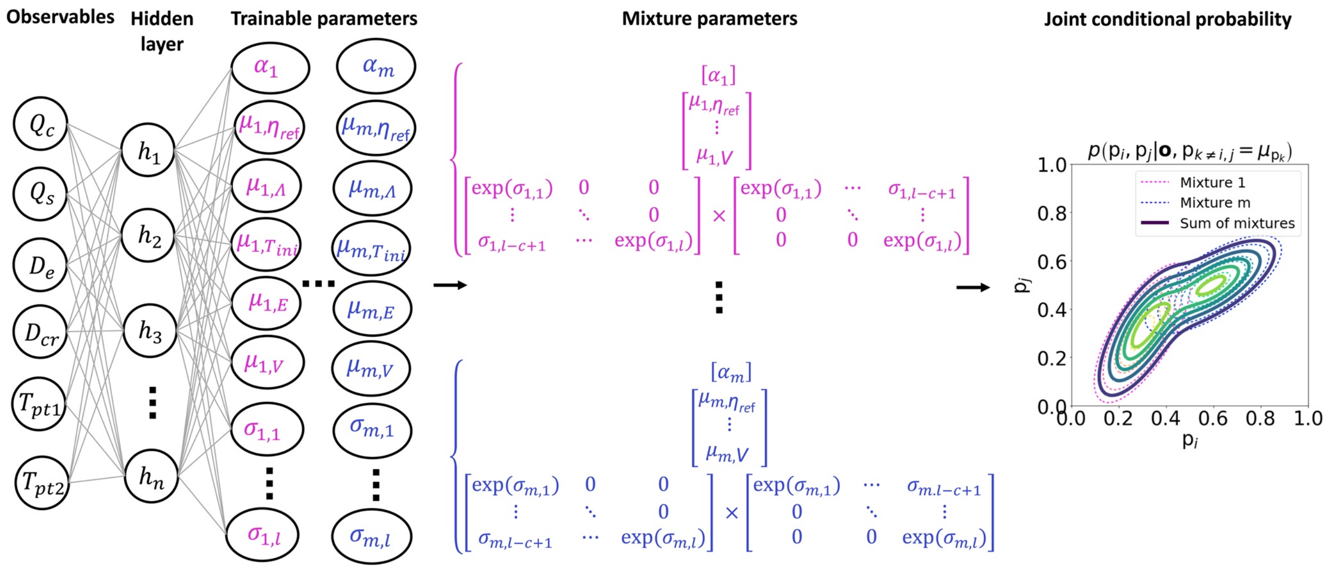

Observables
Hidden layer
Trainable parameters
Mixture parameters
Joint conditional probability
Q_c
Q_s
D_e
D_{cr}
T_{pt1}
T_{pt2}
h_1
h_2
h_3
h_n
α_1
$\mu_{1,\eta_{ref}}$
$\mu_{1,\Lambda}$
$\mu_{1,T_{ini}}$
$\mu_{1,E}$
$\mu_{1,V}$
$\sigma_{1,1}$
$\sigma_{1,l}$
α_m
$\mu_{m,\eta_{ref}}$
$\mu_{m,\Lambda}$
$\mu_{m,T_{ini}}$
$\mu_{m,E}$
$\mu_{m,V}$
$\sigma_{m,1}$
$\sigma_{m,l}$
$[\alpha_1]$
$\begin{bmatrix}\mu_{1,\eta_{ref}}\\ \vdots \\ \mu_{1,V}\end{bmatrix}$
$\begin{bmatrix}\exp(\sigma_{1,1}) & 0 & 0\\ \vdots & \ddots & 0\\ \sigma_{1,l-c+1} & \cdots & \exp(\sigma_{1,l})\end{bmatrix} \times \begin{bmatrix}\exp(\sigma_{1,1}) & \cdots & \sigma_{1,l-c+1}\\ 0 & \ddots & \vdots\\ 0 & 0 & \exp(\sigma_{1,l})\end{bmatrix}$
$[\alpha_m]$
$\begin{bmatrix}\mu_{m,\eta_{ref}}\\ \vdots \\ \mu_{m,V}\end{bmatrix}$
$\begin{bmatrix}\exp(\sigma_{m,1}) & 0 & 0\\ \vdots & \ddots & 0\\ \sigma_{m,l-c+1} & \cdots & \exp(\sigma_{m,l})\end{bmatrix} \times \begin{bmatrix}\exp(\sigma_{m,1}) & \cdots & \sigma_{m.l-c+1}\\ 0 & \ddots & \vdots\\ 0 & 0 & \exp(\sigma_{m,l})\end{bmatrix}$
$p(\mathrm{p}_i, \mathrm{p}_j|\mathbf{o}, \mathrm{p}_{k\neq i,j} = \mu_{\mathrm{p}_k})$
Mixture 1
Mixture m
Sum of mixtures
p_j
p_i
0.0
0.2
0.4
0.6
0.8
1.0

Figure 5. Using a Mixture Density Network (MDNs) to obtain the joint probability distributions of all five input parameters. The input nodes **o** (observables) are connected to the output nodes (trainable parameters) via neurons in hidden layers. For m mixtures and c parameters, the network predicts mixture weights $\alpha \in \mathbb{R}^{m}$, means $\boldsymbol{\mu} \in \mathbb{R}^{m\times c}$ and l non-zeros components of a lower-diagonal matrix $\boldsymbol{\sigma} \in \mathbb{R}^{m\times l}$, where l is the sum of c diagonal components and $\frac{c(c-1)}{2}$ non-diagonal components. These components are used to calculate the mixture of the multivariate normal distributions. The mixture of normal distributions can then be visualized as a 2D joint probability distribution for two parameters (p_i, p_j) at a time, conditioned on the observables o and visualized at the means of the remaining parameters $\mathrm{p}_{k\neq i,j} = \mu_{\mathrm{p}_k}$. In this subsection, we consider all the observables as inputs such as surface heat flux Q_s, elastic lithospheric thickness D_e as well as two points $T_{\mathrm{pt},1}$ and $T_{\mathrm{pt},2}$ from the temperature profile at phase transition locations. 1D, one dimensional; 2D, two dimensional.

and $\frac{c(c-1)}{2}$ non-diagonal components. So, for example, for three mixtures and five parameters the total number of trainable parameters would be $3 \times (1 + 5 + (5\,(5 - 1)/2 + 5)) = 63$.

We construct the covariance matrix in a specific manner to ensure numerical stability during the training procedure. Our code uses Multivariate Normal Full Covariance from the TensorFlow Probability library (Dillon et al., 2017) to define a mixture of high-dimensional probability distributions during the construction of the forward graph. This function uses Cholesky decomposition on the covariance matrix Σ, such that

$$\boldsymbol{\Sigma} = \sigma\sigma^{\mathrm{T}}, \tag{22}$$

where σ is a lower-diagonal matrix. Effectively, we train the MDN to calculate the lower-diagonal matrix σ and then use it to compute the covariance matrix. This ensures that Σ is a positive-definite matrix, a requirement for Cholesky decomposition. Also, before we multiply σ by its transpose, we take the exponential of its diagonal elements to ensure their positivity, as suggested by Kruse (2020).

Then, we can calculate the joint normal distribution $\phi_i\,(\mathbf{p}|\mathbf{o})$ for mixture i:

$$\phi_i(\mathbf{p} \mid \mathbf{o}) = \frac{1}{(2\pi)^{c/2}\left|\boldsymbol{\Sigma}_i\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{p} - \boldsymbol{\mu}_i\right)^{\mathrm{T}} \boldsymbol{\Sigma}_i^{-1}\left(\mathbf{p} - \boldsymbol{\mu}_i\right)\right\}. \tag{23}$$

4. Results and Discussion

4.1. Training Results

We test different combinations of observables to train $\mathcal{F}(\mathbf{o}, \mathbf{w})$ for one parameter at a time. Figure 6 shows how the training works. For given observables, weights of $\mathcal{F}(\mathbf{o}, \mathbf{w})$ are optimized to $\mathbf{w}^*$ as per a stopping criterion of:

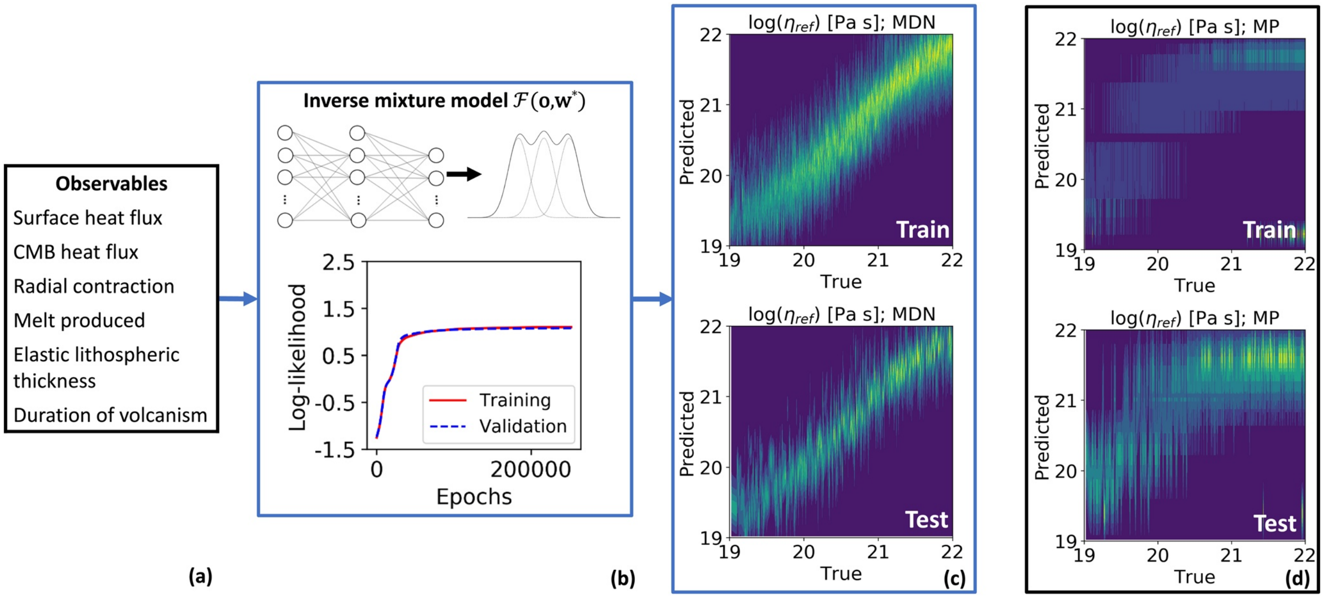


Figure 6. (a) Any number and combination of observables is selected to be input to the MDN. (b) The network is trained until the defined early stopping criteria for the loss function (negative log-likelihood) is met. (c) The trained network is used to obtain the marginal probability of a parameter for given observables and plotted along the *y*-axis for each true value of the parameter along the *x*-axis. (d) A mean-predictor (MP) can be built for comparison by binning the data and obtaining the means and variances in each bin. MDN, mobile density network.

$$\text{train while:}\quad \text{loss}_{\text{validation}}\left(\text{epoch}\right) - \text{loss}_{\text{validation}}\left(\left\lfloor 0.99\text{epoch}\right\rfloor\right) \leq 0 \tag{24}$$

Here, one epoch is when the optimization algorithm has trained over the entire training set. Since epoch is an integer, we take the floor of the product of 0.99 and the number of epochs that have elapsed. We found that a value of 0.00001 for the learning rate delivered optimal results. A higher learning rate led the network to sometimes miss the minimum during optimization. Then, one can visualize the marginal probability distributions of a parameter for each example $p(\mathbf{p} \mid \mathbf{o}) \approx \mathcal{F}(\mathbf{o}, \mathbf{w}^*)$ in the training or the test set, as shown by Atkins et al. (2016). For each true value along the *x*-axis, we plot in Figure 6 the predicted probability distribution functions (PDFs) along the *y*-axis. For a perfectly constrainable parameter, such a contour plot would yield a diagonal line.

As a test to verify if our network predictions are indeed non-trivial, we also plot the PDFs from a so-called mean-predictor (MP). The MP is obtained by binning the data and obtaining the mean and standard deviation of each bin. In Figure 6, we select the mean and the standard deviation for each bin across all observables based on the observable-bin with the minimum standard deviation. It is clear from the PDFs of the reference viscosity η_{ref} in Figure 6, that the MDN outperforms the MP baseline model. In other words, the results of the MDN are non-trivial. The MP can be thought of as a discretized form of a single-gaussian MDN, which looks at each bin to predict a quantity. It provides a very simple baseline to compare the results of the MDN with.

Instead of qualitatively inspecting hundreds of plots such as in Figure 6c to determine the strength of constraints on a parameter for several combinations of network architectures, observables, uncertainties etc., we quantify the strength of constraint using the log-likelihood (the negative of the error function in Equation 20. The higher the log-likelihood value is, the better can a parameter be constrained. This can be seen in Figure 7: row 2, where we plot the probability distributions obtained from the test set for each parameter and show to which value the corresponding log-likelihood value increased to over several epochs (Figure 7: row 1). In rows 3–6, we also plot some of the individual probability distributions (magenta, blue, red) obtained from the MDN, the combined distribution (black), as well as the actual corresponding value indicated by the gray line.

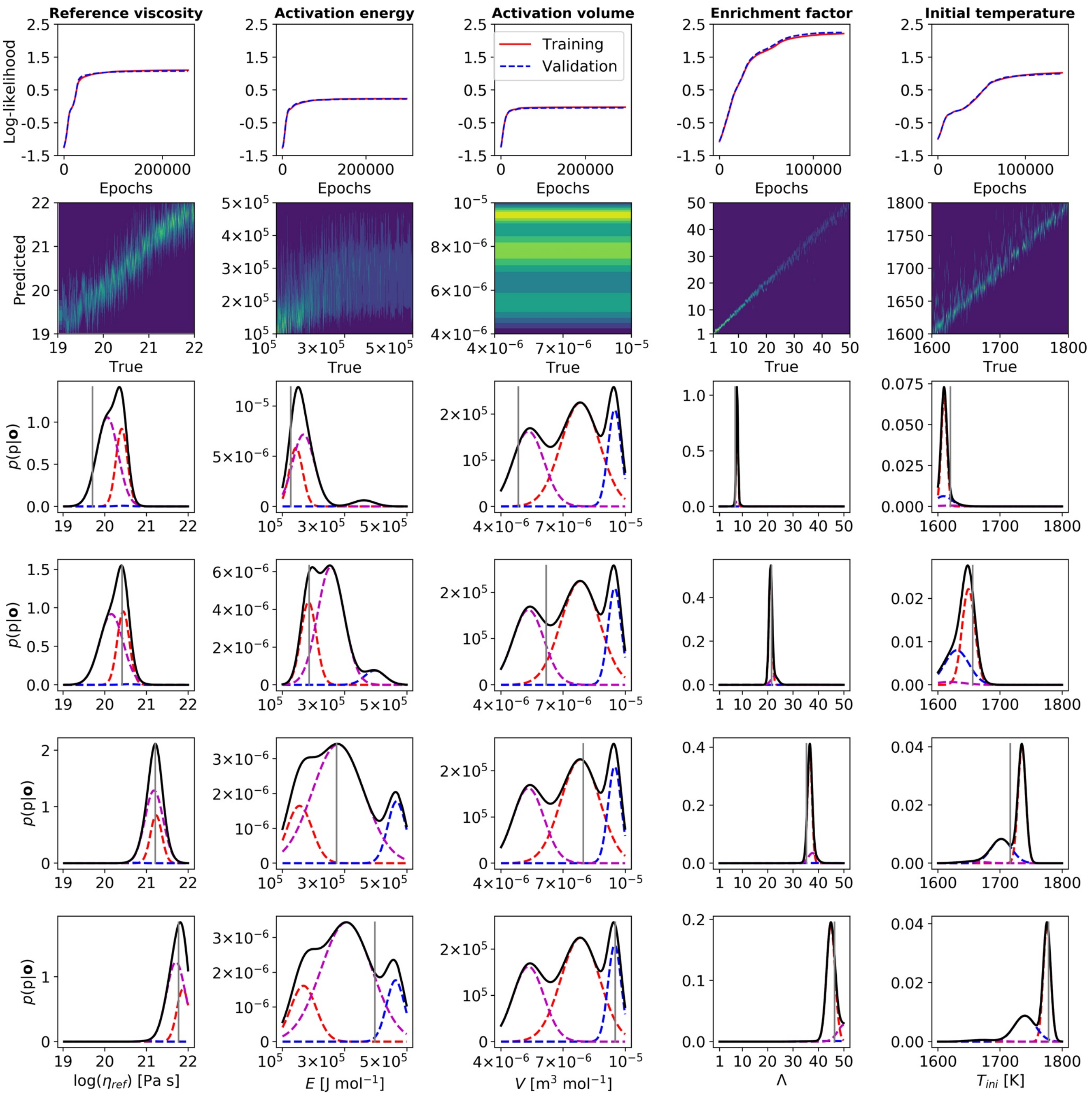


Figure 7. For the case where all observables are available, row 1 shows the evolution of the log-likelihood over epochs. It is the negative of the loss function and the higher it is, the better a parameter can be constrained. In row 2, all the probability distributions from the test set for each of the five parameters are visualized. In rows 3–6 some individual probability distributions (colored in magenta, blue and red) from the three mixtures are plotted as well as the combined (sometimes multi-modal) distribution (colored in black). The real values of each parameter are indicated by the vertical gray line.

With this metric in mind, over the next subsections, we present results of such studies using a fixed network architecture with two hidden layers consisting of a total of 18 neurons and with three mixtures. Upon testing different architectures, we found that the small size of the data set limits us to smaller networks. Further information on architecture selection can be found in the Supporting Information. In the following subsections, we present the results on the test set. Since the training of an MDN is a stochastic process, we repeat the training procedure five times and display the mean and the standard deviation of log-likelihood of all the runs for a particular training setup.

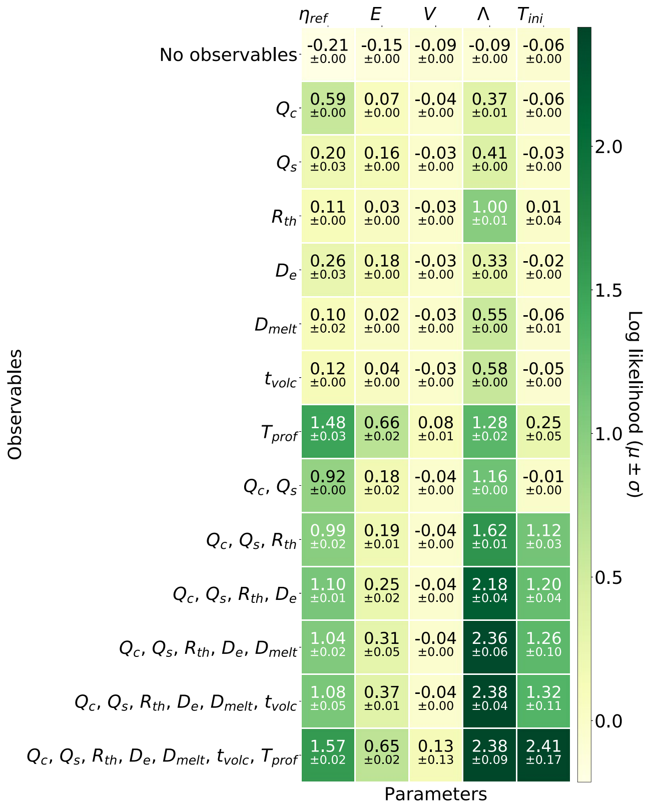


Figure 8. Constraints on each parameter (*x*-axis) for different observables and some selected combinations thereof (*y*-axis) as defined by the log-likelihood. The mean (above) and variance (below) of the log-likelihood for five runs on the test set are given. Here, T_{prof} is an abbreviation for the complete 1D present-day temperature profile. 1D, one dimensional.

4.2. Observational Constraints on Different Parameters

In Figure 8, we plot the log-likelihood of each of the five parameters for different numbers and combinations of observables. Here, we only plot a few cases. A complete list is available in the Supporting Information. We also show the log-likelihood of the prior distributions, or equivalently, for the case when "no observables" are available. The prior distribution is calculated using scikit-learn's Gaussian Mixture Model with three mixtures (see Supporting Information, Pedregosa et al., 2011). Ideally, the prior distribution of each parameter should be flat within the range across which the parameter is varied. While this was the case upon selecting the parameter values for the simulations, the actual distribution is no longer flat because, for some parameter combinations, the simulations did not reach the end time of 4.5 Gyr at which the observables are evaluated. Nevertheless, one can still gauge how much the MDN has learned by subtracting the posterior log-likelihood from the prior log-likelihood.

It is clear that V is unconstrained from any of the observables for our current setup of simulations. Furthermore, T_{ini} cannot be constrained from just the present-day temperature profiles. Indeed, the "thermostat effect" (Tozer,), that is, the tendency for the mantle temperature to converge to similar values due the temperature dependence of the viscosity, is expected to prevent inferences of T_{ini} from the final state of the system (e.g., present-day heat-flux, temperature profile, etc.). However, when the temperature profile and the radial contraction are both available (Figure S2, row 2), T_{ini} can be constrained well. Radial contraction of a planet provides a clue as to what the initial temperature was. Hence, the MDN is able to trace it from the present-day temperature.

E is weakly constrained, while η_{ref} and Λ can be constrained well in this setup. This makes sense in view of the patterns shown in Figure 3. The activation volume V, for example, seems to have no correlation with any of the observables, while E exhibits a weak correlation. The fact that η_{ref} and Λ are well constrained is also a confirmation of the visual patterns and correlations observed in Figure 3.

Figure 8 shows a general trend, namely that adding more observables improves the constraint for except for V. We also note that, with the exception of T_{ini}, adding the full 1D temperature profile to the complete observables vector o improves the constraint on the parameter only slightly. This is because most of the observables are either directly derived from the temperature profile (e.g., Q_{s}) or indirectly depend on it (e.g., D_{melt}). In other words, a combination of all the possible observables helps cover several parts of the temperature profile.

Still, to determine the extent to which constraints on parameters are affected by the inclusion of different parts of the temperature profile (like location and temperatures of the phase transitions), we test the impact of considering as observables different parts of the 1D temperature profiles. In Figure 9, we plot the log-likelihood of each parameter, given different percentages of temperature profile, either starting from the surface (Figure 9a) or from the core-mantle boundary (Figure 9b).

From the row 3 of Figure 8 and row 2 of Figure 9a, we notice that the surface heat flux is a very weak constraint. However, the top 20% of the temperature profile is sufficient for constraining Λ, while E and η_{ref} benefit from larger portions of the temperature profile. For both η_{ref} and Λ, as soon as the temperature profile includes the entire thickness of the upper thermal boundary layer, the constraints dramatically improve. Still, we see that while peripheral parts (such as surface heat flux, CMB heat flux and CMB temperature) of the temperature profiles provide some constraints, more information vastly improves the inference of parameters. This is especially true for E and to some extent for η_{ref}, whose observational signatures seem to be more distributed throughout the mantle rather than being radially concentrated within certain ranges. This

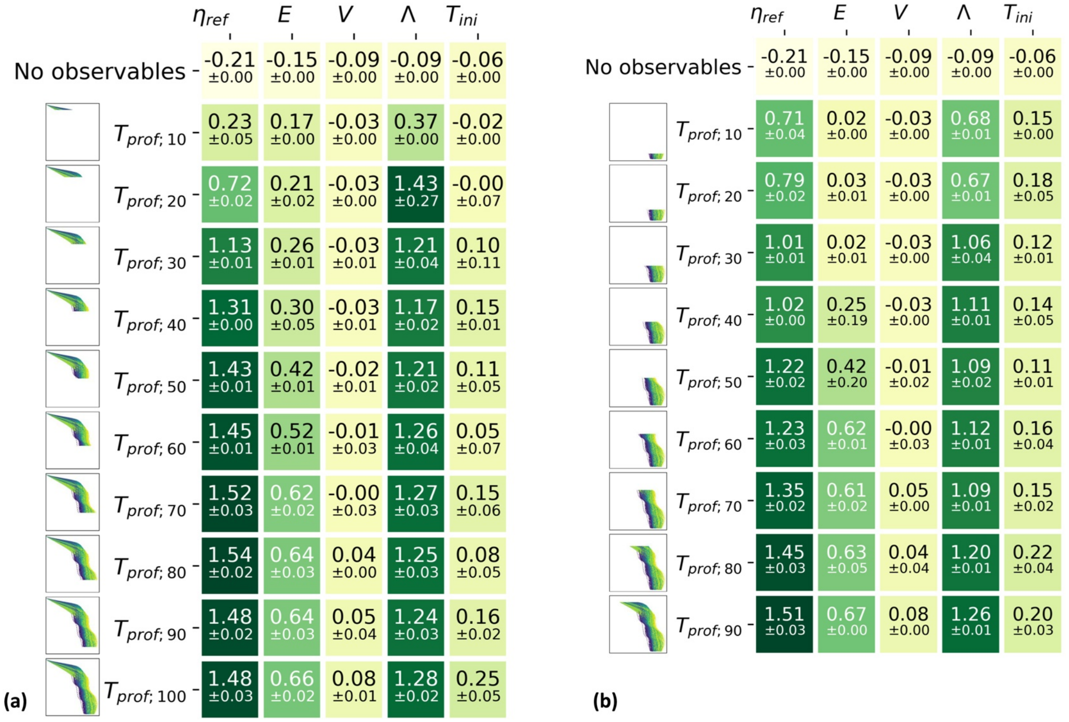


Figure 9. Constraints on each parameter from different parts of the temperature profile. The mean and variance of the log-likelihood for five runs on the test set are given.

additional information can be provided in the form of specific depth-temperature values or other related quantities such as elastic lithospheric thickness, potential temperature and duration of volcanism.

To illustrate more quantitatively the accuracy in the inference of model parameters, we show some selected cases covering a wide range of log-likelihood values from Figure 8 and plot the MDN calculated probability distributions on the test set in Figure 10a. Within each test set, we also display the average standard deviation ($\overline{\sigma}$) of the combined mixture of the probability distributions. Typically, for a normal distribution, ~95% of the data is contained within $\mu \pm 2\sigma$. Hence, in Figure 10b, we also plot the corresponding $\pm 2\sigma$ values divided by the entire range of the given parameter for different log-likelihoods. For the case where only Q_s is available, $\pm 2\sigma$ spans 87% of the entire range of η_{ref} and 74% of the entire range of Λ. However, if all the observables are available, one can constrain η_{ref} within 32% of its range, Λ within 15% and T_{ini} to within 39%. E can be weakly constrained within 80% its entire range, given that all the observables are available.

4.3. Impact of Uncertainty in Observations on Constraints on Parameters

We also quantify the impact of uncertainty in the observations upon constraining the parameters. In Figure 11, we plot the negative log-likelihood of each parameter, given different magnitude of uncertainty in the observations. We do this by adding noise with a normal distribution $\mathcal{N}(0,(\text{noise} \times \sigma_o)^2)$ to the entire data set, where σ_o is the standard deviation of each observable in **o**. This is done for the case where all the observables are available. Upon increasing σ_o, the constraint systematically weakens, albeit with a different rate depending on the parameter. For initial mantle temperature, the constraint is drastically diminished for even a low level of uncertainty of $0.1\sigma_o$. For a potential real-world application, the noise will need to be added according to the proper uncertainty specifications (magnitude, distribution, etc.) of each observable. The noise in observations could potentially be compensated through more data, either through more simulations or through more observables.

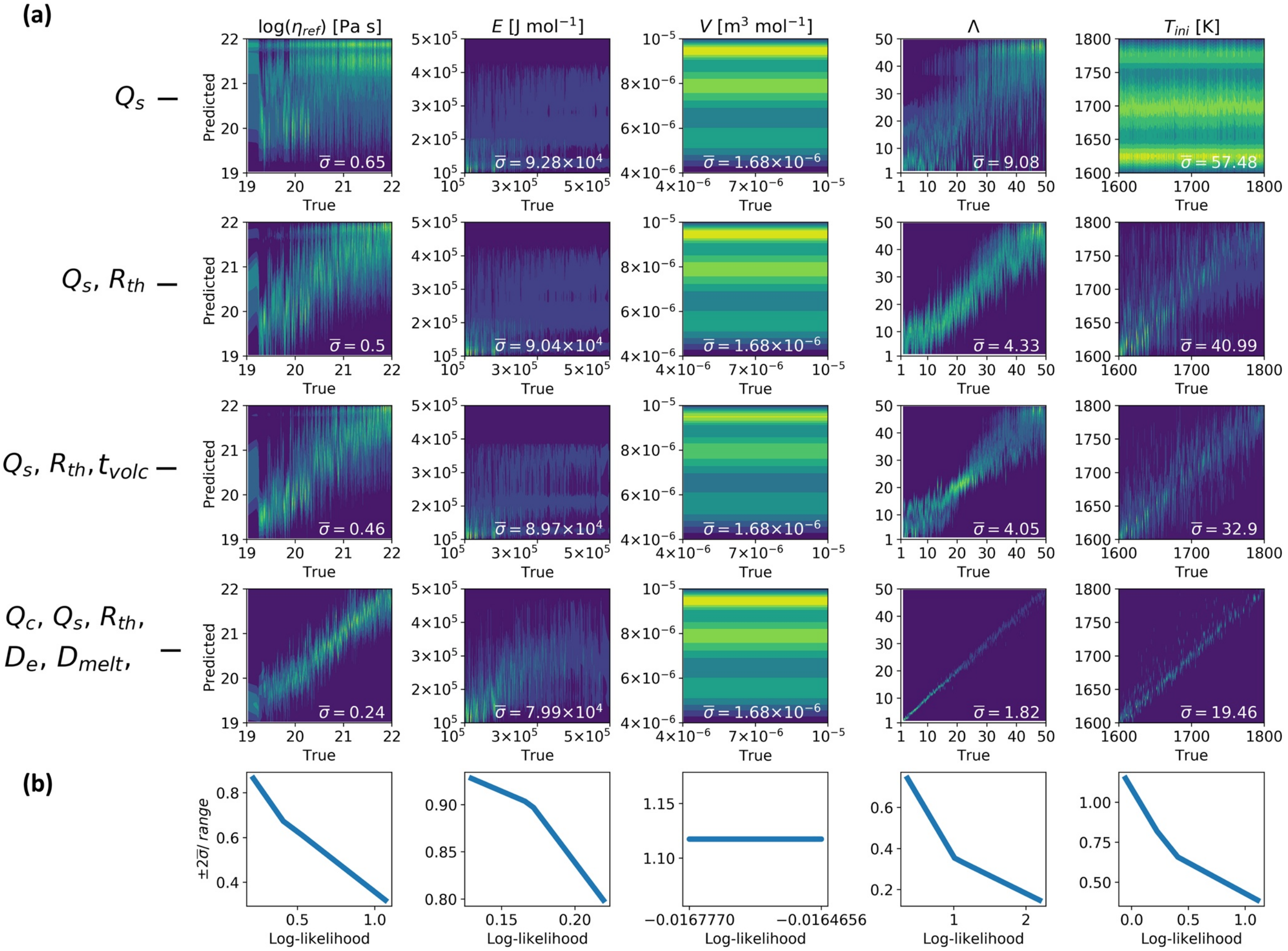


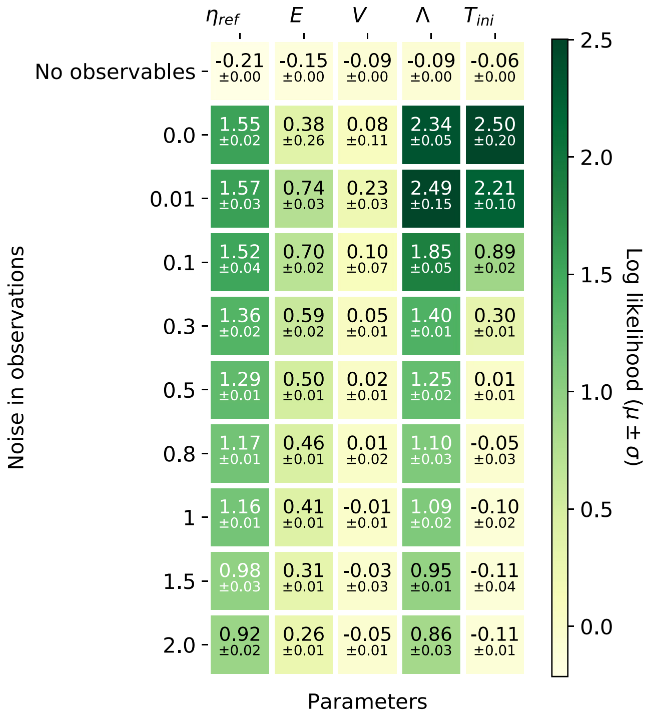
Figure 10. (a) The probability density functions calculated by MDN on the test set for a few selected cases and the average standard deviation of the combined mixture. (b) ±2σ/range of parameter versus log-likelihood for a parameter. For the case where only surface heat flux is available as an observable, reference viscosity can be constrained within 87% of its entire range. With all the observables one can tighten the constraint on reference viscosity to within 32% of its entire range, crustal enrichment to within 15% and initial mantle temperature to within 39%. Activation energy can at best, be weakly constrained to within 80% of its entire range, which activation volume cannot be constrained at all. MDN, mobile density network.

4.4. Availability of Observables and Number of Simulations

We look at two further factors that can impact how well mantle convection parameters can be constrained. In Figure 12, we show how the constraint on different parameters changes with the total number of simulations available (y-axis) and with increasing number of observables (x-axis).

The results are intuitive: more observables result in a tighter constraint. Similarly, more simulations also help tighten the constraints. Nevertheless, the improvements in the log-likelihood are asymptotic from approximately 120 simulations for η_{ref}, E and V and from roughly 500 simulations for Λ and T_{ini}. Clearly, training the MDN with just 56 simulations is not sufficient. The prior log-likelihood of 56 simulations is noticeably lower, suggesting that the distribution of training simulations in this data set is not as representative as the distribution of training simulations in other cases with more simulations. Therefore, improving the distribution of the training data set should improve the prior, as well as the posterior log-likelihood. For Λ, the log-likelihood given all observables increases by 0.45 ± 0.2 in going from 56 to 126 simulations. This is comparable to an increase of 0.34 in prior log-likelihood for the same case. So, the log-likelihood increase in this case can be attributed in large part to the improvement of the distribution of the training data.

However, that is not the only factor. After the underlying training distribution has reached a certain threshold of prior log-likelihood or, equivalently, it has become representative enough then the number of simu-



	η_{ref}	E	V	Λ	T_{ini}
No observables	-0.21 ±0.00	-0.15 ±0.00	-0.09 ±0.00	-0.09 ±0.00	-0.06 ±0.00
0.0	1.55 ±0.02	0.38 ±0.26	0.08 ±0.11	2.34 ±0.05	2.50 ±0.20
0.01	1.57 ±0.03	0.74 ±0.03	0.23 ±0.03	2.49 ±0.15	2.21 ±0.10
0.1	1.52 ±0.04	0.70 ±0.02	0.10 ±0.07	1.85 ±0.05	0.89 ±0.02
0.3	1.36 ±0.02	0.59 ±0.02	0.05 ±0.01	1.40 ±0.01	0.30 ±0.01
0.5	1.29 ±0.01	0.50 ±0.01	0.02 ±0.01	1.25 ±0.02	0.01 ±0.01
0.8	1.17 ±0.01	0.46 ±0.01	0.01 ±0.02	1.10 ±0.03	-0.05 ±0.03
1	1.16 ±0.01	0.41 ±0.01	-0.01 ±0.01	1.09 ±0.02	-0.10 ±0.02
1.5	0.98 ±0.03	0.31 ±0.01	-0.03 ±0.03	0.95 ±0.01	-0.11 ±0.04
2.0	0.92 ±0.02	0.26 ±0.01	-0.05 ±0.01	0.86 ±0.03	-0.11 ±0.01

Figure 11. Constraints on each parameter for different levels of noise added in the case where all observables are available including the temperature profile. The mean and variance of the log-likelihood for five runs on the test set are given.

lations also plays a role. For Λ, the prior log-likelihood only increases by 0.03, when going from 126 to 504 simulations. In contrast, the posterior log-likelihood given all observables increases by 0.74 ± 0.19. This effect is even more pronounced for T_{ini}, where the prior-likelihood of 2,016 simulations is higher than that of 56 simulations by 0.14, while the log-likelihood given all observables improves by 2.98 ± 0.57. Thus, both the quality and quantity of the training examples are important.

It is worth noting that the single log-likelihood number over the entire test set does not provide granular insights into the sub-spaces where we might be lacking data. For example, from Figure 7: row 2, column 1 we notice that the MDN struggles to constrain low values of the reference viscosity. This is because there are fewer simulations run with such values where the mantle is convecting more vigorously.

Nevertheless, within the uncertainties of training MDNs, the constraints on all the parameters improve only asymptotically. We can thus conclude that, for this setup, we have enough data.

4.5. Obtaining the Joint Probability Model

So far, we have limited our focus to how well individual parameters can be constrained from different observables. However, as pointed out by de Wit et al. (2013) and Atkins et al. (2016), the individual 1D probability distributions such as those visualized in Figure 7 are marginal and lack the covariances between parameters (cross-covariances).

de Wit et al. (2013) demonstrated that a higher-dimensional probability distribution can be constructed iteratively by multiplying a lower dimensional marginal distribution with a higher-dimensional conditional distribution. One could, for example, multiply the 1D distribution $p\,(\eta_{\mathrm{ref}}|\mathbf{o})$ with a 2D conditional distribution $p(E|\mathbf{o}, \eta_{\mathrm{ref}})$ to obtain the 2D distribution $p(E, \eta_{\mathrm{ref}}|\mathbf{o})$. In their case of 29 unknown parameters, this approach is reasonable since the number of trainable parameters required to build a joint probability distribution would explode.

For this study, though, where we only vary five parameters, it is feasible and more convenient to train the MDN on the joint five-dimensional probability distribution as described in Section 3.2. For convenience, we use the Keras MDN layer provided by Martin (2018) and replace the multivariate normal distribution whose covariance matrix contains only diagonal elements with the that of a full covariance matrix. Our code can be accessed here: https://github.com/agsiddhant/Inverse_Modelling_Mars_1D. The repository also contains the data set used in this paper. We train the network in Keras API, which is built on top of Tensorflow backend (Chollet et al., 2015).

To ensure computational efficiency in training networks that predict 63 quantities per training example (Section 3.2), compared to only 3 as was the case in previous subsections, we train the MDN in mini-batches of 32 up to 50,000 epochs, while saving the trained model after each epoch only if the loss on the validation set dropped. Predicting 21 times as many quantities as before also means that larger network architectures are helpful. The input to the network consists of all the present-day observables: CMB heat flux (Q_{c}), surface heat flux (Q_{s}), radial contraction (R_{th}), elastic lithospheric thickness (D_{e}), equivalent thickness of the cumulative melt produced (D_{melt}), duration of volcanism (t_{volc}), as well as two temperature points at the reference depth of the phase transitions ($T_{\mathrm{pt},1}$ and $T_{\mathrm{pt},2}$). While, it is unreasonable to expect the full 1D temperature profile for other planets such as Mars, ongoing and future missions could reveal selected temperature-pressure points at discontinuities from seismic data.

A simple trial-and-error approach of training MDN with different number of hidden layers (1–3), hidden units per layer (6–60) and mixtures (3–9) revealed that a network that is large enough, while at the same time not too large to train with a small set of 5,517 training examples, performs the best. Hence, we present

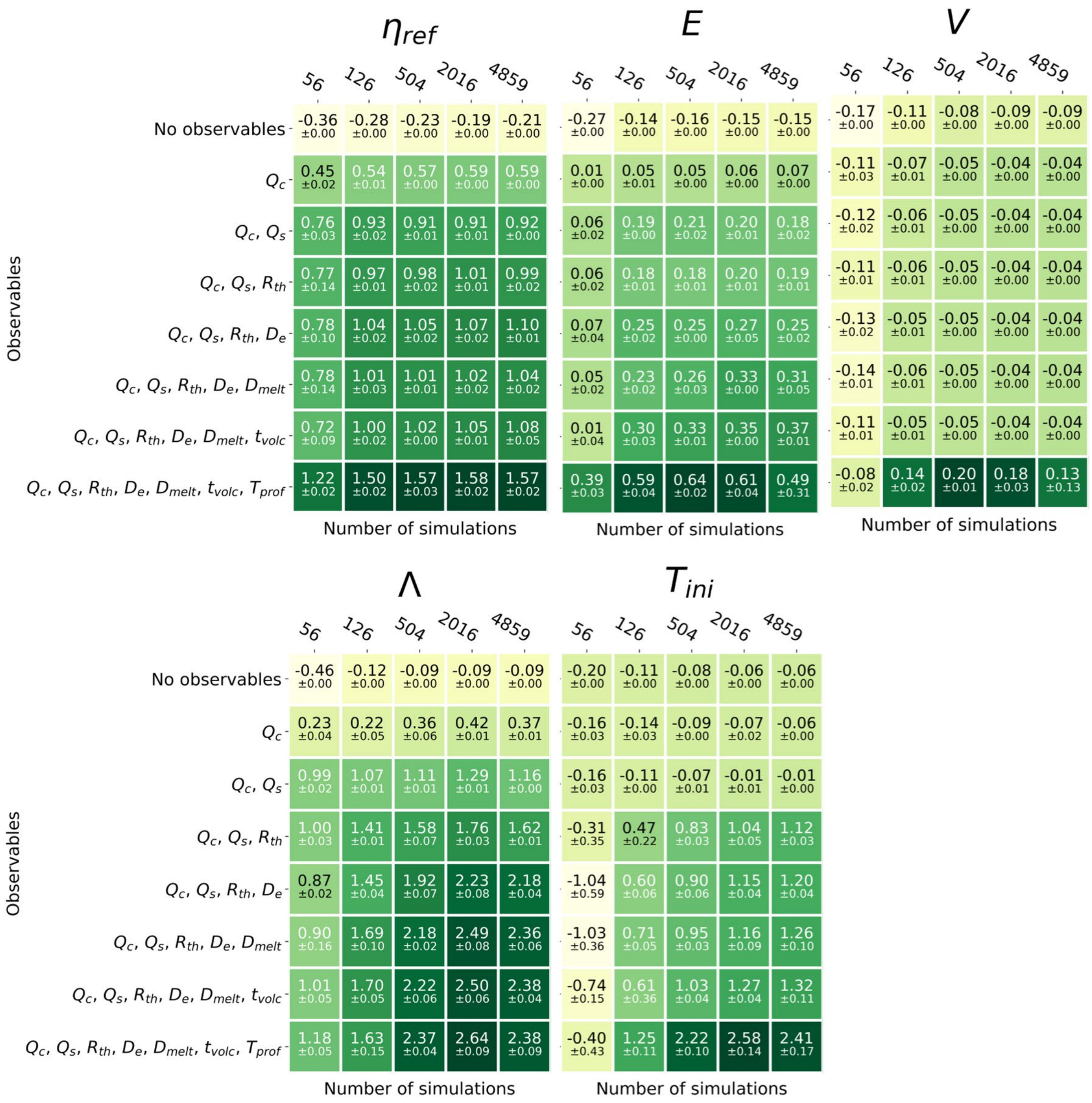


Figure 12. Constraints on all parameters for different observables and varying number of simulations available. The mean and variance of the log-likelihood for five runs on the test set are given.

four examples on the predictions on the test set from a trained MDN with 3 mixtures and hidden units distribution across three hidden layers of [60, 60, 60].

In Figure 13, we plot the log-likelihood values on the entire test set (calculated using the trained MDN). Each point is a log-likelihood value for a specific combination of five parameters. The first thing that is obvious is the lack of data in some "corners" such as low reference viscosity and high activation energy, as expected. The second thing to note would be that for the remaining parameter-space, one sees no obvious patterns. However, upon careful examination one can observe, for example, that in the $E - \eta_{\text{ref}}$ plot, high reference viscosity tends to have higher log-likelihood values (darker dots). The same is true for $\Lambda - \eta_{\text{ref}}$ plot. Yet, it is interesting to note that η_{ref} seems to be more dominant than Λ, since most high log-likelihood values occur at high η_{ref} and low Λ values. This is lightly counter-intuitive, since a more enriched mantle (low Λ) has the same qualitative effect as having a low reference viscosity and thereby should make the mantle convect more vigorously (and more difficult to constrain).

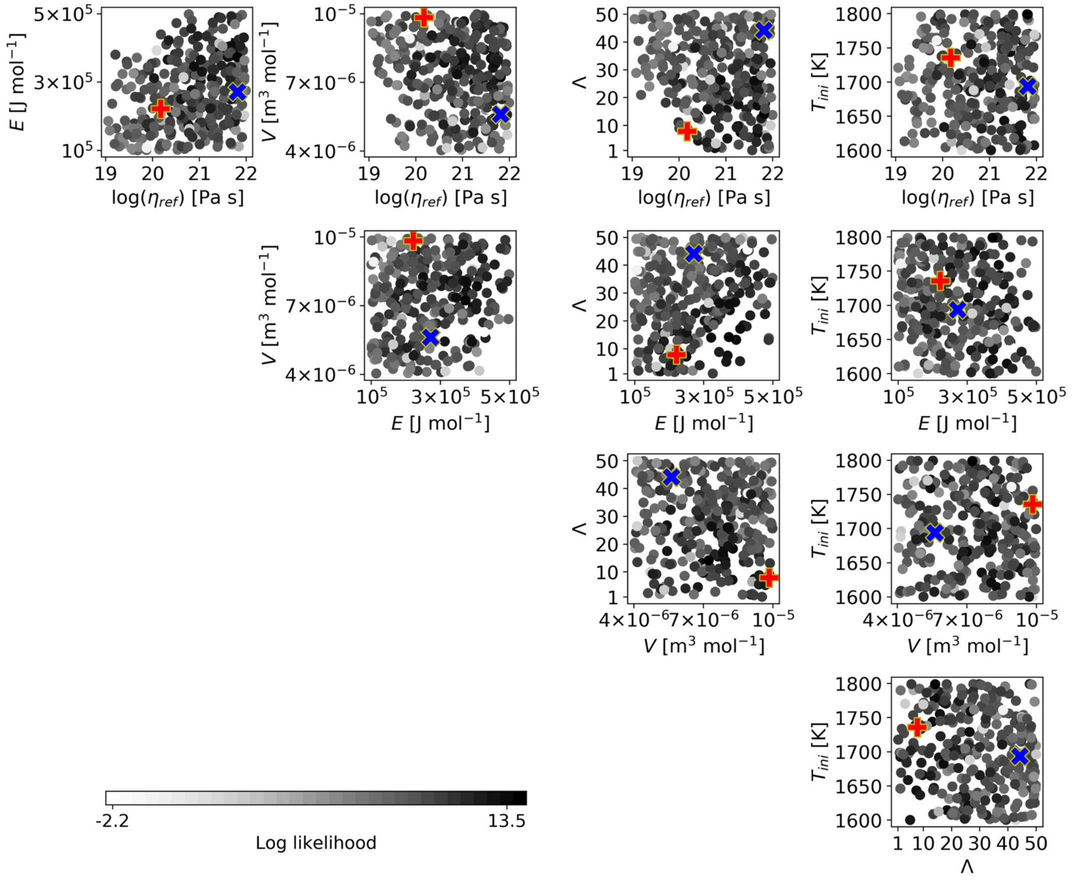


Figure 13. Log-likelihood values on the entire test set, plotted with respect to two parameters at a time. The darker a dot, the higher the log-likelihood value. The predicted probability distributions for example 1 (indicated by a blue cross) and example 2 (indicated by a red plus) are plotted in Figure 14.

We further explore a couple of examples from the test set (indicated by blue and red dots in Figure 13 and plot the joint probability distributions on a $\log_{10}$ scale. We also shade the background by the prior distribution of the training data obtained using a 3-mixture Gaussian Mixture Model for comparison. Since, visualizing a contour plot in more than 2 dimensions is challenging, we plot 2D slices of the joint probability distribution for each combination of parameters in Figure 14. The 2D contours visualized are essentially slices taken from a higher-dimensional space at the predicted mean of each parameter (or each dimension) not represented in that panel. Also, while we do plot the probabilities down to machine-precision, we discarded any mixture with a value below 10^{-8}. This is because Gaussian mixtures below extremely low values can be dominated by the variances and cause visual artifacts in the plots. Fortunately, in both examples, this occurred for only 1 out of the 3 mixtures.

Visually inspecting Figure 14, the predicted probability distributions capture the actual value (marked by a blue cross or red plus sign). It is also possible to examine the covariances among different parameters. For parameters that are difficult to constrain such as E or worse yet, V, it is apparent that the probability distributions are "stretched" along these parameters. In other words, a wide range of these parameters can satisfy an observation. For better-constrained parameters such as η_{ref} and Λ, for example, the probability distributions are smaller "discs". Furthermore, these sometimes demonstrate positive or negative correlations.

Hence, such high-dimensional probability distributions can capture the inter-parameter correlations and the associated degeneracy, providing a more comprehensive picture of all the evolution scenarios that fit given observational constraints.

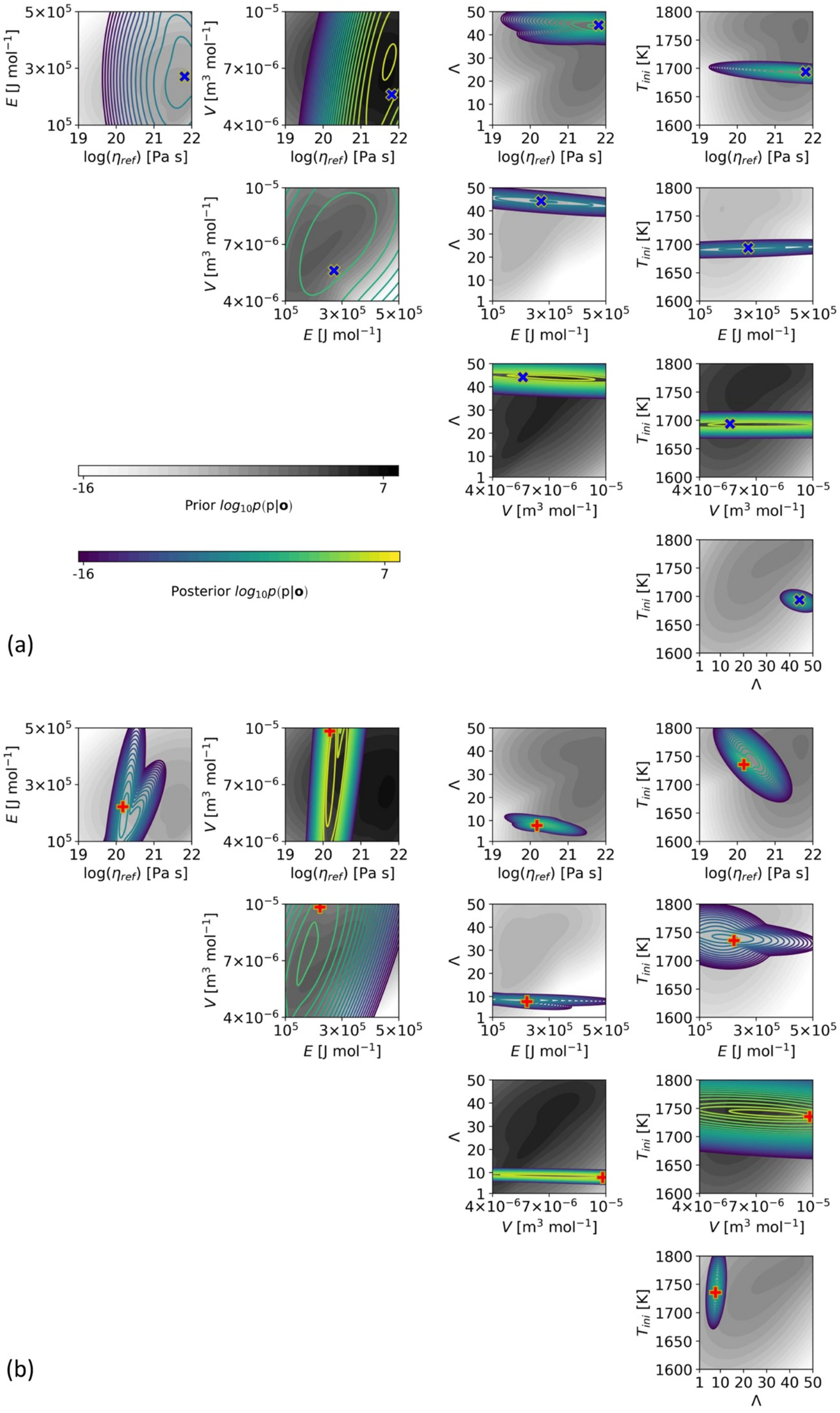


Figure 14. Examples of high-dimensional probability distributions from the test set. The actual values of the parameters are indicated by a blue cross (a) and a red plus sign (b). The background is shaded by the prior log-likelihood, whereas, the contour plot of posterior log-likelihood is given by the viridis colormap.

5. Toward Using Real Data from Mars as Observables

The steps to be taken to invert actual, non-synthetic observables are not straightforward. On the one hand, some observables, such as the amount of accumulated contraction recorded by compressive geological features (Knapmeyer et al., 2006; Mueller & Golombek, 2004; Nahm & Schultz, 2011) or the mantle potential

Table 2
Dimensional Noise Levels Calculated for all Synthetic Observables

Observable	$0.01\sigma_o$	$0.1\sigma_o$	$0.3\sigma_o$	$0.5\sigma_o$	$0.8\sigma_o$	$1.0\sigma_o$
Q_c (mW m^{-2})	0.01	0.1	0.4	0.6	1.0	1.3
Q_s (mW m^{-2})	0.02	0.2	0.7	1.1	1.8	2.2
R_{th} (km)	0.14	1.3	4.1	6.8	10.8	13.5
D_e (km)	0.5	5.4	16.3	27.2	43.4	54.3
D_{melt} (km)	1.0	10.0	29.0	48.3	77.3	96.6
t_{volc} (Gyr)	0.02	0.2	0.6	0.9	1.6	2.0
$T_{pt,1}$ (K)	1.0	10.0	29.5	49.1	78.6	98.3
$T_{pt,2}$ (K)	1.1	10.7	32.2	53.7	86.0	107.0

temperature at a certain time in the past inferred from petrological analyses of meteorites (Filiberto & Dasgupta, 2015), have global character. These can be inverted and interpreted even in terms of 1D parameterized convection models (e.g., Grott & Breuer, 2010; Morschhauser et al., 2011; Thiriet et al., 2018) or using suitably averaged 2D models such as those adopted in our study. On the other hand, several observables useful to constrain the thermal history of Mars are localized in space and time. Particularly relevant are (i) the thickness of the elastic lithosphere associated with the loading of surface features (e.g., Broquet et al., 2020; McGovern et al., 2002; Phillips et al., 2008); (ii) the surface heat flux, which can be inferred from the latter (unfortunately, the efforts of the HP3 experiment to obtain a measurement of the heat flux at the landing site of the NASA mission InSight (Spohn et al., 2018) have not been successful); (iii) the local thickness of the crust, which can be obtained indirectly from gravity and topography data (e.g., Goossens et al., 2017; Wieczorek & Zuber, 2004), but which could be seismically detected in the future also by the InSight mission (Banerdt et al., 2020), possibly along with additional seismic discontinuities bearing information on the interior temperature; (iv) indications of past volcanic activity at specific locations (e.g., Hauber et al., 2011; Werner, 2009). In order to invert these local observations, the MDNs would need to be trained on 3D data generated with simulations that are well representative of the Martian interior. The 3D models of Plesa et al. (2018) would provide a suitable starting point, although creating a 3D data set of comparable size to the 2D one used in this study, also spanning a similar range of parameters, is a significant computational challenge. Yet, it would provide the opportunity to place tighter constraints on key model parameters. Observations of the same quantity (say the elastic lithosphere) at multiple locations on a planet could also help one implicitly capture more information about the convection structures underneath.

Such an inverse study with 3D simulations can theoretically be performed in a similar fashion to this study, although, certain adaptations would be necessary. It is well known that incorporating prior knowledge of the data into the neural network architecture generically leads to both a more accurate and faster training process as well as to better generalization (Mitchell, 1980). For 3D simulations specifically, it is therefore desirable to choose a convolutional architecture (such as a convolutional neural network) to maintain the spatial correlations in quantities predicted by the simulations. Recurrent neural networks (or alternatively masked convolutional neural networks) are suitable candidates to model the time evolution of the simulation. We refer to Goodfellow et al. (2016) for an accessible overview of these architectures.

With the availability of a suitable 3D simulations data set, one will also need to consider the noise associated with measured observables. For each synthetic observable that we considered, Table 2 shows the dimensional values of the corresponding noise levels. Temperature estimates with uncertainties between ±50 and ±100 K, corresponding to noise levels between ~0.5 and $1\sigma_0$, can be considered realistic not only for the Earth (e.g., Boehler, 1996; Katsura et al., 2010), but also for Mars (Filiberto & Dasgupta, 2015). Similar to the temperature, determining the thickness of the elastic lithosphere to within $0.5-1\sigma_0$ at specific locations and the accumulated radial contraction appears to be possible (Grott & Breuer, 2010; Nahm & Schultz, 2011). However, measuring, for example, the surface heat flux with an accuracy of less than a couple of mW/m^2 (i.e., to within $1\sigma_0$) would not be within reach of the HP3 experiment of the InSight mission even if that were successful (the expected uncertainty of the experiment was in fact ±5 mW/m^2 (Spohn et al., 2018)).

While we cannot invert real observables from Mars using our current 2D data set, we provide, here, an estimate of the constraints on all parameters with the assumption of knowing only four observables. Based on the above considerations, we restrict this analysis to the two temperature points, radial contraction and elastic lithosphere thickness. Specifically, for these quantities, we consider a conservatively realistic noise level of $1.0\sigma_o$, as well as noise levels of $0.1\sigma_o$ and $0.5\sigma_o$ for comparison. Figure 15 shows the individual probability distributions along with the average standard deviation for each estimated parameter. For an uncertainty of $1.0\sigma_o$, η_{ref} can approximately be constrained to within 52%, E within 75%, Λ within 54%. V cannot be constrained, and, as expected in light of results in Figure 11, the constraints on T_{ini} are also lost at $1.0\sigma_o$.

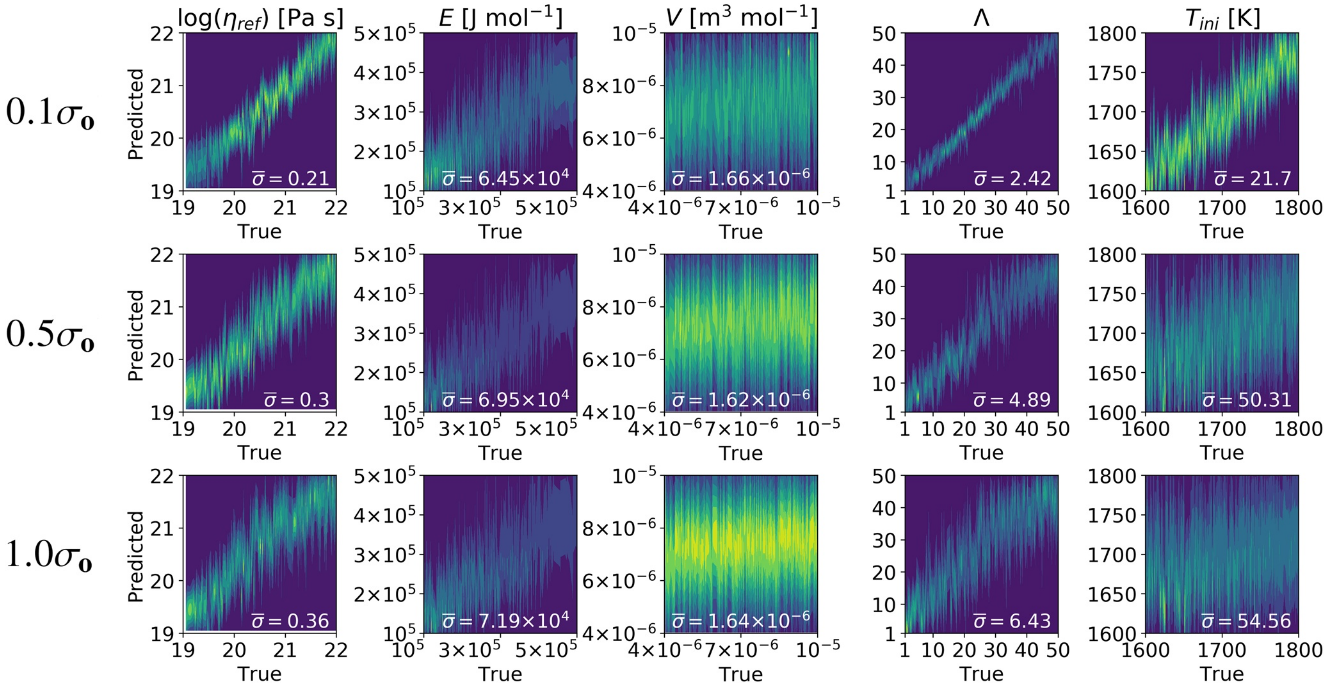


Figure 15. The individual probability distributions from the test-set obtained by training an MDN jointly on all parameters, given four observables: thermally induced radial contraction (R_{th}), elastic lithospheric thickness (D_{e}), and two temperature points at reference depths for the phase transitions ($T_{\mathrm{pt,\,1}}$ and $T_{\mathrm{pt,\,2}}$). Three different noise levels are tested: $0.0\sigma_{\mathrm{o}}$, $0.5\sigma_{\mathrm{o}}$, $1.0\sigma_{\mathrm{o}}$, with the last one being conservatively realistic.

6. Summary and Conclusion

We used MDNs to study the constraints on parameters governing the thermal evolution of Mars. To train the MDNs, we used 6,130 simulations that ran over 4.5 Gyr from a data set of 10,040 evolution simulations with Mars-like parameters, run on a 2D quarter cylindrical grid. We used the MDNs to test different synthetic observables (surface and CMB heat fluxes, elastic lithospheric thickness, radial contraction, duration of volcanism and amount of melt produced) and combinations thereof, to determine how well each of the five parameters can be constrained, as quantified by the log-likelihood (see Figure 7). The reference viscosity (η_{ref}) and crustal enrichment factor (Λ) are well constrained, as can be seen in Figure 8. Initial mantle temperature (T_{ini}) can be constrained if radial contraction is available along with at least some portion of the temperature profile. However, the activation energy of diffusion creep (E) can only be weakly constrained. Activation volume (V) cannot be inferred using any of the observables for the current setup.

We also searched different parts of the temperature profile for observational signatures (see Figure 9). We found that the top 20% of the temperature profile is sufficient for constraining Λ, while for E and η_{ref}, the availability of large portions of the temperature profile is advantageous. Furthermore, successively adding parts of the temperature profile to its peripheral components such as CMB temperature, CMB heat flux and surface heat flux, allows tightening the constraints on η_{ref}, E and Λ. This suggests that using single pressure-temperature point(s) as observables (such as potential temperature) could also help the inversion of mantle convection parameters. This was further evidenced by the tight constraints on initial temperature obtained with the use of the two temperature points at phase-transition depths as observables instead of the complete present-day temperature profile when searching for the joint probability model.

We emulated uncertainty in measuring an observable by adding Gaussian noise to the observables for the case when all observables are available and by training the networks on the noisy data to test how it impacts the constraint on a parameter (see Figure 11). The added noise was quantified by a factor multiplied with the variance of the distribution of the observable (σ_{o}). We found that, in general, small uncertainties (up to $0.01\sigma_{\mathrm{o}}$) in the distribution of the observable are inconsequential to the inference of a parameter (as they can have a regularizing effect on the training of MDNs). With larger uncertainties (say $1\sigma_{\mathrm{o}}$), however, the constraints become weaker. Different parameters also have different sensitivities to the uncertainty in the observations, with the constraint on T_{ini} being the one that is lost most rapidly as uncertainties increase.

We also tested two additional factors that can impact the constraints on all parameters, namely the availability of simulations and availability of observables. We observed that both increasing the number of simulations and increasing the number of observables strengthen the constraints (see Figure 12). The results indicate that already from a few hundred simulations, meaningful results can be obtained with five unknown parameters, which is encouraging in view of extending this approach to 3D simulations. We noted, however, an important caveat: the total number of simulations in the data set does not provide granular insights into the poorly sampled sub-spaces of the data set, such as low values of η_{ref} and high values of E. The inference of parameters in these ranges is limited due to a relative scarcity of data. A generation of simulations that is distributed more uniformly would be desirable in future studies.

Finally, we demonstrated how the joint probability model of all parameters can be obtained from observables and discussed the information gained from this approach: namely the covariances among different parameters

This work for a Mars-like planet builds upon previous MDN studies of ill-conditioned inverse problems such as by de Wit et al. (2013) and Atkins et al. (2016). Several research questions remain open. For example, how will varying more parameters such as number of phase transitions, bounding stress of the elastic lithosphere and the size of the core impact the constraints on thermal evolution? In this study, we varied 5 parameters, but Atkins et al. (2016) varied 59 different parameters in their study of Earth's thermal evolution. That is a plausible explanation for why in Atkins et al. (2016), most parameters such as the reference viscosity were difficult to constrain even when using synthetic observations at times earlier than present day (≤3 Gyr) and despite using reduced representations of the 2D temperature fields. Having more unknowns increases the number of possible combinations of parameters that need to be sampled from. This can also increase the degeneracy of the problem, since more combinations of parameters can lead to the same observation.

This highlights the need for future studies on mantle convection searching for observational signatures in (1) a higher-dimensional space, in (2) fields other than just temperature (such as velocity or density) and in (3) time. (1) We also searched the 1D temperature profiles for observational signatures. However, laterally averaging the 2D temperature fields leads to a loss of information, such as convection structures (plumes and downwelling) that could potentially help constrain parameters like V, since V strongly affects the spatial wavelength of convection. (2) We also limited our focus to temperature profiles and related observables. However, one could just as easily explore other fields and related quantities such as seismic velocities and gravity field. (3) Finally, we only inverted present-day observables. However, one could also include observations at different times, such as petrological constraints on potential mantle temperature and volcanic activity over the planet's history.

Furthermore, machine learning studies like the present one and the one by Atkins et al. (2016) are limited to the datasets used. To make the findings more widely applicable, one could create a more versatile data set that encompasses a broader number of parameters such as number and location of phase transitions and ratio of the radii of the core and the mantle. Furthermore, one could also include variable physical processes such as melting and mode of convection (Shahnas & Pysklywec, 2020).

MDNs and their corresponding log-likelihood function (Equation 13) provide a high-dimensional framework for evaluating how well different observables and combinations of observables can constrain certain parameters. However, the high dimensionality of the problem also leads to a computational challenge. In this study, we performed selected computations. For example, in Figure 11, we only evaluated the case where we had the full data set of 6,130 simulations from which to train and test all the present-day observables available. We also trained the MDN only five times per combination. If one wanted to reduce the variance of the negative likelihood, one could run it, say, 10 times. So, ideally, we would have performed 127 combinations of observables times 5 parameters times 8 sizes of datasets times 9 levels of noise times 10 repetitions times 35 time-steps per simulation. These roughly 16 million combinations would take over 1,000 years to train on a single GPU. The same challenge stands if one wanted to explore all the number and combinations of different temperature-pressure points in the 1D temperature fields, or worse yet, of 2D or 3D fields.

In addition to, or perhaps, instead of doing such machine learning computations on supercomputers, one can also consider more elegant approaches. For example, instead of training the MDN at a given time-step (like 4.5 Gyr in this study), one could just treat time as another variable in the observables vector and train at all time-steps at once. As for identifying the relevant portions of high-dimensional fields used by the model to predict, an interesting avenue of research could be to consider recent methods for explaining neural networks, for example, layer-wise relevance propagation (Bach et al., 2015; Montavon et al., 2019), and extend them to MDNs. The application of machine learning methods such as MDNs in the field of geodynamics has the potential to greatly enhance our understanding of complex ill-posed inverse problems such as the inference of the parameters governing mantle convection.

Data Availability Statement

The data set used to train, test and evaluate the MDNs is available on Github (https://github.com/agsiddhant/Inverse_Modelling_Mars_1D). The repository also contains all the codes used to train the MDNs. All figures in this paper can be reproduced using the codes and the python dictionaries provided.

Acknowledgments

The authors thank the three anonymous reviewers whose comments helped significantly improve previous versions of this work. The authors acknowledge the support of the Helmholtz Einstein International Berlin Research School in Data Science (HEIBRiDS) and the Helmholtz Association (project VH-NG-1017). We also acknowledge the North-German Supercomputing Alliance (HLRN) for providing HPC resources (project id: bep00087). This work was also funded by the German Ministry for Education and Research as BIFOLD - Berlin Institute for the Foundations of Learning and Data (ref. 01IS18025A and ref 01IS18037A). S. Padovan. acknowledges the support of DFG Research Unit FOR 2440 “Matter under planetary interior conditions”. We list the author contributions following the taxonomy by Brand et al. (2015).

Open Access funding is enabled and organized by Projekt DEAL.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). *TensorFlow: Large-scale machine learning on heterogeneous systems*. Retrieved from https://arxiv.org/abs/1603.04467

Agarwal, S., Tosi, N., Breuer, D., Padovan, S., Kessel, P., & Montavon, G. (2020). A machine-learning-based surrogate model of Mars' thermal evolution. *Geophysical Journal International*. *222*, 1656–1670. https://doi.org/10.1093/gji/ggaa234

Atkins, S., Valentine, A. P., Tackley, P. J., & Trampert, J. (2016). Using pattern recognition to infer parameters governing mantle convection. *Physics of the Earth and Planetary Interiors*, *257*, 171–186. https://doi.org/10.1016/j.pepi.2016.05.016

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, *10*(7), e0130140. https://doi.org/10.1371/journal.pone.0130140

Banerdt, W. B., Smrekar, S. E., Banfield, D., Giardini, D., Golombek, M., Johnson, C. L., et al. (2020). Initial results from the InSight mission on Mars. *Nature Geoscience*, *13*(3), 183–189. https://doi.org/10.1038/s41561-020-0544-y

Baumann, T. (2016). Appraisal of geodynamic inversion results: A data mining approach. *Geophysical Journal International*, *207*(2), 667–679. https://doi.org/10.1093/gji/ggw279

Baumeister, P., Padovan, S., Tosi, N., Montavon, G., Nettelmann, N., MacKenzie, J., & Godolt, M. (2020). Machine-learning inference of the interior structure of low-mass exoplanets. *The Astrophysical Journal*, *889*(42). https://doi.org/10.3847/1538-4357/ab5d32

Bishop, C. (1994). *Mixture density networks. Tech. Rep. NCRG/94/004*. Birmingham. Aston University. Retrieved from http://www.ncrg.aston.ac.uk/

Boehler, R. (1996). Melting temperature of the earth's mantle and core: Earth's thermal structure. *Annual Review of Earth and Planetary Sciences*, *24*(1), 15–40. https://doi.org/10.1146/annurev.earth.24.1.15

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: Attribution, contribution, collaboration, and credit. *Learned Publishing*, *28*(2), 151–155. https://doi.org/10.1087/20150211

Breuer, D., & Moore, W. (2015). Dynamics and thermal history of the terrestrial planets, the moon, and io. In G. Schubert (Ed.), Treatise on geophysics (2nd ed.), 10 (pp. 255–305). Oxford: Elsevier. https://doi.org/10.1016/B978-0-444-53802-4.00173-1

Broquet, A., Wieczorek, M. A., & Fa, W. (2020). Flexure of the lithosphere beneath the north polar cap of mars: Implications for ice composition and heat flow. *Geophysical Research Letters*, *47*(5), e2019GL086746. https://doi.org/10.1029/2019GL086746

Chollet, F., et al. (2015). *Keras*. Retrieved from https://keras.io

Christensen, U., & Yuen, D. A. (1985). Layered convection induced by phase transitions. *Journal of Geophysical Research*, *90*(B12), 10291–10300. https://doi.org/10.1029/JB090iB12p10291

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. *Geophysical Journal International*, *195*(1), 408–422. https://doi.org/10.1093/gji/ggt220

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., et al. (2017). *Tensorflow distributions*. Retrieved from https://arxiv.org/abs/1711.10604

Filiberto, J., & Dasgupta, R. (2015). Constraints on the depth and thermal vigor of melting in the Martian mantle. *Journal of Geophysical Research: Planets*, *120*(1), 109–122. https://doi.org/10.1002/2014JE004745

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. Retrieved from http://www.deeplearningbook.org

Goossens, S., Sabaka, T., Genova, A., Mazarico, E., Nicholas, J., & Neumann, G. (2017). Evidence for a low bulk crustal density for mars from gravity and topography. *Geophysical Research Letters*, *44*(15), 7686–7694. https://doi.org/10.1002/2017GL074172

Grott, M., & Breuer, D. (2010). On the spatial variability of the martian elastic lithosphere thickness: Evidence for mantle plumes? *Journal of Geophysical Research*, *115*(E3). https://doi.org/10.1029/2009JE003456

Grott, M., Breuer, D., & Laneuville, M. (2011). Thermo-chemical evolution and global contraction of mercury. *Earth and Planetary Science Letters*, *307*(1), 135–146. https://doi.org/10.1016/j.epsl.2011.04.040

Hüttig, C., Tosi, N., & Moore, W. (2013, 07). An improved formulation of the incompressible navier-stokes equations with variable viscosity. *Physics of the Earth and Planetary Interiors*, *220*, 11–18. https://doi.org/10.1016/j.pepi.2013.04.002

Hauber, E., Brož, P., Jagert, F., Jodłowski, P., & Platz, T. (2011). Very recent and wide-spread basaltic volcanism on mars. *Geophysical Research Letters*, *38*(10). https://doi.org/10.1029/2011GL047310

Hirth, G., & Kohlstedt, D. (2003, 01). Rheology of the upper mantle and the mantle wedge: A view from the experimentalists. *AGU Monograph Series*, *138*, 83–105. https://doi.org/10.1029/138GM06

Hjorth, L. U., & Nabney, I. T. (1999, Sep.). Regularization of mixture density networks. In *1999 Ninth International Conference on Artificial Neural Networks ICANN 99*. (*2*, p. 521–526 vol.2). https://doi.org/10.1049/cp:19991162

Katsura, T., Yoneda, A., Yamazaki, D., Yoshino, T., & Ito, E. (2010). Adiabatic temperature profile in the mantle. *Physics of the Earth and Planetary Interiors*, *183*(1–2), 212–218. https://doi.org/10.1016/j.pepi.2010.07.001

Käufl, P., Valentine, A., de Wit, R., & Trampert, J. (2016). Solving probabilistic inverse problems rapidly with prior samples. *Geophysical Journal International*, *205*(3), 1710–1728. https://doi.org/10.1093/gji/ggw108

King, S. D., Lee, C., van Keken, P. E., Leng, W., Zhong, S., Tan, E., et al. (2010). A community benchmark for 2-D Cartesian compressible convection in the Earth's mantle. *Geophysical Journal International*, *180*(1), 73–87. https://doi.org/10.1111/j.1365-246X.2009.04413.x

Knapmeyer, M., Oberst, J., Hauber, E., Wählisch, M., Deuchler, C., & Wagner, R. (2006). Working models for spatial distribution and level of mars' seismicity. *Journal of Geophysical Research*, *111*(E11). https://doi.org/10.1029/2006JE002708

Kronbichler, M., Heister, T., & Bangerth, W. (2012). High accuracy mantle convection simulation through modern numerical methods. *Geophysical Journal International*, *191*, 12–29. https://doi.org/10.1111/j.1365-246X.2012.05609.x

Kruse, J. (2020). *Training mixture density networks with full covariance matrices*. Retrieved from https://arxiv.org/abs/2003.05739

Magali, J. K., Bodin, T., Hedjazian, N., Samuel, H., & Atkins, S. (2020). Geodynamic tomography: Constraining upper-mantle deformation patterns from Bayesian inversion of surface waves. *Geophysical Journal International*, *224*(3), 2077–2099. https://doi.org/10.1093/gji/ggaa577

Martin, C. (2018). *Keras mixture density network layer*. GitHub. Retrieved from https://github.com/cpmpercussion/keras-mdn-layer

McGovern, P. J., Solomon, S. C., Smith, D. E., Zuber, M. T., Simons, M., Wieczorek, M. A., et al. (2002). Localized gravity/topography admittance and correlation spectra on mars: Implications for regional and global evolution. *Journal of Geophysical Research*, *107*(E12). https://doi.org/10.1029/2002JE001854

Mclachlan, G., & Basford, K. (1988). Mixture Models: Inference and Applications to Clustering (38). https://doi.org/10.2307/2348072

McNutt, M. K. (1984). Lithospheric flexure and thermal anomalies. *Journal of Geophysical Research*, *89*(B13), 11180–11194. https://doi.org/10.1029/JB089iB13p11180

Meier, U., Curtis, A., & Trampert, J. (2007). Global crustal thickness from neural network inversion of surface wave data. *Geophysical Journal International*, *169*(2), 706–722. https://doi.org/10.1111/j.1365-246X.2007.03373.x

Mitchell, T. M. (1980). *The need for biases in learning generalizations*. New Brunswick, NJ: Rutgers University. Retrieved from http://www-cgi.cs.cmu.edu/~tom/pubs/NeedForBias_1980.pdf

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-wise relevance propagation: An overview. In W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, & K.-R. Müller (Eds.) *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 193–209). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6_10

Morra, G., Yuen, D. A., Tufo, H. M., & Knepley, M. G. (2020). Fresh outlook in numerical methods for geodynamics – Part 2: Big data, HPC, education. In D. Alderton, & S. A. Elias (Eds.), *Encyclopedia of Geology* (pp. 841–855). Academic Press. https://doi.org/10.1016/B978-0-08-102908-4.00110-7

Morschhauser, A., Grott, M., & Breuer, D. (2011). Crustal recycling, mantle dehydration, and the thermal evolution of Mars. *Icarus*, *212*(2), 541–558. https://doi.org/10.1016/j.icarus.2010.12.028

Mueller, K., & Golombek, M. (2004). Compressional structures on mars. *Annual Review of Earth and Planetary Sciences*, *32*, 435–464. https://doi.org/10.1146/annurev.earth.32.101802.120553

Nahm, A. L., & Schultz, R. A. (2011). Magnitude of global contraction on mars from analysis of surface faults: Implications for martian thermal history. *Icarus*, *211*(1), 389–400. https://doi.org/10.1016/j.icarus.2010.11.003

Nimmo, F., & Tanaka, K. (2005). Early crustal evolution of mars. *Annual Review of Earth and Planetary Sciences*, *33*, 133–161. https://doi.org/10.1146/annurev.earth.33.092203.122637

Padovan, S., Tosi, N., Plesa, A.-C., & Ruedas, T. (2017). Impact-induced changes in source depth and volume of magmatism on mercury and their observational signatures. *Nature Communications*, 8. https://doi.org/10.1038/s41467-017-01692-0

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.Retrieved from http://scikit-learn.sourceforge.net/

Phillips, R. J., Zuber, M. T., Smrekar, S. E., Mellon, M. T., Head, J. W., & Tanaka, K. L., et al. (2008). Mars north polar deposits: Stratigraphy, age, and geodynamical response. *Science*, *320*(5880), 1182–1185. https://doi.org/10.1126/science.1157546

Plesa, A.-C., Padovan, S., Tosi, N., Breuer, D., Grott, M., Wieczorek, M. A., et al. (2018). The thermal state and interior structure of mars. *Geophysical Research Letters*, *45*(22). 12198–12209. https://doi.org/10.1029/2018GL080728

Plesa, A.-C., Tosi, N., Grott, M., & Breuer, D. (2015). Thermal evolution and urey ratio of mars. *Journal of Geophysical Research: Planets*, *120*(5), 995–1010. https://doi.org/10.1002/2014JE004748

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, & J. L. Mcclelland (Eds.) *Parallel distributed processing* (Vol. 1 pp. 318–362). MIT Press. Retrieved from https://web.stanford.edu/class/psych209a/ReadingsByDate/02_06/PDPVolIChapter8.pdf

Sambridge, M. (1999a). Geophysical inversion with a neighborhood Algorithm-I. Searching a parameter space. *Geophysical Journal International*, *138*(2), 479–494. https://doi.org/10.1046/j.1365-246X.1999.00876.x

Sambridge, M. (1999b). Geophysical inversion with a neighborhood algorithm-II. Appraising the ensemble. *Geophysical Journal International*, *138*(3), 727–746. https://doi.org/10.1046/j.1365-246x.1999.00900.x

Sambridge, M., & Mosegaard, K. (2002). Monte carlo methods in geophysical inverse problems. *Reviews of Geophysics*, *40*(3). https://doi.org/10.1029/2000RG000089

Shahnas, M. H., & Pysklywec, R. N. (2020). Toward a unified model for the thermal state of the planetary mantle: Estimations from mean field deep learning. *Earth and Space Science*. e2019EA000881. https://doi.org/10.1029/2019EA000881

Shahnas, M. H., Yuen, D. A., & Pysklywec, R. N. (2018). Inverse problems in geodynamics using machine learning algorithms. *Journal of Geophysical Research: Solid Earth*, *123*(1), 296–310. https://doi.org/10.1002/2017JB014846

Spohn, T., Grott, M., Smrekar, S., Knollenberg, J., Hudson, T., & Krause, C., et al. (2018). The heat flow and physical properties package (hp^3) for the insight mission. *Space Science Reviews*, *214*(5), 96. https://doi.org/10.1007/s11214-018-0531-4

Stevenson, D. J., Spohn, T., & Schubert, G. (1983). Magnetism and thermal evolution of the terrestrial planets. *Icarus*, *54*, 466–489. https://doi.org/10.1016/0019-1035(83)90241-5

Tackley, P. J. (2008). Modeling 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*(1–4), 7–18. https://doi.org/10.1016/j.pepi.2008.08.005

Thiriet, M., Michaut, C., Breuer, D., & Plesa, A.-C. (2018). Hemispheric dichotomy in lithosphere thickness on mars caused by differences in crustal structure and composition. *Journal of Geophysical Research: Planets*, *123*(4), 823–848. https://doi.org/10.1002/2017JE005431

Tosi, N., Grott, M., Plesa, A. C., & Breuer, D. (2013b). Thermochemical evolution of Mercury's interior. *Journal of Geophysical Research*, *118*(12), 2474–2487. https://doi.org/10.1002/jgre.20049

Tosi, N., & Padovan, S. (2020). Mercury, Moon, Mars: Surface expressions of mantle convection and interior evolution on stagnant-lid bodies. In H. Marquardt, M. Ballmer, S. Cottar, & J. Konter (Eds.) *Mantle convection and surface expressions*. AGU Monograph Series. https://doi.org/10.1002/9781119528609.ch17

Tosi, N., Yuen, A. D., de Koker, N., & Wentzcovitch, M. R. (2013a). Mantle dynamics with pressure- and temperature-dependent thermal expansivity and conductivity. *Physics of the Earth and Planetary Interiors*, *217*, 48–58. https://doi.org/10.1016/j.pepi.2013.02.004

Tozer, D. (1967). Toward a theory of thermal convection in the mantle. In T. Gaskell (Ed.), *The earth's mantle* (pp. 327–353). New York: Academic Press.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transactions on Neural Networks*, *11*(3), 586–600. https://doi.org/10.1109/72.846731

Wänke, H., & Dreibus, G. (1994). Chemistry and accretion history of mars. *Philosophical Transactions of the Royal Society of London, Series A*, *349*(1690), 285–293. https://doi.org/10.1098/rsta.1994.0132

Werbos, P. J. (1982). Applications of advances in nonlinear sensitivity analysis. In R. F. Drenick, & F. Kozin (Eds.) *System modeling and optimization* (pp. 762–770). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/BFb0006203

Werner, S. C. (2009). The global Martian volcanic evolutionary history. *Icarus*, *201*(1), 44–68. https://doi.org/10.1016/j.icarus.2008.12.019

Wieczorek, M., & Zuber, M. (2004). Thickness of the Martian crust: Improved constraints from geoid-to-topography ratios. *Journal of Geophysical Research*, *109*(E1). https://doi.org/10.1029/2003JE002153

Zhong, S., McNamara, A., Tan, E., Moresi, L., & Gurnis, M. (2008). A benchmark study on mantle convection in a 3-D spherical shell using Citcoms. *Geochemistry, Geophysics, Geosystems*, *9*(10). https://doi.org/10.1029/2008GC002048