Applying Machine Learning to Characterize and Extrapolate the Relationship Between Seismic Structure and Surface Heat Flow

Shane Zhang*1,2 and Michael H. Ritzwoller1

1 Department of Physics, University of Colorado Boulder, Boulder, CO 80309, USA.
2 Now at California Institute of Technology, Pasadena, CA 91125, USA.

Abstract

Geothermal heat flow beneath the Greenland and Antarctic ice sheets is an important boundary condition for ice sheet dynamics, but is rarely measured directly and therefore is inferred indirectly from proxies (e.g. seismic structure, magnetic Curie depth, surface topography). We seek to improve the understanding of the relationship between heat flow and one such proxy—seismic structure—and determine how well heat flow data can be predicted from the structure (the characterization problem). We also seek to quantify the extent to which this relationship can be extrapolated from one continent to another (the transportability problem). To address these problems, we use direct heat flow observations and new seismic structural information in the contiguous US and Europe, and construct three Machine Learning models of the relationship with different levels of complexity (Linear Regression, Decision Tree, Random Forest). We compare these models in terms of their interpretability, the predicted heat flow accuracy within a continent, and the accuracy of the extrapolation between Europe and the US. The Random Forest and Decision Tree models are the most accurate within a continent, while the Linear Regression and Decision Tree models are the most accurate upon extrapolation between continents. The Decision Tree model uniquely illuminates the regional variations of the relationship between heat flow and seismic structure. From the Decision Tree model, uppermost mantle shear wavespeed, crustal shear wavespeed and Moho depth together explain more than half of the observed heat flow variations in both the US ($r^2 \approx 0.6$, coefficient of determination), RMSE $\approx 8$ mW/m² (Root Mean Squared Error)) and Europe ($r^2 \approx 0.5$, RMSE $\approx 13$ mW/m²), such that uppermost mantle shear wavespeed is the most important. Extrapolating the US-trained models to Europe reasonably predicts the geographical distribution of heat flow ($\rho = 0.48$ (correlation coefficient)), but not the absolute amplitude of the variations ($r^2 = 0.17$), similarly from Europe to the US ($\rho = 0.66, r^2 = 0.24$). The deterioration of accuracy upon extrapolation is caused by differences between the continents in how seismic structure is imaged, the heat flow data, and intrinsic crustal radiogenic heat production. Our methods have the potential to improve the reliability and resolution of heat flow inferences across Antarctica and the validation and cross-validation procedures we present can be applied to heat flow proxies other than seismic structure, which may

*sdz@caltech.edu
help resolve inconsistencies between existing subglacial heat flow values inferred using different proxies.

Key words: Heat flow; Seismic tomography; Machine Learning; Heat generation and transport; Glaciology; Antarctica

1 Introduction

Surface geothermal heat flow represents the rate of heat loss from the solid Earth. Heat flow observations provide clues to decipher the thermal evolution within the Earth (Jaupart and Mareschal 2015), can inform the exploration for geothermal energy (Blackwell et al. 2013), and provide information about the thermal state of cryospheric realms, such as permafrost (Lachenbruch and Marshall 1986) and continental ice sheets (Pollard et al. 2005; Llubes et al. 2006; Pattyn 2010; Rogozhina et al. 2012; Noble et al. 2020).

Heat flow, in particular, can influence the lower boundary conditions of ice sheets, notably the state of freezing to the subglacial bed and the melt extent, which can strongly affect ice flow. However, direct observations of heat flow beneath the Greenland and Antarctic ice sheets remain rare because of the difficulty and expense of drilling to the base of the ice sheet. Thus, heat flow beneath the Greenland and Antarctic ice sheets imposes a major uncertainty in ice sheet models that aim to predict future melting and sea level changes (Noble et al. 2020). For this reason, indirect inferences of heat flow beneath the Greenland (Shapiro and Ritzwoller 2004b; Greve 2005; Fox Maule et al. 2009; Petrunin et al. 2013; Rezvanbehbahani et al. 2017; Martos et al. 2018; Artemieva 2019; Rezvanbehbahani et al. 2019; Colgan et al. 2022) and Antarctic ice sheets (Shapiro and Ritzwoller 2004b; Fox Maule et al. 2005; Goutorbe et al. 2011; An et al. 2015; Martos et al. 2017; Burton-Johnson et al. 2017; Pollett et al. 2019; Shen et al. 2020; Guimarães et al. 2020; Dziadek et al. 2021; Lösing and Ebbing 2021; Stål et al. 2021; Artemieva 2022; Haeger et al. 2022) have attracted much recent interest.

In both Greenland (Colgan et al. 2022) and Antarctica (Burton-Johnson et al. 2020), inferences of heat flow are largely made through different sets of proxies. These inferences have not yet converged to produce a consistent model of heat flow beneath the major ice sheets. The proxies include seismic structure (Shapiro and Ritzwoller 2004b; Perry et al. 2006b; Shen et al. 2020), magnetic Curie depth (Fox Maule et al. 2005; Martos et al. 2017), surface topography (Artemieva 2022), and a combination of geological and geophysical variables (Goutorbe et al. 2011; Stål et al. 2021). More importantly, many studies that infer heat flow have not yet been subject to validation in which the predicted values are compared with ground truth or observations not included in making the prediction. Such validation is crucial to the evaluation of the extrapolation of inference methods from the region where they are developed (where heat flow measurements may be plentiful) to other regions (where heat flow observations are scarce or absent entirely, such as Greenland and Antarctica).

We investigate the ability to predict surface heat flow using seismic structure as a proxy. We use data from the contiguous US and Europe (section 3) because most of the world’s continental heat flow observations are obtained there (Lucazeau 2019) and high-resolution regional seismic structure has been imaged recently (Shen and Ritzwoller 2016; Lu et al.
We focus on seismic structure as a proxy because seismic tomography provides information at depth, with images of both the crust and mantle, and can thus attribute effects on heat flow to specific depths. Many other proxies cannot do this.

We pose two main questions in this study. First, what is the relationship between seismic structure and surface heat flow in a given region? Specifically, how well is heat flow predicted by seismic structure, which seismic structural variables are most predictive of heat flow, and how is the relationship specified? We refer to this problem as the characterization problem. Second, how well can the relationship between seismic structure and surface heat flow be extrapolated from one continental setting to another? In particular, what aspects of the relationship are similar or dissimilar between various continents and what causes the dissimilarity? We call this problem the transportability problem.

To address aspects of both the characterization and transportability problems, we develop two particular methodological approaches. Our first approach focuses on understanding the relationship between seismic structure and heat flow using Machine Learning models of differing complexity. In the characterization problem, we seek to specify how a specific prediction of heat flow is made from the model using three Machine Learning models: the Linear Regression model, the Decision Tree model, and the Random Forest model, which model the relationship across continental, regional, and local scales, respectively (section 4). This is also essential for the transportability problem, where we seek to understand how the relationship between seismic structure and heat flow differs between continents.

Our second approach holds out certain data to evaluate a specific aspect of a model (section 4.4). In the characterization problem, we are interested in how well heat flow can be predicted by seismic structure in a given region. Therefore, for both the US and Europe, we divide heat flow observations into two disjoint subsets of data: the training and validation datasets. In both continents, we use the training data to train the model, and then validate the trained model on the validation data (section 5.1). We refer to this procedure as validation. In the transportability problem, we are interested in how well the relationship can be extrapolated from one continent to another. Hence, we take models trained in the US and then apply them to Europe, and vice versa (section 5.2). We call this process as cross-validation. Both validation and cross-validation are important tests of the ability of that model to predict heat flow.

We summarize the methodology to tackle the characterization and transportability problems in Fig. 1. We begin by describing the data sources and quality control of heat flow data and seismic data used in this study (section 3). Second, we introduce the three Machine Learning models (Linear Regression, Decision Tree, and Random Forest) used to investigate the relationship between heat flow and seismic structure (section 4). We then present the validation results made in the contiguous US and Europe, and the cross-validation results by applying a model determined on one continent to data from the other continent (section 5). Finally, we discuss the limitations of this work and the potential applications to polar regions and other continents (section 6).
2 Notation

We use geothermal heat flow to refer to the amount of heat per unit time per unit area (J/(s m^2) or more conveniently mW/m^2) transported from the Earth’s interior to its solid surface. Heat flux is also commonly used to refer to the same quantity, but we do not use this term (Burton-Johnson et al. 2020).

The seismic structure is inferred from seismic inversions, the results of which are often referred to as seismic tomography models. Instead, we use seismic structure or structural information or structural variables to avoid confusion with the Machine Learning models. Therefore, here, a model specifies the relationship between seismic structure and heat flow.

We use a hat and a superscript “obs” to denote predicted and observed quantities, respectively. For example, \( \hat{q} \) and \( q^{\text{obs}} \) denote predicted and observed heat flow, respectively.

Here we use validation to mean the splitting of training and validation data on the same continent, which evaluates model accuracy. In contrast, we use cross-validation to denote extrapolating models trained from one continent to another, which evaluates model transportability. In the Machine Learning literature, the cross-validation procedure is often used to mean what we call validation; elsewhere the data may be split into training, validation, and test data, where the test data are used in a process somewhat similar to our cross-validation.

We often use MAD (Median Absolute Deviation) as a robust measure of the variability of any variable \( x_k, k = 1, 2, 3, \ldots \),

\[
\text{MAD}(x) \equiv \alpha \text{Median}(|x_k - \tilde{x}|),
\]

where \( \tilde{x} \equiv \text{Median}(x) \), \( \alpha \approx 1.5 \) such that the MAD of a Gaussian distribution is the same as its standard deviation. Therefore, MAD is similar to standard deviation but less prone to outliers.

3 Data

In this section, we first discuss heat flow observations, focusing on quality control (section 3.1). We then discuss seismic structural variables, aiming to extract those most predictive for heat flow (section 3.2).

3.1 Geothermal heat flow observations

3.1.1 Data Sources

Our primary source of heat flow observations is the New Global Heat Flow (NGHF) compilation (Lucazeau 2019), which is augmented in the US by the SMU Node of the National Geothermal Data System (Blackwell et al. 2013). NGHF is the most recent compilation of global heat flow measurements since Pollack et al. (1993). It contains about 40,000 continental measurements globally. Most heat flow observations in NGHF are from borehole measurements of heat conductivity and temperatures at multiple depths from the surface to a few hundred meters (eq. (2)). It is conventional to define geothermal heat flow, \( q \), as the
conductive flow in the upward direction near the Earth surface:

$$q \equiv k \left. \frac{\partial T}{\partial z} \right|_{z=0}, \quad (2)$$

where $k$ denotes heat conductivity, and $T$ temperature, both varying with depth $z$ in general.

For the contiguous US (Fig. 2a–c), the SMU data augments the NGHF in the Central and Eastern US, especially with bottom hole temperature measurements across parts of the Great Plains and the Midwestern US. The bottom hole temperature measurements can have both good and poor qualities; on the one hand, the measurements can extend to much greater depths ($>1500$ m) than typical geothermal holes (a few hundred meters); on the other hand, conductivity is seldom measured directly, and is often prescribed based on geological information of specific areas (Blackwell and Richards 2004). However, the data sources provide little information about data quality. These measurements are spatially dense enough to compare neighboring locations to one another, which turn out to vary spatially less than traditional geothermal measurements (Fig. 2b).

### 3.1.2 Quality control

Due to the absence of information about the quality of heat flow observations, we perform our own quality control which involves data rejection, imputation, and spatial smoothing. The observed heat flow $q^{\text{obs}}$ based on eq. (2) is the sum of the true heat flow $q_0$ and measurement errors $\epsilon$, and can be written as $q^{\text{obs}} = q_0 + \epsilon$. Moreover, $q_0$ can be approximated as the result of regional background heat flow, $q_R$ (with a spatial scale $R$) perturbed by small-scale variations $\delta q$, such that $q_0 = q_R + \delta q$. Perturbations $\delta q$ may come from paleoclimate, sedimentation and erosion, seismicity, and ground water circulation (Powell et al. 1988). Since our interest is $q_R$, the unwanted part is $\Delta q \equiv \delta q + \epsilon$. Therefore,

$$q^{\text{obs}} = q_R + \Delta q. \quad (3)$$

The resolution of seismic structure used in this study limits $R \sim 100$ km (Shen and Ritzwoller 2016; Lu et al. 2018; Shen et al. 2018), which requires removing small-scale variations in heat flow before investigating the relationship between seismic structure and heat flow. Unfortunately, most values do not have associated quantitative uncertainties $\epsilon$, nor are they corrected for near-surface effects that control $\delta q$. Thus, we apply the following criteria to down-weigh measurements that we believe to have large $\delta q$ or $\epsilon$.

First, we eliminate observations outside the range from 10 mW/m$^2$ to 150 mW/m$^2$. This is because the lower limit of mantle heat flow is typically larger than 10 mW/m$^2$ (Jaupart et al. 2016), while values higher than 150 mW/m$^2$ are commonly associated with geothermal reservoirs (Lachenbruch and Sass 1977). Second, we remove observations made at depths $<100$ m, because shallower depths are affected by climate oscillations with periods shorter than century scales (assuming thermal diffusivity $\kappa \approx 1$ mm$^2$/s$^2$). Third, we compute standard deviations, $s$, for a set of observations within a 10 km radius, and reject all of the observations within this set if $s > 20$ mW/m$^2$. This is because large variations with spatial scales shorter than 10 km are unlikely to be explained by seismic structure with resolution $\sim$100 km. The threshold value of $s$ is significantly larger than typical $\epsilon$. Specifically, among the 1549
continental measurements with $\epsilon$ provided by NGHF, the median of $\epsilon$ is 5 mW/m$^2$ and its MAD is 6 mW/m$^2$.

To remove small-scale variations and to alleviate the spatially uneven sampling of raw observations, we smooth observations twice. First, we average point observations within a 10 km distance on a 10-km grid. Then the median (Fig. 2ad) and MAD (Fig. 2be) are computed for these local means across a 100 km radius, producing a heat flow map on a 100-km grid.

In the US, heat flow observations from the Snake River Plain (Brott et al. 1981) and South Dakota (Gosnold 1990) are known to be dominated by hydrothermal circulations. Therefore, we replace heat flow observations with 90 mW/m$^2$ for the Snake River Plain (Brott et al. 1981), and 60 mW/m$^2$ for South Dakota (Gosnold 1990), which are more representative of regional heat flow values.

3.1.3 Spatial patterns of heat flow observations

The US heat flow shows a west-east dichotomy, which reflects fundamentally distinct tectonic regimes of the western and eastern US. This dichotomy is basic for any model to reproduce heat flow from seismic structure. Specifically, the western US has higher values and the eastern US has lower values, with intermediate values in the Great Plains (Fig. 2a). Regionally, high anomalies are observed in the Basin and Range, and the Rio Grande rift; while low anomalies are shown in the Cascadia subduction zone, the Colorado Plateau, and much of the eastern US except near the Appalachian Mountains. The central goal of our model is to reproduce these regional-scale variations.

We compute MAD values of heat flow observations within a 100 km radius. Large MAD values of heat flow mostly correspond to high values in the western US (Fig. 2b). Given this radius of 100 km is much larger than the natural variability of heat flow, MAD values measure not only observation uncertainty, but also spatial variability. In particular, large MAD values can be caused by averaging across values of contrast such as along the Cascade Range, while small MAD values across the central US may suggest either the high quality of the bottom hole temperature measurements there or the uniformity in prescribed heat conductivity.

European heat flow observations in NGHF are mostly based on a compilation by Cermak and Rybach (1979). Heat flow in Europe shows a northwest-southeast dichotomy, divided by the Trans-European Suture Zone (Fig. 2d). High heat flow values are observed in the Massif Central and the Upper Rhine Plain, east of the Dinarides and Hellenides, and in Anatolia, while low values are measured in the Baltic Shield, the Russian Platform, the Apennines, and the Dinarides and Hellenides. The North Germain Plain has intermediate heat flow.

The MAD values of heat flow are generally larger in Europe than the US, although the US data have more exceptionally large values (Fig. 2e). This means that the European heat flow observations have larger spatial variability than the US data on a 100 km scale. Large MAD values of European heat flow also correlate with high heat flow values, especially along the Alps.
3.2 Seismic structure

We focus on seismic structure as a proxy for heat flow for the following reasons. First, seismic structure is sensitive to temperature and composition as functions of depth which fundamentally determine heat flow (Shapiro and Ritzwoller 2004a). Second, a significant correlation has been observed between seismic structure and heat flow (Pollack et al. 1993; Röhm et al. 2000; Shen et al. 2020). Third, seismic observations provide one of the highest-resolution images of the solid Earth structure.

3.2.1 Sources of information

The US seismic structural information is from Shen and Ritzwoller (2016) (Fig. 5a–c), which represents the state of the art in continental seismic tomography. This information uses the seismic data from the EarthScope USArray, including Rayleigh wave dispersion and amplitudes (ellipticity), and body wave converted phases (receiver functions). In particular, Rayleigh waves are measured from both earthquakes and ambient noise, which complements earthquakes at higher frequencies to constrain crustal structure. At each target location, a seismic shear wavespeed profile $V_S$ is parameterized with three “layers” from the surface to 150 km depth: a sediment layer in which $V_S$ increases linearly with depth, a crustal layer where $V_S$ varies smoothly with depth, and a mantle layer where $V_S$ also changes smoothly with depth. Discontinuities can exist between the layers. Posterior distributions of profile parameters are presented, which are determined via a Bayesian approach with Markov Chain Monte Carlo sampling. The inferred seismic shear wavespeed is actually $V_{SV}$, but we refer to it as $V_S$ here.

The European seismic tomography result is from Lu et al. (2018, Fig. 5d–f), which is based on data mainly from the AlpArray. They adopt a similar methodology as Shen and Ritzwoller (2016) so we only discuss the main differences. First, they use Rayleigh waves only from ambient noise and only measure group speed. Second, their model is parameterized with four constant layers from the surface to 80 km depth: a sedimentary layer, an upper crust, a lower crust, and an uppermost mantle. Third, they construct a posterior distribution at each location with a Bayesian approach, but their final model is a linear inversion of the posterior mean. These differences in data types, parameterization, and inversion strategies can manifest as inconsistencies between the seismic structural variables, which we discuss in sections 3.2.2 and 3.2.3.

3.2.2 Selection of seismic structural variables

While the seismic structure is a continuous function of depth at each location, we deliberately choose a few seismic variables to characterize a seismic profile for two reasons. First, seismic structural information can differ systematically from diverse methods by various groups, so they might not be directly comparable. Second, seismic inversions suffer from trade-offs between the parameters, so we seek a set of seismic structural variables that correlate the least with each other.

To identify a few variables that are representative of the whole structure, we perform a hierarchical clustering with Ward variance minimization (Ward Jr. 1963). For both the US and Europe seismic structural variables, we identify three clusters that correspond to the
crust, the crust-mantle transition zone, and the uppermost mantle (Fig. 4). Thus, we use the shear wavespeed at 15 km depth to represent the crust, shear wavespeed at 65 km depth for the uppermost mantle, and the Moho depth for the crust-mantle boundary (Fig. 5). The visual similarity between seismic structure and heat flow is striking for $V_S$ at 65 km depth (Fig. 5b). These three variables essentially characterize a seismic structural profile with two constant layers. This small number of variables leads to a simple model, which is more likely to be interpretable and transportable than a model involving a large number of variables.

3.2.3 Standardization of seismic structural variables

The statistical distributions of these three seismic structural variables are significantly different between the US and Europe (Fig. 5g–i). Thus, we standardize a seismic structural variable by subtracting its spatial mean across each continent and then dividing it by its spatial standard deviation. The mean and standard deviation account for the systematic differences (e.g. due to parameterization) and relative variations (e.g. due to varied resolution) of the seismic structural variables, respectively. For example, increasing the $Q$ used in the seismic inversion can systematically decrease the inverted $V_S$. The standardization thus helps reduce the error of $Q$ models used in the inversion. These standardized variables are part of the input to the models, corresponding to $v_i$ in eq. (4).

4 Methods

In this section, we first discuss the heat flow and seismic data, and how a Machine Learning model relates them (section 4.1). We then describe three specific Machine Learning models (section 4.2), i.e., the Linear Regression (section 4.2.1), Decision Tree (section 4.2.2), and Random Forest models (section 4.2.3). We then discuss how we estimate uncertainties (section 4.3), and use validation (section 4.4.1) and cross-validation (section 4.4.2) to evaluate model accuracy and transportability, respectively.

4.1 Model specification, data weighting, and misfit

We now quantitatively define how we use a Machine Learning model to relate heat flow with seismic structure. Let $r_i$ be a location on the Earth’s surface, and $\hat{q}(r_i)$ be the heat flow prediction at location $r_i$, and $v(z, r_i)$ be 1-D local seismic structure as a function of depth $z$ beneath location $r_i$, and $q^{obs}(\{r_j\}_{j=1}^{N_{obs}})$ be a total of $N_{obs}$ heat flow observations at locations $\{r_j, j = 1, 2, \ldots, N_{obs}\}$ and $w(\{r_j\}_{j=1}^{N_{obs}})$ for their associated weights (uncertainties). We define the relationship between heat flow and seismic structure as a Machine Learning model $\hat{f}$ that predicts heat flow $\hat{q}(r_i)$ based on seismic structure $v(z, r_i)$ and heat flow observations $q^{obs}(\{r_j\}_{j=1}^{N_{obs}})$ with associated weights $w(\{r_j\}_{j=1}^{N_{obs}})$,

$$\hat{q}_i = \hat{f}(v_i, q^{obs}, w; \Theta). \tag{4}$$

Here, $\Theta$ denotes the parameters in the model, and to avoid clutter, $\hat{q}_i \equiv \hat{q}(r_i)$, $v_i \equiv v(z, r_i)$, $q^{obs} \equiv q^{obs}(\{r_j\}_{j=1}^{N_{obs}})$, $w \equiv w(\{r_j\}_{j=1}^{N_{obs}})$. 

8
In eq. (4), we have specified \( q^{\text{obs}} \) as the heat flow on a 100 km grid (section 3.1), and \( v(z, r) \) as the three seismic structural variables after standardization: the \( V_S \) at 15 km depth, \( V_S \) at 65 km depth, and the Moho depth (section 3.2).

Now we specify heat flow weights \( w \), the model parameters \( \Theta \), and data misfit. We assign a quality estimate, which we refer to as a weight \( w \), to each point of the heat flow map as follows. The median heat flow values are weighed by both the associated MAD, \( w_{\text{MAD}} \), and the misfit from a smooth model prediction, \( w_{\text{misfit}} \), such that the total weight is the product of these two weights,

\[
w = w_{\text{MAD}} \cdot w_{\text{misfit}}.
\]

Both \( w_{\text{MAD}} \) and \( w_{\text{misfit}} \) for the \( i \)-th heat flow median value are defined as

\[
w_i = \min \left( 1, \left( n \frac{\text{MAD}(x)}{x_i} \right)^2 \right),
\]

where we choose \( n = 2 \). If the data are normally distributed, then this choice of \( n \) will down-weight the outer 5% of data. For \( w_{\text{MAD}} \), the variable \( x \) denotes the MAD of heat flow within a 100 km radius, \( x_i = \text{MAD} \left\{ q^{\text{obs}}_j : |r_j - r_i| < 100 \text{ km} \right\} \). For \( w_{\text{misfit}} \), the variable \( x \) denotes the absolute difference between heat flow observations and (smooth) predictions, \( x_i = |\hat{q}_i - q^{\text{obs}}_i| \). The smooth model prediction is from a Decision Tree with a level of 3 (section 4.2.2). The final weights are presented in Fig. 2c. The median heat flow values and their associated weights are the input data to models, corresponding to \( q^{\text{obs}} \) and \( w \) in eq. (4), respectively. These weights are used to define the misfit between model predictions and heat flow observations (section 5).

Each Machine Learning model is specified by model parameters \( \theta \) and hyper-parameters \( \tilde{\theta} \), which together constitute all the parameters of a model, \( \Theta = \{ \theta, \tilde{\theta} \} \) (eq. (4)). Model parameters are derived via fitting model predictions to data, while model hyper-parameters are defined by the user and cannot be inferred from data fitting. We will specify the parameters and hyper-parameters of each of the three models in the following sections, i.e., the Linear Regression model (section 4.2.1), the Decision Tree model (section 4.2.2), and the Random Forest model (section 4.2.3).

We use the weighted root mean squared error (RMSE) to quantify data misfit for all models,

\[
\text{RMSE}(\tilde{q}) \equiv \left( \sum_{i=1}^{N_{\text{obs}}} w_i' (\tilde{q}_i - q^{\text{obs}}_i)^2 \right)^{1/2},
\]

where \( w_i' \) denotes the normalized weight, \( w_i' = w_i / \sum_{k=1}^{N_{\text{obs}}} w_k \), and \( N_{\text{obs}} \) denotes the number of heat flow observations defined on a 100 km grid.

### 4.2 Machine Learning models across a hierarchy of complexity

We choose these three models because they model heat flow across a hierarchy of spatial scales, from continental scale (Linear Regression), to regional scale (Decision Tree), to local scale (Random Forest). In this context, the complexity of the models is essentially equivalent to their spatial resolution and the number of model parameters (which may be different in
other contexts). We implement the models with scikit-learn (Pedregosa et al. 2011), using the LinearRegression, DecisionTreeRegressor, and RandomForestRegressor modules, respectively.

### 4.2.1 Linear Regression

As a baseline large-scale reference, we consider the Linear Regression model,

\[ \hat{q}_i = c_0 + \sum_{j=1}^{N_v} c_j v_{i,j}, \] (8)

where \( \hat{q}_i \equiv \hat{q}(r_i) \), \( r_i \) denotes spatial locations, \( v_{i,j} \) denotes the \( j \)-th seismic variable of \( v(r_i) \) (the \( V_S \) at 15 km depth, \( V_S \) at 65 km depth, and the Moho depth), \( N_v = 3 \) denotes the number of seismic structural variables, and \( c_j \) are the coefficients that define a line (\( j = 0, 1, 2, 3 \)). For the Linear Regression model, the model parameters \( \theta \) are \( c_j \), and there are no model hyper-parameters \( \tilde{\theta} \).

The Linear Regression model can be straightforwardly interpreted based on its coefficients, \( c_j \). However, the coefficients are fit to the whole area, i.e., they are the same for the entire continent. Thus, the Linear Regression model cannot distinguish regional differences in the relationship (e.g., tectonically active versus stable). One could address this issue by fitting a Linear Regression model to each region, but this regionalization information needs to be input \textit{a priori}, and cannot be derived from the model itself. The Decision Tree and Random forest models can overcome this issue by obtaining regionalization \textit{a posteriori}.

### 4.2.2 Decision Tree

Physically, we would expect a tectonically active region to have high temperature and low seismic wavespeed, and thus high heat flow. Conversely, we may relate high heat flow with such variables. Such regional information has to be provided to a Linear Regression model \textit{a priori}, which motivates us to develop the Decision Tree model to obtain regionalization \textit{a posteriori}.

A Decision Tree model (e.g., Breiman et al. 2017) splits the seismic structures into two disjoint subsets by a seismic structural variable \( v^{(i)} \) and an associated threshold \( v_*^{(i)} \). For each subset, the weighted average of heat flow associated with the seismic structures makes the prediction of heat flow. The variable and threshold are chosen to minimize the misfit between heat flow predictions and observations. This splitting is repeated for each subset and so on until a stop criterion is met. Each splitting path from start to end makes a decision rule. The model parameters \( \theta \) are splitting variables and associated thresholds, \( v_*^{(i)} \), while we choose the model hyper-parameter \( \tilde{\theta} \) as the maximum level of the tree, \( L \). We provide a concrete example of applying the Decision Tree model to real data in section 5.1.2.

### 4.2.3 Random Forest

A Random Forest is an ensemble of Decision Trees and is designed to avoid over-fitting (Breiman 2001). This is achieved by using a random subset of both heat flow observations and seismic structural variables to train each tree, and then combining all the trees for
prediction. The model parameters are the splitting variables and associated thresholds for each tree, while the model hyper-parameter is chosen by us to be the total number of trees.

Because a Random Forest typically consists of more than tens of trees, it operates somewhat like a black box where the explicit linkage from seismic structure to heat flow is highly complicated. It is, therefore, not effective at illuminating how a specific heat flow prediction is made from seismic structural variables.

### 4.3 Uncertainty estimation

Both the seismic structure information and heat flow observations have errors. We denote uncertainties in heat flow predictions from these two sources as $\epsilon_v$ and $\epsilon_q$, respectively. Errors in seismic structural information can arise from biases in seismic measurements and non-uniqueness in seismic inversions. To estimate $\epsilon_v$, we first randomly perturb the seismic variables assuming Gaussian distributions with standard deviations of about 1% (0.05 km/s), 2% (0.1 km/s), and 10% (4 km) for the $V_S$ at 15 km, $V_S$ at 65 km, and Moho depth, respectively. These standard deviations are estimated from both random and systematic errors evaluated from Shen and Ritzwoller (2016). For example, a lower $Q$ has a larger impact on the inverted $V_S$ and thus $Q$ affects the mantle more than the crust because the crust has a higher $Q$ than the mantle. Then we compute the standard deviation of predictions from 1000 such perturbations and use it as a prediction uncertainty. For the Linear Regression models, the uncertainty can also be linearly propagated from seismic variable uncertainty, which we estimate to be about 7.5 mW/m$^2$.

Heat flow observations in the US and Europe also have errors, which can result from local or non-steady state processes such as paleoclimate variations, sedimentation and erosion, seismicity, and ground water or magma convection (Powell et al. 1988). Unfortunately, most heat flow observations do not have associated quality information such as measurement errors and site characterization. Therefore, we use the spatial variability of heat flow (MAD) as a proxy for its error. In both the US and Europe, the spatial MAD values of heat flow observations within a 100 km radius are about 10 mW/m$^2$. For the Linear Regression model, this propagates to an error of about 0.3 mW/m$^2$ in heat flow estimates: $\epsilon_q \approx 0.3$ mW/m$^2$. The model has a reduced chi-squared misfit of about 0.9, which is close to 1 and thus justifies using MAD to approximate for error. This uncertainty is also small for the Decision Tree and Random Forest models, because they make predictions by averaging a large number of heat flow observations (typically > 100).

Therefore, we conclude that $\epsilon_v \gg \epsilon_q$ and thus identify $\epsilon_v$ as the primary uncertainty. We present an example of estimating $\epsilon_v$ where both the data and models are from the US (Fig. 9). As expected from error propagation, the Linear Regression (LR) predictions have an almost constant uncertainty of about 7.5 mW/m$^2$. Uncertainties for the Decision Tree (DT) predictions are larger for intermediate heat flow values near boundaries, because a slight change of seismic variable can activate a different decision rule with a distinct heat flow prediction. Uncertainties are higher for higher heat flow for both the Decision Tree and Random Forest (RF) models, where heat flow is also more variable and therefore the model predictions are more sensitive to input variations.
4.4 Evaluation of model accuracy and transportability

4.4.1 Validation for model accuracy

To optimize model hyper-parameters and validate prediction uncertainties, we should train the model on one set of data (the training set) and then validate on a separate set of data (the validation set). Therefore, we partition the heat flow observations into an 80% training set and a 20% complementary validation set. We then train a model on the training data, and compare its predictions to the validation data. This process is repeated using a total of 20 different random partitions of the data.

4.4.1.1 Fit to validation data

To determine the accuracy or predictive power of the models, we evaluate the goodness of fit to the heat flow data using the coefficient of determination, \( r^2 \), which measures the proportion of observed heat flow variations predicted by each model:

\[
r^2(\hat{q}) = 1 - \frac{\sum_{i=1}^{N_{\text{obs}}} w'_i (q_{\text{obs}} - \hat{q}_i)^2}{\sum_{i=1}^{N_{\text{obs}}} w'_i (q_{\text{obs}} - \bar{q}_{\text{obs}})^2} = 1 - \left( \frac{\text{RMSE}}{\sigma_{\text{obs}}} \right)^2,
\]

where \( \bar{q}_{\text{obs}} \) denotes the average value across the heat flow map, \( w'_i \) denotes the normalized weight (section 3.1.2), \( \sigma_{\text{obs}} \) denotes the spatial standard deviation of \( q_{\text{obs}} \), RMSE is defined in eq. (7), and \( r^2 \in (-\infty, 1] \). In particular, if \( \hat{q}_i = q_{\text{obs}} \), then \( r^2 = 0 \); if \( \hat{q}_i = q_{\text{obs}} \), then \( r^2 = 1 \). It is instructive to put in some realistic numbers. Assume \( \sigma_{\text{obs}} = 15 \text{ mW/m}^2 \), which is close to both the US and European data (Fig. 3). Thus, \( r^2 = 0 \), 0.5, 0.75 for RMSE = 15, 11, 7.5 mW/m\(^2\) respectively. Since \( r^2 \) is a linear function of the mean squared error (RMSE\(^2\)), their relative variations with model hyper-parameters are the same. Therefore, using either \( r^2 \) or least squares will yield the same choices of optimal model hyper-parameters.

The \( r^2 \) on validation data provides an estimate of model accuracy. For example, each error bar in Fig. 6 denotes the mean and standard deviation of \( r^2 \) from validating the Decision Tree model with a specific max level \( L \). The relative \( r^2 \) for different hyper-parameters (e.g. max level \( L \) for a Decision Tree) serves as a basis for choosing the optimal hyper-parameter, which we elaborate below.

4.4.1.2 Hyper-parameter choices

We illustrate how hyper-parameters are chosen with the Decision Tree models (Fig. 6). The structure of a Decision Tree is controlled by its maximum level \( L \). The training performance of a Decision Tree increases monotonically with \( L \), which is simply caused by a better fit to the data by adding model parameters. In contrast, the validation score of a tree first increases with \( L \) until \( L = 5 \) but then plateaus/decreases, as a result of transitioning from under-fitting to over-fitting. Moreover, the validation scores are more variable than training scores, probably because a larger fraction of data are repeated in training. Thus, we choose \( L = 3 \) as a compromise between fitting the data while keeping the model interpretable.
particular, increasing the level from 3 to 4 will approximately double the number of decision rules from $\sim 8$ to $\sim 16$. We choose the number of trees to be 10 for the Random Forest model (Fig. S2).

### 4.4.2 Cross-validation for model transportability

To evaluate the transportability of a model, we apply each of the US-trained models to the European data, and vice versa. The results from this cross-validation are discussed in section 5.2. We introduce another metric besides $r^2$ to measure the geographical coherence between heat flow predictions and observations. This is because $r^2$ measures the overall fit (eq. (9)), and thus includes the amplitudes of the observations as well as the geographical pattern. We introduce the Pearson’s correlation coefficient $\rho$, which is less sensitive to the absolute amplitudes. Specifically, $\rho$ measures the linear (in)dependence between predictions $\hat{q}$ and observations $q^{\text{obs}}$,

$$\rho(\hat{q}) = \frac{\text{Cov}(\hat{q}, q^{\text{obs}})}{\sigma \sigma^{\text{obs}}},$$

where Cov denotes the covariance, $\hat{\sigma}$ and $\sigma^{\text{obs}}$ denote the spatial standard deviation of predicted and observed heat flow, respectively, and $\rho \in [-1, 1]$. In particular, if the predictions are linearly related to the observations, $\hat{q}_i = kq^{\text{obs}}_i + b$ where $k, b$ are constants and $k > 0$, then $\rho = 1$, while $r^2$ can be low if $k \neq 1$ or $b \neq 0$. If $\hat{q}_i$ is a constant, then $\rho = 0$, while $r^2 = 0$ only for the mean heat flux $\hat{q}_i = \bar{q}^{\text{obs}}$, and $r^2 < 0$ otherwise.

### 5 Results

In this section, we first address the characterization problem using data from the same continent (section 5.1), which includes evaluation of model accuracy from validation (section 5.1.1) and the resulting improvement in understanding of the relationship between seismic structure and heat flow (section 5.1.2). Then we tackle the transportability problem by applying the US-trained model to the European data, and vice versa (section 5.2). This includes evaluation of model transportability by model cross-validation (section 5.2.1) and the resulting insight gained about challenges in extrapolating models between continents (section 5.2.2).

#### 5.1 Characterization within the same continent

We first develop our method in the best-case scenario, where both heat flow and seismic structural data are from the same continent. This is because we expect greater consistency between the training and validation data in the US than between the US and European data. More consistent data will produce more accurate models.

##### 5.1.1 Comparison of model accuracy

The model statistics are summarized in Fig. 7a for the US data alone. The Random Forest model has the best scores, but its validation performance ($r^2 = 0.67 \pm 0.03$) is much worse than its training performance ($r^2 = 0.95 \pm 0.00$), which indicates over-fitting (Fig. 8e). The validation performance of the Decision Tree model is comparable ($r^2 = 0.62 \pm 0.03$) to
that of the Random Forest, but its training score \( r^2 = 0.68 \pm 0.01 \) is much closer to its own validation score, which suggest that it is less prone to over-fitting than the Random Forest model. The Linear Regression model performs slightly worse \( r^2 = 0.51 \pm 0.04 \) than the Decision Tree and Random Forest models. For different models, the validation \( \rho \) are relatively similar to \( r^2 \) while absolutely higher, ranging from 0.7 to 0.8 (Fig. 7c). The relative performance of the models is similar using the European data alone (Fig. 7b). However, both \( r^2 \) and \( \rho \) values are generally lower, which suggests that heat flow can be better explained by seismic structure in the US than in Europe.

We also present model accuracy and spatial variability in map form (Fig. 8). For each of the three models, our 20%/80% (validation/training) splitting of training and validation data with 20 folds (section 4.4.1) means that every location on average is predicted four times as validation data. We randomly choose a unique prediction for each location (Fig. 8a–c), and compare them with observations (Fig. 8d). The Linear Regression predictions are the most smooth (Fig. 8a), because the coefficients are fit to the whole area and thus are not sensitive to subsetting of the data (Fig. S3). In contrast, the Decision Tree predictions are more accurate but also are more spatially variable (Fig. 8b), which is caused by its hard splits at threshold values of seismic structural variables, and its weak sensitivity to data subsetting the data. The Random Forest predictions are the most spatially variable (Fig. 8c), because of its strong sensitivity to data subsetting, as it attempts to fit all local data variations. These models exhibit similar spatial variability for the European data alone (Fig. 10). However, the European predictions generally have larger differences with observations. Thus, the European predictions are less accurate, consistent with their lower statistics (Fig. 7).

In summary, the Decision Tree model fits the validation data equally well as the Random Forest model and slightly better than the Linear Regression model. Moreover, the Decision Tree model illuminates the regional variations of the relationship between heat flow and seismic structure, while the other two models cannot. Therefore, we focus the following insights on the region-specific relationship from the Decision Tree model (section 5.1.2).

### 5.1.2 Insights from the Decision Tree model

We use three levels in the Decision Tree \( (L = 3) \), and retrain a Decision Tree model using all data from the US alone to discover the most general relationship between heat flow and seismic structure in the US (Figs 11 to 13). To avoid over-fitting, we require that each of the two parts from a splitting has at least 5% of the total data. We also repeat the process to build a Decision Tree using all the European data (Figs 14 to 16).

In the US, a single splitting of a Decision Tree reveals the west-east dichotomy of heat flow (Fig. 13a), which manifests as a bimodal distribution of \( V_S \) at 65 km depth (Fig. 12 bottom row). As we further grow the tree by increasing its level \( L \), we start to see regional variations finer than the west-east dichotomy (Fig. 13bc). Specifically, across the Western US, the highest heat flow values are correlated with very slow uppermost mantle wavespeeds in the Basin and Range, and the Rio Grande Rift (Fig. 12ab). High heat flow values surrounding the Colorado Plateau are characterized by a slow crustal wavespeed (Fig. 12c), while relatively low heat flow values within the Colorado Plateau correlate with fast/intermediate crustal wavespeed (Fig. 12d). In the Eastern US, heat flow values are generally lower than those in the West and are associated with fast uppermost mantle wavespeed (Fig. 12e–g).
Relatively high heat flow values are associated with either a thick crust in the Great Plains and Appalachian Mountains (Fig. 12f), or slow crustal wavespeed in the Mississippi Embayment (Fig. 12c). The lowest heat flow values in the Eastern US correlate with regions having thin crust and fast crustal wavespeed (Fig. 12g). Most large misfits are correlated with large spatial variability of heat flow observations even before these regions are down-weighted (Fig. 2b). These include undershooting (predictions smaller than observations) at the Basin and Range, the Rio Grande Rift, and South Dakota; and the overshooting at Yellowstone. These regions also host hydrothermal systems (Lachenbruch and Sass 1977; Gosnold 1990). Large MAD values and residuals at the Southern Appalachians are probably caused by the high radiogenic heat production concentrated in the uppermost crust there (Perry et al. 2006a). Prediction overshoots at the Cascade Range, which is undergoing active subduction and is therefore tectonically distinct from elsewhere in the US.

In Europe, we find that the northeast-southwest dichotomy of heat flow is anti-correlated with uppermost mantle $V_S$, roughly separated by the Trans-European Suture Zone between the Precambrian East European Craton and the Phanerozoic Southwest European orogens (Fig. 16). Moreover, the highest heat flow is correlated with a very slow uppermost mantle (e.g. Massif Central, east of Dinarides and Hellenides) or a thin crust (e.g. Upper Rhine Plain; Fig. 15a). The lowest heat flow values are characterized by a fast uppermost mantle, and a thick crust (the Baltic Shield, Russian Platform, Apennines, and Dinarides and Hellenides; Fig. 15gh). Intermediate heat flow at the North German Plain is mostly anti-correlated with crustal thickness (Fig. 15de). Similar to the US data, large misfits often correlate with large MAD values of heat flow observations (Fig. 2d). These include the undershooting at France and Anatolia; and the overshooting at the Apennines.

To quantify the relative importance of seismic structural variables, we compute the variance reduction from each seismic structural variable, normalizing the total reduction as unity (Gini importance). For the US data, this shows that the uppermost mantle contributes about 80% of the variance reduction, while the crust $V_S$ and Moho depth each accounts for around 10% (Fig. 17). For the European data, uppermost mantle $V_S$ still dominates while crustal $V_S$ is secondary. However, the Moho depth is more important than in the US. A comparison of the decision rules suggest that the Moho depth is positively related with heat flow in the US (Fig. 12fg), but negatively in Europe (Fig. 12f–h). This opposite trend is consistent with the Linear Regression model, where the coefficients for the Moho depth have opposite signs and different relative magnitudes (Fig. S3).

These tree structures, rules, and maps from growing the number of levels of the Decision Tree illustrate the unique advantage of the Decision Tree model. It tends to split the whole area into regions corresponding with geological provinces, despite no such geographical information being given. Moreover, it can yield a specific relationship between heat flow and seismic structure for each region. These region-specific relationships constitute our answer to most of the characterization problem, which we summarize below (section 5.1.3).

5.1.3 Summary for the characterization problem

We find that seismic structure can explain more than half of the observed heat flow variations in both the US and Europe (Figs 7, 8 and 10). Uppermost mantle wavespeed is the primary predictor of heat flow, while crustal wavespeed and Moho depth are secondary
The region-specific relationship between seismic structure and heat flow is explicitly characterized by several decision rules (Figs 11 to 16). However, we note two differences using the US and European data. First, the US heat flow observations are better explained by seismic structure than the European observations. Second, the Moho depth plays a more important role for the European data than the US data. We discuss potential causes for the unexplained heat flow variations (> 30%) in section 6.1.1 and for these two differences in section 6.1.2.

5.2 Extrapolating models from one continent to the other

In continents where heat flow observations are rare or absent (such as Antarctica), training the model is difficult with only local heat flow observations. Therefore, the scenario in section 5.1 is inapplicable, and the model has to be trained on other continents where heat flow observations exist. To evaluate how our models will perform in this more challenging case, we apply the US-trained models to the European data, and vice versa.

5.2.1 Comparison of model transportability

We apply each of the three models trained in the US to the European data and summarize their metrics in Fig. 18a. The Linear Regression and Decision Tree models have the best overall fit (r² ≈ 0.2), while the Random Forest model is significantly worse (r² = −0.05). As discussed above (section 5.1.1), the Random Forest model is prone to over-fitting, as suggested by small scale variations in its heat flow predictions (Fig. 19c). The relative performance of the models in geographical coherence (ρ) is similar to performance in overall fit (r²). However, the absolute values of ρ are about 0.3 larger than r² for all three models (ρ ≈ 0.5 for the Decision Tree and Linear Regression models, and ρ ≈ 0.3 for the Random Forest model). This means the prediction of the geographical patterns of heat flow is better than the overall fit to the data, which is degraded by the models’ limitations in fitting amplitudes. These model comparisons are consistent with applying the Europe-trained models to the US data (Fig. 18b). However, the values of both metrics are higher for extrapolating from Europe to the US than from US to Europe (r² increases about 0 to 0.2, ρ increases about 0.2 to 0.3). This is consistent with better model accuracy using the US data alone than the European data alone (section 5.1.1). For both the US and European data, the misfit values are significant worse than using data from the same continent (Fig. 7) for all three models (r² reduces about 0.2, 0.4, and 0.6 for the Linear Regression, Decision Tree, and Random Forest models, respectively). This suggests that model transportability deteriorates with increasing spatial resolution.

We also present model transportability in map form (Fig. 19), where we compare predictions (Fig. 19a–c) with observations (Fig. 19d). The Linear Regression and Decision Tree model predictions have similar spatial variability as observations, while the Random Forest model predictions show noisy local variations. Compared to predictions using the European data alone (Fig. 10), the geographical patterns are similar, but the absolute amplitudes tend to be under-predicted from all three models. These models exhibit similar spatial variability when the Europe-trained models are applied to the US data (Fig. 20). However, the absolute amplitudes tend to be over-predicted, compared to the US-trained models (Fig. 8).
These metrics and maps suggest that the geographical patterns of heat flow observations can be reasonably predicted by the Decision Tree and Linear Regression models (and to a much lesser extent by the Random Forest model), but the absolute amplitudes cannot by any of the three models. Because the Decision Tree model can uniquely identify region-specific relationships between seismic structure and heat flow, we use the Decision Tree model to discuss the differences in extrapolating the relationship from one continent to the other in the next section (section 5.2.2).

5.2.2 Insights from the Decision Tree models

The Decision Tree model trained with the US data alone is able to delineate the northeast-southwest European tectonic dichotomy based on uppermost mantle wavespeed, where high heat flow is related to a slow uppermost mantle in the Massif Central, east of the Dinarides and Hellenides, and Anatolia (Fig. 21a). Relative high heat flow is associated with slow crustal wavespeed for the rest of southwest Europe (Fig. 21b). The US-trained model, however, mistakenly assigns higher heat flow at the eastern Baltic Shield and the Russian Platform than the western Baltic Shield (Fig. 21c). This is because the eastern Baltic Shield has a thick crust and similar seismic structure to the Central US (higher heat flow), while the western Baltic Shield has a thinner crust and similar seismic structure to the eastern US (lower heat flow). Moreover, both models trained in the US (Fig. 21) and Europe (Fig. 16) cannot fit the exceptionally high heat flow in France. While Cenozoic volcanism can explain high values at the Massif Central, relatively high values in the Paris Basin are not well understood (Cermak and Rybach 1979).

The Decision Tree model trained with the European data alone can identify the west-east US tectonic dichotomy based on uppermost mantle wavespeed. In this dichotomy, high heat flow is related to a slow uppermost mantle in the western US, while low heat flow is related to a fast uppermost mantle in the eastern US (Fig. 22a). Relative high heat flow is associated with a relatively thin crust in the Mississippi Embayment, while relative low heat flow is associated with a relatively thick crust in the eastern US (Fig. 22b). The Europe-trained model, however, wrongly assigns higher heat flow at Florida than the Central US (Fig. 22c). This is because Florida has a relatively thin crust and seismic structure similar to the North Germany Basin (higher heat flow), while the Central US has a thick crust and similar seismic structure to the Baltic Shield (lower heat flow). Furthermore, models trained either in the US (Fig. 21) or Europe (Fig. 16) cannot fit the exceptionally low heat flow in the Cascade Range region, which is undergoing active subduction. These comparisons provide key insights for the transportability problem, which we summarize below (section 5.2.3).

5.2.3 Summary for the transportability problem

We find that the geographical coherence of heat flow can be reasonably reproduced from one continent to the other (Fig. 18). In particular, the Europe-trained models can better predict the US data, than vice versa. Exceptions are mostly for relatively low heat flow values. The absolute amplitudes of variations, however, are under-predicted in Europe when using the US models (Fig. 21), and over-predicted in the US when using the European models (Fig. 22). We discuss potential reasons for these differences in section 6.1.2.
the implications for heat flow inferences across Antarctica in section 6.2.

6 Discussion

The validation and cross-validation of our method against existing heat flow observations in the US and Europe provide insights about relating seismic structure to heat flow (section 5) but also expose several limitations of our method (section 6.1). Both the insights and limitations have implications for predicting heat flow across Antarctica (section 6.2) and validating heat flow proxies other than seismic structure (section 6.3).

6.1 Limitations of our methods

6.1.1 Unexplained heat flow variations within a continent

Even using data from the same continent, a significant fraction of heat flow variations are not explained by seismic structure in either the US ($r^2 < 0.7$) or Europe ($r^2 < 0.6$; Fig. 7). This means that dissimilar seismic structure can be linked to similar heat flow values and vice versa. We attribute the unexplained 30% to 40% variations mainly to heterogeneous distribution of crustal radiogenic heat production. This is because heat production can dominate heat flow, but its relationship with seismic structure is weak (Jaupart et al. 2016). For example, an empirical relationship has been proposed by Pollack and Chapman (1977) which attributes $\sim 40\%$ of observed heat flow to upper crustal radiogenic sources and the remaining $\sim 60\%$ to any deeper source (radiogenic sources in the lower crust, heat flow from the mantle, etc.). This empirical 40%/60% partition was established from several provinces including the Basin and Range, the Eastern US, and the Baltic Shield (Pollack and Chapman 1977).

6.1.2 Differences between the US and Europe data

For predictions within each continent (Fig. 7) or between the two continents (Fig. 18), the US heat flow is better predicted than the European heat flow. This suggests that either heat flow is more directly related with seismic structure in the US than Europe, the quality of information about heat flow and/or seismic structure is better in the US than those in Europe, or crustal radiogenic heat production is stronger in Europe than in the US.

When extrapolating models between the US and Europe, we find that the absolute amplitudes are over-predicted in the US (Fig. 20) and under-predicted in Europe (Fig. 19). This is because the statistical distributions of heat flow are different in the US and Europe, where the European data show more variability (higher highs, lower lows) despite Europe’s smaller area (Fig. 3). We are not certain whether this is caused by intrinsic differences in heat flow between the continents, or by errors in the measurements of heat flow. We do find that absolute amplitudes can be better predicted by standardizing the heat flow distributions on each continent separately. Specifically, for heat flow data on each continent, we remove the mean and divide by the standard deviation. The standard deviation of heat flow across Europe is about 30% larger than across the US. We find that this standardization improves
the prediction of absolute amplitudes and reduce the overall fit ($r^2$ increase about 0.1 to 0.2; Fig. S6).

However, the standardization does not change the geographical coherence ($\rho$). This means that similar seismic structure can be linked with different relative heat flow values, which violates the operational assumption that underlies the use of seismic structure as a heat flow proxy: that similar seismic structural information should be related with similar heat flow values. A notable example lies in Northeast Europe, and its counterpart in the Central and Eastern US, where the Moho depth is negatively and positively related with heat flow, respectively. Assuming heat flow observations are reliable, this may be caused by differences in the US and Europe resulting from (1) the seismic structural imaging methods or (2) crustal radiogenic heat production.

(1) Significant trade-offs between the Moho depth and uppermost mantle $V_S$ may exist in the European seismic structural information (Lu et al. 2018). We note that the uppermost mantle $V_S$ and the Moho depth are positively correlated in the Apennines, the Dinarides, and the Baltic Shield (Fig. 5ef), at length scales which we believe are below the resolution of the seismic data. This suggests that these structures are not independently estimated in the European seismic structural information by Lu et al. (2018). Transportability would improve if seismic structure is estimated similarly for both continents, for example by using the same types of data, methods, and parameterization.

(2) As discussed in section 6.1.1, the relationship between crustal radiogenic heat production and seismic structure is probably weak (Jaupart et al. 2016). A log-linear relationship between heat production and $V_P$ has been discovered, but the relationship varies between different rock types, geologic ages, and geological provinces (Rybach and Buntebarth 1984; Hasterok and Webb 2017). A correlation between geological age and heat flow and crustal heat production is controversial due to a competing effect of processes decreasing heat flow such as erosion, radioactive decay and cooling, and increasing processes such as plutonism, volcanism and collisions (Rao et al. 1982; Perry et al. 2006c; Jaupart et al. 2016). These geological differences can thus cause different heat production in the US and Europe, linking similar seismic structure with dissimilar heat flow. In the future, geological information (rock type, age, and province) could complement seismic structural information to better constrain heat production and thus heat flow predictions between continents.

6.2 Implications for heat flow inference across Antarctica

We suggest that the US data (heat flow, seismic structural information) should be given more weight than the European data because heat flow data can be predicted from seismic structure more accurately in the US than in Europe. In addition, the seismic data and parameterization of Shen et al. (2018) in Antarctica is the same as those used by Shen and Ritzwoller (2016) in the US. In contrast, the European seismic structure imaged by Lu et al. (2018) is based on a different approach. Since the statistical distribution of heat flow across Antarctica is not currently known, however, the European data could be incorporated to improve predictions of the absolute amplitudes of heat flow. The difference of about 30% in the variability of heat flow between Europe and the US is a valuable guide to the expected differences in the absolute amplitudes of heat flow predictions for Antarctica. Although different Machine Learning models produce similar results, we recommend the Decision Tree
model for its transportability and geologically relevant regionalization. Because the uppermost mantle $V_S$ is the primary predictor of heat flow, future work to improve the imaging of the uppermost mantle structure is most fruitful. We suggest that crustal radiogenic heat production is a prominent systematic uncertainty, which can cause at least 30% of heat flow variations to be unpredicted by seismic structure. Current knowledge of crustal radiogenic heat production in Antarctica is highly localized, most of which is near the periphery of Antarctica where outcrops occur (Sanchez et al. 2021). Heat production is likely to continue to be a fundamental factor that limits the accuracy of heat flow inference across Antarctica from both seismic structure and other proxies for years to come. In the future, geological information can complement seismic structural information to better constrain heat production and thus heat flow across Antarctica. The inference of heat flow across Antarctica is pursued in our companion work (Zhang and Ritzwoller, “Transporting Heat Flux From the US and Europe to Antarctica Guided by Regional Seismic Structure”, in preparation).

6.3 Validation of proxies other than seismic structure

Our validation and cross-validation methodology can also be applied to validate proxies other than seismic structure. As an example, we consider the magnetic Curie depth below (Fox Maule et al. 2005; Martos et al. 2017), but this procedure is also applicable to other heat flow proxies such as surface topography (Artemieva 2022). Heat flow inferences based on magnetic anomalies within a region begin by estimating the Curie depth $Z_b$. Then they typically assume an exponential decay of heat production with depth $z$: $H(z) = H_0 \exp(-z/h_r)$, where $H_0$ is the surface heat production, and $h_r$ is the characteristic depth of heat production. Assuming 1-D steady state, the following relationship has been obtained between the Curie depth $Z_b$, Curie temperature $T_c$, and surface heat flow $\hat{q}$ (e.g. Turcotte and Schubert 2002):

$$\hat{q} = -k \frac{T_c}{Z_b} - H_0 h_r \left[ 1 - \frac{h_r}{Z_b} \left( 1 - \exp \left( -\frac{Z_b}{h_r} \right) \right) \right],$$

(11)

where $k$ denotes heat conductivity.

For most places in the world and especially Greenland and Antarctica (Sanchez et al. 2021), heat production $H(z)$ is poorly known and can disobey an exponential function (Lachenbruch and Sass 1977; Jaupart et al. 2016), and perhaps to a lesser extent so are $k, T_c$. However, because magnetic anomaly data are also available in the US and Europe (e.g. Maus et al. 2009), heat flow observations from these two continents can be used to validate this relationship. In particular, the characterization on the same continent can determine the predictive power of this relationship and calibrate the model parameters such as $k, H_0, h_r$ (section 5.1). Moreover, the transportability of this relationship can be measured by cross-validation between the US and Europe (section 5.2).

7 Conclusion

This study is motivated by the importance of estimating heat flow beneath polar ice sheets, and the observed inconsistency between previous heat flow estimates inferred from
different proxy data, particularly for Antarctica (section 1). Therefore, we aim to characterize the relationship between seismic structure and surface heat flow in a given region (the characterization problem), and to evaluate the transportability of this relationship from one continental setting to another (the transportability problem). To address both problems, we directly compare predictions and observations of surface heat flow using data from the US and Europe. We apply quality control on heat flow observations by data rejection, imputation and smoothing (section 3.1). We identify three principal seismic structural variables (uppermost mantle $V_S$, crustal $V_S$, and Moho depth) to represent seismic structure in both the US and Europe, using a clustering analysis (section 3.2). We find that further complications in the seismic structure do not improve the ability to predict heat flow appreciably. Our analysis focuses on using three Machine Learning models of varying degrees of complexity to determine the relationship between these three structural variables and surface heat flow (section 4).

To address the characterization problem, we train the Machine Learning models with the training data and then validate the models with the validation data within each continent separately (US, Europe). We find that uppermost mantle $V_S$ is the strongest predictor of heat flow and is considerably more important than crustal $V_S$ and Moho depth (section 5.1). However, crustal $V_S$ and Moho depth are relatively more important in tectonically stable regions (i.e., eastern US and northern Europe). Together, these three variables explain more than half of heat flow observations ($r^2$ about 50% to 60% in Europe and the US, respectively). We attribute the unexplained heat flow variations (> 30%) mainly to crustal radiogenic heat production. This is because heat production can greatly affect heat flow but may not be reflected in seismic structure (Jaupart et al. 2016).

To address the transportability problem, we apply the models trained on the US data to predict heat flow across Europe, and vice versa (section 5.2). While performing this cross-validation between the two continents, we find that the Decision Tree and Linear Regression models are more transportable than the Random Forest model. The Decision Tree and Linear Regression models can reasonably reproduce the geographic distribution of heat flow, but to a lesser extent reproduce the absolute amplitudes of heat flow variations. The US heat flow is better predicted from the Europe-trained models, than vice versa.

Given the recent emergence of high-resolution seismic structural information for Antarctica (e.g. Shen et al. 2018; Lloyd et al. 2020), the methodology presented in this paper promises to improve the reliability and resolution of inferences of heat flow across Antarctica. We suggest that the US data should be given more weight than the European data in extrapolating models of the relationship between seismic structure and heat flow, because the data are better fit in the US. Moreover, the parameterization of seismic structure we have in the US is more similar to the seismic structural model we have for Antarctica. However, the European data can be incorporated to improve predictions of the absolute amplitudes of heat flow variations. We recommend the use of the Decision Tree model for its greater transportability and correlation with geological provinces. Uppermost mantle $V_S$ emerges from the Decision Tree model as the primary predictor of heat flow. We suggest that crustal radiogenic heat production is a prominent systematic uncertainty, which can cause at least 30% of the heat flow variations.

Our methods can also be applied to validate distinct models of the relationship between heat flow and proxies other than seismic structure (e.g. Fox Maule et al. 2005; Goutorbe et
This validation is crucial to help resolve existing inconsistencies between heat flow inferences for Antarctica based on different proxies (section 6.3). This improvement of the inference of heat flow beneath Antarctica, and the potential resolution of existing inconsistencies, would help to improve constraints on the subglacial thermal boundary conditions, and thereby lead to more accurate predictions of the future evolution of ice sheets and their contributions to future sea level change.

Data Availability

Geothermal heat flow observations are based on NGHF (Lucazeau 2019), and the SMU Node of the National Geothermal Data System (http://geothermal.smu.edu/gtda; Blackwell et al. 2013). The US seismic structural information is from Shen and Ritzwoller (2016), and the European structure is from Lu et al. (2018). Machine Learning models are developed with scikit-learn (Pedregosa et al. 2011), and the trained models are available on Zenodo: https://doi.org/10.5281/zenodo.12173897.

Funding

This research is supported by NSF grant EAR-1943112 at the University of Colorado Boulder.

Acknowledgements

We thank the Editor, Andrew Valentine, and reviewers Claire Perry and Anonymous for constructive comments. We acknowledge helpful discussions with Billy Shinevar, Kyle Goodrick, and Derek Mease. We thank Francis Lucazeau, SMU Geothermal Lab, Weisen Shen, and Yang Lu for sharing the data specified in the Data Availability section.

References


Figures

Figure 1: Overview of the logical flow of this paper. Data, methods, and results are color-coded in green, blue, and red, respectively. The US data (heat flow, seismic structure) are split into training and validation data. The Machine Learning models are trained with the training data and then validated with the validation data for accuracy. The models specify the relationship between heat flow and seismic structure which addresses the characterization problem. The US-trained models are then cross-validated against the European (EU) data to tackle the transportability problem. The Europe-trained models are also extrapolated to the US data by switching the role of the US and European data.
Figure 2: **Heat flow observations.** (a)–(c) Heat flow observations are smoothed onto a 100 km grid to match seismic structural resolution across the contiguous US. (a) The median values of heat flow within a 100 km radius, \( q^{\text{obs}} \), are shown. (b) The MAD (Median Absolute Deviation) values of heat flow within a 100 km radius are shown, which is a robust measure of spatial variability (eq. (1)). (c) The weights, \( w \), of median values are shown, such that regions with large MAD or misfit to a smooth model are down-weighted (eq. (5)). The median values, \( q^{\text{obs}} \), and weights, \( w \), are part of the input data to the Machine Learning models (eq. (4)). (d)–(f) Similar to (a)–(c) but for European heat flow.
Figure 3: **Distributions of heat flow observations.** Histograms of heat flow values for the US data (blue) and European data (orange) from the smoothed maps (Fig. 2a & d). The European data are distributed more broadly (higher highs, lower lows), which affects the prediction of the absolute amplitudes of heat flow when extrapolating between the two continents.
Figure 4: Clustering of seismic structural variables. Cluster analysis is performed to identify principal independent seismic structural variables and determine if they are the same in the US and Europe. (a) The arrangement (dendrogram) of seismic structural variables for the US based on hierarchical clustering (section 3.2.2). The seismic structural variables (x-labels) are color-coded by depth. For example, Vs.015 denotes shear wavespeed at 15 km depth. The distances between clusters (vertical axis) are measured based on Ward’s method (Ward Jr. 1963). (b) Similar to (a) but for European seismic variables. For both the US and European data, three variables (Vs.015, Vs.065, and Moho) represent the three clusters for the crust (blue), crust-mantle boundary (green), and uppermost mantle (red). These three variables, \( v \), are part of the input to the Machine Learning models (eq. (4)).
Figure 5: The three representative seismic structural variables identified by the cluster analysis (Fig. 4).

(a)–(c) The US seismic structural variables from Shen and Ritzwoller (2016): (a) $V_s$ at 15 km depth ($V_{s015}$), (b) $V_s$ at 65 km depth ($V_{s065}$), and (c) Moho depth (Moho). (d)–(f) Similar to (a)–(c) except for the European seismic structure of Lu et al. (2018). The geographic distribution of seismic structure and heat flow (Fig. 2ad) are similar, particularly for (b) & (e), $V_s$ at 65 km depth. (g)–(i) Histograms of (g) $V_{s015}$ at 15 km depth, (h) $V_{s065}$ at 65 km depth, and (i) Moho depth. The distributions are significantly different between the US and Europe.
Figure 6: Choosing the maximum level of the Decision Tree (DT) models. (a) The $r^2$ (data misfit) is plotted as a function of maximum tree level $L$. The $r^2$ measures the proportion of observed heat flow variations predicted by the model (eq. (9)). Error bars denote mean ± standard deviation for the training (blue) and validation (orange) data in the US, respectively (section 5.2). The model is trained to fit the training data but not the validation data, so the training data are fit better with fit increasing continuously with $L$. (b) Similar to (a) except for the European data, which are fit worse than the US data. This is part of the basis for our choice of $L = 3$ (star) as a trade-off between model accuracy (high $r^2$) and simplicity (low $L$).
Figure 7: Comparison of model accuracy based on validation within a continent.

(a) Based on the US data: $r^2$ (data misfit) under validation (orange) and training $r^2$ (blue) are shown for the Linear Regression (LR), the Decision Tree (DT) ($L = 6$), and the Random Forest (RF) models (300 trees). $r^2$ measures the proportion of observed heat flow variations predicted by each model (eq. (9)). Error bars denote mean ± standard deviations from 20-fold validation (section 5.2), and the numeric values are annotated for the validation data (numbers in parentheses denote standard deviations). (b) Similar to (a) but for European data, which are fit worse under validation than the US data. (c) & (d) Similar to (a) & (b) except for the metric $\rho$ (geographical coherence).
Figure 8: Comparison of model accuracy based on validation with the US data alone. (a) The Linear Regression (LR) model is trained with the training data and validated on the validation data. The predictions using the validation data are shown in (a), which should be compared with (d) observations (same as Fig. 2a). (b) & (c) Similar to (a) but for (b) the Decision Tree (DT) model ($L = 3$) and (c) the Random Forest (RF) model. The Linear Regression model predictions are the most smooth but least accurate, while the Random Forest model predictions are the most accurate but have small scale variations, which we interpret as noise. The Decision Tree model strikes a balance between accuracy and spatial resolution.
Figure 9: **Uncertainties for predictions in Fig. 8.** (a) & (b) The Linear Regression (LR) predictions have a constant uncertainty of about 7.5 mW/m². (c) & (d) Uncertainties for the Decision Tree (DT) predictions are larger for intermediate heat flow values near boundaries. (e) & (f) Uncertainties generally increase with heat flow for the Random Forest (RF) model.
Figure 10: Comparison of model accuracy based on validation with the European data alone. Similar to Fig. 8, but based on the European data alone. The relative accuracy and spatial resolution for different model predictions are similar to those using the US data alone, but the European predictions generally have larger differences from observations and are thus less accurate.
Figure 11: The Decision Tree model trained using US data alone. The structure of the Decision Tree model trained using the US data alone, with a maximum level of three: \( L = 3 \). The nodes (boxes) are color-coded by level \( L \), where blue, orange, and green denote \( L = 1, 2, 3 \), respectively, for the internal nodes, and the leaf nodes (terminal boxes) are color-coded in red. In the internal nodes, the first row shows the splitting criterion, \( v \leq v^* \), where \( v \) and \( v^* \) denote the splitting variable and its threshold, respectively. The unit is km for Moho depth, and km/s for crustal and uppermost mantle \( V_s \). The standardized threshold values are shown in parentheses. The second row presents the predicted heat flow \( \hat{q} \) and the percentage of the total data included in the prediction. The two arrows below an internal node represent a split of the parent data into two parts, where the left and right arrows denote the satisfaction \( (v \leq v^*) \) and violation \( (v > v^*) \) of the splitting criterion, respectively. In the seven leaf nodes, only the predicted heat flow \( \hat{q} \) is shown. Each leaf node yields a decision rule and is identified with a letter label (a)–(g). These seven heat flow values (79, 76, 70, 62, 60, 53, 49 mW/m\(^2\)) constitute the predicted map of heat flow (Fig. 13c). These decision rules are further illuminated in Fig. 12.
Figure 12: Decision rules of the US-trained Decision Tree model in Fig. 11. Each column denotes a decision rule (a unique path from the root node to one leaf node), ordered by decreasing heat flow predictions $\hat{q}$ from left to right. Each row represents one of the three seismic structural variables, increasing with depth from top to bottom. The statistical distributions of the variables are shown as black histograms. For each decision rule, the 3-D seismic variable space is successively split by the value ranges of each of the three variables, with the order of splitting color-coded in blue, orange, and green for $L = 1, 2, 3$ (section 4.2.2), which is consistent with the node colors in Fig. 11. Note that the distributions are often split at valleys or peaks, which tend to correlate with geological boundaries.
Figure 13: Increasing complexity of the Decision Tree models as the number of levels ($L$) is increased. Heat flow predictions are plotted for Decision Tree models with maximum levels of (a) 1, (b) 2, and (c) 3. The unique advantage of the Decision Tree model is that increased complexity with levels tends to split the whole area into regions corresponding with geological provinces (despite no such geographical information given), and it can yield a specific relationship between heat flow and seismic structure for each region.
Figure 14: Decision tree model trained using the European data alone. Similar to Fig. 11 except that the heat flow data and seismic structural information are from Europe. The decision rules (labeled (a)–(h)) are illustrated in Fig. 15.
Figure 15: Decision rules of the Europe-trained Decision Tree model in Fig. 14. Similar to Fig. 12 except for the European data alone. Note that the Moho depth plays a more important role than in the US-trained Decision Tree model (Fig. 12).
Figure 16: Increasing complexity of the Decision Tree models as the number of levels (L) is increased using the European data alone. This figure is similar to Fig. 13 except for the European data. This figure should be compared with observations in Fig. 2d, and applying the US-trained models in Fig. 21. Note that geologically relevant provinces are successively identified by growing the tree.
Figure 17: **Importance of seismic structural variables from the Decision Tree (DT) models.** Normalized variance reduction (Gini importance) is shown for the Decision Tree models based on the US data alone (blue; Fig. 11) and the European data alone (orange; Fig. 14). Uppermost mantle $V_S$ is the most important and the Moho depth is more important for the European data than the US data. These findings agree with the Linear Regression coefficients (Fig. S3).
Figure 18: Comparison of model transportability by cross-validation between the US and European data. (a) The $r^2$ (overall fit, eq. (9)) and $\rho$ (geographical coherence, eq. (10)) are shown, applying the US-trained models to the European data. The $r^2$ of the Random Forest (RF) model is negative, i.e., worse than using the average European heat flow as a constant predictor. The Linear Regression and Decision Tree models are comparable and significantly more transportable than the Random Forest model. (b) Similar to (a) but for applying the Europe-trained models to the US data. The relative performance between the models is similar to applying the US-trained models to the European data. This figure should be contrasted with Fig. 7 where data are within a continent.
U.S. → Europe

Figure 19: Comparison of model transportability by applying the US-trained models to European data. Similar to Fig. 10, except the models are trained with the US data. This figure should be contrasted with the predictions of the Europe-trained models on the European data in Fig. 10. The geographic patterns of heat flow are reasonably reproduced, but the absolute amplitudes are not.
Figure 20: Comparison of model transportability by applying the Europe-trained models to the US data. Similar to Fig. 8, except the models are trained with the European data. This should be contrasted with the predictions of the US-trained models on the US data in Fig. 8. The geographic patterns of heat flow are reproduced better than the absolute amplitudes, both are better reproduced than applying the US-trained models to Europe (Fig. 19).
Figure 21: Extrapolating the US-trained Decision Tree models to the European data. This figure is similar to Fig. 13, but the models are trained with the US data (Fig. 11). This should be contrasted with the predictions of the Decision Tree models trained from the European data (Fig. 16). The geographic patterns of heat flow are better reproduced with increasing the level $L$, but the absolute amplitudes are under-predicted.
Figure 22: Extrapolating the Europe-trained Decision Tree models to the US data. Similar to Fig. 21 but the models are trained with the European data (Fig. 14). The geographical patterns are similar to the predictions from the US-trained Decision Tree models, but the absolute amplitudes are over-predicted (Fig. 13).