ORIGINAL RESEARCH **Open access**

A Machine Learning Framework for Predicting Geothermal Heat Flow from Sparse Geological Data

Burak Demir[1*], Elif Arslan[2], Mehmet Kaya[1], Selin Kurt[2]

Abstract

In the appraisal of geothermal reservoirs, geothermal heat flow (GHF) is a vital parameter. Mounting interest surrounds the use of machine learning models to infer GHF from Kriging-generated geological datasets, thereby sidestepping the high costs associated with classical measurement procedures. Yet the spatial configuration of certain geological properties presents intricate data structures and gaps, underscoring the demand for a methodology that surpasses standard Kriging in both accuracy and efficiency. In the present contribution, we describe an innovative machine-learning-driven framework designed to infer GHF from scattered geological observations. More precisely, gaps in the geological records are filled through a machine learning algorithm—MissForest. By capitalizing on the spatial interdependencies among geological variables (such as upper crust thickness, Moho depth, and rock type), the MissForest approach achieves higher imputation performance (R^2 = 0.90 at a 20% missing rate) than classical Kriging (R^2 = 0.84). Machine learning regression models are subsequently fitted on the completed records to encode the functional dependence of GHF on geological predictors. Our leading model achieves a modest 10.18% error when predicting GHF across diverse settings, improving upon earlier reports. In addition, the framework yields credible global GHF estimates, offering fresh insights into the global distribution of geothermal endowments and their potential for exploitation.

Keywords Machine learning, Geothermal heat flow, Geothermal energy, Data imputation, AI for science

*Correspondence:
Burak Demir
burak.demir@metu.edu.tr
[1] Department of Artificial Intelligence in Electrical Engineering, Faculty of Engineering, Middle East Technical University, Ankara, Turkey
[2] Department of Smart Grid Intelligence, Faculty of Electrical Engineering, Istanbul Technical University, Istanbul, Turkey

Introduction

Recognized as a renewable energy source [1-3], geothermal resources provide notable constancy [4] and carry considerable weight in the pursuit of carbon peaking and neutrality [5-8]. Serving as a principal benchmark for geothermal potential, geothermal heat flow (GHF) quantifies the thermal flux conveyed from the planetary interior upward to the crustal surface [9, 10]. Its determination normally relies on the scrutiny of borehole-derived temperature profiles coupled with thermal conductivity data [11]. That procedure, however, incurs substantial expense and requires significant resources, leading to an irregular distribution of observational sites [12]. That irregular scattering, in turn, compels the search for a faithful and efficient means of estimating GHF across territory yet to be surveyed. Interpolation algorithms—most notably Kriging—currently predominate due to their operational simplicity [13]. Even so, their trustworthiness is conditioned by the spatial density of available measurements [14]. Where geological feature samples are thinly distributed, elementary interpolation rarely faithfully reconstructs subsurface characteristics and their spatial gradients at unsampled coordinates.

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

Over the past several years, machine learning methodologies have penetrated deeply into energy-domain applications, spanning building energy performance evaluation [15, 16], the identification of energy-relevant materials [17], demand forecasting for intelligent power grids [18], carbon profit estimation [19], network governance [20], and the tuning of ORC installations [21]. In contrast to established techniques, machine learning-based models carry the dual virtues of modest computational burden and freedom from a priori data assumptions [14]. This renders machine learning an attractive route for tackling GHF prediction. To illustrate, Rezvanbehbahani *et al*. [22] constructed a gradient-boosted regression tree (GBRT) predictor that achieved comparatively high fidelity (R^2 = 0.75) relative to a linear regression baseline (R^2 = 0.40). Their outcomes confirm that GHF can be reliably cast as a response to geological controls such as crustal thickness, magnetic anomaly fields, and gravity signatures. Aiming for further gains, later investigations have trialed an array of learning architectures—among them a coupled support vector machine-deep neural network (SVM-DNN) hybrid [13], random forests [23], and multi-layer perceptrons (MLP) [24]—ultimately documenting a well-calibrated model reaching R^2 = 0.79.

Although considerable headway has been achieved, two core obstacles persist in harnessing machine learning for GHF estimation. The first pertains to the design of an accurate gap-filling strategy. While efforts to date have relied heavily on publicly available geological compilations [12, 13, 22, 23], the reality of fragmentary coverage has frequently prompted the use of Kriging interpolation to complete data [22, 25]. Kriging is straightforward and practical, yet it struggles to replicate the signatures of structurally complex geological terrain, notably within mountainous belts and volcanic provinces, potentially eroding the fidelity with which sparse observations can reproduce the true data structure. The second obstacle concerns safeguarding the model's portability. Existing investigations remain largely anchored to individual nations or geographic sectors. Within the Chinese context, for example, machine-learning-driven GHF inference has been circumscribed to select basins and regions: the Songliao Basin [12], the Bohai Bay Basin [13], and the North China domain [11]. Accordingly, there is a clear impetus to develop a GHF forecasting tool with strong generalizability and suitable for real-world industrial deployment.

In the present study, we advance a novel machine-learning-centered framework for GHF estimation, as schematically portrayed in **Figure 1.** Drawing upon publicly accessible sources, we first assembled a compendium comprising thousands of entries that pair GHF observations with 23 geological covariates. From this compendium, machine learning imputers are deployed to reconstruct missing feature values, outperforming the classical Kriging procedure. Using the filled dataset, machine learning regressors are then tuned to approximate the deterministic linkage between the geological covariates and GHF. Ultimately, the fully trained machine learning architectures are exercised to credibly extrapolate global geothermal heat flow to geographic positions absent from the original archive.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

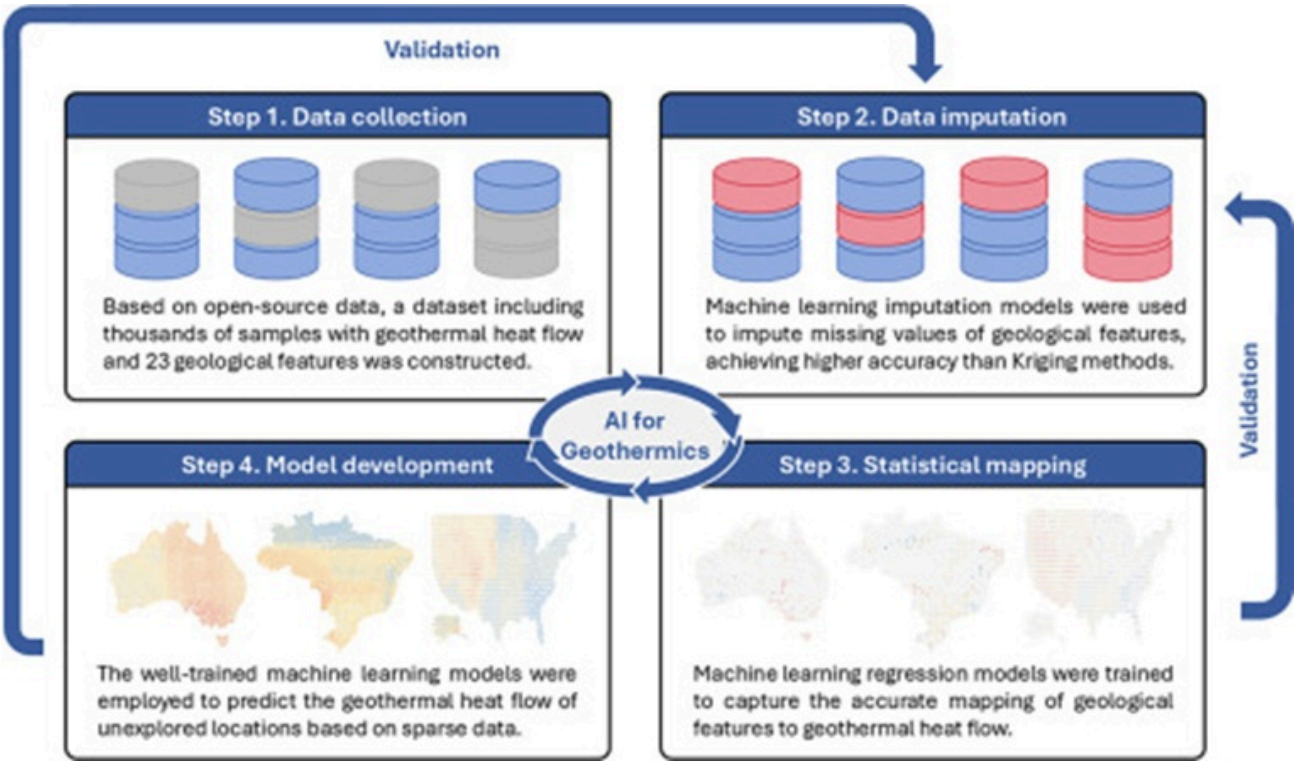

Figure 1. Machine learning-based framework for predicting geothermal heat flow. The featured framework unfolds across four phases: data gathering, data completion, statistical mapping, and model construction, embodying the workflow of "AI for Geothermics".

Data

Geothermal heat flow

The GHF compilation analyzed here was sourced from Rezvanbehbahani *et al*. [22]. After purging extreme outliers—defined as values generally above 200 mW/m²—the raw GHF observations were resampled to a 0.5° × 0.5° grid, producing a total of 4016 recorded GHF points [23, 26]. In terms of statistics, the measurements range from 16.95 mW/m² to 183.96 mW/m², with a mean of 60.73 mW/m² and a standard deviation of 18.11 mW/m². As illustrated in **Figure 2a,** territories with comparatively low heat flow are clustered in Central and Northern Asia and in the eastern portion of North America. By contrast, elevated flux signatures characterize western North America, Europe, and Northern Africa, and a minor cluster is also evident within Central Australia. Turning to the United States, **Figure 2b** indicates that heat flow is relatively subdued across the Southeast and intensifies towards the Northwest, reaching particularly high magnitudes in the vicinity of Nevada, Utah, and Colorado. In Brazil, GHF levels are somewhat higher across the southern, northeastern, and western regions. In contrast, it stays comparatively low across the interior and northern domains—prominently the Amazon basin—as conveyed by **Figure 2c. Figure 2d** shows that across Australia, the more energetic GHF signals are positioned in eastern and central zones, in contrast to the weaker signatures that dominate the west, collectively revealing an overarching west-to-east intensification.

Furthermore, **Figure 2** highlights the geographic footprint of the GHF record. The global coverage is markedly uneven, featuring dense clusters of measuring locations in Europe and the south-central band of North America, while South America, Africa, and Oceania remain thinly represented. More specifically, throughout the USA, observations are relatively plentiful in the west and south but far more scattered in the northeastern quarter. In Brazil, sampling sites are concentrated in the southeast and the northern fringe, leaving the central belt with sparser representation. Over Australia, the data coverage is, on the whole, quite thin, with somewhat greater density hugging the shoreline and fewer points

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

penetrating the interior. This spatially irregular character of the GHF observations underscores the demand for efficient and dependable imputation methodologies, an issue taken up later in the discussion.

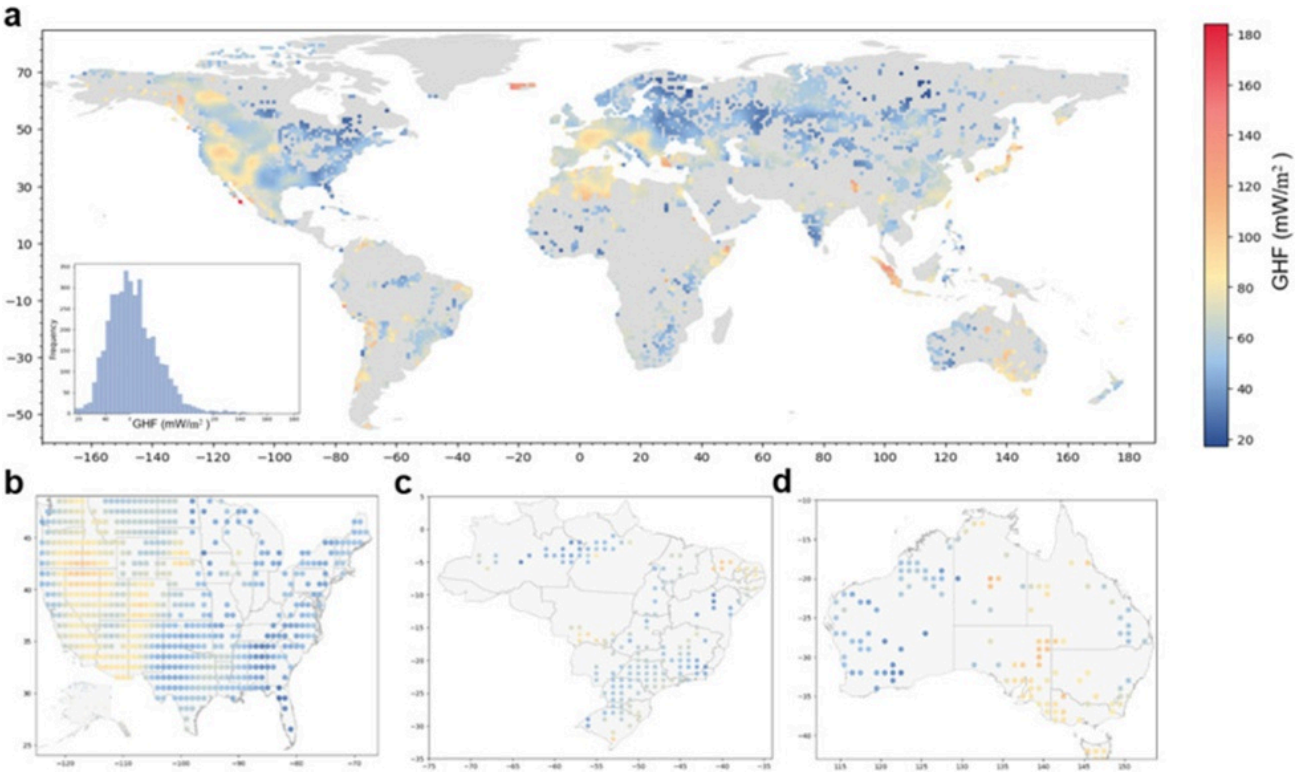

Figure 2. The distribution map and frequency distribution histogram of GHF measurements. (a) Global GHF distribution. (b) GHF distribution in certain regions of the United States. (c) GHF distribution in Brazil. (d) GHF distribution in Australia. The concentration of points indicates the density of actual sampling locations, while the color of the points represents the GHF values at those locations.

Geological features

The birth of geothermal systems is deeply intertwined with the assembly, modification, and reworking of the lithosphere, along with the thermal state and heat-transport history of nearly the entire crustal column, and responds acutely to tectonic stress configurations and subsurface temperature fields [27]. Moreover, GHF magnitudes are tied to numerous geological and geophysical parameters, including crustal thickness, depth undulations, gravitational anomaly patterns, and radiogenic heat generation. GHF can likewise be approximated through geophysical proxies or physical rock-property measurements [12]. Drawing on the contributions of Rezvanbehbahani *et al*. [22] and Al-Aghbary *et al*. [23], 23 geological descriptors harvested from multiple sources were adopted as the independent variables (labeled X for ease of reference in the analyses that follow). At the same time, GHF was assigned as the dependent variable (labeled y), as summarized in **Table 1.**

Table 1. Mathematical representation and data source of the input and output variables.

Mathematical representation	Input/Output variable	Data source	Data type
X1	Timing of the most recent thermotectonic event	[28]	Categorical
X2	Upper crustal thickness	[29]	Continuous

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

X3	Lower crustal thickness	[29]	Continuous
X4	Density anomaly in the upper mantle	[30]	Continuous
X5	Distance from mid-ocean ridge	[31]	Continuous
X6	Distance from subduction trench	[31]	Continuous
X7	Distance to transform fault ridge	[31]	Continuous
X8	Distance to nearest five volcanoes	[32]	Continuous
X9	Distance to hotspot location	[33]	Continuous
X10	Distribution of heat-producing provinces	[32]	Continuous
X11	Total crustal thickness	[34]	Continuous
X12	Proximity to young rift zones	[35]	Continuous
X13	Structure of upper mantle seismic velocity	[36]	Categorical
X14	Lithology (rock classification)	[37]	Categorical
X15	Geological age	[38]	Continuous
X16	Magnetic field anomaly	[39]	Continuous
X17	Moho discontinuity depth	[40]	Continuous
X18	Lithosphere–asthenosphere boundary depth	[41]	Continuous
X19	Bouguer gravity anomaly	[42]	Continuous
X20	Global surface elevation/topography	[43]	Continuous
X21	Geoid height variation	[44]	Continuous
X22	Bouguer gravity anomaly (alternative dataset)	[44]	Continuous
X23	Free-air gravity anomaly	[44]	Continuous
y	Geothermal heat flow	[45]	Continuous

Note: Following classical mathematical convention, this study uses X to denote the input variables and y to denote the output variable.

The geological sampling coordinates used here span the entire planet and exhibit a highly non-uniform spatial distribution. Much like the GHF distribution described above, the worldwide geological attributes likewise exhibit considerable heterogeneity. It is noteworthy, however, that while the continuous-type variables **(Figures 3a and 3b)** retain a normal-like spread reminiscent of GHF, this distributional form is not clearly mirrored among the categorical-type variables **(Figures 3c and 3d)**. Securing improved prediction accuracy requires preprocessing categorical features. In addition, the assembled dataset brings together diverse feature modalities that exhibit no obvious spatial correlation structure, thereby posing substantial interpretive challenges.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

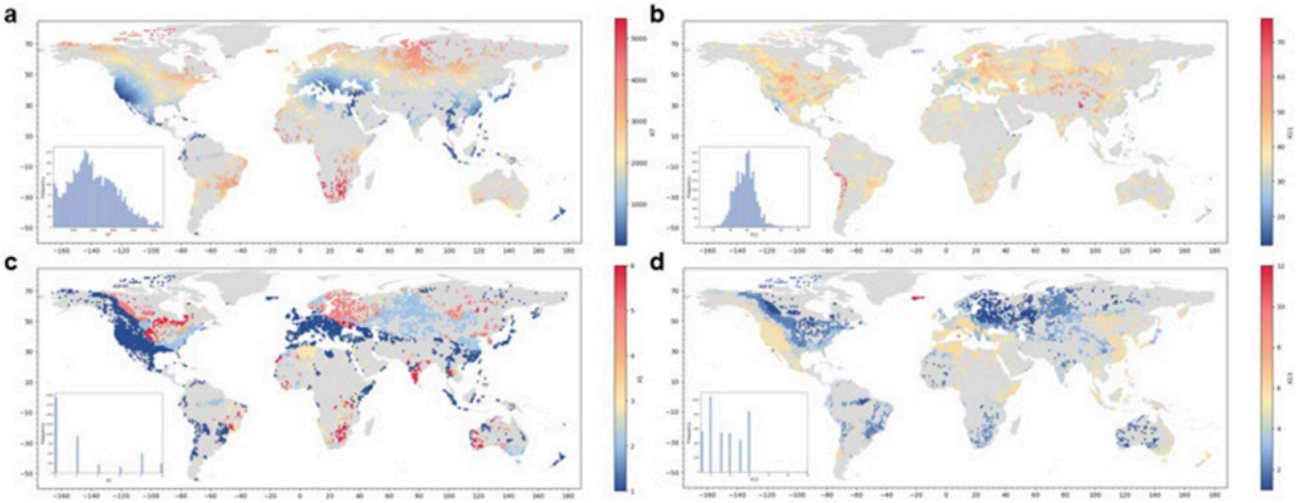

Figure 3. The distribution map and frequency distribution histogram of geological features: (a) distance to transform ridge (X7, continuous type), (b) crustal thickness (X11, continuous type), (c) age of last thermotectonic event (X1, categorical type), and (d) upper mantle velocity structure (X13, categorical type). The concentration of points indicates the density of actual sampling locations, while the color of the points represents the feature values at those locations.

Materials and Methods

Missing data imputation models

Loss and inaccuracy of data are persistent challenges in geological data assembly, stemming from discrepancies in spatial resolution and coordinate systems across disparate sources [46]. Previous investigations relied heavily on the Kriging interpolation formalism to address these gaps; yet, a comparative assessment and deliberate tuning of the key parameter (e.g., the semi-variogram) within the Kriging workflow remain lacking, and alternative infilling approaches outside the Kriging family have not been systematically explored. The present work juxtaposes the interpolation effectiveness of the Kriging paradigm using four separate semi-variogram structures (linear, power, spherical, and exponential). It also appraises two machine-learning-driven imputation algorithms (K-nearest neighbors and MissForest) alongside the ordinary Kriging interpolation technique, using R^2 and RMSE as evaluation criteria.

Kriging

Kriging interpolation is a commonly used method for reconstructing missing geological feature entries, first developed in the 1960s [47, 48]. The method primarily operates in the spatial domain and does not draw on auxiliary external variables. Rather, it deduces a value at a predetermined location within a region by leveraging a pre-characterized semi-variogram and observations from nearby sites [49-52]. In its typical implementation, Kriging interpolation starts by calculating the separation length and the semi-variance between each pair of measurement locations, then fitting a semi-variogram model that captures the functional relationship between distance and semi-variance. The distances separating the prediction target from the known sites are then determined, and the corresponding semi-variances are inferred through the fitted model. Optimal weighting coefficients are subsequently resolved on the strength of these semi-variances. Last, the unknown-site estimate is formed as the weighted linear combination of the neighboring observations. The predicted quantity at position x_i is expressed as:

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

$$p_i = \sum_{i=1}^{n} \lambda_{i,} y_j \quad (1)$$

Where n is the number of neighboring points; $\lambda_{i,j}$ is the weight point i assigned to neighboring point j, which is determined by the semi-variogram; y_j is the measured value at neighboring point j.

Selection of the semi-variogram model is a decisive step, as it fundamentally shapes the weighting scheme within the Kriging framework and, by extension, the fidelity of the interpolated surfaces [53]. Ordinary Kriging is broadly recognized as an optimal linear unbiased estimator. In the present study, four distinct semi-variogram models—linear, power, spherical, and exponential—are applied to assess the ability of ordinary Kriging to fill data gaps.

K-nearest neighbors

As a prototypical data-driven strategy, K-nearest neighbors (KNN) reconstructs missing entries at a given location by forming a weighted average of the contributions from its k most proximal neighbors [54, 55]. Concretely, the algorithm begins by measuring the separation between the query instance and every other instance in the dataset, using metrics such as Euclidean or Manhattan distance. With a user-supplied k value in hand, the model then isolates the k samples closest to the query, thereby reducing the distances from the randomly selected set of k neighbors to the target. The final estimate for the missing value is obtained by taking a weighted mean across the k neighboring observations. There are conceptual overlaps between the KNN approach and the Kriging interpolation framework insofar as both techniques estimate a value at an unmeasured coordinate by weighting nearby points. Where they fundamentally diverge, however, is in how neighbor proximity is conceptualized. Within the Kriging paradigm, distance is purely a matter of geographic coordinates, determined solely from latitude and longitude. KNN, by comparison, uses the entire set of available features to quantify the distance between any two records. The weighting philosophy, too, departs sharply: rather than letting the semi-variogram dictate the weights as Kriging does, KNN adopts the inverse of the computed distance as the weighting factor. The expression utilized in this investigation is as follows:

$$\lambda_{i,j} = \frac{1}{d_{ij}} \quad (2)$$

Where $\lambda_{i,j}$ is the weight point i assigned to neighboring point j; $d_{i,j}$ indicates the distance from point i to neighboring point j. Because KNN imputation reliably preserves local data patterns while offering flexibility, adaptability, and straightforward interpretability, it has found broad adoption in data interpolation contexts [56, 57]. For this study, the neighborhood size was configured at 5.

MissForest

MissForest operates as an iterative gap-filling mechanism rooted in the random forest algorithm [58]. At the outset, every vacant slot is provisionally populated with the mean (for numeric variables) or the mode (for categorical ones), after which the column m containing the fewest missing entries is identified. That column m is then cast as the dependent variable, with the other columns serving as independent variables for those rows where column m holds genuine readings, and a random forest is trained on this subset. The fitted forest is subsequently used to overwrite the placeholder values initially inserted in column m. The routine then proceeds to the column with the next-smallest count of missing observations, designates it as the new dependent variable, and repeats the cycle until the convergence criteria are satisfied [59]. Cast

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

as a spatial imputation tool, random forests impose only light data preparation demands and remain comparatively unfazed by aberrant values, attributes that render them effective for high-dimensional datasets [60].

Machine learning regression models

With the fully populated dataset at hand, machine learning regression models were pressed into service to capture the deterministic linkage through which the independent variables (geological descriptors) map onto the dependent variable (geothermal heat flow). To home in on an appropriate regression architecture, the present work juxtaposes five candidate algorithms—LGBM, SVR, linear regression, random forest, and XGBoost—each receiving dedicated treatment in the ensuing paragraphs. Concurrently, a hyperparameter optimization suite (Optuna) was employed to identify the most performant parameter configurations, and five-fold cross-validation was used to mitigate overfitting during training [61, 62].

Linear regression

Linear regression ranks among the most foundational instruments in the analytical toolkit. The method formulates the response as a linear combination of the predictors and tunes it by minimizing the sum of squared differences between observed and predicted quantities [63]. The prediction is rendered as:

$$\hat{y} = \hat{w}^T X \quad (3)$$

Where $\widehat{y}$ is the predicted value; $\widehat{w}$ is the p-dimensional model coefficient vector. In the present case, model parameters are estimated via the least squares principle, specified as:

$$\hat{w} = \mathrm{argmin}_w \| y - Xw \|_2^2 + \lambda P(w) \quad (4)$$

Where X is the matrix of features; P(w) depends on the specific regularization technique used.

Support vector regression

Support vector regression (SVR) broadens the Support Vector Machine framework. Its central tenet is to erect a family of hyperplanes in a high-dimensional embedding such that the largest gap between the hyperplane and any training sample is kept as small as possible [64]. The SVR model's efficacy is acutely sensitive to the configuration of its hyperparameters, the nature of the loss function, the choice of kernel, and the magnitude of the error penalty term [65]. SVR has demonstrated noteworthy proficiency in time-series and nonlinear estimation tasks and has been fruitfully applied across a spectrum of engineering problems [66]. The regression equation is given by:

$$\hat{y} = \sum_{i=1}^{n} a_i K(x, x_i) + b \quad (5)$$

Where $\hat{y}$ denotes the predicted value; a_i is the Lagrange multipliers; $K(x, x_i)$ is the kernel function; b is the bias term; n is the number of training samples. In this study, the adopted kernel is the radial basis function (RBF), delineated as:

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

$$K(x, x_i) = \exp\left(- \frac{\|x - x_i\|^2}{2\sigma^2} \right) \quad (6)$$

Where σ denotes the standard deviation; $\|x - x_i\|^2$ is the Euclidean distance between x and x_i.

Random forest

Random forests constitute an ensemble learning framework composed of numerous decision trees. For each tree segment, split the data at every internal node by selecting the predictor and splitting threshold that yield the greatest information gain. It continues recursive partitioning until a preset maximum depth is reached, the data are fully exhausted, or some other manually imposed stopping condition is met. When presented with an unfamiliar instance, the tree traces a path beginning at the root and descending until a terminal leaf is reached; that leaf's value is then emitted as the tree's prediction. The random forest construct can generate hundreds or even thousands of such trees, each acting as a distinct regression function, and ultimately delivers the simple average of all individual tree outputs with the dual ambition of lifting predictive accuracy and reining in overfitting [67]. By pooling the outputs of diverse trees, random forests can shrink variance, though a trade-off in the form of somewhat inflated bias can occasionally surface. The training protocol for random forests involves random sampling of the data, which means that certain observations may appear multiple times during fitting, while others may be entirely omitted [68]. The random forest paradigm is comparatively tolerant of outlying values and can handle large numbers of input variables with a modest parameter load and without requiring predictor dropping. However, it does entail greater model complexity and longer training times [65].

Gradient boosting regression trees

Both XGBoost and LightGBM serve as instantiations of the gradient boosting regression trees (GBRT) paradigm. GBRT follows a forward-accumulating ensemble logic, where new base learners are appended sequentially along the gradient direction to chip away at the remaining residuals [13]. The GBRT framework has been corroborated in several GHF regression inquiries [11, 22, 25]. XGBoost, a tree-based ensemble architecture, is widely endorsed as a go-to option for predictive modeling on tabular data [69]. In contrast to random forests, XGBoost augments the ensemble incrementally, correcting deficiencies left by the current set of base models by incorporating additional base learners, thereby boosting overall model accuracy [70]. LightGBM is another tree-centric gradient boosting framework that combines gradient-based one-sided sampling and exclusive feature bundling into the GBRT workflow [71]. Unlike the depth-first expansion strategy used by XGBoost, LightGBM proceeds via a leaf-wise growth heuristic. LightGBM places lighter demands on memory and achieves higher training throughput, offering distinct advantages for large-volume datasets.

Analytical procedures

Measures of predictive accuracy

For this work, the source dataset was split into training and test sets at 6:4. Four indicators were used to assess the models' predictive performance (R^2, n-RMSE, MAE, MAPE). These indicators are extensively adopted across the machine learning domain [13, 22], and their formulas are given as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \quad (7)$$

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

$$\mathrm{n-RMSE} = \frac{n}{\sum_{i=1}^{n} y_i} \times \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}} \quad (8)$$

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n} \quad (9)$$

$$MAPE = \left(\frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \right) \times 100\% \quad (10)$$

Where y_i stands for the observed GHF; $\hat{y}_i$ stands for the forecasted GHF; $\overline{y}$ stands for the average of all observed GHF readings; $\sum_{i=1}^{n} y_i$ stands for the aggregate of all observed GHFs; n stands for the quantity of output observations.

Measures of association

Spearman's correlation coefficient provides a robust and versatile means of examining monotonic relationships in data that are both nonlinear and non-normally distributed. Owing to the nonlinear properties of the dataset under study, this coefficient was chosen to quantify association strength, with the following formulation:

$$\rho_{X,\,Y} = \frac{\mathrm{cov}(X,\,Y)}{\sigma_X \sigma_Y} \quad (11)$$

Where $\rho_{X,\,Y}$ denotes the Spearman's rank correlation coefficient; cov(X, Y) denotes the covariance of the rank-transformed variables; σ_X and σ_Y denote the standard deviations of the rank-transformed variables.

Findings and Interpretation

Data exploration and statistical summaries

Scatter plots, heatmaps, and bar graphs were used to probe the relationships between single geological attributes and GHF. As depicted in **Figure 4**, the degree of association between individual geological attributes and GHF is broadly low. A straightforward linear regression fails to capture this interrelation **(Figure 4a).** Even though a few attribute pairs display robust mutual ties (e.g., X2 and X3), most lack linear association. **Figures 4b and 4c** provide further depiction of Spearman's rank correlation assessment. The results indicate low correlations among individual geological attributes and no notable correlation between the attributes and GHF. Therefore, machine learning algorithms are necessary to extract hidden structures.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

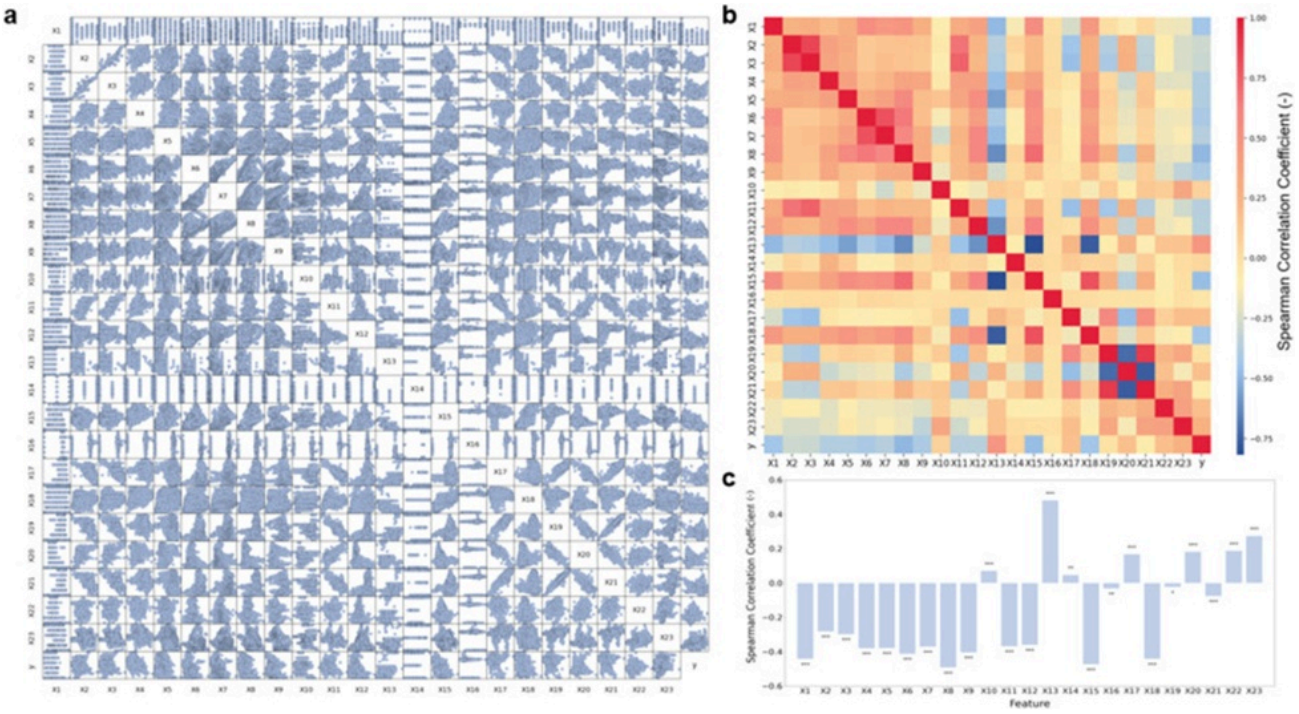

Figure 4. Association analysis between geophysical attributes and GHF. (a) Scatter plots among single geophysical attributes and between geophysical attributes and GHF. (b) Correlation heatmap among geophysical attributes and GHF. A shift toward yellow tones signals a weakening correlation. (c) Histograms of correlations between single geophysical attributes and GHF. Within this chart, stars mark the statistical significance thresholds of Spearman correlation outcomes, with additional stars implying higher significance.

Beyond that, while the overarching correlation between GHF and the various attributes does not reach statistical significance, a finer look at **Figure 4b** discloses that GHF is, for the most part, inversely related to most attributes. Specifically, it shares a degree of positive correlation with X13, X17, X20, X22, and X23, and an inverse correlation with X1, X8, X15, and X18. These patterns are further quantified and visualized in **Figure 4c.** Additionally, the correlation coefficients linking GHF to X16 and X19 are exceedingly small, approaching zero.

Imputation of absent geological attributes

Prior investigations have not undertaken a systematic comparison or selection of semi-variograms within Kriging-based interpolation frameworks. This study adopts four separate semi-variogram formulations to calibrate the Kriging interpolation scheme. Before error computation, column-wise normalization is performed to offset the influence of disparate scales among geological attributes. The outcomes **(Tables 2 and 3)** reveal that the choice of semi-variogram markedly shapes the Kriging interpolation results. Among the candidate models, the spherical variant yields the best performance, with test-set RMSE consistently below 0.11. The linear variant, by contrast, registers comparatively weaker outcomes. Such divergence may stem from the spherical formulation's prioritization of proximal point weighting, which amplifies the contribution of nearby known observations to missing-value estimates relative to the linear formulation.

Table 2. Imputation accuracy expressed via R2, reported as mean ± standard deviation.

Algorithm	10 %	20 %	30 %	40 %

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

KRIGING (linear)	0.6233±0.0047	0.6333±0.0125	0.6133±0.0047	0.6167±0.0094
KRIGING (power)	0.6667±0.0170	0.6733±0.0047	0.6433±0.0125	0.6467±0.0047
KRIGING (spherical)	0.8533±0.0094	0.8433±0.0125	0.8500±0.0082	0.8400±0.0000
KRIGING (exponential)	0.8000±0.0141	0.7633±0.0047	0.7733±0.0094	0.7600±0.0082
KNN	0.9067±0.0047	0.9000±0.0000	0.8867±0.0047	0.8300±0.0000
MissForest	0.9167±0.0047	0.9033±0.0047	0.8933±0.0047	0.8800±0.0000

Table 3. Imputation accuracy expressed via RMSE, reported as mean ± standard deviation.

Algorithm	**10 %**	**20 %**	**30 %**	**40 %**
KRIGING (linear)	0.1578±0.0013	0.1559±0.0027	0.1595±0.0014	0.1586±0.0022
KRIGING (power)	0.1487±0.0031	0.1466±0.0005	0.1528±0.0019	0.1531±0.0013
KRIGING (spherical)	0.0986±0.0028	0.1025±0.0042	0.0993±0.0018	0.1027±0.0006
KRIGING (exponential)	0.1149±0.0034	0.1254±0.0023	0.1215±0.0028	0.1253±0.0023
KNN	0.0785±0.0021	0.0810±0.0004	0.0858±0.0013	0.1063±0.0004
MissForest	0.0750±0.0010	0.0789±0.0012	0.0841±0.0016	0.0890±0.0010

As addressed in **Figure 5**, stacking ordinary Kriging interpolation under its optimal configuration (spherical model) against two machine learning alternatives (KNN and MissForest) demonstrates that both machine learning approaches outclass Kriging interpolation. The KNN technique bears a partial resemblance to Kriging interpolation, as it, too, resolves missing outputs by averaging neighboring points based on distance. The two diverge, nonetheless, in how they compute distances and assign weights. Whereas Kriging interpolation relies exclusively on latitude and longitude to fix positions, KNN draws upon the full set of available attributes when appraising locations. Doing so permits a more nuanced recognition of geological contrasts among far-flung data points, thereby remedying the deficiencies that imputation methods exhibit in sparsely sampled zones. The edge MissForest has lies in its ability to handle high-dimensional inputs and perform recursive refinement, resulting in superior imputation fidelity. Relative to the baseline Kriging approach (spherical function, 20% missingness), KNN cuts the RMSE by 0.0215 (20.98%), and MissForest cuts it by 0.0236 (23.02%)—both representing meaningful gains.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

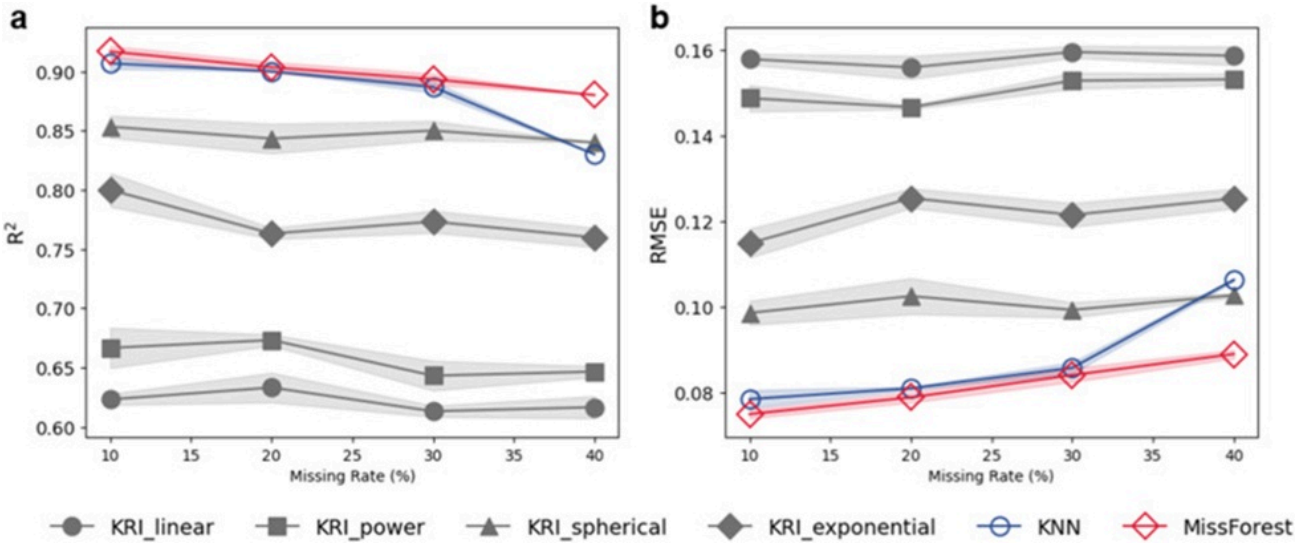

Figure 5. Imputation accuracy gauged through R2 (a) and RMSE (b) for Kriging-based methods, KNN, and MissForest across varying input-data missingness levels: 10 %, 20 %, 30 %, and 40 %. The solid trace marks the mean, while the shaded band marks the standard deviation.

Accordingly, the MissForest algorithm was applied to impute missing entries, yielding a complete dataset at a 20% missingness rate. These infilled records then underpinned the subsequent predictive modeling. As shown in **Figure 6**, the newly synthesized portions clearly conform to the distributional patterns of the original attributes. Scatter plots juxtaposing genuine and synthetic readings confirm that the synthetic values faithfully reproduce the genuine readings. With respect to attributes such as X5, X9, X15, and X23, the imputation outputs post-R2 values in excess of 0.95, further corroborating MissForest's ability to reconstruct the absent segments of geological descriptors faithfully.

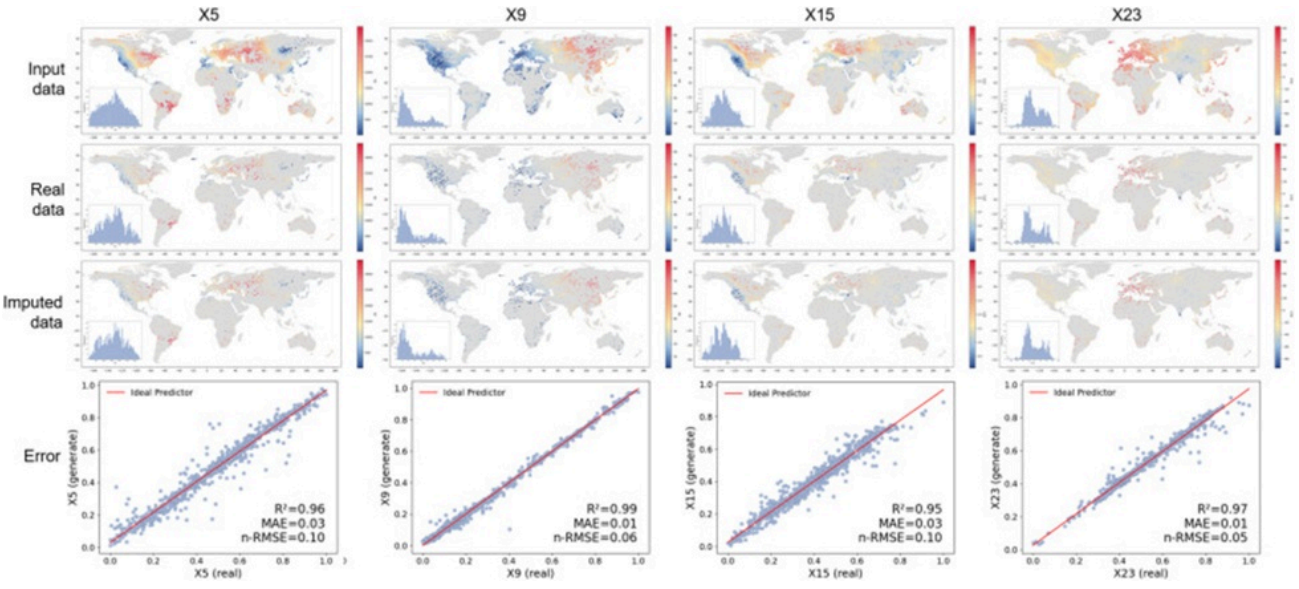

Figure 6. MissForest interpolation outcome on geological attributes at 20 % missingness, illustrated through attributes X5, X9, X15, and X23. Map pixel hues represent the attribute magnitude at the respective location.

Estimation of geothermal heat flow

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

The fully infilled dataset, encompassing 23 geological attributes and geospatial coordinates (longitude and latitude), served as the input variable suite for the GHF forecasting model to improve estimation precision. This work recognizes that, although GHF variation along a fixed longitude (or latitude) does not boil down to a purely linear function, geographic relative positioning nonetheless exerts considerable sway over GHF distribution patterns. The estimation outputs indicate that, compared with linear regression (R2 = 0.40, n-RMSE = 0.23), the remaining four machine learning models all deliver stronger predictive fidelity for GHF. In particular, LGBM (R2 = 0.75, n-RMSE = 0.15) and XGBoost (R2 = 0.74, n-RMSE = 0.15) produce the most faithful estimates. The closeness of their outputs likely reflects the fact that both algorithms derive from the GBRT architecture and thus operate through comparable predictive logic **(Figure 7).**

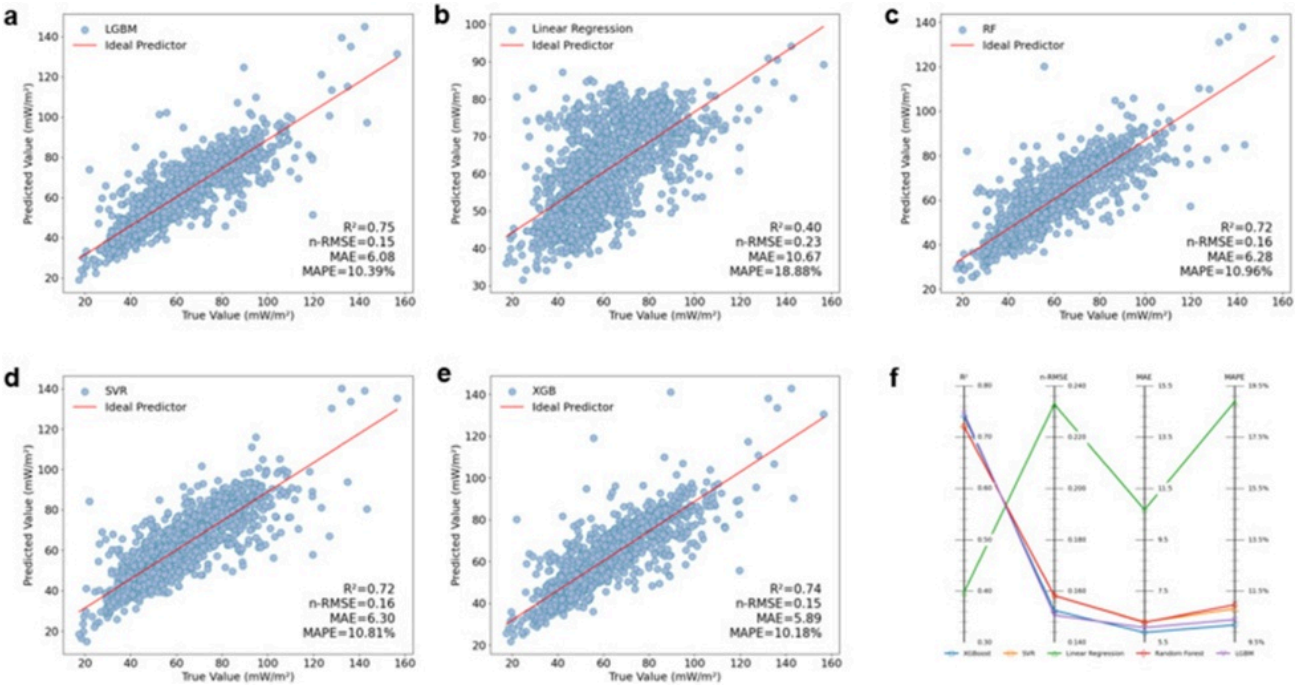

Figure 7. Estimation accuracy of LGBM (a), linear regression (b), random forest (c), support vector machine (d), and XGBoost (e). (f) Side-by-side comparison of peak accuracy across the different algorithms.

Spatial mapping of GHF across targeted regions

The central aim of this segment is to assess whether the model can reliably project localized GHF with only a modest volume of regional training data. The assembled dataset spans territories across Australia, Brazil, and the United States, providing broad continental coverage alongside marked contrasts in sampling density, thereby considerably bolstering the referential value of the investigation. Deploying the best-performing model configuration (XGBoost) on these geographically segmented data reveals a clear pattern: predictive accuracy peaks in zones characterized by denser measurement networks, exemplified by the United States (R^2 = 0.82, n-RMSE = 0.10), while still maintaining respectable fidelity under sparse sampling regimes, as evidenced by the Australian case (R^2 = 0.70, n-RMSE = 0.16).

Translating the XGBoost model's output into mapped visualizations **(Figure 8)** shows that the algorithm has successfully replicated GHF's underlying spatial architecture. The model projections suggest that, within the randomly partitioned test set examined in this work, Brazil exhibits a pattern of relatively elevated GHF concentrated in three principal zones: the southernmost reaches, the western flank, and the northeastern corridor—with particular prominence in the southern highland complex, the Andean-adjacent western margins, and portions of the northeastern coastal strip—while comparatively depressed values characterize the country's interior and northern expanse, above all across the Amazon

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

Basin and neighboring districts. Turning to Australia, the model recovers a spatial gradient in which GHF is higher across the eastern and central interior and declines toward the western margin (Western Australia), consistent with an overarching west-to-east intensification trend. Regarding the United States, GHF readings remain relatively subdued across southeastern sectors and systematically rise toward northwestern latitudes, with markedly elevated measured values clustered around the tri-state region encompassing Nevada, Utah, and Colorado. Taken collectively, the XGBoost model demonstrates the ability to capture the macroscale GHF distribution morphology across all three national contexts, with performance scaling positively with measurement density. Of the three countries evaluated, the United States yields the strongest predictive metrics, achieving R^2 and n-RMSE figures of 0.82 and 0.10, respectively.

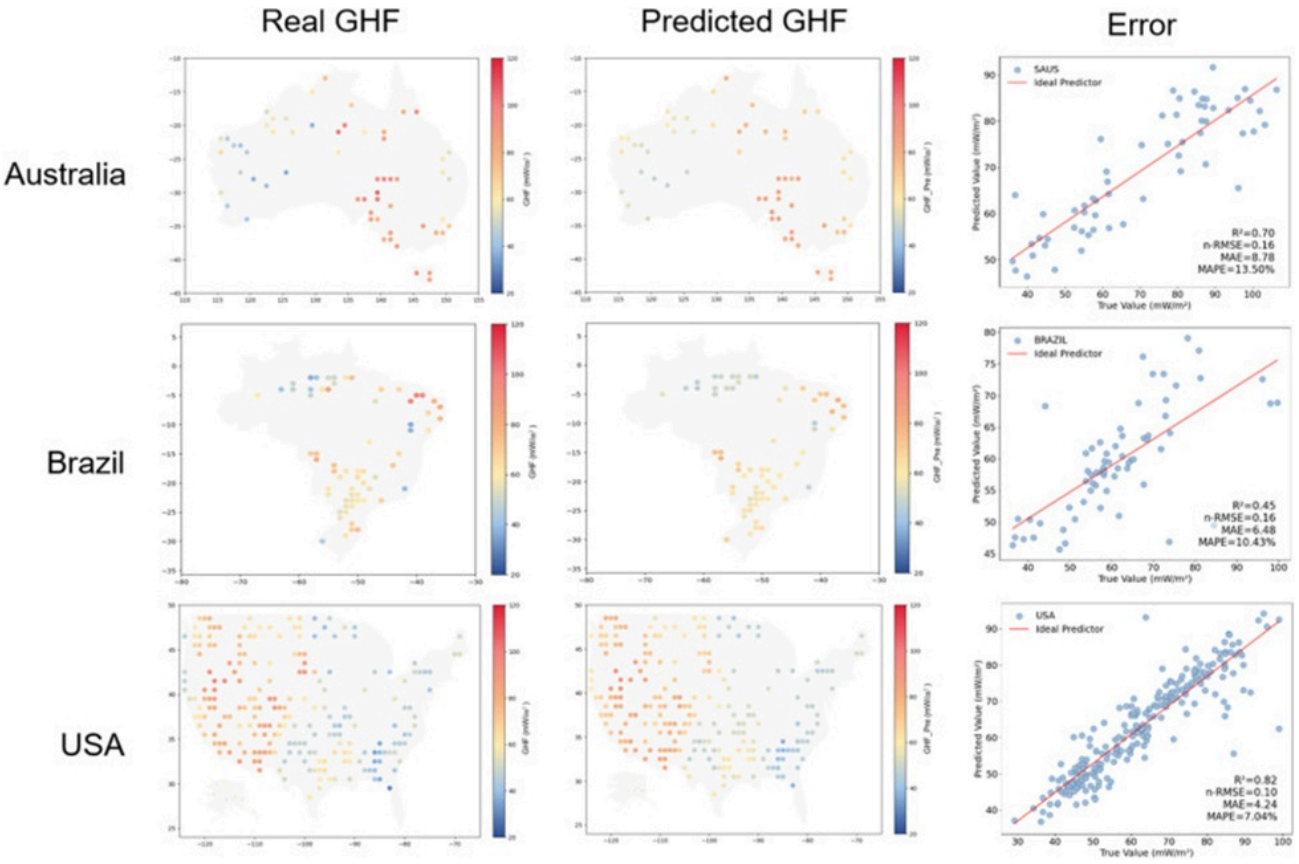

Figure 8. Spatial juxtaposition of observed and predicted GHF across Australia, Brazil, and the USA. The scatter plots depict the degree of alignment between recorded readings and model-generated estimates.

Leveraging a compact subset of predictor variables (X21, X22, X23) in conjunction with geospatial coordinates (latitude and longitude), synthetic records were generated to augment and complete the source dataset. These synthesized data, combined with the previously trained XGBoost architecture, were subsequently used to extrapolate GHF estimates across broad spatial extents, deliberately omitting locations where direct measurements were available. The resulting predicted value surfaces enable visualization of continental-scale tendencies in GHF distribution.

As shown in **Figure 9**, the model credibly reproduces Australia's characteristic longitudinal progression of GHF from west to east, with comparatively lower values in Western Australia. Likewise, it demonstrates competence in recovering Brazil's defining GHF spatial signatures, namely suppressed values across northern latitudes juxtaposed against amplified readings throughout the southern highland belt. Within the United States, the predictive surface resolves considerably finer spatial detail, including the pronounced GHF maxima centered on Nevada, Utah, and Colorado. These regionally resolved projections, in concert with the recovered broad-scale GHF spatial trends, underscore the substantive practical value of the predictive framework advanced in this study. It serves as a reliable benchmark for future GHF exploration and

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

resource development initiatives, particularly in geographic settings where in-situ measurement infrastructure remains underdeveloped.

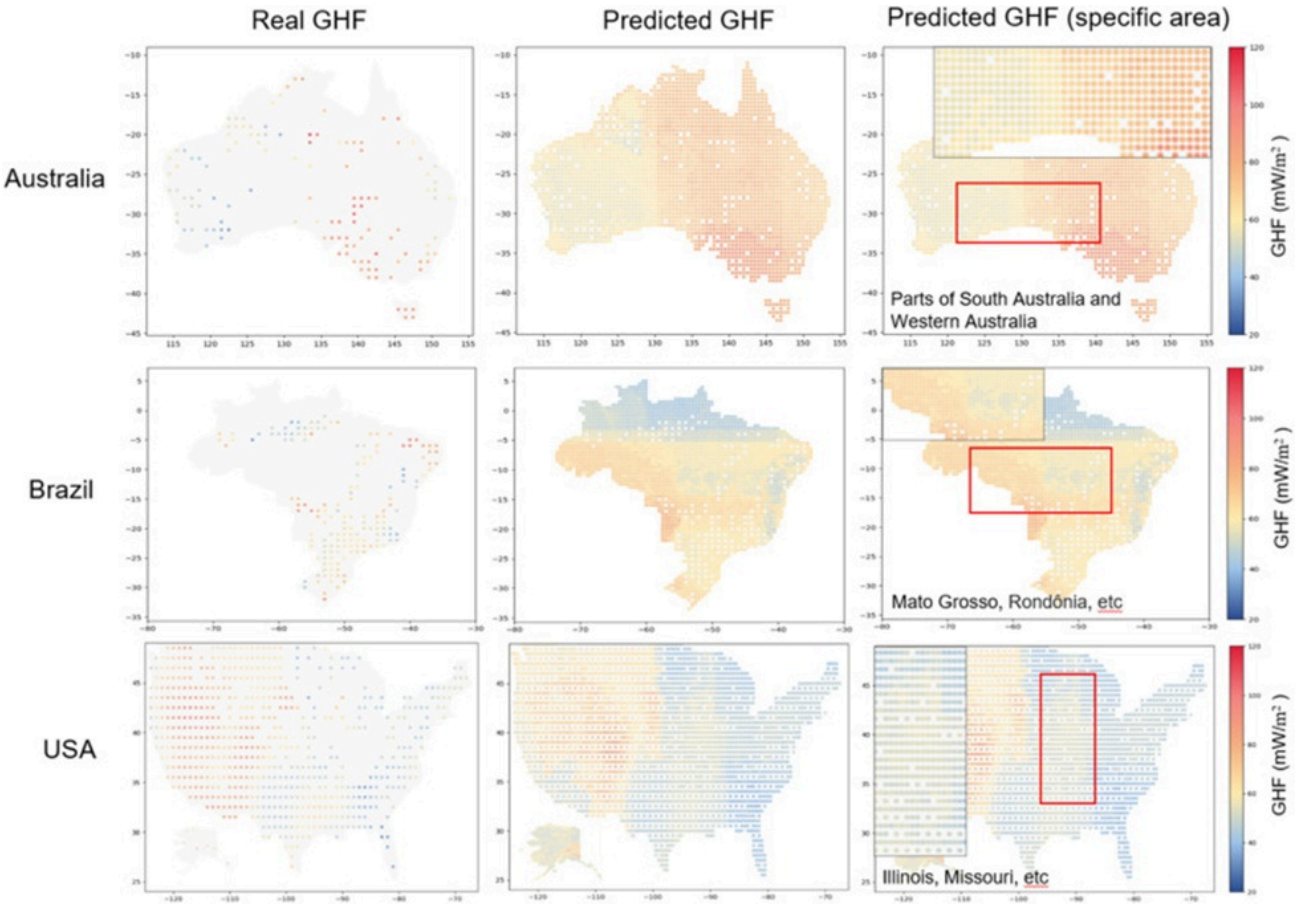

Figure 9. Comparative illustration of GHF distribution trends between actual observations and model-generated estimates across Australia, Brazil, and the USA. Projections are confined to areas lacking true GHF readings, with GHF magnitude encoded through color.

Concurrently, the GHF spatial trend map constructed via this modeling pipeline—wherein a limited yet complete ensemble of geological attributes reconstructs the full dataset and estimates GHF at unsampled coordinates—delivers actionable intelligence for future geothermal prospecting campaigns. As highlighted in **Figure 9**, the higher-resolution GHF predictive surface brings into focus prospective geothermal development zones across southern Australia (straddling portions of South Australia, Western Australia, and their transboundary regions), central-western Brazil (encompassing Mato Grosso, Rondônia, and adjoining territories), and the central-eastern United States (spanning Illinois, Missouri, and neighboring areas). Notably, such prospective signatures remain entirely obscured within the original GHF observational map due to inherent limitations in data coverage. Accordingly, the model-derived GHF predictive map not only faithfully reproduces observed spatial trends at instrumented sites but also provides critical reconnaissance insights into untapped geothermal potential in severely data-impoverished regions.

Conclusion

The present study has put forward a machine learning-grounded workflow for the estimation of global GHF drawing upon geographically sparse geological observations. A worldwide compilation comprising 23 geological predictor variables was first assembled as the analytical foundation. Working from this assembled data resource, the sensitivity of Kriging

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

interpolation fidelity to alternative semi-variogram formulations was evaluated, followed by the operationalization of machine-learning-driven imputation algorithms (KNN, MissForest) to reconstruct missing data entries. The empirical results show that across the range of missingness rates investigated herein (10%- 40%), the Kriging variant with a spherical semi-variogram consistently outperforms competing semi-variogram parameterizations. Beyond this, both machine learning-based imputation approaches (KNN and MissForest) outperform all Kriging model realizations, irrespective of the selected semi-variogram, with MissForest delivering the optimal balance between stability and reconstruction precision (R^2 = 0.90, RMSE = 0.0789, evaluated at a 20% missingness level). Within the suite of predictive modeling algorithms (linear regression, LGBM, random forest, XGBoost, and SVR) fitted to the MissForest-completed dataset, XGBoost and LGBM jointly achieved the highest predictive performance, with an R^2 of 0.75 and an n-RMSE of 0.15, thereby closely approximating the empirically observed GHF records. Following model training, the optimized architecture was extended to predict GHF across geographic domains with no direct measurement coverage. The resulting GHF predictive surface demonstrates close concordance with known spatial distributions, including elevated GHF across Brazil's southern highlands, contrasted with depressed readings throughout its northern territories; the systematic west-to-east intensification observed across Australia; and the northwest-high versus southeast-low GHF configuration characteristic of the United States. Validation assessments conducted in unmonitored regions substantiate the pronounced transferability and broad practical applicability of the modeling framework developed in this investigation. The insights emerging from this body of work are anticipated to furnish an economically efficient and scientifically robust reference base to support the global assessment and sustainable harnessing of geothermal energy resources.

Future research may pursue integrating advanced machine learning paradigms to enhance further both the predictive accuracy and interpretability of global geothermal heat flow (GHF) models. In particular, the deployment of Transformer-based architectures holds potential to more faithfully represent explicit spatial dependency structures. In contrast, generative modeling frameworks—including diffusion-based formulations and generative adversarial networks (GANs)—may offer streamlined pathways for imputing incomplete geological feature sets. In parallel, graph neural networks (GNNs) offer a compelling methodological avenue for explicitly encoding geospatial relational structures within the predictive architecture. To facilitate a deeper understanding of variable contributions, Shapley additive explanations (SHAP) could be leveraged to systematically quantify the relative importance ascribed to individual input features, thereby informing more judicious feature selection and bolstering the transparency of model operations. The synergistic integration of these computational methodologies with domain-grounded geological knowledge promises to yield GHF estimates that are simultaneously more robust and more physically interpretable.

Acknowledgements

None

Conflict of interest

None

Financial support

This work is financially supported by the National Natural Science Foundation of China (Grant No. 52576233), Shenzhen Science and Technology Innovation Commission (KCXST20221021111609024, JCYJ20220818103010021) and the Hetao Shenzhen-Hong Kong Science and Technology Innovation Cooperation Zone Project (No. HZQSWS-KCCYB-2024016).

Ethics statement

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

None

Received: 19 Dec 2021 Revised: 01 Mar 2022 Accepted: 18 Mar 2022
Published online: 15 November 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Guiling W, Wei Z, Jiyun L, Wenjing L, Zhiming L, Wanli W. Evaluation of geothermal resources potential in China. Acta Geosci Sin. 2017;38(4):449-59.

2. Horne R, Genter A, McClure M, Ellsworth W, Norbeck J, Schill E, et al. Enhanced geothermal systems for clean firm energy generation. Nat Rev Clean Technol. 2025;1(2):148-60.

3. Wang Y, Li C, Zhao J, Wu B, Du Y, Zhang J, et al. The above-ground strategies to approach the goal of geothermal power generation in China: state of art and future researches. Renew Sustain Energy Rev. 2021;138:110557.

4. Wang Y, Du Y, Wang J, Zhao J, Deng S, Yin H. Comparative life cycle assessment of geothermal power generation systems in China. Resour Conserv Recycl. 2020;155:104670.

5. Liu S, Taleghani AD, Ji K. An advanced closed-loop geothermal system to substantially enhance heat production. Energy Convers Manag. 2024;322:119168.

6. Zhu J, Hu K, Lu X, Huang X, Liu K, Wu X. A review of geothermal energy resources, development, and applications in China: current status and prospects. Energy. 2015;93:466-83.

7. Jolie E, Scott S, Faulds J, Chambefort I, Axelsson G, Gutiérrez-Negrín LC, et al. Geological controls on geothermal resources for power generation. Nat Rev Earth Environ. 2021;2(5):324-39.

8. Ma Y. Deep geothermal resources in China: potential, distribution, exploitation, and utilization. Energy Geosci. 2023;4(4):100209.

9. Burton-Johnson A, Dziadek R, Martin C. Geothermal heat flow in Antarctica: current and future directions. Cryosph Discuss. 2020;2020:1-45.

10. Oh H, Akar S, Beckers K, Bonnema E, Vivas C, Salehi S. Techno-economic feasibility of geothermal energy production using inactive oil and gas wells for district heating and cooling systems in Tuttle, Oklahoma. Energy Convers Manag. 2024;308:118390.

11. Xu S, Ni C, Hu X. Predicting terrestrial heat flow in North China using multiple geological and geophysical datasets based on machine learning method. Energies. 2023;16(4):1620.

12. Bai L, Li J, Zeng Z, Tao D. Prediction of terrestrial heat flow in Songliao basin based on deep neural network. Earth Space Sci. 2023;10(12):e2023EA003186.

13. He J, Li K, Wang X, Gao N, Mao X, Jia L. A machine learning methodology for predicting geothermal heat flow in the Bohai bay basin, China. Nat Resour Res. 2022;31(1):237-60.

14. Ishitsuka K, Kobayashi Y, Watanabe N, Yamaya Y, Bjarkason E, Suzuki A, et al. Bayesian and neural network approaches to estimate deep temperature distribution for assessing a supercritical geothermal system: evaluation using a numerical model. Nat Resour Res. 2021;30(5):3289-314.

15. Seyedzadeh S, Rahimian FP, Glesk I, Roper M. Machine learning for estimation of building energy consumption and performance: a review. Vis Eng. 2018;6(1):5.

16. Robinson C, Dilkina B, Hubbs J, Zhang W, Guhathakurta S, Brown MA, et al. Machine learning approaches for estimating commercial building energy consumption. Appl Energy. 2017;208:889-904.

17. De Luna P, Wei J, Bengio Y, Aspuru-Guzik A, Sargent E. Use machine learning to find energy materials. Nature. 2017;552(7683):23-7.

18. Hafeez G, Alimgeer KS, Khan I. Electric load forecasting based on deep learning and optimized by heuristic algorithm in smart grid. Appl Energy. 2020;269:114915.

19. Xu Y, Dai Y, Guo L, Chen J. Leveraging machine learning to forecast carbon returns: factors from energy markets. Appl Energy. 2024;357:122515.

20. Hussein NAHK, Abouessaouab MB. Optimizing energy efficiency in smart grids using machine learning algorithms: a case study in electrical engineering. Khwarizmia. 2023;2023:113-20.

21. Chitgar N, Hemmati A, Sadrzadeh M. A comparative performance analysis, working fluid selection, and machine learning optimization of ORC systems driven by geothermal energy. Energy Convers Manag. 2023;286:117072.

22. Rezvanbehbahani S, Stearns LA, Kadivar A, Walker JD, van der Veen CJ. Predicting the geothermal heat flux in Greenland: a machine learning approach. Geophys Res Lett. 2017;44(24):12-271.

23. Al-Aghbary M, Sobh M, Gerhards C. A geothermal heat flow model of Africa based on random forest regression. Front Earth Sci. 2022;10:981899.

24. Ouzzaouit LA, Ouassif BE, Idri A, Qarbous A, Barich A, Weisenberger TB, et al. Geothermal flow in Northern Morocco: a machine learning approach. J Afr Earth Sci. 2023;205:104995.

25. Lösing M, Ebbing J. Predicting geothermal heat flow in Antarctica with a machine learning approach. J Geophys Res. 2021;126(6):e2020JB021499.

26. Colgan W, Wansing A, Mankoff K, Lösing M, Hopper J, Louden K, et al. Greenland geothermal heat flow database and map (version 1). Earth Syst Sci Data Discuss. 2021;2021:1-50.

27. Wang S, Lu C, Nan D, Hu X, Shao J. Geothermal resources in Tibet of China: current status and prospective development. Environ Earth Sci. 2017;76(6):239.

28. USGS. Geologic province and thermo-tectonic age maps. 1997 [cited 2026 Jul 2]. Available from: https://earthquake.usgs.gov/data/crust/maps.php

29. Bassin C. The current limits of resolution for surface wave tomography in North America. Eos Trans AGU. 2000;81:F897.

30. Kaban MK, Schwintzer P, Reigber C. A new isostatic model of the lithosphere and gravity field. J Geod. 2004;78(6):368-85.

31. Coffm MF, Gahagan LM, Lawver LA. Present-day plate boundary digital data compilation. Tectonophysics. 1998;183:243-64.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

32. Goutorbe B, Poort J, Lucazeau F, Raillard S. Global heat flow trends resolved from multiple geological and geophysical proxies. Geophys J Int. 2011;187(3):1405-19.

33. Anderson DL. The complete hot spot. 2016 [cited 2026 Jul 2]. Available from: http://www.mantleplumes.org/CompleateHotspot.html

34. Laske G, Masters G, Ma Z, Pasyanos M. Update on CRUST1. 0—a 1-degree global model of earth's crust. Geophys Res Abstr. 2013;15:2658.

35. Sengör AC, Natal'in BA. Rifts of the world. In: Ernst RE, Buchan KL, editors. Mantle plumes: their identification through time. Boulder: Geological Society of America; 2001. p. 389-482. (Special Paper; vol. 352).

36. Shapiro NM, Ritzwoller MH. Monte-Carlo inversion for a global shear-velocity model of the crust and upper mantle. Geophys J Int. 2002;151(1):88-105.

37. Hartmann J, Moosdorf N. The new global lithological map database GLiM: a representation of rock properties at the earth surface. Geochem Geophys Geosyst. 2012;13(12):Q12004.

38. Poupinet G, Shapiro NM. Worldwide distribution of ages of the continental lithosphere derived from a global seismic tomographic model. Lithos. 2009;109(1-2):125-30.

39. Maus S, Sazonova T, Hemant K, Fairhead JD, Ravat D. National geophysical data center candidate for the world digital magnetic anomaly map. Geochem Geophys Geosyst. 2007;8(6):Q06017.

40. Reguzzoni M, Sampietro D, Sansò F. Global Moho from the combination of the CRUST2. 0 model and GOCE data. Geophys J Int. 2013;195(1):222-37.

41. Pasyanos ME, Masters TG, Laske G, Ma Z. LITHO1. 0: an updated crust and lithospheric model of the earth. J Geophys Res. 2014;119(3):2153-73.

42. Balmino G, Vales N, Bonvalot S, Briais A. Spherical harmonic modelling to ultra-high degree of Bouguer and isostatic anomalies. J Geod. 2012;86(7):499-520.

43. Amante C, Eakins BW. ETOPO1 arc-minute global relief model: procedures, data sources and analysis. Boulder: National Geophysical Data Center; 2009.

44. Ince ES, Barthelmes F, Reißland S, Elger K, Förste C, Flechtner F, et al. ICGEM–15 years of successful collection and distribution of global gravitational models, associated services, and future plans. Earth Syst Sci Data. 2019;11(2):647-74.

45. Gosnold W. The global heat flow database. International Heat Flow Commission; 2011 [cited 2026 Jul 2]. Available from: https://www.heatflow.und.edu:80/data.html

46. Lei M, Labbe A, Wu Y, Sun L. Bayesian kernelized matrix factorization for spatiotemporal traffic data imputation and kriging. IEEE Trans Intell Transp Syst. 2022;23(10):18962-74.

47. Matheron G. Principles of geostatistics. Econ Geol. 1963;58(8):1246-66.

48. Oriani F, Stisen S, Demirel MC, Mariethoz G. Missing data imputation for multisite rainfall networks: a comparison between geostatistical interpolation and pattern-based estimation on different terrain types. J Hydrometeorol. 2020;21(10):2325-41.

49. Carrera-Hernández JJ, Gaskin SJ. Spatio temporal analysis of daily precipitation and temperature in the basin of Mexico. J Hydrol. 2007;336(3-4):231-49.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

50. Shtiliyanova A, Bellocchi G, Borras D, Eza U, Martin R, Carrère P. Kriging-based approach to predict missing air temperature data. Comput Electron Agric. 2017;142:440-9.

51. Goovaerts P. Geostatistics for natural resources evaluation. New York: Oxford University Press; 1997. 483 p.

52. Fagandini C, Todaro V, Tanda MG, Pereira JL, Azevedo L, Zanini A. Missing rainfall daily data: a comparison among gap-filling approaches. Math Geosci. 2024;56(2):191-217.

53. Khan M, Almazah MM, EIlahi A, Niaz R, Al-Rezami AY, Zaman B. Spatial interpolation of water quality index based on ordinary kriging and universal kriging. Geomat Nat Hazards Risk. 2023;14(1):2190853.

54. Lin Q, Bao X, Li C. Deep learning based missing data recovery of non-stationary wind velocity. J Wind Eng Ind Aerodyn. 2022;224:104962.

55. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520-5.

56. Alnowaiser K, Alarfaj AA, Alabdulqader EA, Umer M, Cascone L, Alankar B. IoT based smart framework to predict air quality in congested traffic areas using SV-CNN ensemble and KNN imputation model. Comput Electr Eng. 2024;118:109311.

57. Li J, Guo S, Ma R, He J, Zhang X, Rui D, et al. Comparison of the effects of imputation methods for missing data in predictive modelling of cohort study datasets. BMC Med Res Methodol. 2024;24(1):41.

58. Aguilera H, Guardiola-Albert C, Serrano-Hidalgo C. Estimating extremely large amounts of missing precipitation data. J Hydroinformatics. 2020;22(3):578-92.

59. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. Bioinformatics. 2012;28(1):112-8.

60. Zhao W, Ma J, Liu Q, Dou L, Qu Y, Shi H, et al. Accurate prediction of soil heavy metal pollution using an improved machine learning method: a case study in the Pearl river delta, China. Environ Sci Technol. 2023;57(46):17751-61.

61. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: a next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019 Jul 4-8; Anchorage, AK, USA. New York: ACM; 2019. p. 2623-31.

62. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825-30.

63. Yu C, Yao W. Robust linear regression: a review and comparison. Commun Stat-Simul Comput. 2017;46(8):6261-82.

64. Balabin RM, Lomakina EI. Support vector machine regression (SVR/LS-SVM)—an alternative to neural networks (ANN) for analytical chemistry? Comparison of nonlinear methods on near infrared (NIR) spectroscopy data. Analyst. 2011;136(8):1703-12.

65. Wang LA, Zhou X, Zhu X, Dong Z, Guo W. Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. Crop J. 2016;4(3):212-9.

66. Chen Y, Xu P, Chu Y, Li W, Wu Y, Ni L, et al. Short-term electrical load forecasting using the support vector regression (SVR) model to calculate the demand response baseline for office buildings. Appl Energy. 2017;195:659-70.

67. Li Y, Zou C, Berecibar M, Nanini-Maury E, Chan JCW, Van den Bossche P, et al. Random forest regression for online capacity estimation of lithium-ion batteries. Appl Energy. 2018;232:197-210.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press

68. Wang L, Zhang Y, Yao Y, Xiao Z, Shang K, Guo X, et al. GBRT-based estimation of terrestrial latent heat flux in the Haihe river basin from satellite and reanalysis datasets. Remote Sens. 2021;13(6):1054.

69. Shwartz-Ziv R, Armon A. Tabular data: deep learning is not all you need. Inf Fusion. 2022;81:84-90.

70. Shahdi A, Lee S, Karpatne A, Nojabaei B. Exploratory analysis of machine learning methods in predicting subsurface temperature and geothermal gradient of Northeastern United States. Geotherm Energy. 2021;9(1):18.

71. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: a highly efficient gradient boosting decision tree. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in neural information processing systems 30 (NIPS 2017). 2017 Dec 4-9; Long Beach, CA, USA. Red Hook: Curran Associates; 2017. p. 3146-54.

© 2022 The Author(s). | Published by Institute of Smart Energy Systems Press