Exploratory Analysis of Machine Learning Methods in Predicting Subsurface Temperature and Geothermal Gradient of Northeastern United States

Arya Shahdi (aryashahdi@vt.edu)
Virginia Polytechnic Institute and State University College of Engineering https://orcid.org/0000-0002-2328-5147
Seho Lee
Virginia Polytechnic Institute and State University
Anuj Karpatne
Virginia Polytechnic Institute and State University
Bahareh Nojabaei
Virginia Polytechnic Institute and State University

Research

Keywords: Renewable Energy, Geothermal Energy, Machine Learning, XGBoost, Subsurface temperature, geothermal gradient

Posted Date: January 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-131433/v1

Title: Exploratory Analysis of Machine Learning Methods in Predicting Subsurface Temperature and Geothermal Gradient of Northeastern United States

Keywords: Renewable Energy, Geothermal Energy, Machine Learning, XGBoost, Subsurface temperature, geothermal gradient

Authors: Arya Shahdi[*], Ph.D (corresponding author); Seho Lee[*]; Anuj karpatne[*], Ph.D; Bahareh Nojabaei[**], Ph.D

* Department of Computer Science at Virginia Tech

** Department of Mining and Mineral Engineering at Virginia Tech

Abstract

Geothermal scientists have used bottom hole temperature data from extensive oil and gas well datasets to generate heat flow and temperature-at-depth maps to locate potential geothermally active regions. Considering that there are some uncertainties and simplifying assumptions associated with the current state of physics-based models, in this study, the applicability of several machine learning models is evaluated for predicting temperature-at-depth and geothermal gradient parameters. Through our exploratory analysis, it is found that XGBoost results in the highest accuracy for subsurface temperature prediction with average mean-absolute-error and root-mean-square-error of 3.19[℃] and 4.94[℃], respectively. Furthermore, we apply our model to regions around the sites to provide 2D continuous temperature maps at three different depths using XGBoost model, which can be used to locate prospective geothermally active regions. We also validate the proposed XGBoost and DNN models using an extra dataset containing measured temperature data along the depth for fifty-eight wells in the state of West Virginia. Accuracy measures show that machine learning models are highly comparable to the physics-based model and can even outperform the thermal conductivity model. Also, a geothermal gradient map is derived for the whole region by fitting linear regression to the XGBoost predicted temperatures along the depth. Finally, thorough our analysis, the most favorable geological locations are suggested for potential future geothermal developments.

Introduction

Bottom hole temperature (BHT) measurements have largely been used for mapping subsurface temperatures for geothermal resource analysis across the United States (Blackwell and Richards, 2010; Frone and Blackwell, 2010; Stutz et al., 2012; Tester et al., 2006). BHT data are predominantly provided by oil and gas wells, where maximum temperature is usually reported at the final drilled depth. In 2011, Blackswell et al. (Blackwell and Richards, 2010) incorporated BHT data in northeastern United States with stratigraphic information (Childs, 1985), and used a simple thermal conductivity model to generate surface heat flux and temperature-at-depth maps. Jordan et al. (Jordan, T.E., 2016), conducted a thorough analysis to explore the associated risks and potentials of prospective geothermal resources in New York, Pennsylvania and West Virginia States. Even though most geothermally active regions are located in the west of United States (near Earth's tectonic plate boundaries), Jordan et al. (Jordan, T.E., 2016) showed that the stored energy in the low-temperature geothermal regions in the northeast could be utilized for many direct-use applications. Even though Snyder et al. (Snyder et al., 2017) illustrated that myriad industrial and residential direct-use applications of geothermal energy could result in reduction of electricity consumption, there are not many geothermal sites in northeastern states due to a high financial risk. Heat flux and temperature-at-depth are two most important geothermal parameters, which have extensively been investigated through physics-based models.

In the previous geothermal studies, the generalized thermal conductivity model has been adopted to compute the heat flow associated with BHT data points (Blackwell and Richards, 2010; Frone and Blackwell, 2010; Jordan, T.E., 2016; Jordan, 2015; Stutz et al., 2012; Tester et al., 2006). Initially, the measured bottom hole temperature is corrected (through various available correlations(Deming, 1989)) and, then, is used to calculate the temperature gradient through the following relation:

$$\left(\frac{dT}{dz}\right) = \frac{BHT - T_{surf}}{z} \tag{1}$$

Next, the geological formation thickness and thermal conductivity values are interpolated at the point's latitude and longitude mainly from the data reported in Correlation of Stratigraphic Units of North America (COSUNA)(Childs, 1985). Then, average thermal conductivity is calculated between surface and well's depth(Stutz et al., 2012). Finally, the heat flux is calculated through the following equation:

$$Q_s = \bar{k}\left(\frac{dT}{dz}\right) \quad (2)$$

Obviously, the above formula is oversimplified and only represents the main theoretical framework of the physics-based model, which is used in geothermal energy studies. Despite physics-based model's long-time applicability, they all have some underlying assumptions that could result in uncertainties and, therefore, inaccurate predictions. Some of the assumptions are explained by Stutz et al.(Stutz et al., 2012) and Blackwell et al.(Blackwell and Richards, 2010) In particular, there is no easy-to-use method to independently measure the heat flux parameter; it is only approximated through the thermal conductivity model using the BHT data as shown in Equation (2).

In addition to the geothermal energy industry, subsurface temperature is an extremely important parameter in oil and gas industry (Bassam et al., 2010; Forrest et al., 2005; Khan and Raza, 1986; P.L. Moses (Core Laboratories Inc.), 1961). Characteristics of hydrocarbons are greatly dependent on the temperature and they require to be approximated to be used in reservoir and drilling simulations. In practice, it is common to use geothermal gradient maps to obtain the geothermal gradient value at the desired location and then calculate the subsurface temperature at the depth of interest (Forrest et al., 2005; Khan and Raza, 1986). In this study, we provide an alternative solution of using machine learning methods for predicting subsurface temperature using BHT data from more than 20,750 oil and gas wells in northeastern United States. Furthermore, the physics-based and machine learning models are compared through an extra dataset containing vertical temperature profile of fifty-eight wells in the state of West Virginia. Finally, we provide the geothermal gradient map using the validated XGBoost model for the northeast region of United States.

Case of study

Marcellus formation is one of the highest potential hydrocarbon prospects in the United States, which is located throughout the northern Appalachian Basin. For several decades, thousands of wells have been drilled in this region which contain, at least one temperature measurement (usually at the final depth). For our analysis, we have used a dataset with corrected bottom hole temperature (BHT), surface temperature, API, latitude, longitude and geological setting information (including layer thickness and conductivity) and many other information from 20,750 oil and gas wells in the

northeast (**Error! Reference source not found.**). This dataset has been developed and reported as part of a DOE funded research grant led by Cornell University (Jordan, 2015). In **Error! Reference source not found.**, the scatter points are referred to the well locations and the shaded area depicts the region where temperature predictions will be provided.

For preprocessing, we removed the outliers (101 data points) using the heat-flow parameter (outside the three standard-deviation with values larger than zero). We primarily used heat-flow parameter for outlier removal because its histogram shape is very close to normal distribution (**Error! Reference source not found.**).

Table 1 provides the summary of some important parameters after outlier removal. Bottom-hole temperature values have already been corrected (CorrBHT) using the drilling fluid and geological setting information (Jordan, 2015).

Table 1: Statistical summary of important parameters after outlier removal

	Surface temperature	**Depth**	**Corrected BHT**	**Heat Flow**
Unit	°C	m	°C	mW/m^2
Mean	12.4	1154	37	49
std	1.8	459	13.2	13.4
min	8.8	43	10.2	0.2
25%	10.6	868	28.9	41.57
50%	12.1	1129	34.5	47.91
75%	14.3	1358	42.8	55.26
max	15.6	6541	146.9	130.21

In addition to the above parameters, geological layer's thickness and conductivity values (between surface and basement) are provided. These values have originally been interpolated through COSUNA data points using the Kriging method (Jordan, 2015).

We also exclusively gathered data for additional fifty-eight wells across West Virginia (annotated on **Error! Reference source not found.**). In this dataset, for each well, temperature profile is provided within a depth interval (with mean and standard deviation of 1167 and 511 meters, respectively). We obtained this dataset from West Virginia Geological and Economical Survey (West Virginia Geological and Economical Survey, n.d.) and primarily used it for comparing our

results with those from physics-based model. We refer to this new source as temperature-profile dataset throughout this paper. Among the fifty-eight wells, bottom hole temperature points of 11 wells already exist in the first dataset (20,750 wells). The rest are new wells which have been used to compare the physics-based model with the machine learning methods.

The reported temperatures in the temperature-profile dataset are prone to errors and we were required to correct them. Even though there are myriad temperature-correction methods, we decided to use the correction methodology reported by Jordan et al.(Jordan, 2015) in order to be consistent with their method. This allowed us to compare our results to those reported by the physics-based model in Jordan et al.(Jordan, 2015). Unfortunately, we could not find the complete information to reiterate their model for well temperature correction. Nevertheless, we decided to use Kth Nearest Neighbor regression model to estimate the corrected temperature values. Fortunately, in the first dataset, raw and corrected BHT data were available and we could use them to train our interpolation model. After performing cross-validation, we found that the mean-absolute-error of the interpolation model was only 1.5℃. Next, we used the trained KNN regression model to correct the temperature profile data in the new dataset.

Methodology

Machine learning models

In this section, we provide a thorough summary of the machine learning models that have been used in this study to estimate subsurface temperature and geothermal gradient. We mainly used regression models with different characteristics including Deep Neural Network (DNN), Ridge regression (R-reg) and decision-tree-based models (e.g., XGBoost and Random Forest). In Table 2, we include the features and the label which have been used in the machine learning models.

Table 2: Detailed information about the label and features

Variable number	Name	Unit	Source	Description	Type
1	BHTCorr	[°C]	well log report	Corr bottom-hole temperature	Label
2	LatDegree	-	well log report	Lat degree of the well's location	Feature
3	LongDegree	-	well log report	Long degree of the well's location	Feature

4	MeasureDepth	[m]	well log report	The depth where BHT is recorded	Feature
5	SurfTemp	[°C]	Annual average temperature	Surf temperature at the well's location	Feature
6 to 55	KH	[W/(°K)]	Interpolated from the data reported in Correlation of Stratigraphic Units of North America (COSUNA)	Multiplication product of each geological layer's thickness with its corresponding thermal conductivity	Feature

The following chart illustrates the developed machine learning pipeline which has been used throughout this study. We used bootstrapping method for train-test data split. At each split, 90% of data were

Ridge Regression. In our dataset, there are uncertainties associated with the bottom hole temperature measurements and it is believed that noise is present in the dataset. We used Ridge regression as one of the candidate machine learning models since it is robust to overfitting by introducing a penalty (also known as L2 Regularization) on the model's complexity(Hoerl and Kennard, 1970). Baruque et al.(Baruque et al., 2019) successfully used Ridge regression for a geothermal application where heat exchanger energy was predicted using time series readings of several sensors. The goal is to find the model's parameters which minimizes the objective function.

$$\hat{\theta}^{ridge} = \underset{\theta}{\operatorname{argmin}}(\|y - X\Theta\|_2^2 + \alpha\|\Theta\|_2^2) \quad (3)$$

where hyperparameter α is a positive number that specifies the trade-off between the OLS and regularization terms. In our implementation, we initially scaled the data and then used the grid-search method in the log space to search for the best alpha.

Gradient Boosting and Random Forest. In this study, class imbalance is present within our dataset since the majority of BHT data correspond to the shallower wells (< 3000 m). On the other hand, the deeper wells contain valuable information with higher temperature values which should not be removed (or be considered as outliers). We mainly used ensemble-based algorithms (like Random Forest and XGBoost) because they are believed to work relatively well in a case where class imbalance exists (Galar et al., 2011). In addition, tree-based models usually improve the accuracy by decreasing the variance in the prediction. Gradient Boosting and Random Forest are both tree-based methods which have been successfully applied in variety of industries. In

geothermal energy, Gul et al.(Gul et al., 2019) predicted the formation temperature using drilling fluid data using extreme gradient boosting and random forest models with satisfactory accuracy.

Single decision tree is often referred to as a weak classifier as it can be susceptible to overfitting(Ho, 1998). Random Forest builds an ensemble of multiple decision trees (weak classifiers) in parallel and takes the mean of the predictors for the prediction. Furthermore, during the ensemble construction, random features or columns are dropped while learning every decision tree, so that every tree is de-correlated from other trees as much as possible. Gradient Boosting, on the other hand, builds decision trees in a sequential manner. Gradient Boosting keeps adding decision trees at every step, making a fine separation in space to predict the response variable (Chen and Guestrin, 2016). Every new step considers the previous steps which result in accuracy improvement after each iteration. XGBoost is a library that allows Gradient Boosting to be run in parallel in terms of computing.

Deep Neural Network (DNN). One of the other candidate machine learning models that we have used is Deep Neural Network (DNN). Bassam et al.(Bassam et al., 2010) was among the first studies that evaluated the application of a shallow ANN in formation temperatures in geothermal wells. In that study, collected BHT logs (during long-shut-in times) have been used for training and validation.

Deep Neural Networks is a type of model that attempts to capture the true relationship of input and output. DNN algorithm can capture high non-linearity through a network of nodes and connections. We tried different DNN architectures and finally picked a four-layer DNN as illustrated in **Error! Reference source not found.**. In the input layer, the number of nodes is the same as feature numbers followed by two hidden layers where each layer contains fifty nodes. Arrows correspond to connections among nodes and are associated with learnable edge weights.

For the last neuron at the output layer, the weighted responses from the neurons at the second hidden layer are fed into a linear activation function and the final prediction for temperature is obtained. In **Error! Reference source not found.**, one neuron of the hidden layer is illustrated with the given inputs.

Feature space interpolation

Temperature-at-depth maps have extensively been used in geothermal energy studies to illustrate the temperature distribution at a given depth. In this study, we also provide temperature-at-depth maps at different depths in the northeastern United States. This allows investors to have another source of temperature prediction map for any potential future development. In addition, the new machine learning temperature maps can be compared to those from the thermal conductivity model to locate the similarities and differences. A simple concave hall algorithm was used to obtain a tight boundary around the given data points. In order to avoid sharp edges, we derived average values for the boundary points and then implemented the algorithm (shaded area in **Error! Reference source not found.**). We initially used an online source-code (Dwyer, n.d.) and made major modifications to meet our project's needs.

For constructing the subsurface temperature prediction map, the features should be available within different locations (with varying latitude and longitude). Therefore, we interpolated the required features throughout the northeastern region using a Gaussian kernel weighted k-nearest neighbor regression model. We chose KNN regression method as it is expected to perform well in our region of interest due to high concentration of wells. We used cross-validation for hyperparameter tuning (k and kernel width) using 20,750 data points.

Result and discussion

In this study, our primary objective is to evaluate the applicability of machine learning models to predict subsurface temperature at any given latitude, longitude and depth by merely using geological information and surface temperature. Even though we only used single temperature measurement points (bottom-hole temperatures) for training, we observed that the machine learning models performed well in predicting underground temperatures. Among the machine learning models, XGBoost outperformed other models and provided more accurate results. Furthermore, we obtained temperature-versus-depth data for fifty-eight wells and compared them with XGBoost, DNN and physics-based model's predictions. The results show that machine learning models predictions were in close agreement with the measured data.

Feature importance analysis

Knowing the predictive power of the features can help us identify the most important factors affecting the subsurface temperature parameter. In our analysis, we initially scaled the input features to avoid any bias related to variables. Next, we used XGBoost and Random Forest to identify the most important features. Both methods recognized the depth, layer-29 and layer-27 among top five important features. In **Error! Reference source not found.**, seven most important features in XGBoost are shown. In the x-axis, "Layer-#" is referred to multiplication product of thickness and conductivity of the desired layer which is available in the dataset for each well. More detailed information about the important layers can be found in Table 3.

Table 3: Important geological layer's information

	Average thickness	**Average conductivity**	**Average layer's depth**
Layer-13	180	2.6	1633
Layer-16	50	2.6	1838
Layer-24	65	2.7	2366
Layer-27	84	2.4	2736
Layer-29	41	3.5	2864
Layer-32	88.6	2.7	3295

Temperature-at-depth result analysis

After training and tuning hyperparameters, we evaluated the accuracy of each model using the test data for 10 bootstrapped samples. As you can see in the two following boxplots, XGBoost performs the best among other three machine learning models.

We, then, used the trained models to predict subsurface temperature at three different depths ($Z = 1000, 2000, 3000$ meters) in the northeastern United States. In **Error! Reference source not found.**, temperature predictions are plotted using XGBoost models. We used KNN regression (with k = 8) for temperature interpolation for physics-based predictions.

Temperature profile prediction

In our analysis, we decided to use the corrected temperature-profile dataset (described in section 3) to evaluate XGBoost and DNN accuracies against the thermal conductivity model. We used KNN regression model to interpolate temperature profile predictions for the physics-based model at new locations. Fortunately, predicted temperature profiles from the thermal conductivity model were available in our main dataset and we used those to train our interpolation model for predicting the temperature profile of the new wells. In the following schematic, we illustrate the procedure that we have used to compare predictions from machine learning and the physics-based models.

After analyzing the results, the mean-absolute-errors of XGBoost, DNN, and Physics-based models were calculated to be 7.28, 7.34, and 8.46 respectively for the fifty-eight wells. These numbers show that machine learning models can be comparable, in terms of accuracy, to the physics-based thermal conductivity model. It is important to note that we have used multiple interpolations to be able to perform such comparison (**Error! Reference source not found.**). Therefore, there are some level of uncertainties associated with the reported numbers.

For illustration purposes, we include six temperature profile predictions (in **Error! Reference source not found.**) which are fair representatives of the remaining cases. Among all plots, we could see that the thermal conductivity model performs relatively better in tracking the true temperature data in 11.3 and 11.4. On the hand, both XGBoost and DNN models provide more accurate results in 11.1 and 11.6. Nevertheless, there are some cases where all models fail to track the actual data. For example, in plot 11.2, we could see that neither physics-based nor machine learning models predict the temperature profile accurately. Temperature profile prediction plots of other wells are included in our GitHub repository(Shahdi and Lee, n.d.). Among machine learning predictions, DNN and XGBoost predictions follow very similar trends even though DNN curves are smoother and have less variation with depth. This is expected because decision-tree-based models tend to show such discrete predictive behavior when used for regression.

In Table 3, we include more information about the wells that are shown in **Error! Reference source not found.**. The shown plots are from the wells that are close to at least one of the wells in the main dataset. This is important because it shows that the interpolated temperature values for

the physics-based predictions are reliable and close to those reported by the original study (Jordan, 2015).

Table 4: Corresponding details about the wells that are shown in **Error! Reference source not found.**. Distance column is referred to the distance from the test well to the closest well in the main dataset

Plot #	API	Distance [km]
1	4710300645	0.26
2	4700502148	0.22
3	4709501963	0.22
4	4700502167	0.5
5	4701304647	0.34
6	4705900805	3.27

Geothermal gradient map

It is very popular to use geothermal gradient maps to predict the subsurface temperature at the desired location. In this study, we provide the geothermal gradient map for the northeastern United States.

Similar to the plots (shown in **Error! Reference source not found.**), we generate temperature profile predictions for 28,000 locations across the region and then fit a linear regression line to the temperature data for each location. Through our analysis, we found that the fitted lines accurately represent the predicted temperatures with average R^2 of 0.973. The reported slopes are equal to the associated geothermal gradients and are illustrated in **Error! Reference source not found.**.

Some areas in West Virginia and New York states show high values for temperature gradient. In the next graph, areas with geothermal gradient higher than 27.5$\left[\frac{^\circ C}{km}\right]$ are annotated in the map. We cautiously suggest these machine-learning guided prospective regions for future geothermal developments.

Next, we calculated the mean absolute errors between the geothermal gradients predicted using different models (e.g., physics-based, XGBoost and DNN) and measured temperatures for the temperature-profile dataset.

Table 5: Average mean-absolute-errors and standard deviations (with unit of $\left[\frac{^{\circ}C}{km}\right]$) for physics-based, XGBoost and DNN model predictions in compare to the measured temperature data.

Model	MAE	STD
Physics	6.4	5.5
XGBoost	5.6	5.9
DNN	6.9	6.2

Conclusion

The goal of this paper is to highlight the importance and applicability of machine learning methods in producing reliable predictions of important geothermal parameters from the rich volumes of data available from geothermal sites. It is critical to understand that this paper does not claim to prove that machine learning models are ubiquitously superior to conventional physics-based models in geothermal energy research. In this study, we explored the applicability of four machine learning models in predicting subsurface temperatures in northeastern United States using bottom-hole temperature data and geological information from 20,750 wells. It was shown that XGBoost outperformed all other models, with only 3.19 [°C] mean absolute error. Furthermore, we compared the predictions from machine learning and physics-based models to the measured temperature data obtained from an extra dataset with fifty-eight wells in the state of West Virginia and showed that XGBoost can successfully predict the temperature at different depths. Lastly, we provided a geothermal gradient map for the corresponding region which can be used as a quick tool to calculate the underground temperature at any desired location and depth. In the map, east of West Virginia along with southern region of New York state show the highest potential.

We believe that this study provides a complementary analysis for geothermal energy exploration for future investments. Furthermore, oil and gas industry can benefit tremendously from this paper too. The presented machine learning models can be incorporated in reservoir and drilling simulators for more accurate subsurface temperature predictions, and consequently, more reliable fluid properties characterization.

Availability of data and materials

Complete information about the data resources and source-codes are provided in a GitHub repository(Shahdi and Lee, n.d.). The source codes associated with each of the figures (in the

manuscript) are specified in the “README.txt”. In addition, we provide the exact locations where we obtained the data which are used in the paper.

Acknowledgements

We thank the departments of computer science and mining and minerals engineering to provide all required resources to complete this research paper.

Authors’ contributions

Arya Shahdi: Conceptualization, Methodology, Data curation, Writing original draft preparation, Software, Validation - Seho Lee: Software, Investigation, Visualization, Validation - Anuj Karpatne and Bahareh Nojabaei: Supervision, Writing- Reviewing and Editing. The final manuscript is approved by all authors.

Competing interest

We (the authors) declare that there are not competing interest associated with the research.

Funding

This work was funded by the department of Mining and Minerals Engineering at Virginia Tech with no additional outside funding.

Bibliography

Baruque B, Porras S, Jove E, Calvo-Rolle J. Geothermal heat exchanger energy prediction based on time series and monitoring sensors optimization. Energy 2019;171:49–60. https://doi.org/https://doi.org/10.1016/j.energy.2018.12.207.

Bassam A, Santoyo E, Andaverde J, Herná Ndez JA, Espinoza-Ojeda OM. Estimation of static formation temperatures in geothermal wells by using an artificial neural network approach. Comput Geosci 2010;36:1191–9. https://doi.org/10.1016/j.cageo.2010.01.006.

Blackwell D, Richards M. New geothermal resource map of the northeastern US and technique for mapping temperature at depth. Sacramento, CA: 2010.

Chen T, Guestrin C. XGBoost: A scalable tree boosting system. Proc. 22nd acm sigkdd Int. Conf. Knowl. Discov. data Min., vol. 13-17- August-2016, Association for Computing

Machinery; 2016, p. 785–94. https://doi.org/10.1145/2939672.2939785.

Childs OE. Correlation of Stratigraphic Units of North America--COSUNA. Am Assoc Pet Geol Bull 1985;69:173–80. https://doi.org/https://doi.org/10.1306/AD461C73-16F7-11D7-8645000102C1865D.

Deming D. Application of bottom-hole temperature corrections in geothermal studies. Geothermics 1989;18:775–86. https://doi.org/https://doi.org/10.1016/0375-6505(89)90106-5.

Dwyer K. Concave hull - Python code n.d.;30. https://doi.org/10.1063/1.1347984.

Forrest J, Marcucci E, Scott P. Geothermal gradients and subsurface temperatures in the northern gulf of mexico. Gulf Coast Assoc Geol Soc Trans 2005;55:233–48.

Frone Z, Blackwell D. Geothermal map of the northeastern United States and the West Virginia thermal anomaly. Geotherm Resour Counc Annu Meet 2010;34:GRC1028668.

Galar et al. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man, Cybern Part C (Applications Rev 2011;42:463–84. https://doi.org/10.1109/TSMCC.2011.2161285.

Gul S, Aslanoglu V, Tuzen M, Senturk E. Estimation of bottom hole and formation temperature by drilling fluid data: a machine learning approach. 44th Work. Geotherm. Reserv. Eng., Stanford, CA: 2019.

Ho TK. The random subspace method for constructing decision forests. IEEE Trans Pattern Anal Mach Intell 1998;20:832–44. https://doi.org/10.1109/34.709601.

Hoerl AE, Kennard RW. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics 1970;12:55–67. https://doi.org/10.1080/00401706.1970.10488634.

Jordan, T.E. et al. Low temperature geothermal play fairway analysis for the Appalachian Basin: Phase 1 Revised Report November 18, 2016. Ithaca, NY: 2016. https://doi.org/10.2172/1341349.

Jordan T. et al. Appalachian Basin play fairway analysis: thermal quality analysis in low-temperature geothermal play fairway analysis (GPFA-AB) [data set]. Cornell Univ 2015. https://gdr.openei.org/submissions/638.

Khan MA, Raza HA. The role of geothermal gradients in hydrocarbon exploration in Pakistan. J

Pet Geol 1986;9:245–58. https://doi.org/10.1111/j.1747-5457.1986.tb00388.x.

P.L. Moses (Core Laboratories Inc.). Geothermal gradients. Drill. Prod. Pract., New York: American Petroleum Institute; 1961.

Shahdi A, Lee S. GitHub repository n.d. https://bit.ly/3qRnqcD (accessed December 13, 2020).

Snyder DM, Beckers KF, Young K. Update on geothermal direct-use installations in the United States. Forty-Second Work. Geotherm. Reserv. Eng., Stanford, CA: Stanford University; 2017, p. 13–5.

Stutz G, Williams M, Frone Z, Reber T, Blackwell D, Jordan T, et al. A well by well method for estimating surface heat flow for regional geothermal resource assessment. Thirty-Seventh Work. Geotherm. Reserv. Eng. Stanford Univ. SGP-TR-194, Stanford: 2012.

Tester J, Anderson B, Batchelor A, Blackwell D, DiPippo R, et al. The future of geothermal energy. Impact of Enhanced Geothermal Systems (EGS) on the United States in the 21st Century. DOE Geotherm. Progr. Work., Washington, D.C: 2006.

West Virginia Geological and Economical Survey. n.d.;30. https://doi.org/10.1063/1.1347984.

Figures

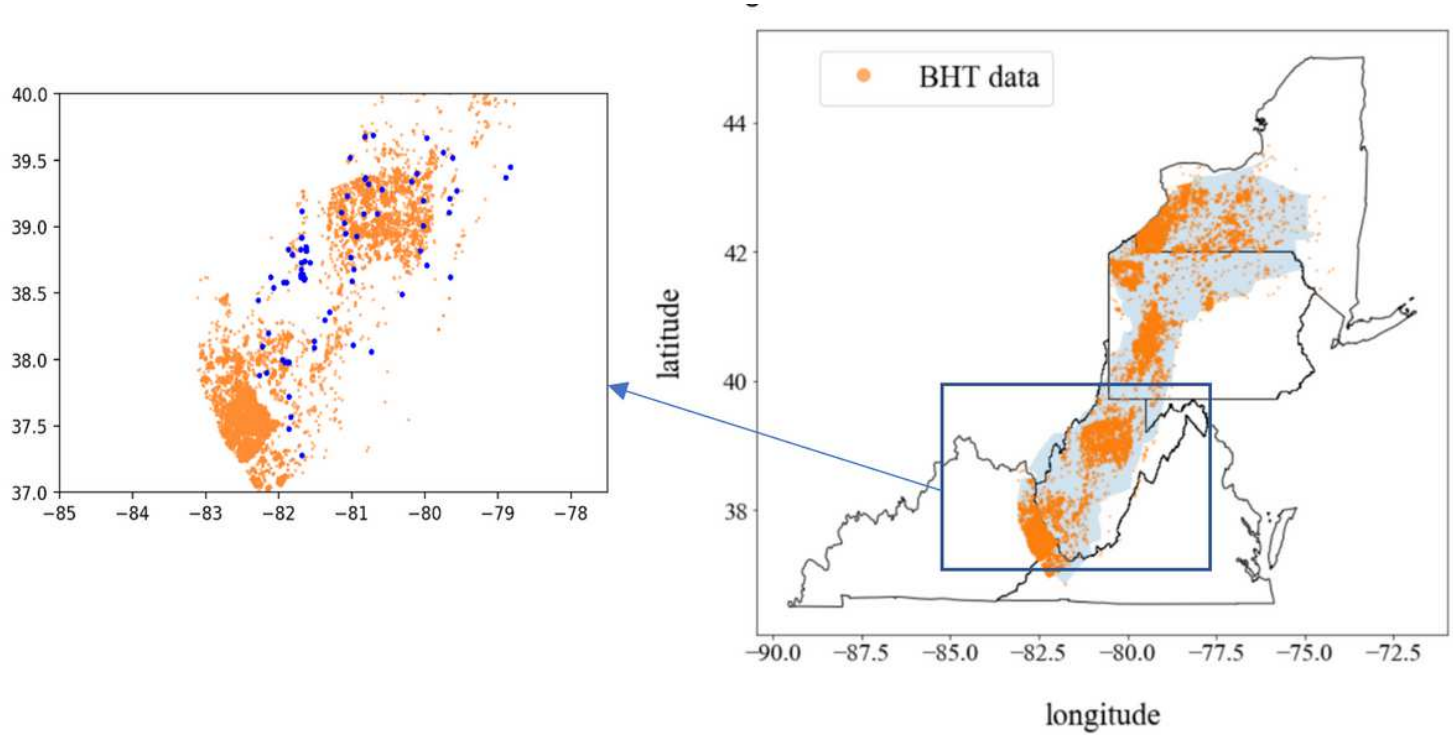


Figure 1

Right plot represents the spread of oil and gas wells in the first dataset (containing 20,750 BHT data points). In the left plot, the locations of the fifty-eight newly obtained wells (with full temperature profile) are annotated using the blue color.

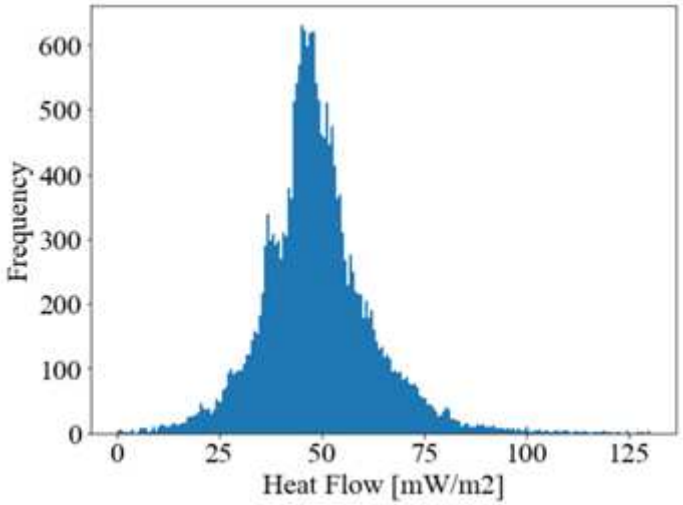


Figure 2

Heat-flow histogram after outlier removal.

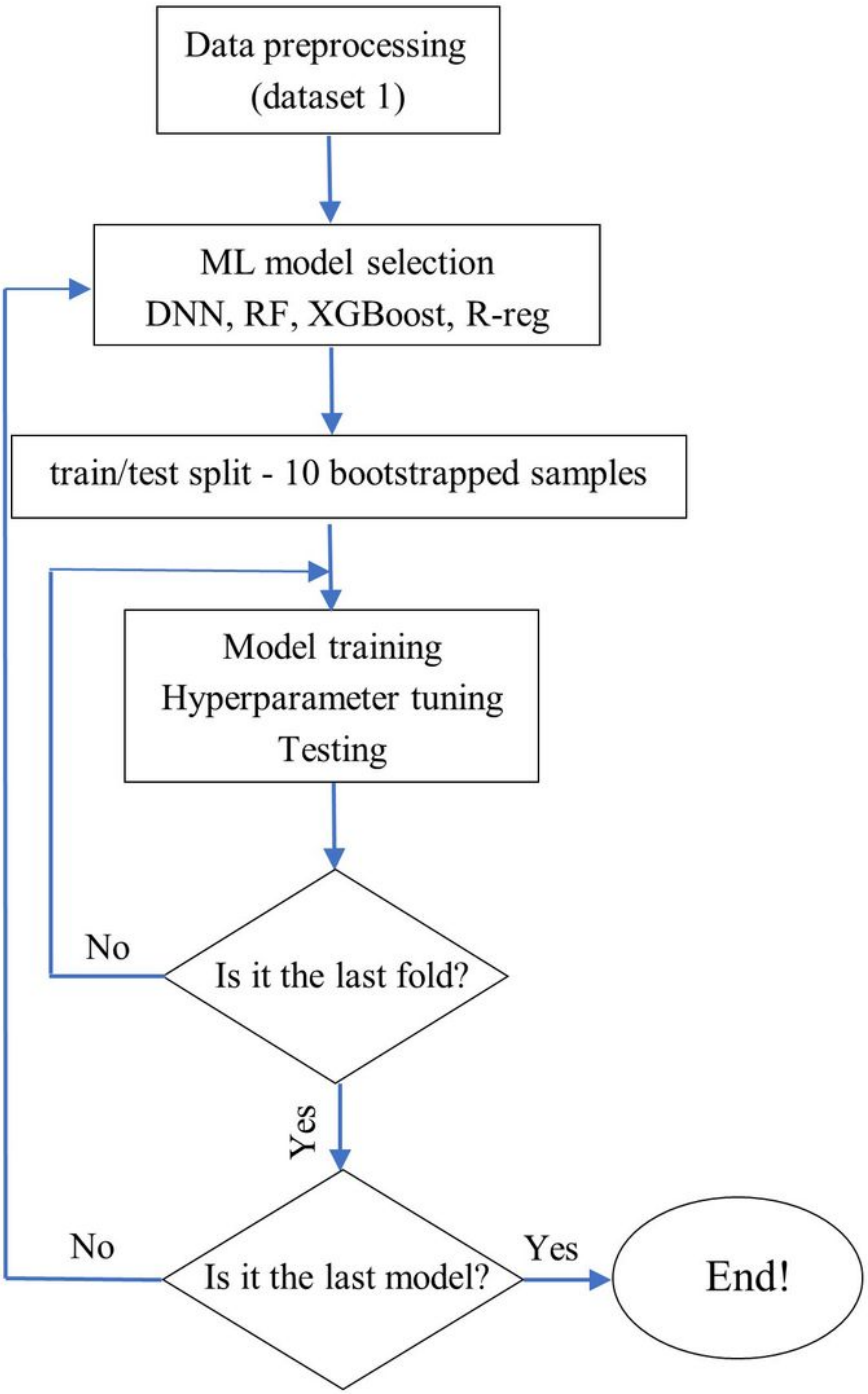


Figure 3

Developed machine learning pipeline

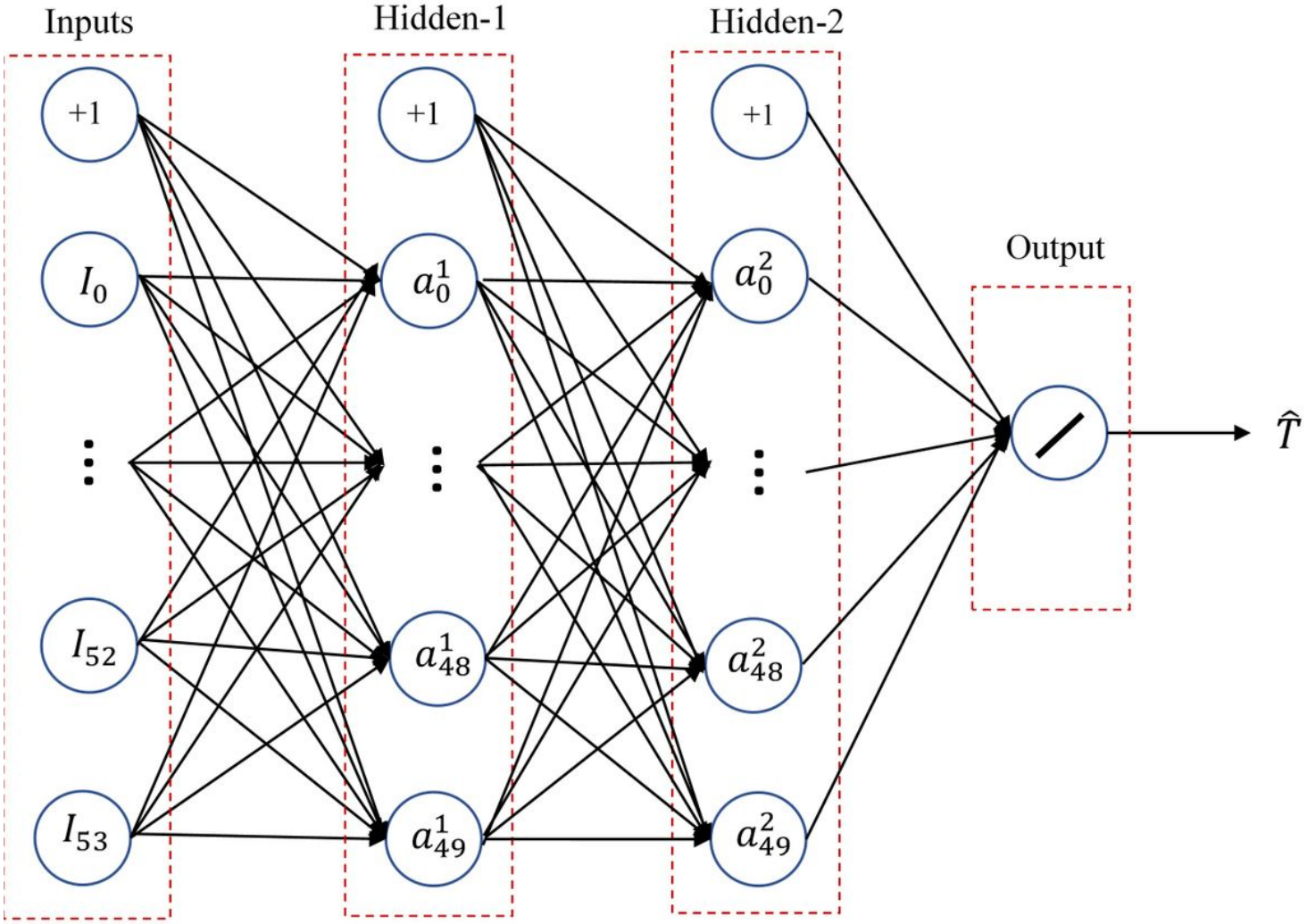


Figure 4

Deep Neural Network architecture for subsurface temperature prediction

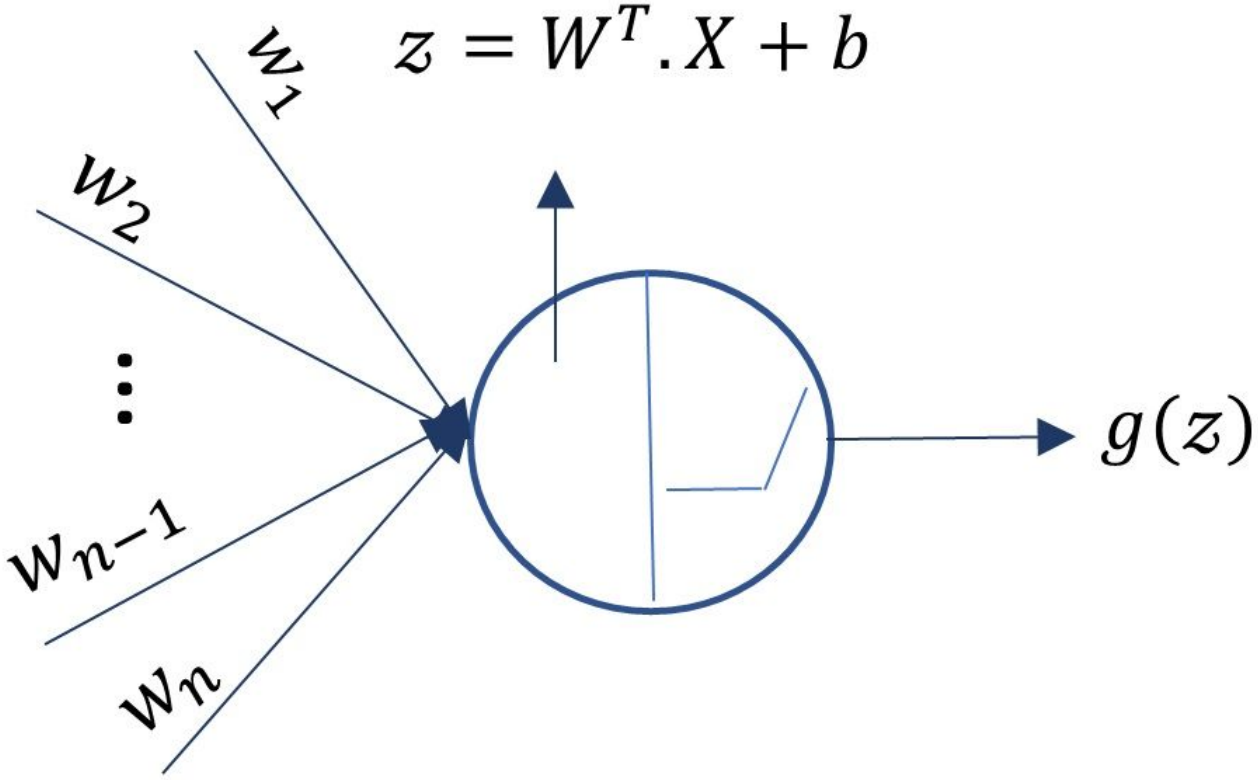


Figure 5

Single neuron illustration

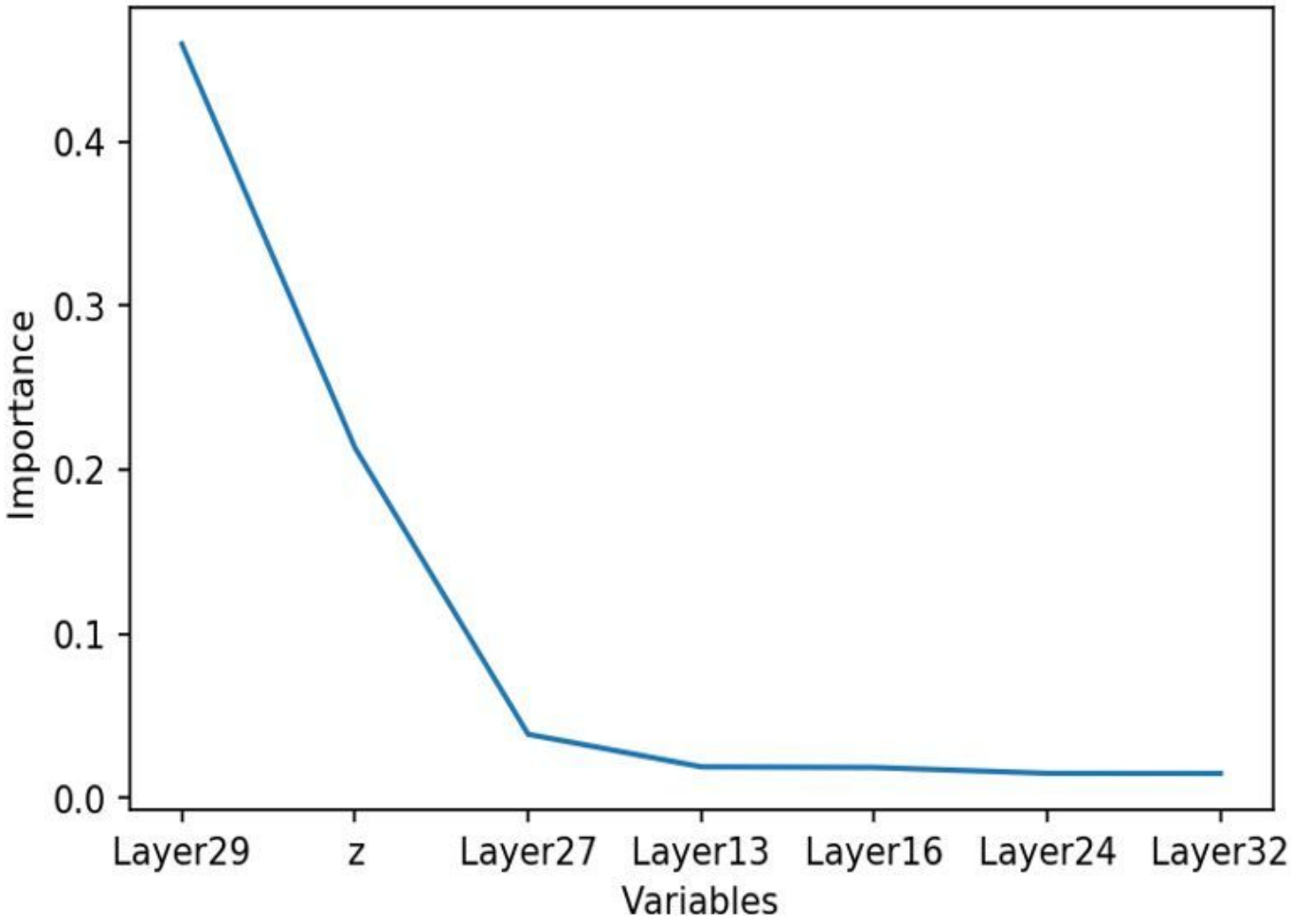


Figure 6

Important parameters identified by XGBoost machine learning model

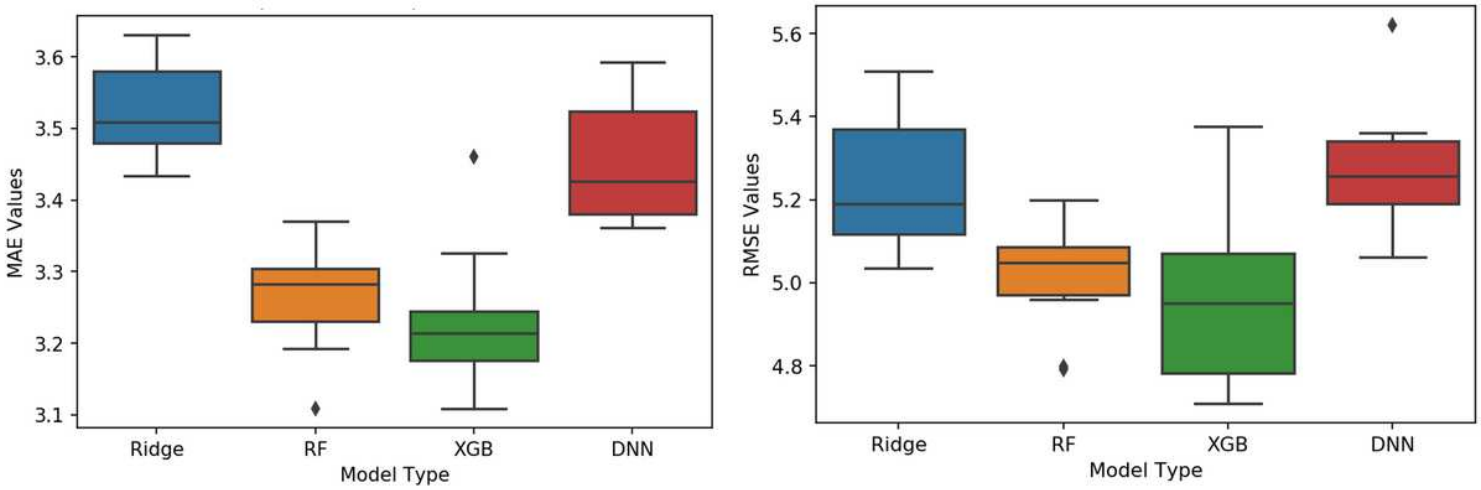


Figure 7

Accuracy comparison between four machine learning models

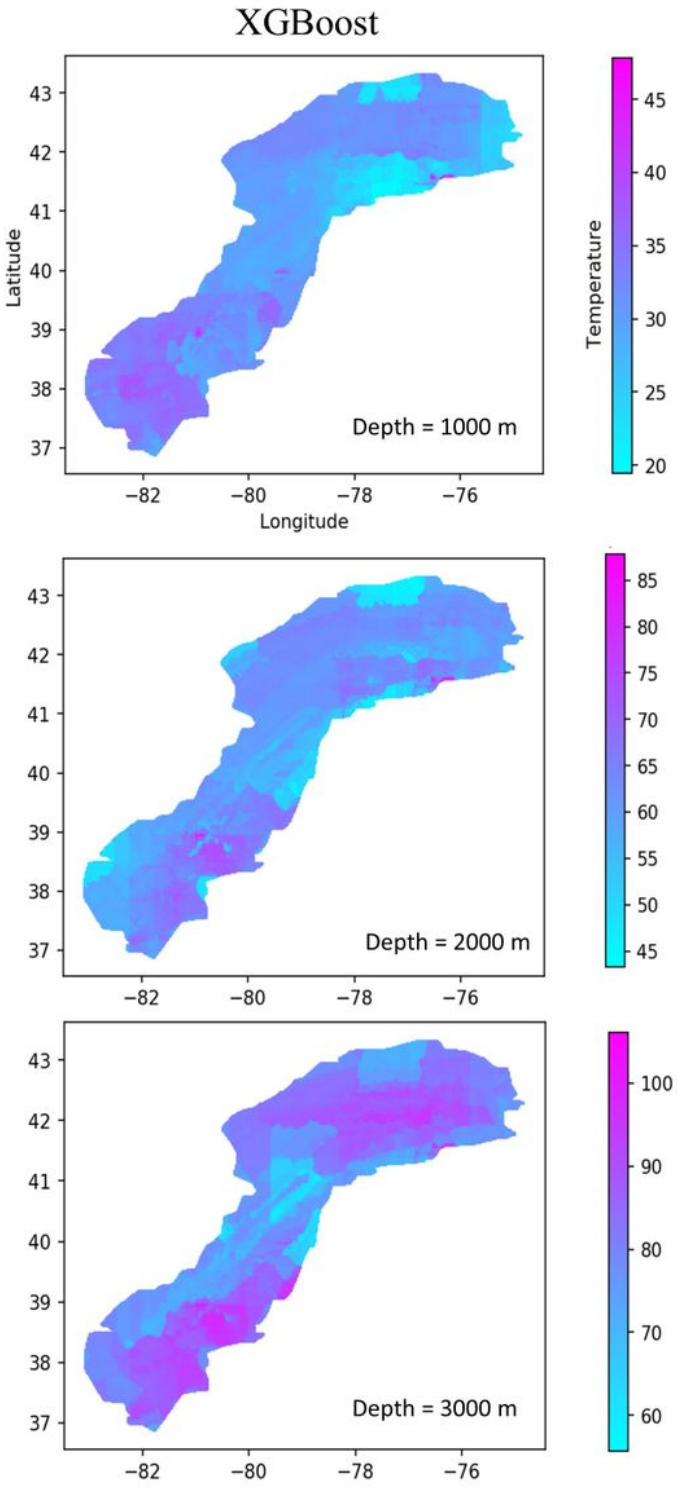
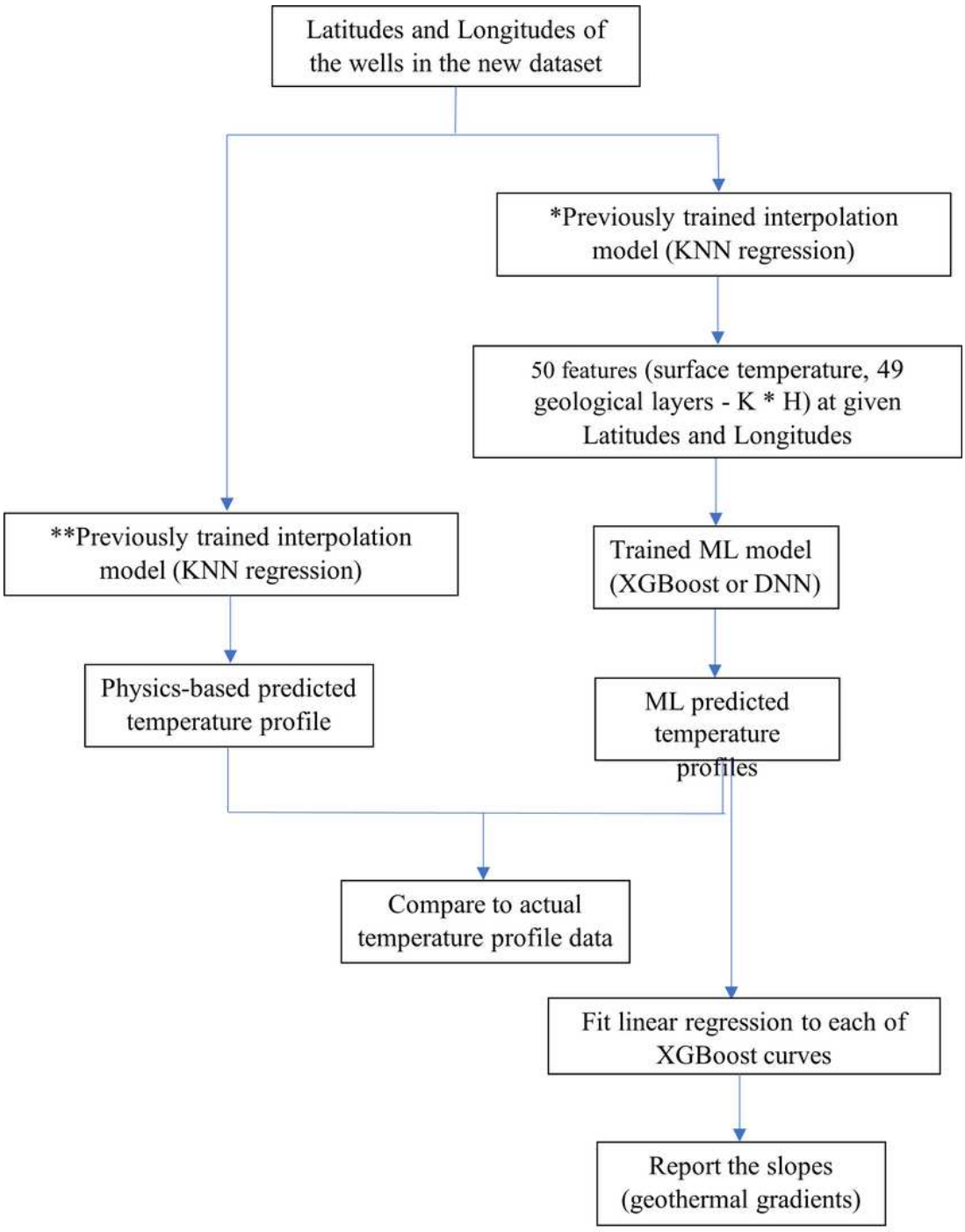


Figure 8

Temperature map at three different depths using XGBoost model



*Interpolation model is trained using the feature information in the main dataset
**Interpolation model is trained using the predicted temperature profiles at 20,750 locations from thermal conductivity model

Figure 9

Figure 9

Followed procedure for comparing predictions from physics-based and machine learning models

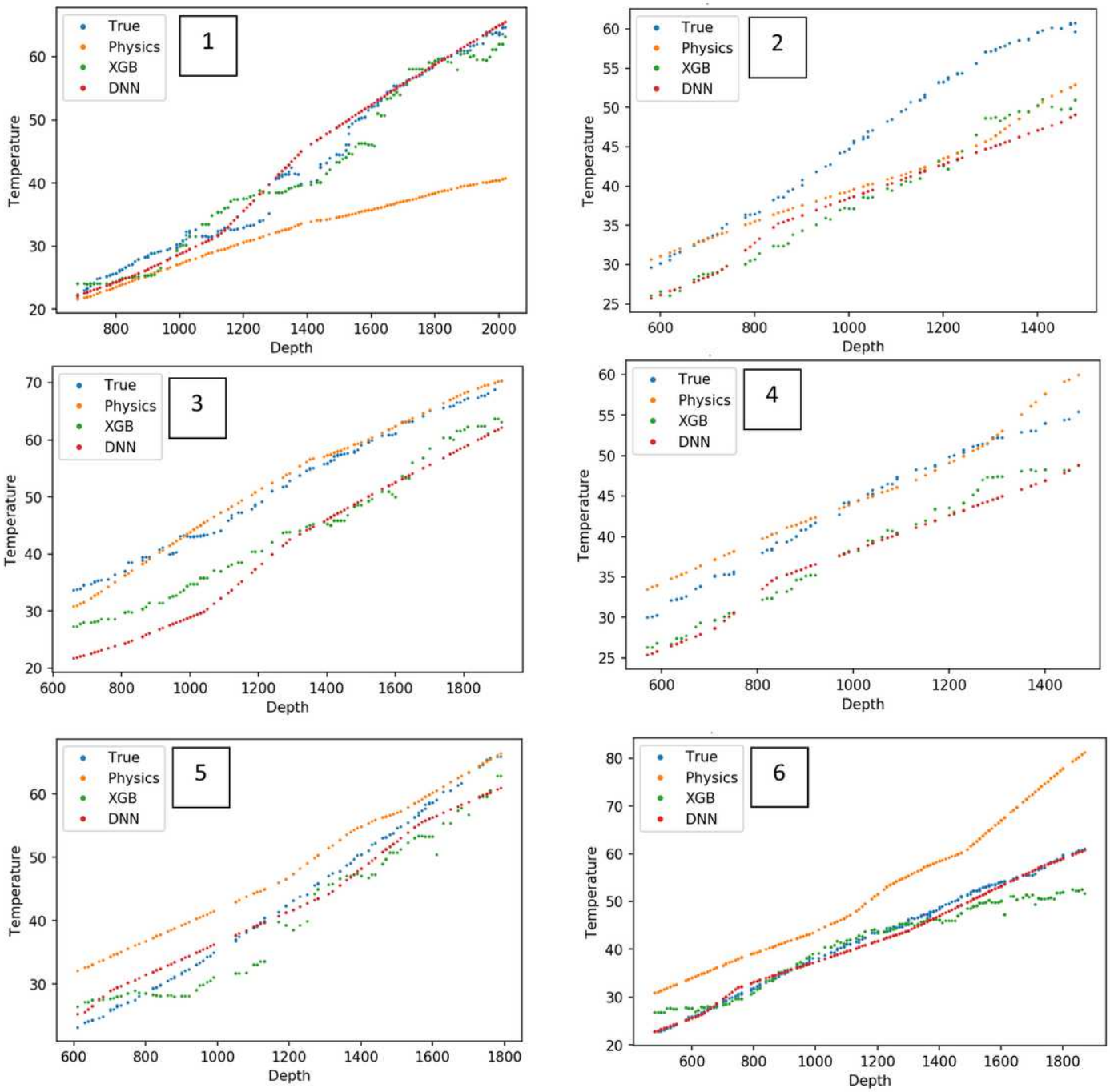


Figure 10

Temperature profile predictions using thermal conductivity, XGBoost and DNN models versus measured data. The units are [°C] and [m] for temperature and depth, respectively.

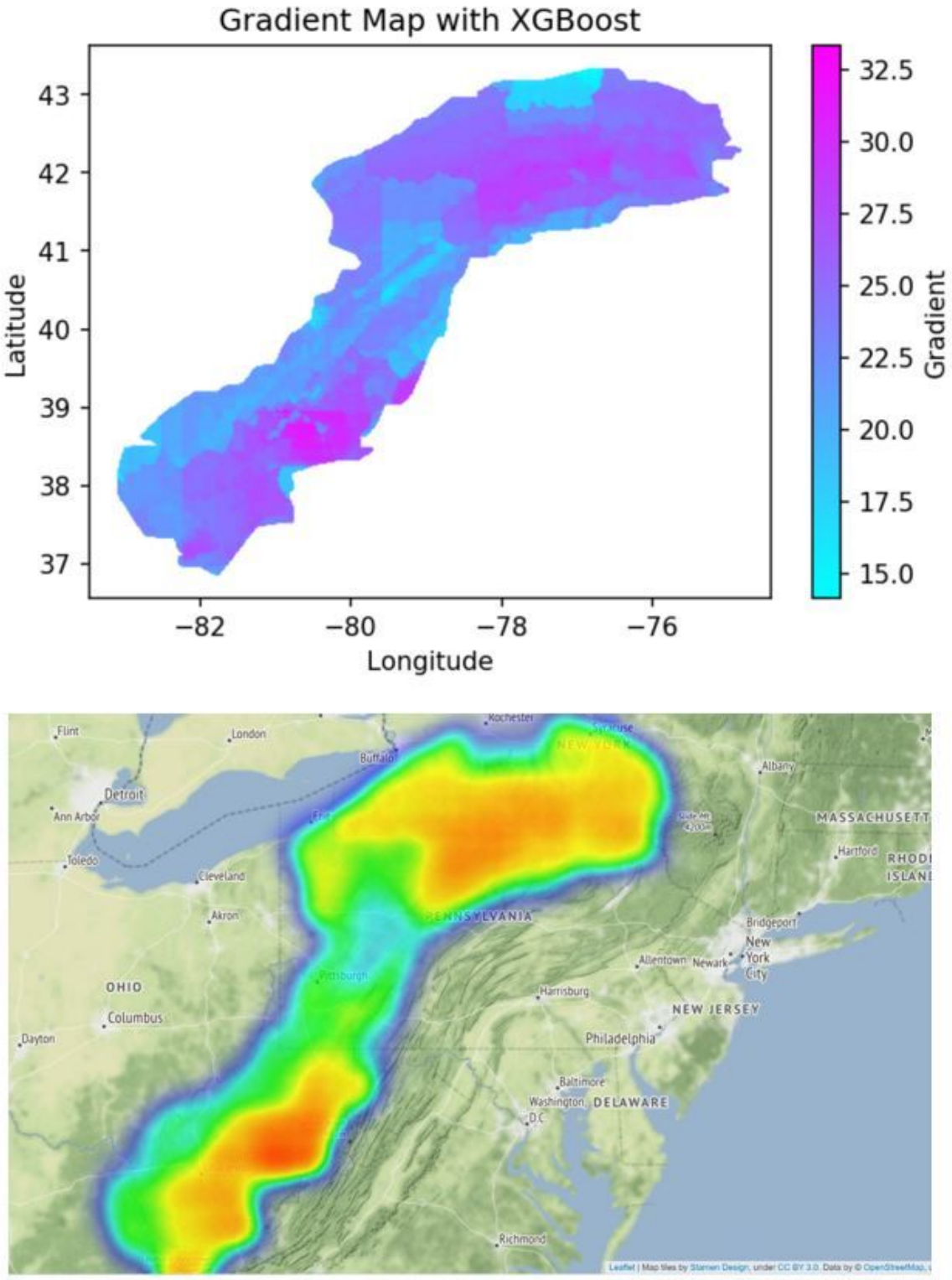


Figure 11

Geothermal gradient map using XGBoost model. The gradient has the unit of [°C/km].

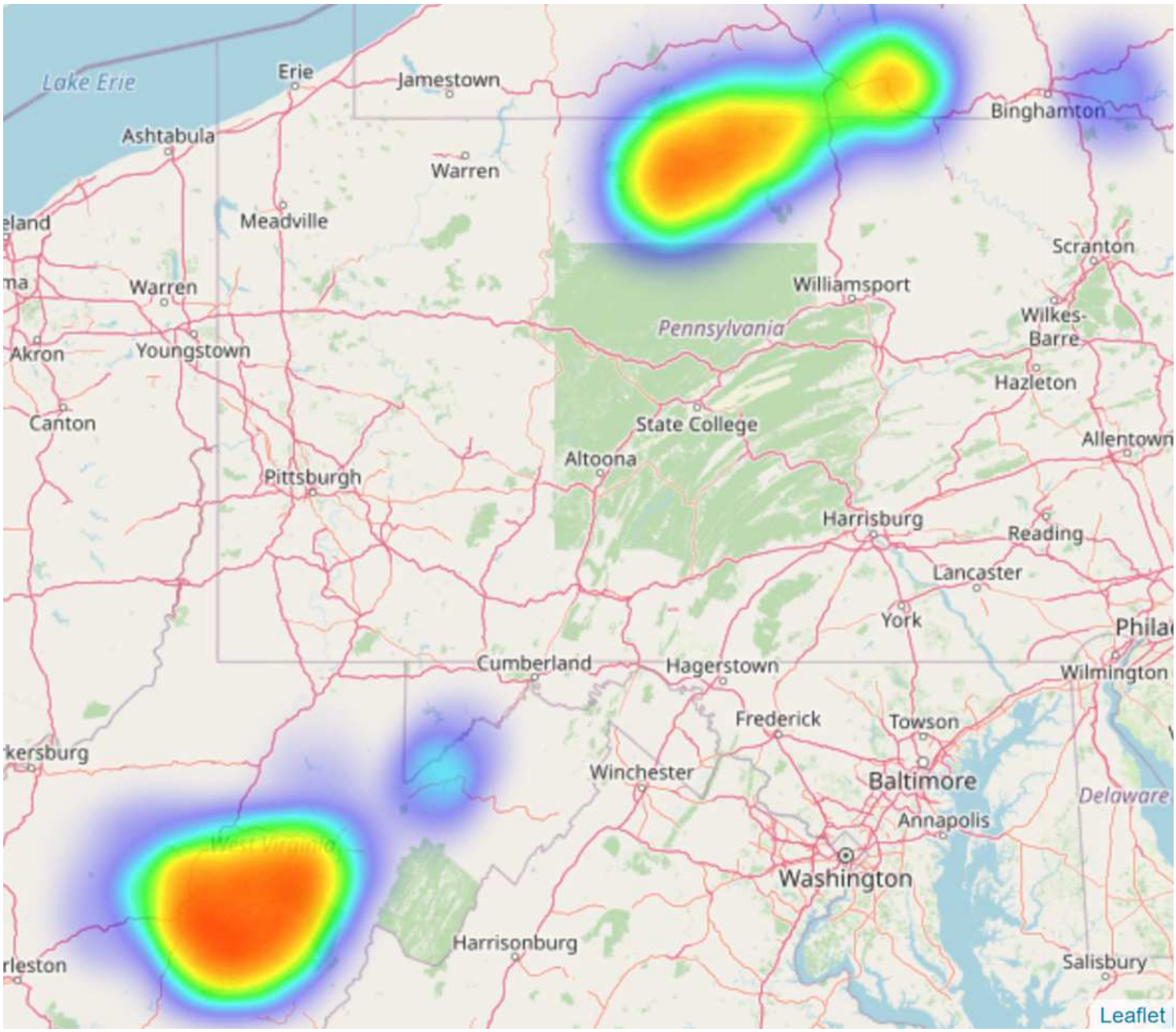


Figure 12

Regions with subsurface temperature gradient higher than 27.5 [°C/km]. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.