Predicting Large Hydrothermal Systems

Stanley P. Mordensky[1], Erick R. Burns[1], Jacob DeAngelo[2], John J. Lipor[3]

[1]U.S. Geological Survey, Portland OR 97201, USA

[2]U.S. Geological Survey, Moffett Field CA 94035, USA

[3]Portland State University, Portland OR 97201, USA

Keywords

geothermal, supervised machine learning, heat flow, regression, play fairway analysis, PFA, INGENIOUS

ABSTRACT

We train five models using two machine learning (ML) regression algorithms (i.e., linear regression and XGBoost) to predict hydrothermal upflow in the Great Basin. Feature data are extracted from datasets supporting the INnovative Geothermal Exploration through Novel Investigations Of Undiscovered Systems project (INGENIOUS). The label data (the reported convective signals) are the difference between the background conductive heat flow and the well heat flow. The reported convective signals contain outliers that may affect upflow prediction, so the influence of outliers is tested by constructing models for two cases: 1) using all the data (i.e., -91 to 11,105 mW/m^2), and 2) truncating the range of labels to include only reported convective signals between -25 and 200 mW/m^2. Because hydrothermal systems are sparse, models that predict high convective signal in smaller areas better match the natural frequency of hydrothermal systems. Preliminary results demonstrate that XGBoost outperforms linear regression. For XGBoost using the truncated range of labels, half of the high reported signals are within < 3 % of the highest predictions. For XGBoost using the entire range of labels, half of the high reported signals are within < 13 % of the highest predictions. Although this implies that the truncated regression is superior, the all-data model better predicts the locations of power-producing systems (i.e., the operating power plants are in a smaller fraction of the study area given by the highest predictions). Even though the models generally predict greater hydrothermal upflow for higher reported convective signals than for lower reported convective signals, both XGBoost models consistently underpredict the magnitude of higher signals. This behavior is attributed to low resolution/granularity of input features compared with the scale of a hydrothermal upflow zone (a few km or less across). Trouble estimating exact values while still reliably predicting high versus low convective signals suggests that an alternate strategy such as ranked ordinal regression (e.g., classifying into ordered bins for low, medium, high, and very high convective signal) might fit better models, because doing so reduces problems introduced by outliers while preserving the property of larger versus smaller signals.

1. Introduction

The U.S. Geological Survey is developing a geothermal assessment update for the Great Basin. As part of these efforts and in support of the INnovative Geothermal Exploration through Novel Investigations Of Undiscovered Systems project (INGENIOUS; e.g., Ayling et al., 2022a), DeAngelo et al. (2022) produced a map representing the estimated conductive heat flow for the region that allowed for differences between the modeled conductive heat flow and the heat flow measurements from the wells used to produce the model of conductive heat flow (Fig. 1). We term these differences as the reported convective signal. Larger reported convective signals are assumed to be indicative of convective hydrothermal upflow. In total, the heat flow model by DeAngelo et al. (2022) created 3,869 convective signals ranging in value from -91 to 11,105 mW/m^2 (Figs. 1,2).

Past conventional hydrothermal energy assessments used classification strategies to construct favorability maps (e.g., the presence or absence of a hydrothermal system; e.g., Williams and DeAngelo, 2008), but these assessment workflows suffered from not having confirmed sites with known hydrothermal systems (i.e., reliable negatives) for use during model fitting. The reported convective signals from DeAngelo et al. (2022) allow for the use of regression strategies to fit using example sites with low and high signals that can then predict the magnitude of the differences from conductive heat flow imparted by hydrothermal systems.

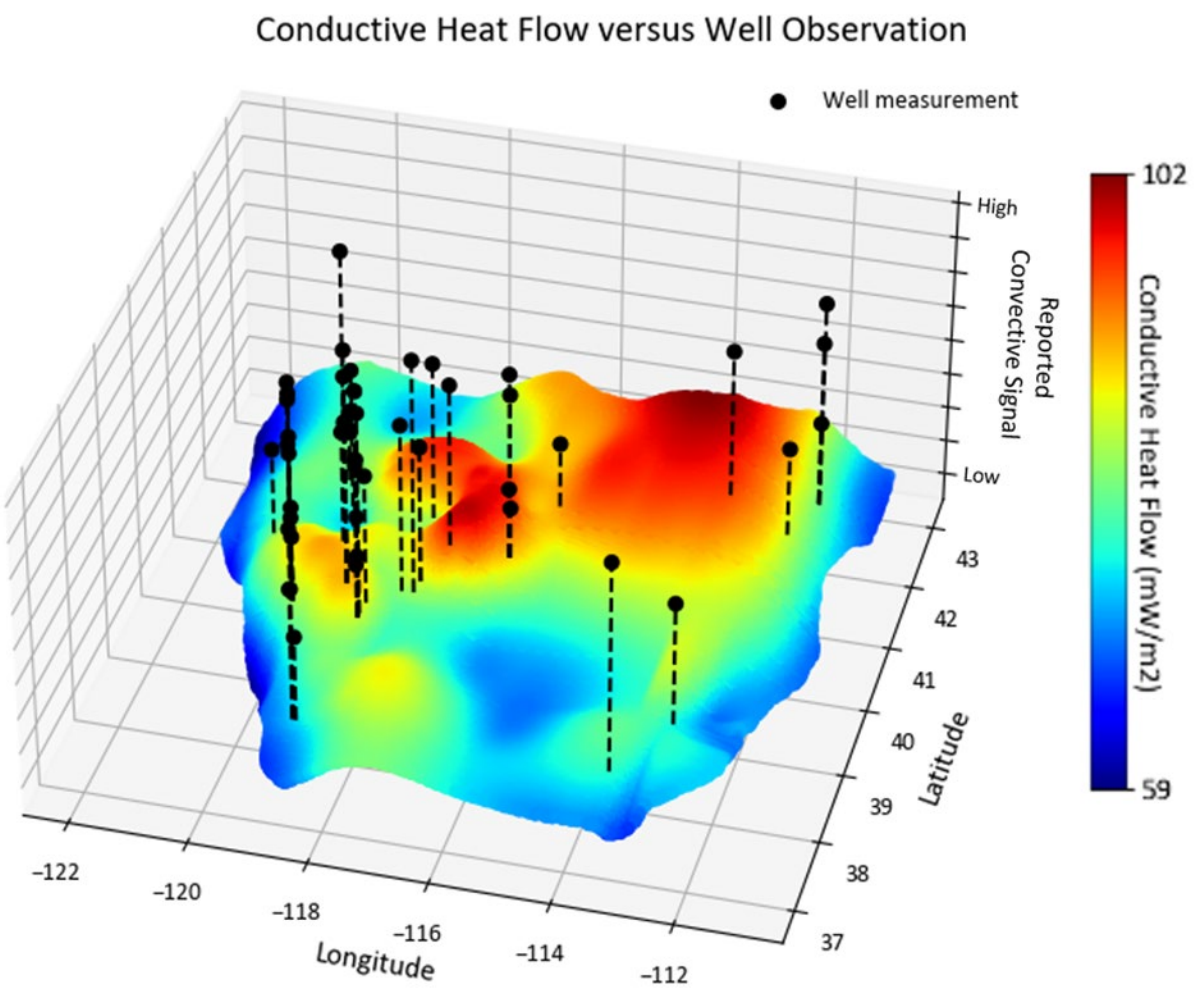

Figure 1. Oblique perspective of the conductive heat flow surface modeled by DeAngelo et al. (2022). Black points represent heat flow measurements from wells. The dashed lines represent the convective signal (i.e., difference in heat flow between the modeled conductive heat flow and the well measurements). The z-axis is not to scale. The reported convective signals depicted here are only a small subset of high reported convective signals from DeAngelo et al. (2022).

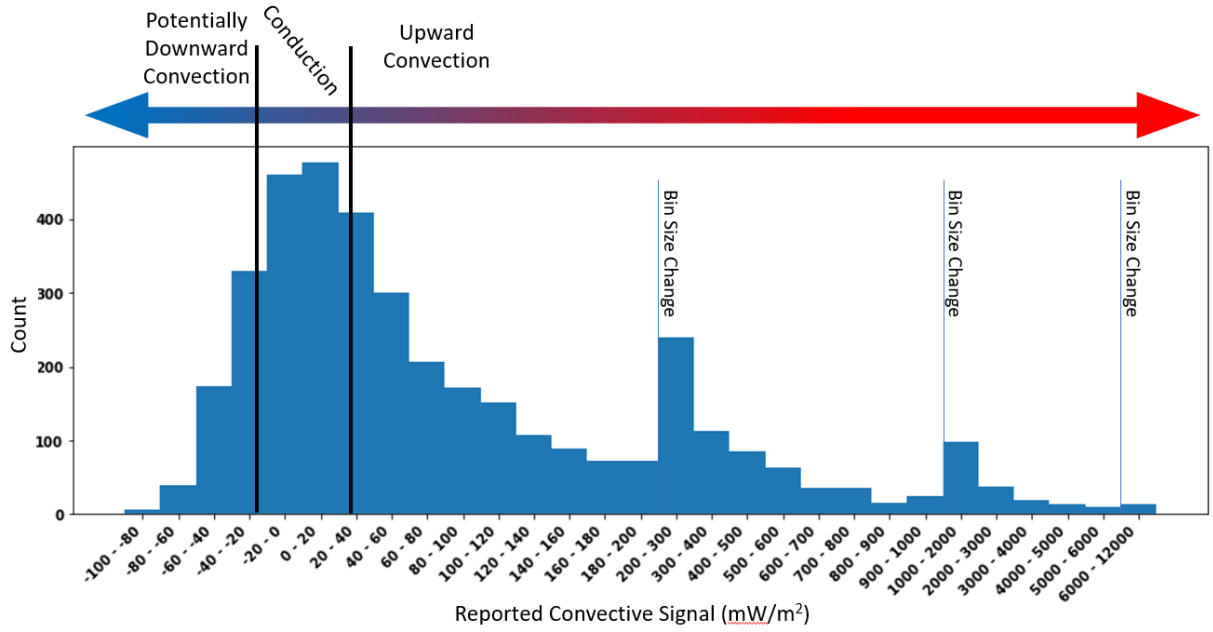

Figure 2. Distribution of reported convective signals by DeAngelo et al. (2022). Note that bin size changes at 200 mW/m², 1,000 mW/m², and 6,000 mW/m². Modified from Figure 4 of DeAngelo et al. (2023).

In general, the upper limit for regional heat flow across the western United States is often roughly estimated as 35 to 80 mW/m^2, where heat flow > 50 mW/m^2 above regional trends suggests some form of hydrothermal component (e.g., Burns et al., 2015). Although some volcanic complexes (e.g., Yellowstone Volcano, WY; Medicine Lake Volcano, CA) require higher heat flow estimates (e.g., > 150 mW/m^2), regional heat flow nonetheless remains bound in the hundreds of mW/m^2 at most hydrothermal systems. Yet, some of the well data used by DeAngelo et al. (2022) report heat flow values in excess of thousands of mW/m^2, thereby suggesting that these high measured heat flow values are from actively convecting hydrothermal systems. Reciprocally, the heat flow within several hundreds of mW/m^2 above conductive heat flow more likely represents heat flow measurements from wells in the vicinity of convective hydrothermal upflow. In the context of regressing a convective signal, the influence of these two different types of heat flow measurements remains unknown.

Herein, we present our ongoing research describing methods to regress a convective signal for hydrothermal upflow across the Great Basin. We compare models produced by two different algorithms (i.e., linear regression and XGBoost [Chen and Guestrin, 2016]) using the entire range of reported convective signals as labels and datasets related to INGENIOUS as features. To remove the potential influence of heat flow measurements from actively convecting systems, we also implement a third modeling approach using XGBoost and a truncated range of reported convective signals (-25 to 200 mW/m^2).

2. Methods

We implement three modeling approaches (referred to as Linear Regression, All-Data XGBoost, and Truncated-Data XGBoost) to create five new machine learning (ML) models that predict the magnitude of convective hydrothermal upflow across the Great Basin using the reported convective signals from DeAngelo et al. (2022) as the label data and 16 datasets supporting

INGENIOUS as the feature data. The first approach (Linear Regression) uses linear regression and the entire range of the reported convective signals to fit a single model. The second approach (All-Data XGBoost) uses XGBoost and the entire range of reported convective signals to fit two models. The third approach (Truncated-Data XGBoost) uses XGBoost and only reported convective signals between -25 and 200 mW/m^2 to fit two models. This approach presumes that: 1) extremely negative convective signals (< -25 mW/m^2) represent a process other than no convection or potential hydrothermal convection, perhaps possible downward convection; and 2) extremely positive label values (> 200 mW/m^2) represent different conditions than convective hydrothermal upflow in the general vicinity (e.g., potentially fault-driven fluid pathways).

In the remainder of this section, we detail the selection, preprocessing, and exploration of the data, describe the training approaches and why each XGBoost approach requires two models, and conclude with measures of feature importance.

2.1 Labeled Data

We infer three general components in the values for the reported convective signals. The first component is defined by the wells having no or a low convective signal in DeAngelo et al. (2023). These low convective signals have a symmetrical distribution about a mean of nearly 0 mW/m^2 with a standard deviation of nearly 25 mW/m^2 (see Section 3.2 in DeAngelo et al., 2023); hence, we define wells with a reported convective signal within 25 mW/m^2 (i.e., one standard deviation) of 0 mW/m^2 as having a low reported convective signal. We define wells with a reported convective signal two standard deviations greater than 0 mW/m^2 (i.e., > 50 mW/m^2) as having a high reported convective signal. Likewise, we regard wells with values of 25 to 50 mW/m^2 as having an intermediate reported convective signal.

We use the INGENIOUS grid of 250-m-by-250-m cells across most of the Great Basin and INGENIOUS study area (Ayling et al., 2022c) so that there are 7,814,099 cells that serve as examples. Of these grid cells, 3,869 have a reported convective signal (i.e., a label). To account for bias potentially imparted by the smoothly varying feature data, we remove labeled examples with a low reported convective signal that are within a specified distance to a labeled example with a high reported convective signal; we choose a distance of 4 km with consideration for the scale at which structural perturbations influence permeability (e.g. faults; Barbour, 2015; Xue et al., 2016). The remaining examples without heat flow measurements serve as unlabeled examples.

2.2 Feature Data

For the feature data, we use select datasets that support the INGENIOUS project (Table 1; e.g., Ayling et al., 2022a; Ayling et al., 2022b; DeAngelo et al., 2022; Glen et al., 2022; Peacock and Bedrosian, 2022; Kreemer and Young, 2023). We standardize each feature (i.e., subtract the mean and divide by the standard deviation of each dataset) to bring each feature to the same unitless scale. The cumulative distribution of values from each dataset is compared to the cumulative distribution of values at the sites with reported convective signals from DeAngelo et al. (2022), allowing for an evaluation of sample bias (e.g., are only high heat flow areas sampled?).

Table 1. Features and their data sources.

Feature	Reference
Conductive Heat Flow	DeAngelo et al. (2022)
Distance to Nearest Quaternary Fault	Ayling et al. (2022a)
Distance to Nearest Quaternary Magmatic Activity	Ayling et al. (2022b)
Magnetic Field	Glen et al. (2022)
Isostatic Gravity	Glen et al. (2022)
Depth-to-Basement	Glen et al. (2022)
Shear Strain Rate	Kreemer and Young (2023)
Dilation Strain Rate	Kreemer and Young (2023)
Second Invariant of Strain-Rate Tensor	Kreemer and Young (2023)
Independent Seismic Density (> 2M, ≤ 30-km depth, n = 200, α = 0.05)	Kreemer and Young (2023)
Foreshock/Aftershock Seismic Density (> 2M, ≤ 30-km depth, n = 200, α = 0.05)	Kreemer and Young (2023)
Electrical Surface Conductance (2 - 12 km)	Peacock and Bedrosian (2022)
Electrical Middle Crust Conductance (12 - 20 km)	Peacock and Bedrosian (2022)
Electrical Lower Crust Conductance (20 - 50 km)	Peacock and Bedrosian (2022)
Electrical Upper Mantle Conductance (50 - 90 km)	Peacock and Bedrosian (2022)
Electrical Mantle Conductance (90 - 200 km)	Peacock and Bedrosian (2022)

In order to understand the correlative relationships of the reported convective signals and features, we examine the Pearson and Spearman correlation coefficients. The Pearson and Spearman correlation coefficients provide measures of correlation between the labels and features (see generally Lee Rodgers and Nicewander, 1988). The Pearson correlation coefficient provides a measure of linear correlation between features. The Spearman correlation coefficient provides a measure of correlation between the ranked values of features. Both have a minimum and maximum of negative one and one, corresponding to negative and positive correlation, respectively. The greater the absolute correlation coefficient, the greater the degree of correlation.

2.3 Three Modeling Approaches

We apply three modeling approaches (Linear Regression, All-Data XGBoost, and Truncated-Data XGBoost) to predict convective signals across the Great Basin. Below, we provide more details about these approaches.

2.3.1 Linear Regression

We select linear regression for its simplicity and linearity; that is, linear regression provides a baseline against which to compare more complex approaches. We define multiple linear regression in Equation 1 as

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ 1 & \vdots & \vdots & & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{bmatrix}, \quad (1)$$

where y_n is the label of the n^{th} example, $x_{n,m}$ is the m^{th} feature value of the n^{th} example, β_m is the weight (i.e., fitting parameter) of the m^{th} feature. Linear regression aims to fit the labels y_n as a linear function of the n feature vectors (or examples). Let the i^{th} example be x_i with corresponding label y_i and $\hat{y}_i$ be the corresponding prediction; linear regression then minimizes the following equation for root mean square error (RMSE), provided in Equation 2 as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|^2}. \quad (2)$$

2.3.2 XGBoost

We select XGBoost (Chen and Guestrin, 2016) because Mordensky et al. (2023b) identified XGBoost as a superior ML algorithm when fitting models using geothermal data from the 2008 U.S. Geological Survey Geothermal Resource Assessment (Williams and DeAngelo, 2008; Williams et al., 2008) due to its boosted tree-based architecture, which allows for a non-linear predictor without the need for excessively large datasets. The XGBoost algorithm functions by sequentially adding estimators (i.e., decision trees) to an ensemble, where the goal of each new estimator is to account for the errors of the current ensemble. The final prediction is then a weighted combination of each estimator in the ensemble.

Mean absolute error (MAE) is a common error metric that is insensitive to outliers and is defined in Equation 3 as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \quad (3)$$

in which n is the number of cells, i is the i$^{\text{th}}$ cell, y is the label, and $\hat{y}$ is the prediction. However, MAE is undifferentiable at zero and, therefore, cannot be used as a loss function with XGBoost. To obtain additional robustness to outliers, we select the Pseudo-Huber function (Huber, 1964) as the loss function for the XGBoost approaches because the Pseudo-Huber function is not as sensitive to outliers as RMSE but remains differentiable across its entirety. The Pseudo-Huber function is given by Equation 4 as

$$PL_\delta = \frac{\delta^2}{n}\sum_{i=1}^{n}\sqrt{1 + \left(\frac{y_i - \hat{y}_i}{\delta}\right)^2}, \quad (4)$$

in which PL_δ is the error by the Pseudo-Huber loss function, n is the number of labeled examples, i is the i$^{\text{th}}$ cell, y is the label, $\hat{y}$ is the prediction, and δ is a user-set parameter. In the Pseudo-Huber loss function, we set δ equal to 1. By setting δ to 1, differences between the reported convective

signals from DeAngelo et al. (2023) and the predicted convective signals from the XGBoost models that are less than 1 produce error values similar to that of RMSE and differences greater than 1 produce error values similar to that of MAE. Hence, the Pseudo-Huber loss function with $\delta = 1$ is differentiable about zero while remaining insensitive to outliers. A similar outlier-robust loss could be used for linear regression. However, the relationship between examples and labels is inherently non-linear (see Section 3.2 Feature Data), we hypothesize that a more robust loss would provide minimal improvement to the performance of linear regression.

For hyperparameter optimization, we tune the number of estimators, maximum depth of each estimator, learning rate, and number of leaves per node across 120 train-test splits. We fit final models using all the labeled examples per that approach and the median hyperparameter values from the 120 train-test splits using the U.S. Geological Survey high-performance computer DENALI (Falgout et al., 2021).

To check and prevent against overfitting in the final models, we develop a variant of the low progress method for early stopping (see generally Tian and Zhang, 2022) by comparing the change in testing loss and training loss per new estimator in the 120 train-test splits. During the addition of early estimators, the rates of improvement in the training and testing data are similar, but eventually, the rate of improvement in the testing data slows compared to the rate of improvement in the training data, indicating that overfitting is beginning. Specifically, we define early improvement as the improvement between estimators 1 and 2, and we compare that initial ratio to the ratio from each sequential estimator as the Relative Reduction in Slope (RRiS) given in Equation 5 as:

$$\text{RRiS} = \frac{\dfrac{MAE_{i-1}^{testing} - MAE_{i}^{testing}}{MAE_{i-1}^{training} - MAE_{i}^{training}}}{\dfrac{MAE_{1}^{testing} - MAE_{2}^{testing}}{MAE_{1}^{training} - MAE_{2}^{training}}}, \quad (5)$$

where $MAE^{testing}$ is the testing loss, $MAE^{training}$ is the training loss, and i is the ith estimator. Conceptually, we say that the model is beginning to overfit when the ratio of improvement of the testing to training data is some defined fraction of the initial improvement of this ratio from the addition of sequential estimators. We implement early stopping before the ith estimator at which the median RRiS from the 120 train-test splits falls below that defined fraction. In practice, we choose two fractions, an upper fraction of ½ and a lower fraction of ¼, respectively denoted as $RRiS_{1/2}$ and $RRiS_{1/4}$, resulting in two final models per XGBoost approach. Producing these two variations of final models for a single approach allows for a comparison of the predictive skill resulting from the change in model complexity. The goal is that the final model at $RRiS_{1/2}$ might be slightly underfit while the final model at $RRiS_{1/4}$ might be slightly overfit, allowing an analysis of robustness of model estimates (e.g., do both models substantially agree over most of the area?). The two critical fractions are verified based on plots of the 120 train-test splits to ensure that there is a reasonable confidence that the range of produced models are slightly under- and over-fit.

2.4 Feature Importance

We evaluate feature importance for each of the modeling approaches using two algorithm-agnostic measures of feature importance (MAE sensitivity and SHapely Additive exPlanation [SHAP] values; Lundberg and Lee, 2017) with the final models.

MAE sensitivity analysis functions by randomly shuffling the values of a single feature while the other features remain unshuffled, using the model to make new predictions, and then comparing these new predictions with the predictions from the originally unshuffled data and the impact on MAE. By sequentially completing this process through all the features, sensitivity analysis gauges the magnitude of the contribution of each feature toward a prediction.

SHAP values operate similarly to sensitivity analysis at a conceptual level but with some fundamental differences. The SHAP function varies values for every possible combination of feature sets, whereas sensitivity analysis sequentially shuffles only one feature at a time. Also, SHAP measures the differences between predictions and does not rely on a specific performance metric. More specifically, every sample for every feature with consideration for every combination of features is assigned a SHAP value that is the difference between the original and permutated predictions, and the sample SHAP values are then averaged by feature to provide the mean feature SHAP values (Lundberg and Lee, 2017).

3. Results

In this section, the label and feature data used for fitting are presented, the optimal hyperparameters for XGBoost are reported, and the predicted convective signals from all five models are given and compared. Lastly, feature importance is provided.

3.1 Label Data

Pre-processing reduces the total number of labeled examples from the initially available 3,869 reported convective signals. After removing examples with low reported convective signals within 4 km of a high reported convective signal, 3,275 convective signals remain. The remaining reported convective signals span -91 to 11,105 mW/m^2 and have a right-skewed distribution with a median of 262 mW/m^2 and a standard deviation of 784 mW/m^2 (Fig. 3). After truncating the data to only include reported convective signals between -25 and 200 mW/m^2, only 2,156 convective signals remain. The convective signals for the truncated range still have a right-skewed distribution but with a median of 46 mW/m^2 and a standard deviation of 56 mW/m^2.

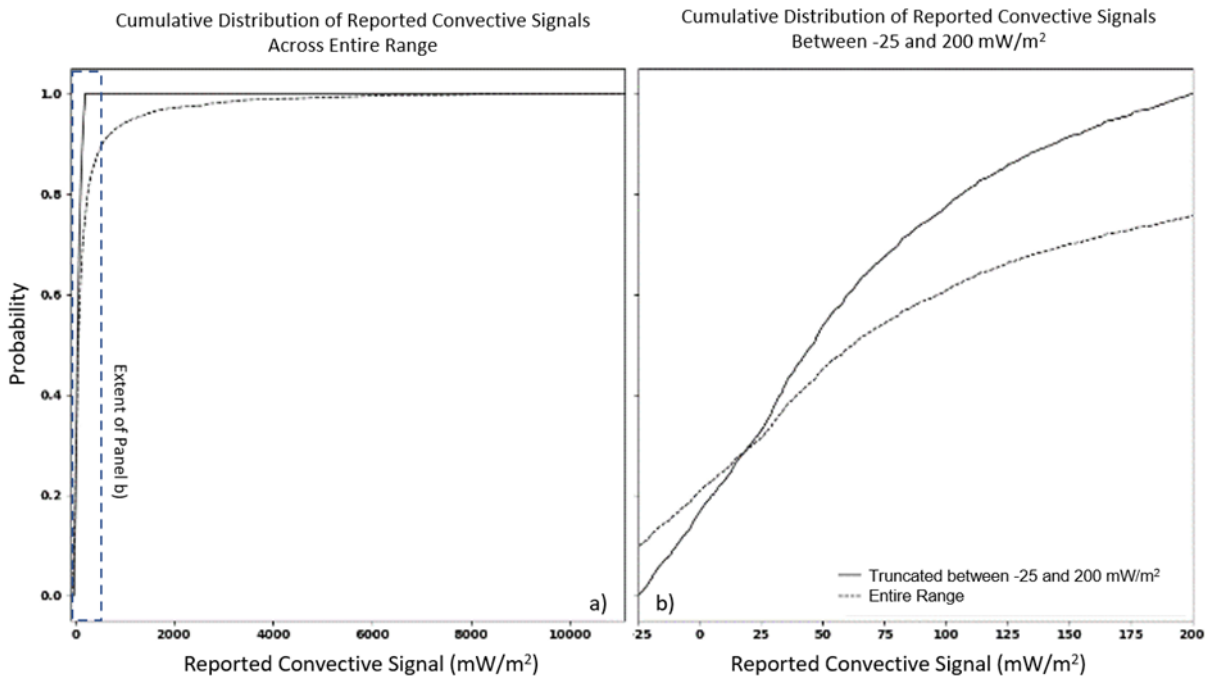

Figure 3. Cumulative distribution functions (CDFs) for reported convective signals: a) over the entire range of reported convective signals; and b) over the range of reported convective signals used for Truncated-Data XGBoost. Dashed curve: CDF for reported convective signals for Linear Regression and All-Data XGBoost; Solid line: CDF for reported convective signals for Truncated-XGBoost. Dashed box in a) provides extent of b).

3.2 Feature Data

The feature data at the reported convective signals generally cover the same range of values as that for the unlabeled examples (Fig. 4) with values having a low rate of change per unit distance. That is, feature maps generally appear smooth despite the resolution of the INGENIOUS grid (i.e., 250-m by 250-m; Fig. 5); however, the seismic density data do not share this characteristic and, instead, have small areas of elevated values.

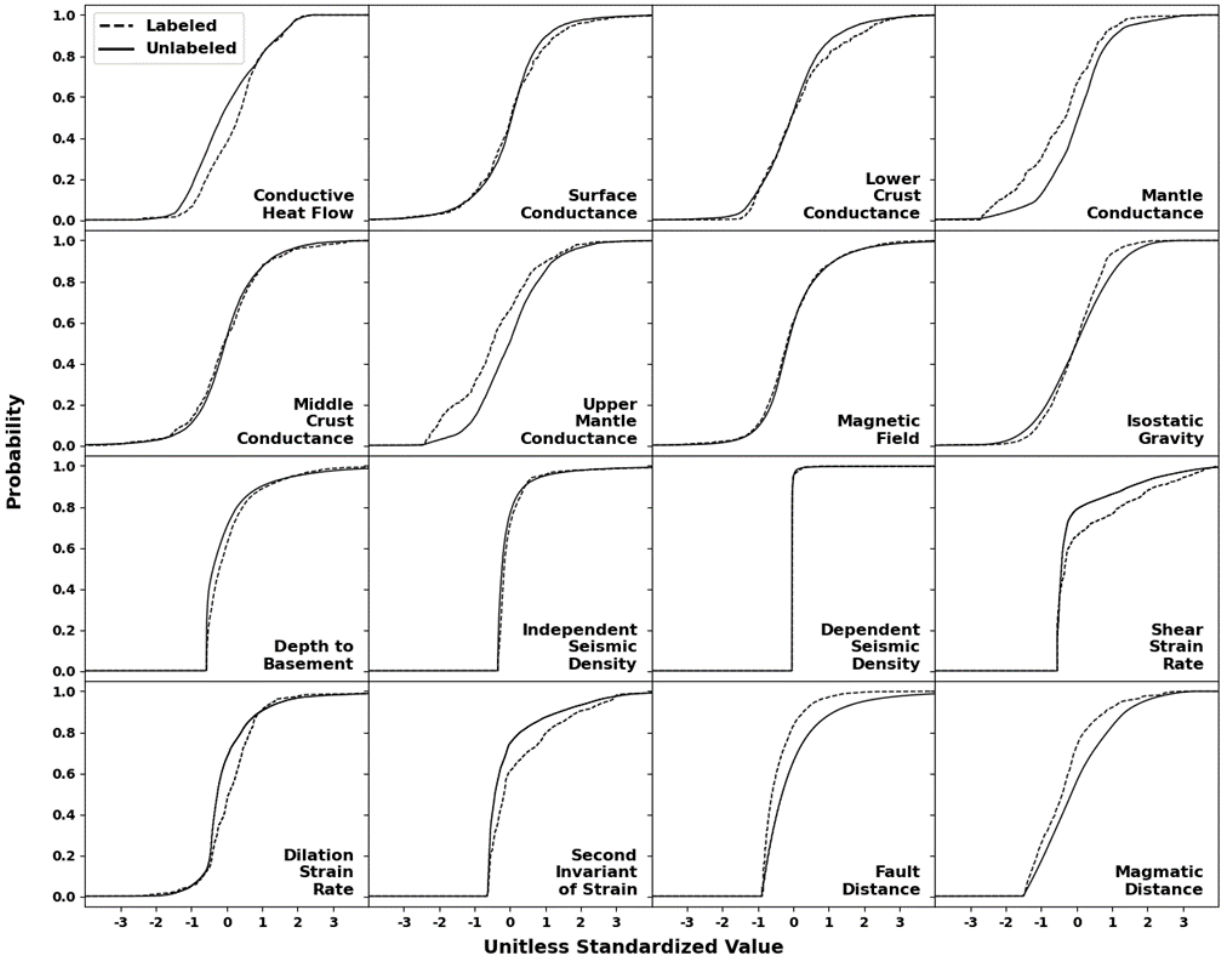

Figure 4. Cumulative distribution functions (CDFs) of standardized unlabeled examples (solid line) compared with CDFs of standardized labeled examples (i.e., examples with reported convective signals; dashed line). Steeper slopes on a CDF indicate a greater density of examples with that feature value. Shallower slopes on a CDF indicate a lower density of examples with that feature value. Differences between lines indicate sample bias relative to the input feature distribution. For example, thermal gradient wells used to construct the heat flow maps from DeAngelo et al. (2022) preferentially sample regions with higher conductive heat flow (upper left panel).

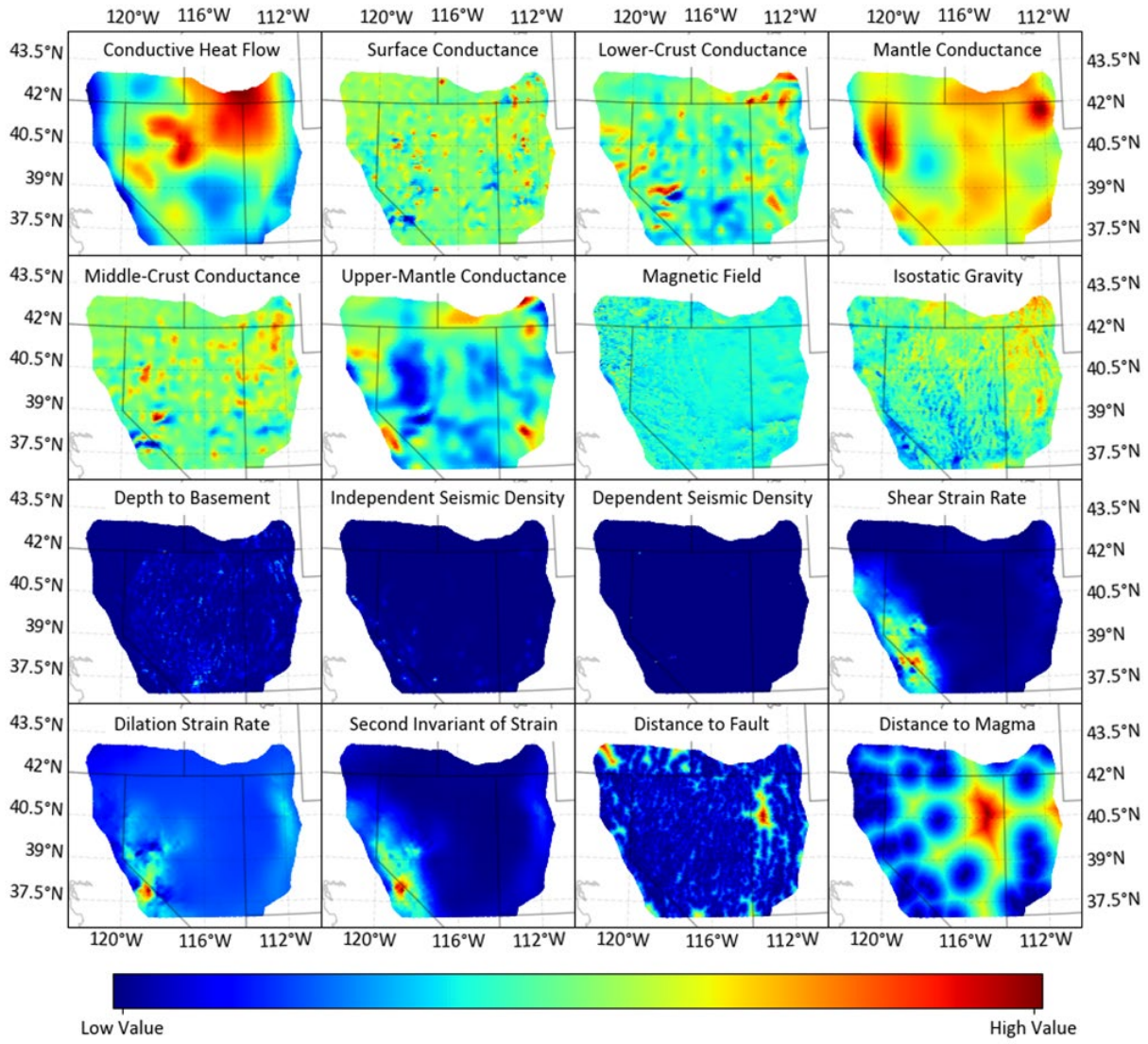

Figure 5. Standardized feature maps for the 16 features from the datasets supporting INGENIOUS. The extent of the study is defined by the complete overlap of the different features. The base map has been made using data from Natural Earth.

In general, most features share moderate or strong correlation with at least one other feature. There are three groups of correlated variables: 1) features from geodetic methods, seismic derivatives, and distance to nearest Quaternary magmatic activity; 2) conductance features; and 3) features from geophysical methods (Fig. 6). The reported convective signals have low correlation with any feature.

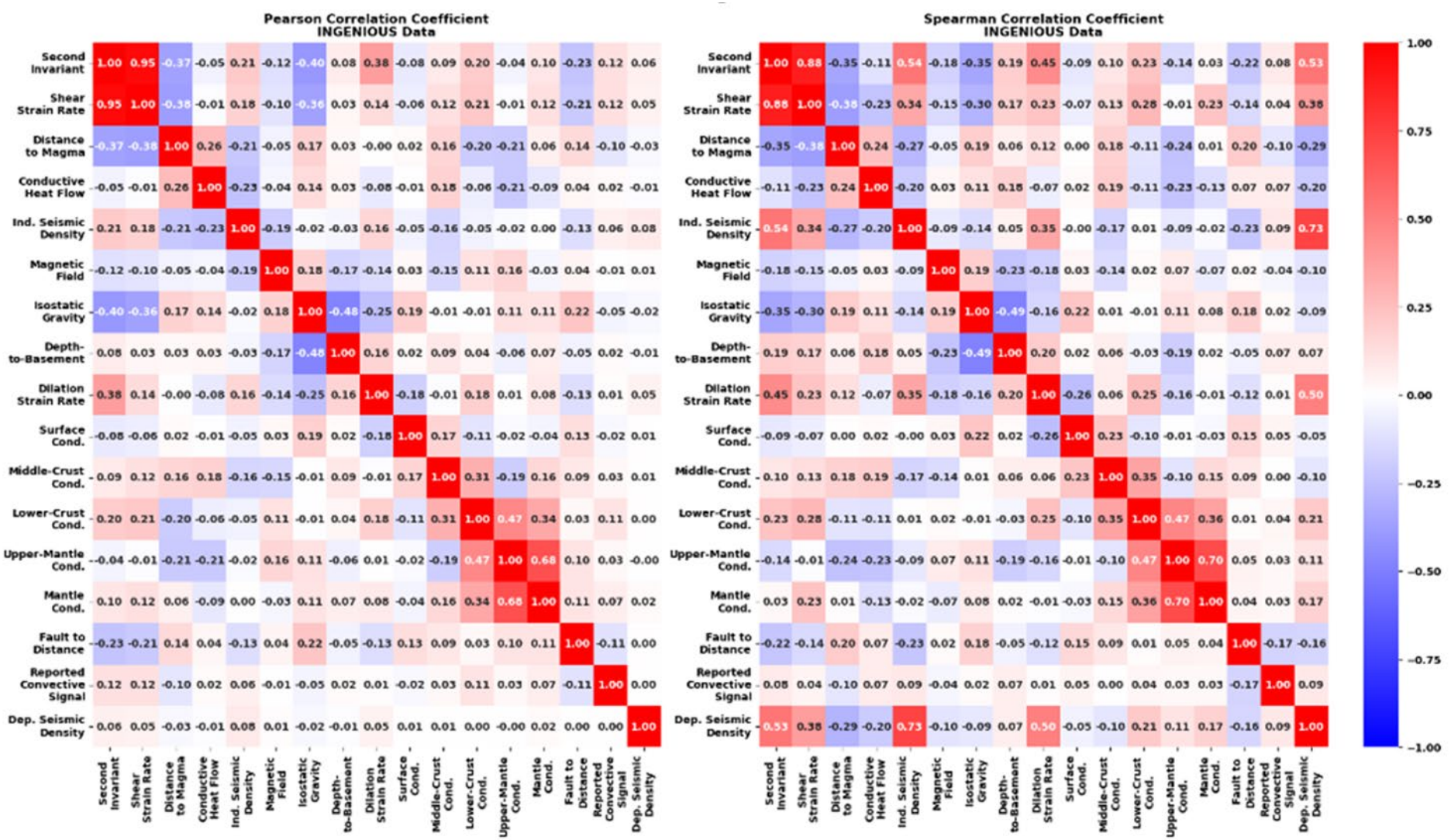

Figure 6. Pearson and Spearman correlation coefficients between input features. Red corresponds to positive correlation coefficients. Blue corresponds to negative correlation coefficients. Fitting from features with extremely high correlation (e.g., 0.95) may negatively impact model performance (e.g., Mordensky et al., 2023a). Abbreviations: Ind. – Independent; Dep. – Dependent; Cond. – Conductance.

3.3 Hyperparameters and Predictions for Convective Hydrothermal Upflow

In this section, we provide optimal hyperparameters and then evaluate model performance by comparing reported and predicted convective signals. Prediction maps detail the geospatial distribution of predicted convective signals. Lastly, we report feature importance.

The median optimal hyperparameter values from the 120 train-test splits are provided in Table 2. The median i^{th} estimator for early stopping at $RRiS_{1/2}$ and $RRiS_{1/4}$ from the 120 train-test splits are provided in Table 2 and Fig. 7. All-Data XGBoost uses fewer estimators (5 and 8 for $RRiS_{1/2}$ and $RRiS_{1/4}$, respectively) than Truncated-Data XGBoost (14 and 24 for $RRiS_{1/2}$ and $RRiS_{1/4}$, respectively), but the estimators for All-Data XGBoost are more complex (18 nodes deep) than Truncated-Data XGBoost (12 nodes deep).

Table 2. Hyperparameters for the XGBoost models.

	i Estimators	Max Depth	Learning Rate	Max Leaves	i^{th} Estimator for Early Stopping at $RRiS_{1/2}$	i^{th} Estimator for Early Stopping at $RRiS_{1/4}$
All-Data XGBoost	15	18	0.05	1	5	8
Truncated-Data XGBoost	40	12	0.05	1	14	24

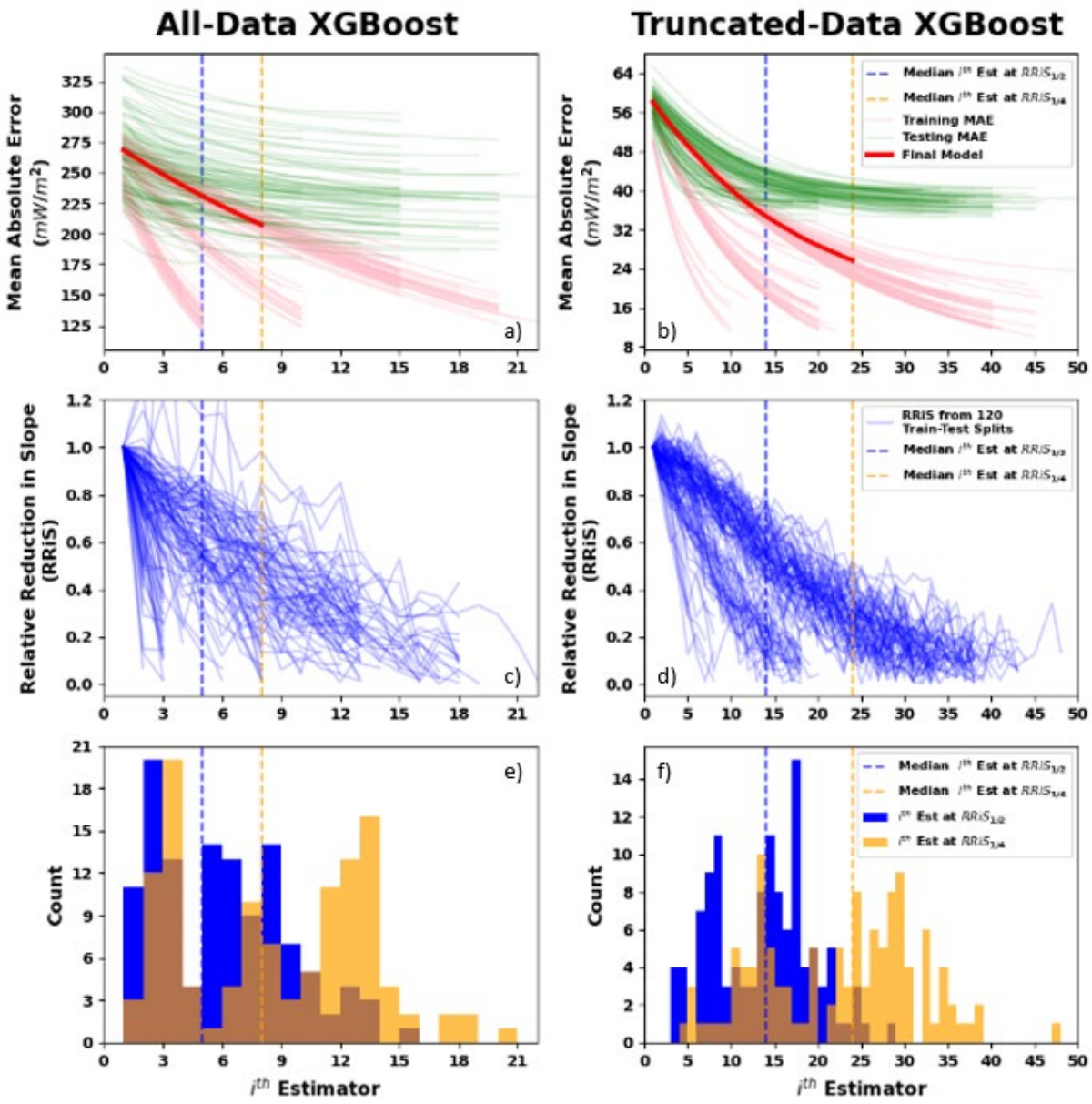

Figure 7. Loss-vs-Estimator relationships depicting workflow to identify when to engage early stopping. Training loss (pink) and testing loss (green) per estimator are provided in a) and b). The RRiS (Eq. 5) in c) and d) provides the relative changes in slope from a) and b), respectively, as estimators are added in each of the approaches. The i^{th}-estimator intersect at which the RRiS of a train-test split is at half of its initial RRiS (i.e., $RRiS_{1/2}$; blue bins) and a quarter of its initial RRiS (i.e., $RRiS_{1/4}$; orange bins) from the 120 train-test splits are provided in e) and f). Overlap of the blue bins for $RRiS_{1/2}$ and orange bins for $RRiS_{1/4}$ is depicted as brown. The median i^{th}-estimator intersect for these distributions (i.e., when early stopping is employed) for $RRiS_{1/2}$ (blue dashed line) and $RRiS_{1/4}$ (orange dashed line) overlay each subplot. Abbreviation: Est. – Estimator.

All the modeling approaches underpredict the highest reported convective signals. Linear Regression is the worst performing of the three approaches with its predictions having a roughly Gaussian distribution about a value of 260 mW/m^2 (Fig. 8), whereas the two XGBoost approaches

consistently predict high convective signals as high and low reported convective signals as low (Figs. 9, 10). Of the two XGBoost approaches, All-Data XGBoost predicts the highest convective signals (>2000 mW/m^2) and Truncated-Data XGBoost never predicts above 125 mW/m^2. The more-complex models (i.e., with *i* estimators corresponding to RRiS$_{1/4}$) predict higher than the less complex model variants (i.e., with *i* estimators corresponding to RRiS$_{1/2}$).

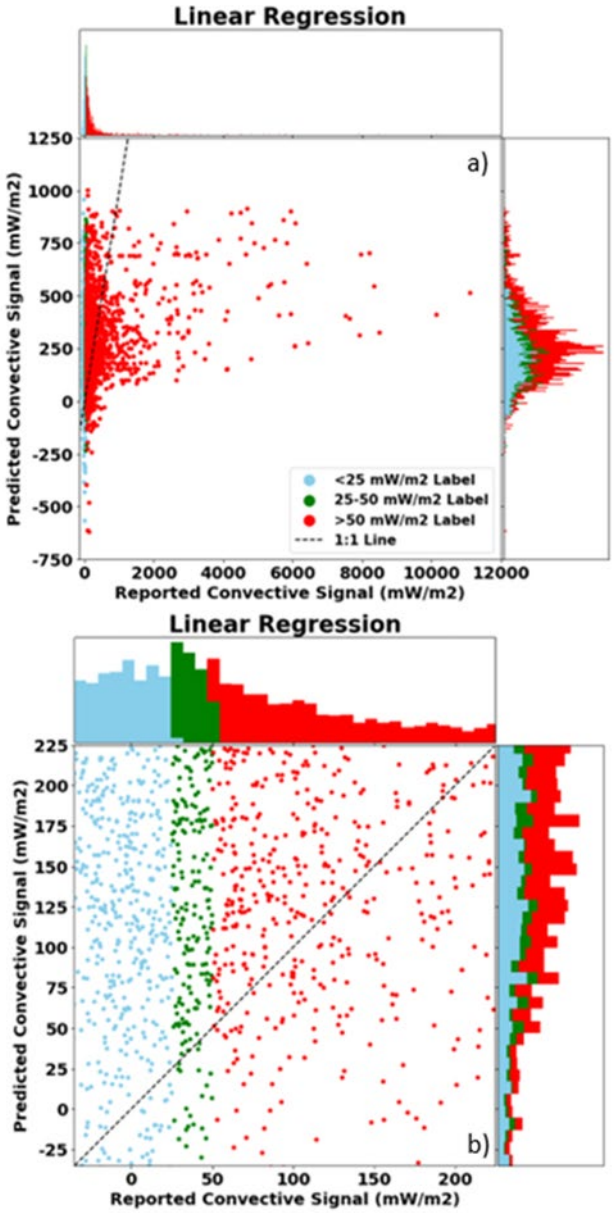

Figure 8. Cross-plots and stacked marginal histograms for reported convective signals and predicted convective signals from the approach for Linear Regression. Marginal histograms provide distribution to corresponding axis. Top plot (a) depicts the entire range of the reported convective signals. Bottom plot (b) is provided for comparison to predictions using the narrower range of reported convective signals with Truncated-Data XGBoost (Fig. 10).

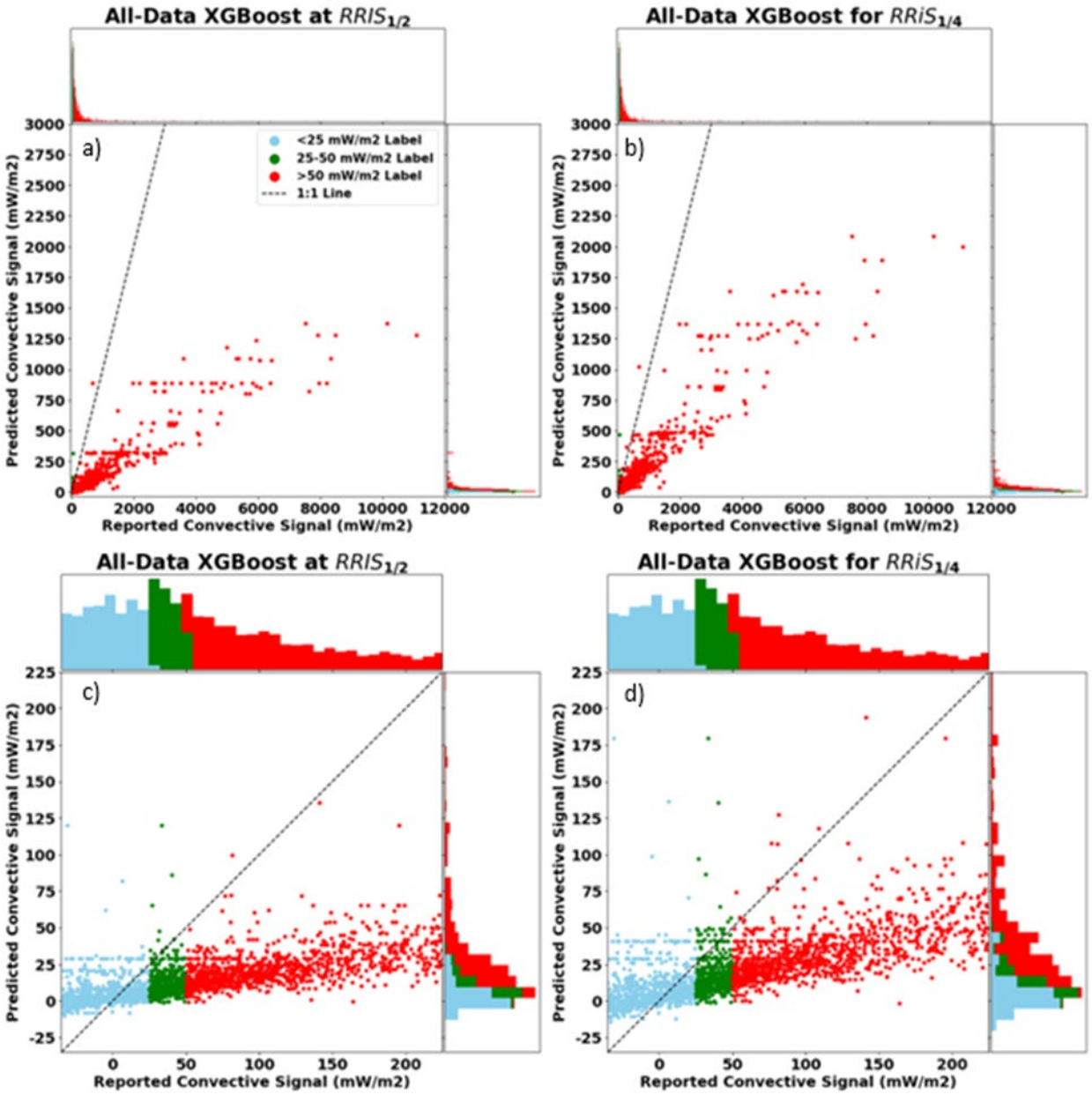

Figure 9. Cross-plots and stacked marginal histograms for reported convective signals and predicted convective signals from All-Data XGBoost for RRiS$_{1/2}$ (a,c) and for RRiS$_{1/4}$ (b,d). Top plots (a,b) depict the entire range of the reported convective signals. Bottom plots (c,d) are provided for comparison to predictions using the narrower range of reported convective signals with Truncated-Data XGBoost (Fig. 10).

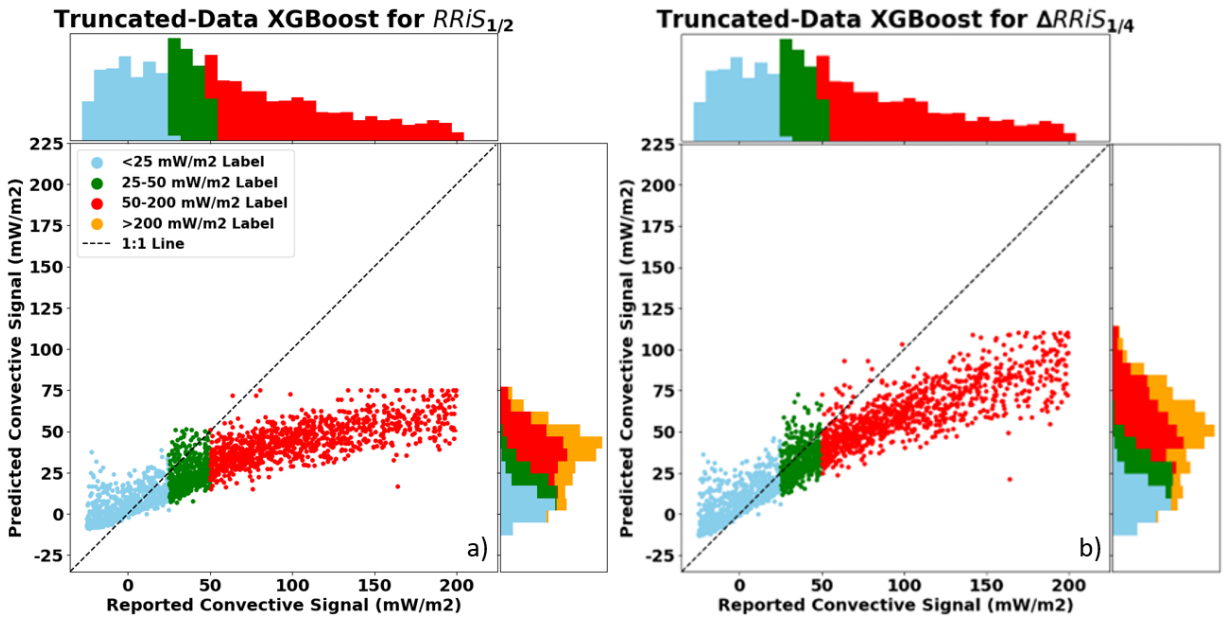

Figure 10. Cross-plots and stacked marginal histograms for reported convective signals and predicted convective signals from Truncated-Data XGBoost. Predictions for reported connective signals > 200 mW/m^2 are only depicted in the marginal histogram as orange bins to allow for a detailed perspective of predictions over the range of convective signals with which this model trained.

Because none of the approaches examined predict reasonably matching values for high convective signals (Figs. 8, 9, 10), the prediction maps are presented by categorizing the predicted convective signals according to the categorical distinctions from their corresponding reported convective signals (Fig. 11; Table 3). More specifically, we refer to predictions as having a low predicted convective signal when the predictions are less than the median prediction for examples with reported convective signals within 25 ± 5 mW/m^2, a high predicted convective signal when the predictions are greater than the median prediction for examples with reported convective signals within 50 ± 5 mW/m^2, and an intermediate predicted convective signal when the predictions are between the bounds for the high predicted convective signal and low predicted convective signal. In doing so, Linear Regression and All-Data XGBoost predict similar proportions (i.e., roughly 33%) of the study area as having a high predicted convective signal and Truncated-XGBoost predicts the smallest percentage (i.e., roughly 16%) of the study area as having a high predicted convective signal (Table 3). The All-Data and Truncated-Data XGBoost approaches have greater granularity than Linear Regression (Fig. 11).

Table 3: Prediction values defining low convective signal, intermediate convective signal, and high convective signal per approach and the corresponding predicted percent of study area.

Model	Bound between Low and Intermediate Predicted Convective Signal (mW/m^2)	Bound between Intermediate and High Predicted Convective Signal (mW/m^2)	Percent Area with Low Predicted Convective Signal	Percent Area with Intermediate Predicted Convective Signal	Percent Area with High Predicted Convective Signal
Linear Regression	232	244	65.1	2.7	32.3
All-Data $XGBoost_{1/2}$	6	13	48.9	18.5	32.6
All-Data $XGBoost_{1/4}$	9	18	46.8	18.7	34.5
Truncated-Data $XGBoost_{1/2}$	16	30	54.0	29.6	16.4
Truncated-Data $XGBoost_{1/4}$	22	40	51.8	32.7	15.6

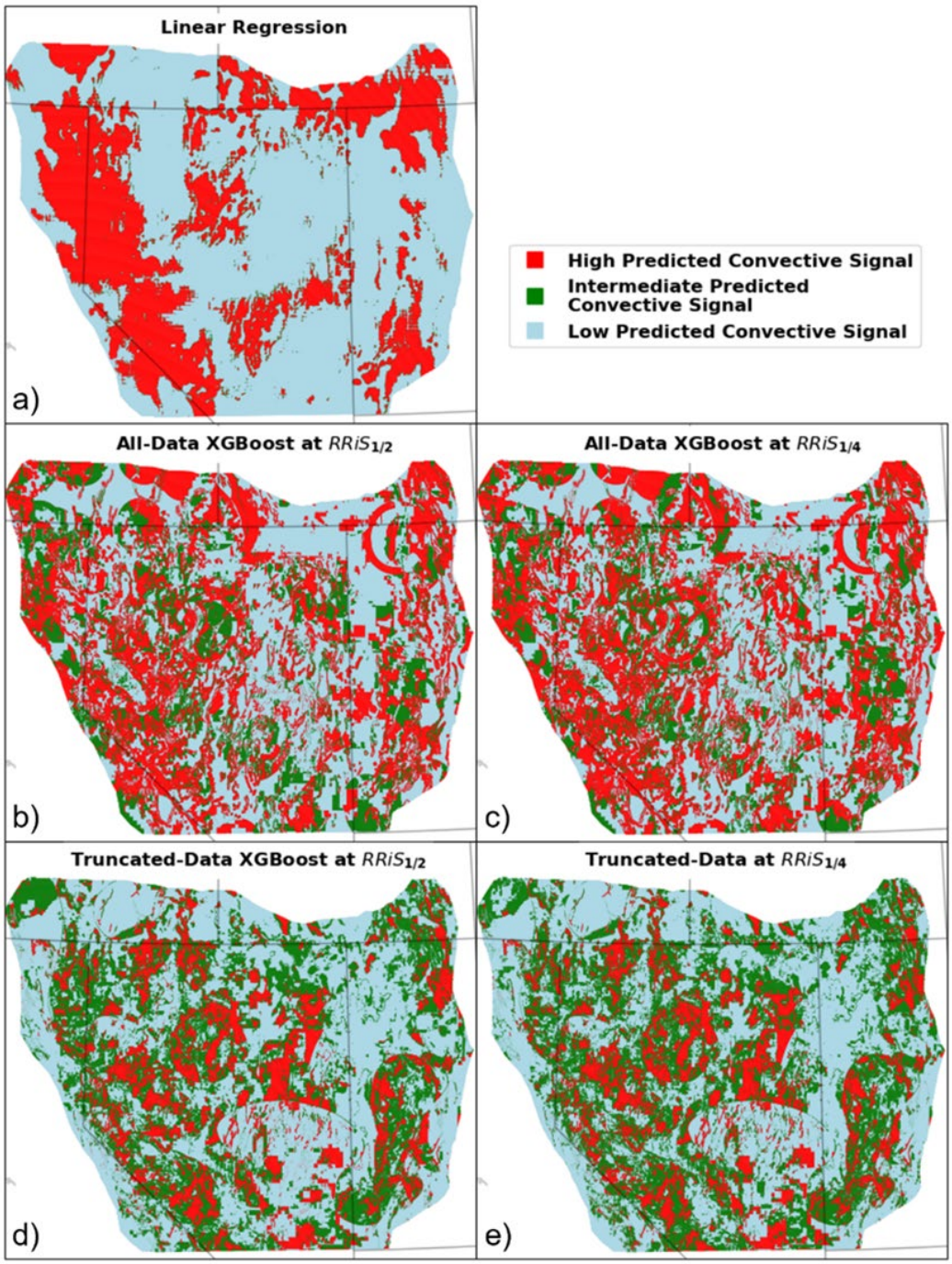

Figure 11. Prediction maps from using Linear Regression (a), All-Data XGBoost (b, c), and Truncated-Data XGBoost (d, e). Predictions are depicted as Low Predicted Convective Signal (blue), Intermediate Predicted Convective Signal (green), and High Predicted Convective Signal (red) because the absolute predictions suggest a strong bias in both approaches (i.e., predicted convective signals are consistently less than reported convective signals; see Table 3 for categorical boundary thresholds); although, high reported convective signals are still predicted as high and low reported convective signals are still predicted as low. The base map has been made using data from Natural Earth. Higher-resolution maps are available in Appendix A.

3.4 Feature Importance

The most important features predominantly vary by the selection of the algorithm (Fig. 12). The shear strain rate and second invariant to the strain rate are the most important features for linear regression. Conductive heat flow and distance to nearest Quaternary fault are the two most important features for All-Data XGBoost, with the second invariant to the strain rate and distance to nearest Quaternary magmatic activity being roughly equal as the third most important feature. Conductive heat flow, distance to nearest Quaternary fault, and distance to nearest Quaternary magmatic body are the most important features for Truncated-Data XGBoost.

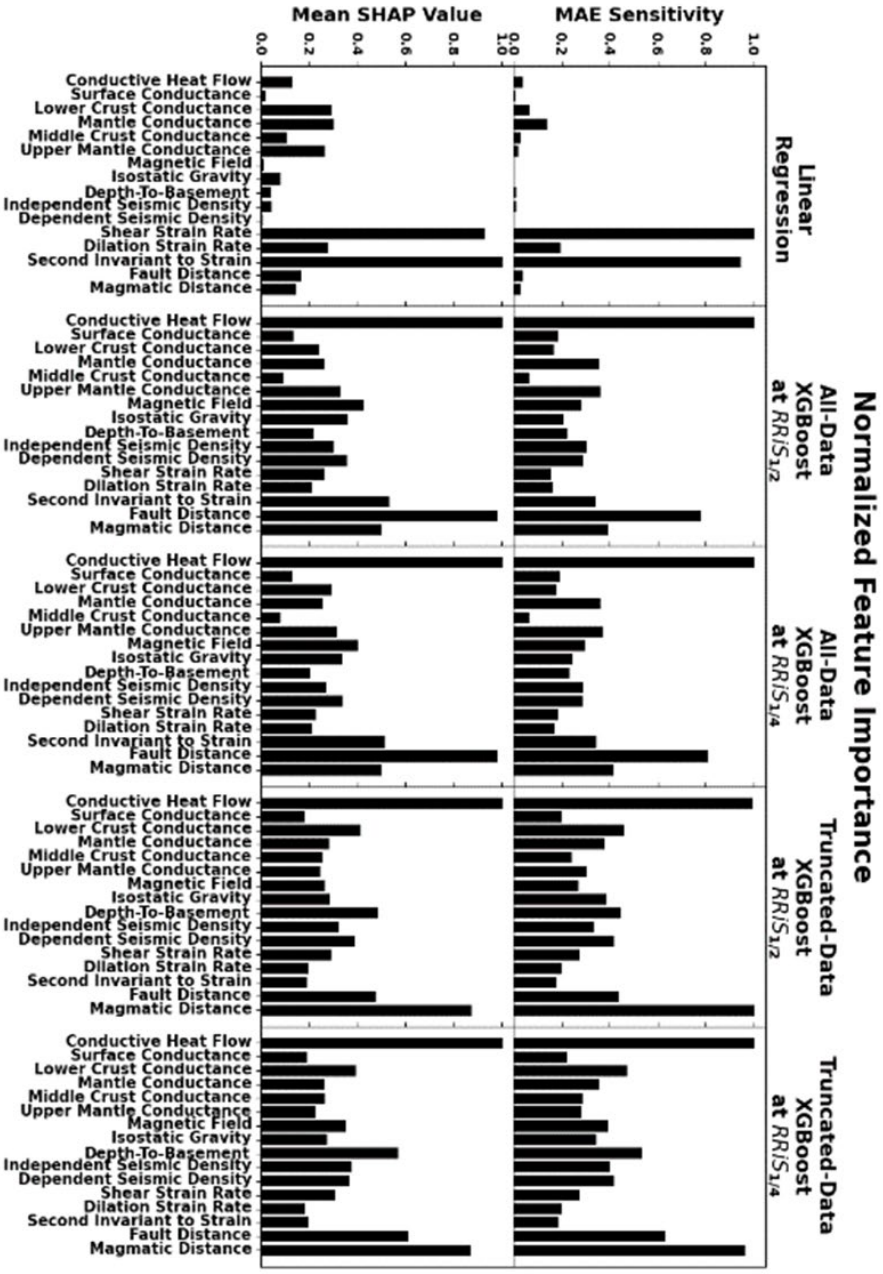

Figure 12. Min-max, 0-to-1 normalized feature importance for final models.

4. Discussion

All models generally express strong bias by underpredicting high reported convective signals, but the approaches using XGBoost perform better at predicting high convective signals as high and low convective signals as low (Figs. 9, 10) than linear regression (Fig. 8). This general observation of predictive behavior is consistent with the relative feature importance (Fig. 12), which suggests that the selection of the ML algorithm had a greater effect than the choice of which range of reported convective signals to use for fitting. However, identifying the best-performing XGBoost approach varies by which measures of performance are considered.

In terms of minimizing the area predicted as having a high convective signal, Truncated-Data XGBoost is the best-performing approach (Fig. 11; Table 3). Truncated-Data XGBoost also has greater separation between the distributions of predictions for high and low reported convective signals (Fig. 13). More specifically, when predicting for the entire study area, Truncated-Data XGBoost predicts half of the examples with high reported convective signals in the top 3 % of predictions, whereas All-Data XGBoost requires the top 13 % of predictions to include half of the examples with high reported convective signals (Fig. 13). Yet, when evaluating model performance by an ability to predict convective hydrothermal upflow at operating geothermal power plants, the All-Data XGBoost approach outperforms the Truncated-Data XGBoost approach (Fig. 14), suggesting that valuable information was lost by removing labels with values > 200 mW/m^2 from Truncated-Data XGBoost.

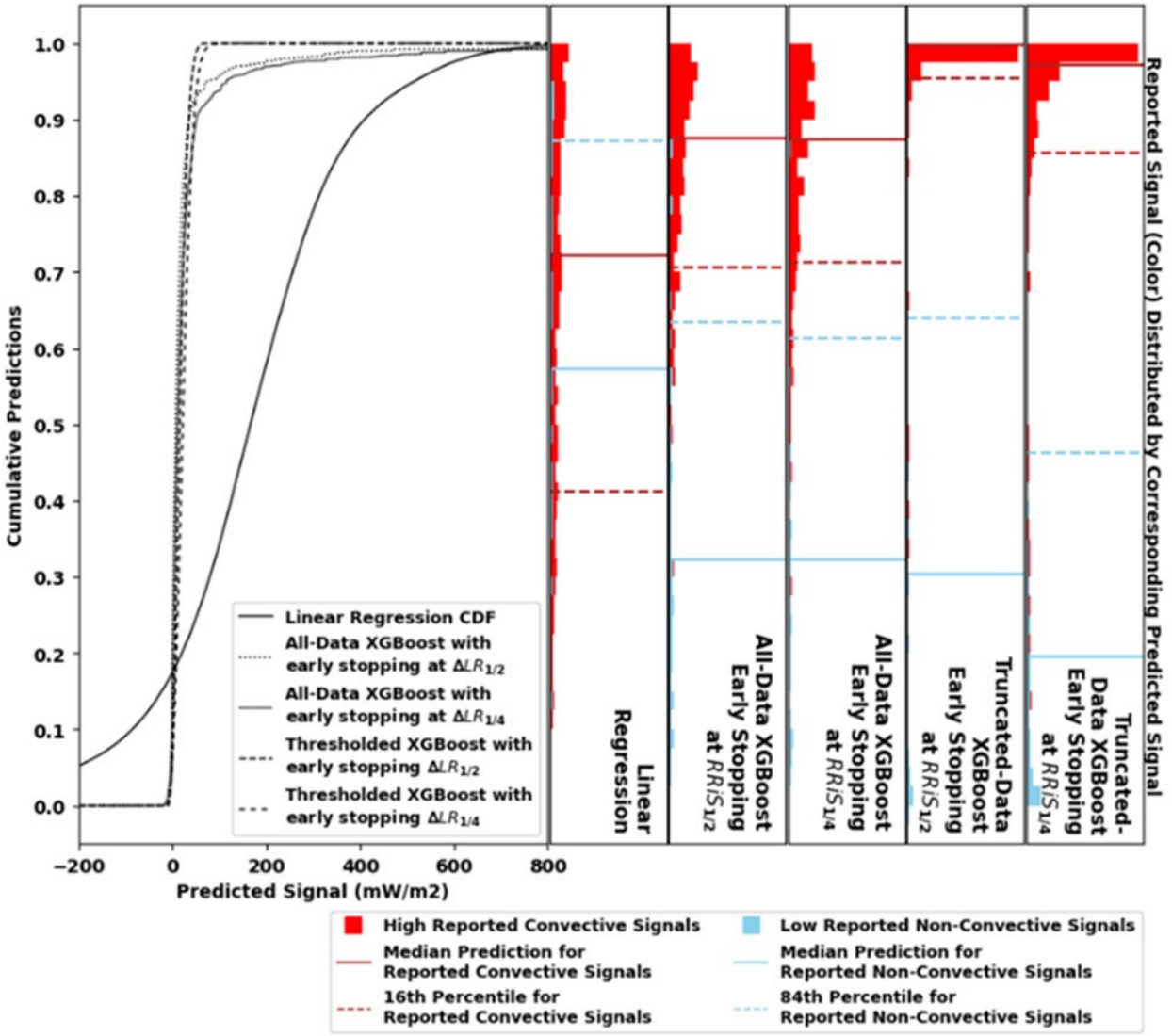

Figure 13. Cumulative distribution functions of predicted convective signals for all examples (labeled and unlabeled) with distributions of corresponding reported convective signals provided in marginal histograms (i.e., red: high reported convective signal, blue: low reported convective signal). Solid red lines depict the median high predicted convective signals. Dashed red lines depict the 16th percentile (i.e., one standard deviation or one sigma below the median value in a normal distribution) prediction for high reported convective signals. Solid blue lines depict the median prediction for low reported convective signals. Dashed blue lines depict the 84th percentile (i.e., one standard deviation or one sigma above the median value in a normal distribution) prediction for low reported convective signals. Bin size is 0.025 on a unitless 0-to-1 scale.

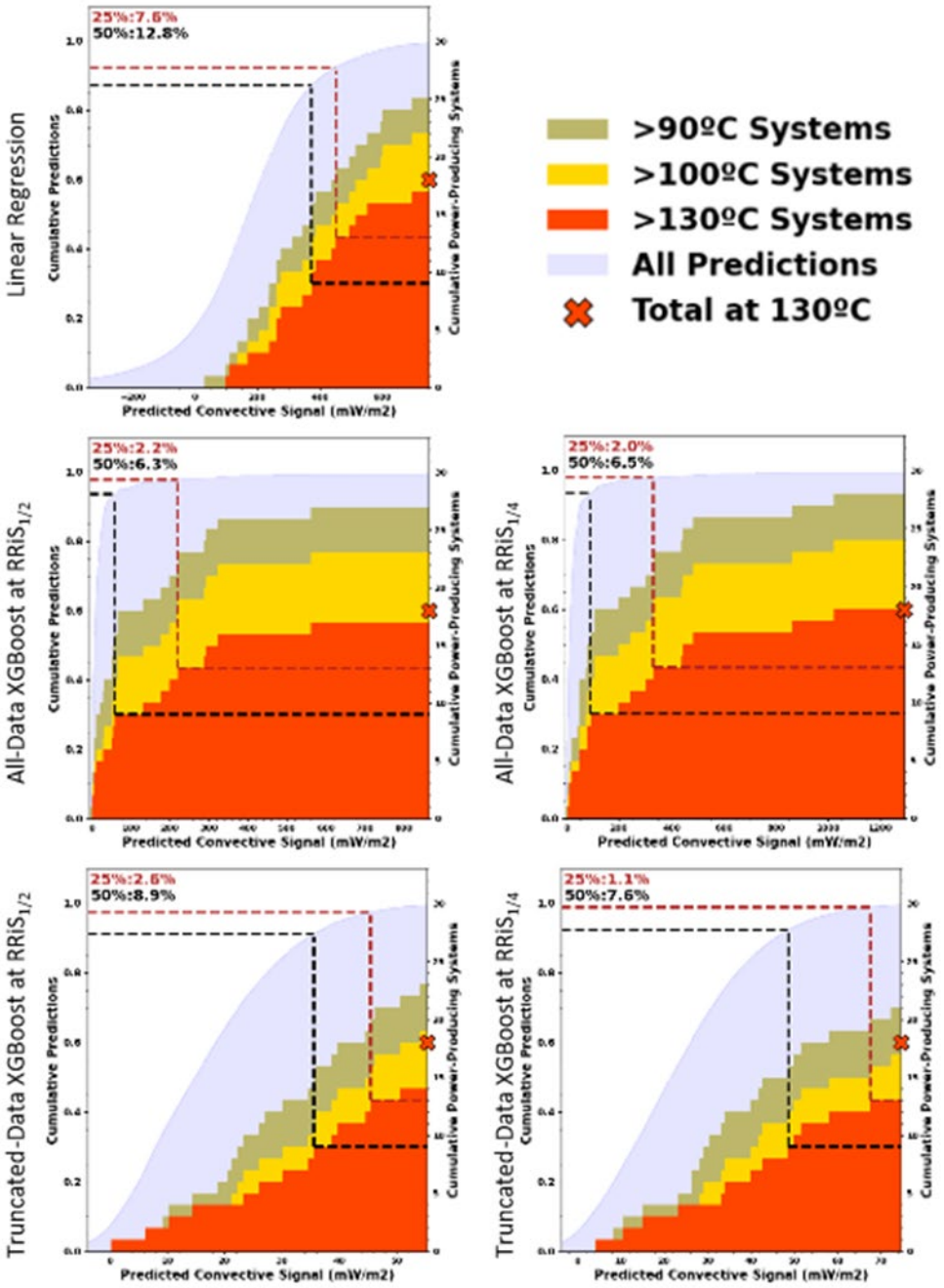

Figure 14. Cumulative distribution functions of power-producing hydrothermal systems respective to model predictions. Black and red dashed lines depict the percentile of highest predictions (left y-axis; analogous to the percent of the study area) required to capture 50 % (9) and 25 % (13), respectively, of the 18 power plants with reported > 130˚C temperature (right y-axis). Similarly, the black and red text in upper left of each subplot corresponds to the dashed lines by color and reports the percentile of highest predictions required to capture 50 % (9) and 25 % (13), respectively, of the 18 power plants with reported > 130˚C temperature. The range of the x-axes are defined by bounding 2.5% and 99% of the total predictions specific to the approach depicted. The power production data are from Faulds et al. (2021) and available through Mlawsky and Ayling (2021). Red X on right y-axis marks the total number (18) of powerplants operating at > 130˚C. MW – Megawatts power capacity.

Like with performance, neither XGBoost approach has consistent behavior to suggest one approach is less susceptible to overfitting as new estimators are added between $RRiS_{1/2}$ and $RRiS_{1/4}$. In terms of minimizing the percent of the study area predicted as having a high convective signal, adding estimators to Truncated-Data XGBoost spreads the distribution of predictions for reported high convective signals to relatively lower predictions (i.e., from a median high reported convective signal at the 99.77^{th} percentile prediction to the 97.11^{th} percentile prediction; Fig. 13) more than the addition of new estimators to All-Data XGBoost (i.e., from a median high reported convective signal at the 87.61^{th} percentile prediction to the 87.39^{th} percentile prediction), suggesting that All-Data XGBoost is less sensitive to the addition of new estimators. Yet, All-Data XGBoost appears more sensitive to the addition of new estimators than Truncated-Data XGBoost in terms of percent change of the study area that was predicted as low for $RRiS_{1/2}$ and high for $RRiS_{1/4}$ (Fig. 15). As additional estimators are added between $RRiS_{1/2}$ and $RRiS_{1/4}$, the absolute predictions in both XGBoost approaches generally increased (Figs. 9, 10, 13), but the majority (i.e., > 99 %) of the resulting changes to the categorical predictions, like those depicted in Fig. 11 and detailed in Table 3, were only from one category of predicted convective signal to the next category of predicted convective signal (i.e., from a low predicted convective signal to an intermediate predicted convective signal or from an intermediate predicted convective signal to a high predicted convective signal). Although < 1 % of the study area changes from a predicted low convective signal to a predicted high convective signal with either XGBoost approach, Truncated XGBoost expresses greater stability in generalization with only 0.03 % of the study area changing from a predicted low to a predicted high, whereas 0.62 % of the study area changes from a predicted low to a predicted high with All-Data XGBoost (Fig. 15). Hence, the high outlying reported convective signals in All-Data XGBoost are likely biasing predictions as the model complexity of that approach increases.

The hypothesis that the outliers are impacting bias in more complex All-Data XGBoost models is supported by optimizing early stopping using validation data. Although we use testing data to optimize early stopping in this study, we note that optimizing early stopping with testing data is not considered a best practice. To address this concern, we could optimize early stopping using validation data. Preliminary results find that optimizing early stopping with validation data instead of testing data reduces model complexity more in the All-Data XGBoost approach than in the Truncated-Data XGBoost approach. The greater simplification of model complexity in the All-Data XGBoost approach may be related to the high outlying reported convective signals used in that approach. Because the subset of examples used for optimization decreases from 20% to 16% of the labeled data (i.e., from the testing data to the validation data), the outlying label values may be forcing fewer estimators in the All-Data XGBoost approach to prevent overfitting.

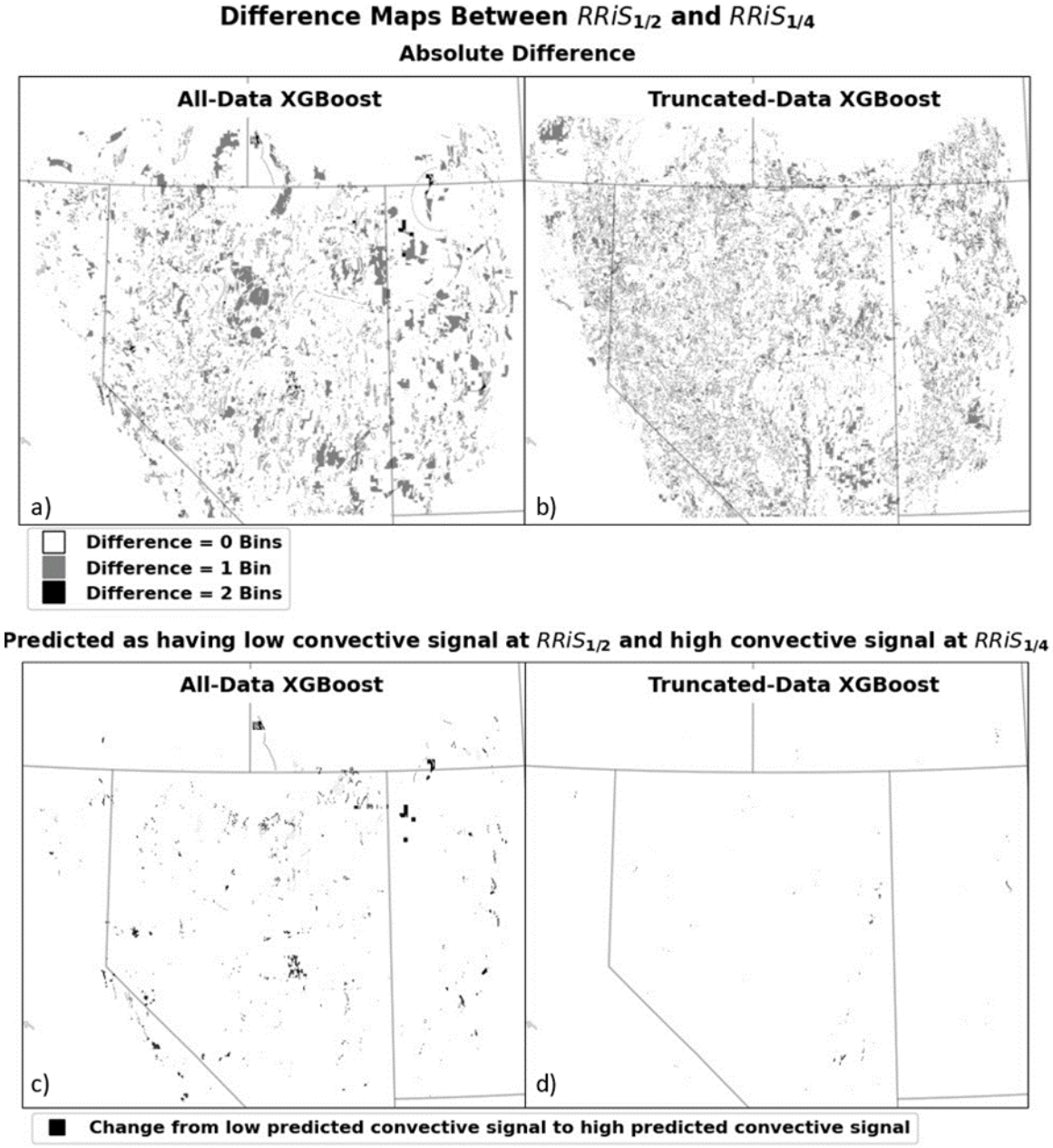

Figure 15. Maps depicting differences in categorical predictions of convective signal between models trained with estimators corresponding to $RRiS_{1/2}$ and $RRiS_{1/4}$. The absolute ordered differences are provided in the top plots (a,b). The bottom plots provide locations predicted as low convective signal at $RRiS_{1/2}$ but as high convective signal at $RRiS_{1/4}$. The base map has been made using data from Natural Earth.

The comparison of the different supervised ML algorithms strongly demonstrates that non-linear approaches (e.g., XGBoost) hold some skill for predicting convective signal, but the performance tradeoff resulting from the removal of examples with high outlying convective signals likely impairs model performance (Figs. 13, 14). Hence, regression may not be the appropriate class of algorithm to predict strong convective signals. Simultaneously, supervised ML classification

implicitly does not distinguish between high and very high convective signals (e.g., Mordensky et al., 2023b). Therefore, we propose using ordinal regression as a compromise between regression and classification. By binning the convective signals, we can isolate high convective signals from very high convective signals; thereby allowing models that convey information from all the convective signals but without the bias of outlying label values.

A more discerning approach during feature selection would also likely improve model performance. Hitherto, the models we have presented and discussed used feature data without regard for their correlative relationships. However, Mordensky et al. (2023a) showed that having few labeled examples, like when working with geoscience data, emphasizes the importance of using as few features as possible in order to maximize model performance. The high correlation between the features in this study (Fig. 6) means that every feature may not be benefitting model performance, and there likely is opportunity to omit the less informative features in favor of other new, more informative features (e.g., the first derivative of isostatic gravity or the magnetic field).

5. Conclusion

In this study, we fit three models to predict the magnitude of hydrothermal upflow across the Great Basin using reported convective signals from DeAngelo et al. (2022) as labels, datasets supporting INGENIOUS as features, and three machine learning approaches (i.e., linear regression using the entire range of reported convective signals, XGBoost using the entire range of reported convective signals, and XGBoost using reported convective signals truncated to a specified range [-25 to 200 mW/m^2] so that large outliers are excluded). Linear regression offers only limited meaningful predictive skill; however, the XGBoost approach using the truncated range of reported convective signals performs the best at predicting high known convective signals as high whereas the XGBoost approach fit using all the data performs the best at predicting hydrothermal systems with power production. The duality of these two XGBoost approaches and their performance measures suggests that very high convective signals have valuable information, but the outlying nature of these very high convective signals imparts bias. Therefore, our results suggest using a supervised machine learning algorithm that allows for the use of very high convective signals but reduces their potential bias (e.g., ranked ordinal regression) to predict the magnitude of convective hydrothermal upflow.

Acknowledgements

This work was supported by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy (EERE), Geothermal Technologies Office (GTO) under Contract No. DEAC02-05CH11231 with Lawrence Berkeley National Laboratory, Conformed Federal Order No. 7520443 between Lawrence Berkeley National Laboratory and the U.S. Geological Survey (Award Number DE-EE0008105), and Standard Research Subcontract No. 7572843 between Lawrence Berkeley National Laboratory and Portland State University. Additional support for John Lipor was provided by the National Science Foundation awards NSF CRII CIF-1850404 and NSF CAREER CIF-2046175. Support for Jake DeAngelo and Erick Burns was provided by the U.S. Geological Survey Energy Resources Program. INGENIOUS is supported by the U.S. Department of Energy - Geothermal Technologies Office under award DE-EE0009254 to the University of Nevada, Reno. Any use of trade, firm, or product names is for descriptive purposes

only and does not imply endorsement by the U.S. Government. We thank the U.S. Geological Survey Advanced Research Computing (ARC) group for their assistance in using DENALI. We thank Jared Peacock and Joshua Rosera for their review of this manuscript.

Appendix A – Higher-Resolution Prediction Maps

Appendix A provides higher-resolution prediction maps for all the models produced in this study.

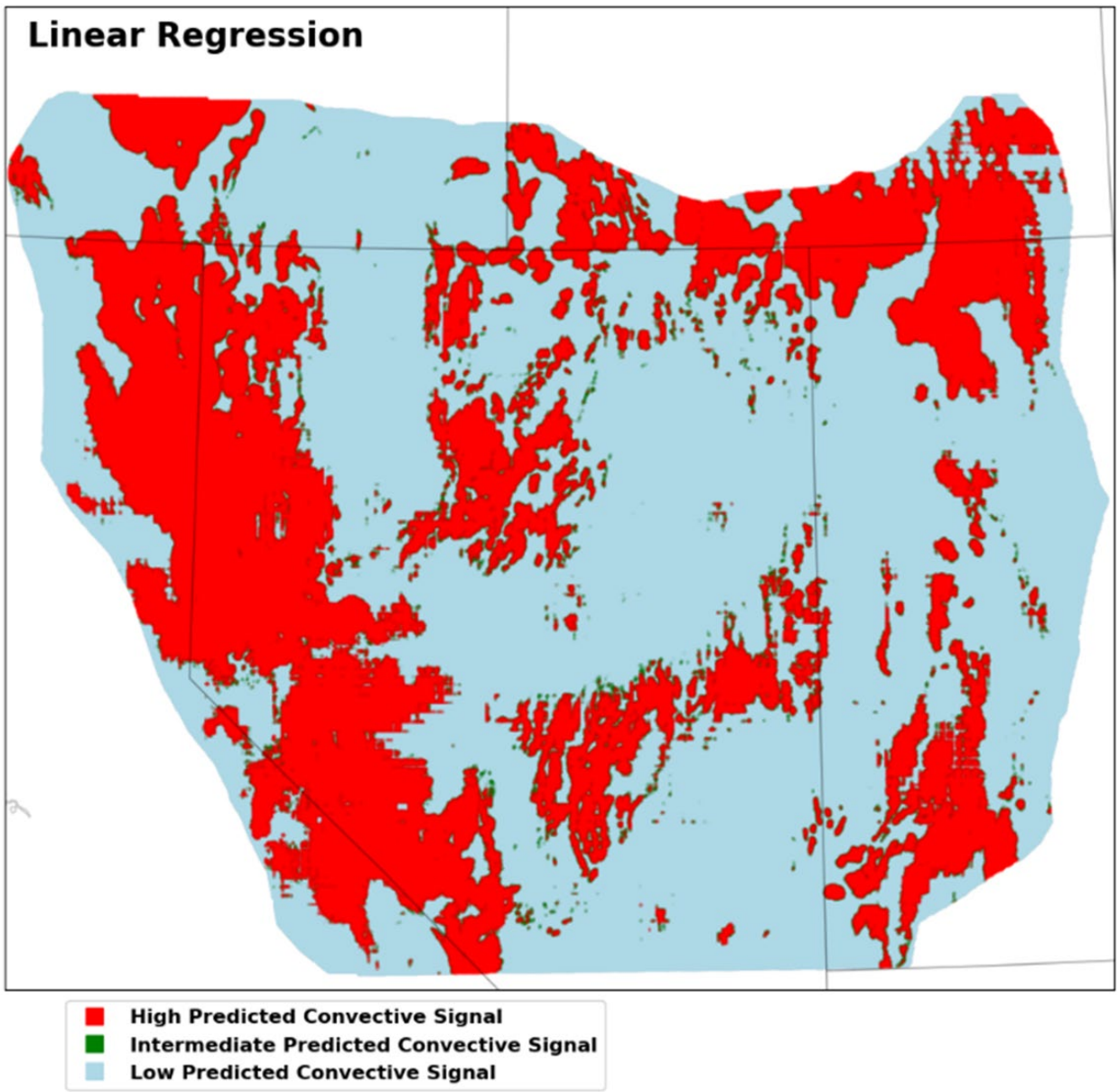

Figure A1: Predicted signals from conductive heat flow using linear regression and the entire range of convective signals. The base map has been made using data from Natural Earth.

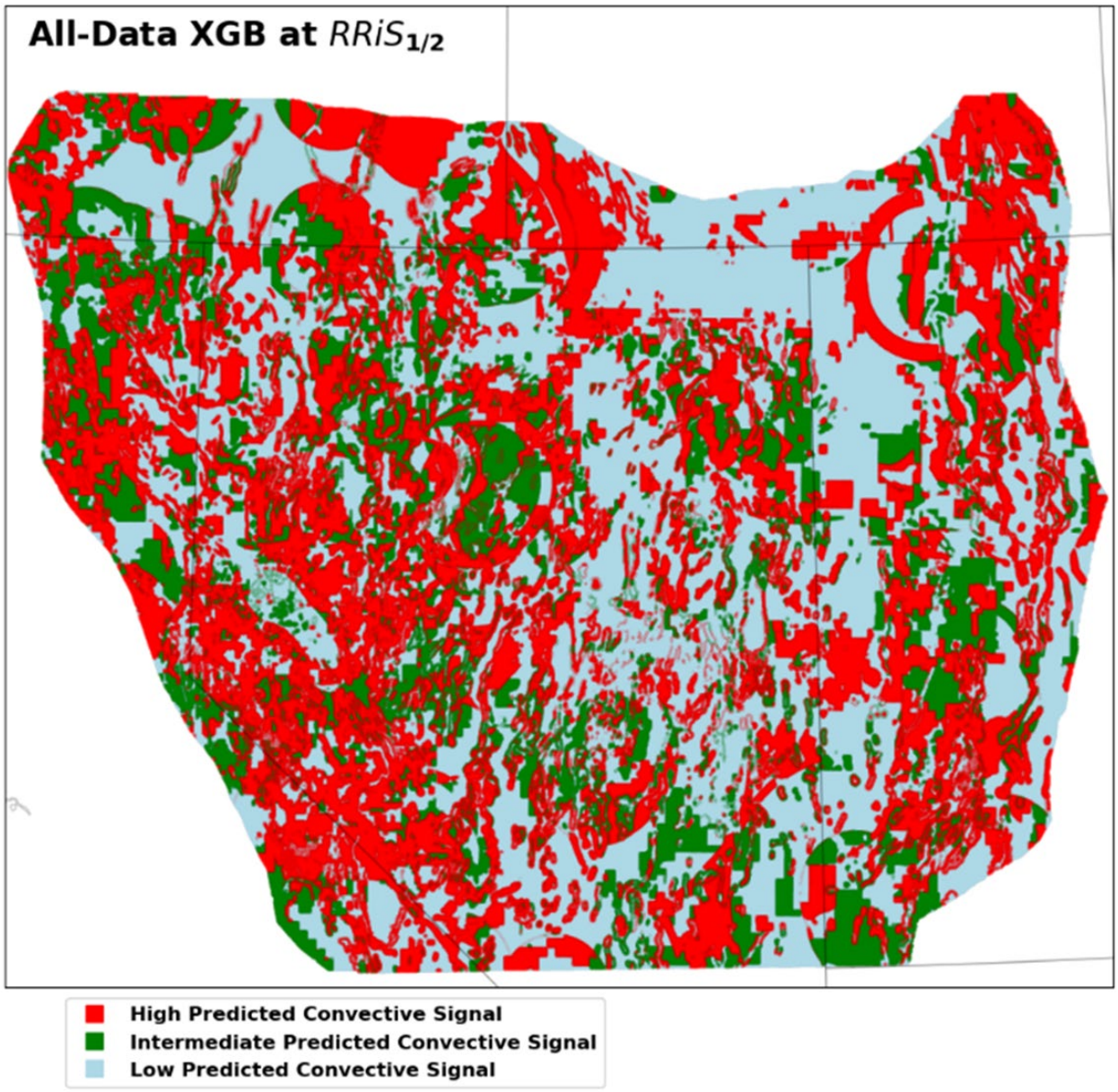

Figure A2: Predicted signals from conductive heat flow using XGBoost and the entire range of convective signals and $RRiS_{1/2}$ early stopping. The base map has been made using data from Natural Earth.

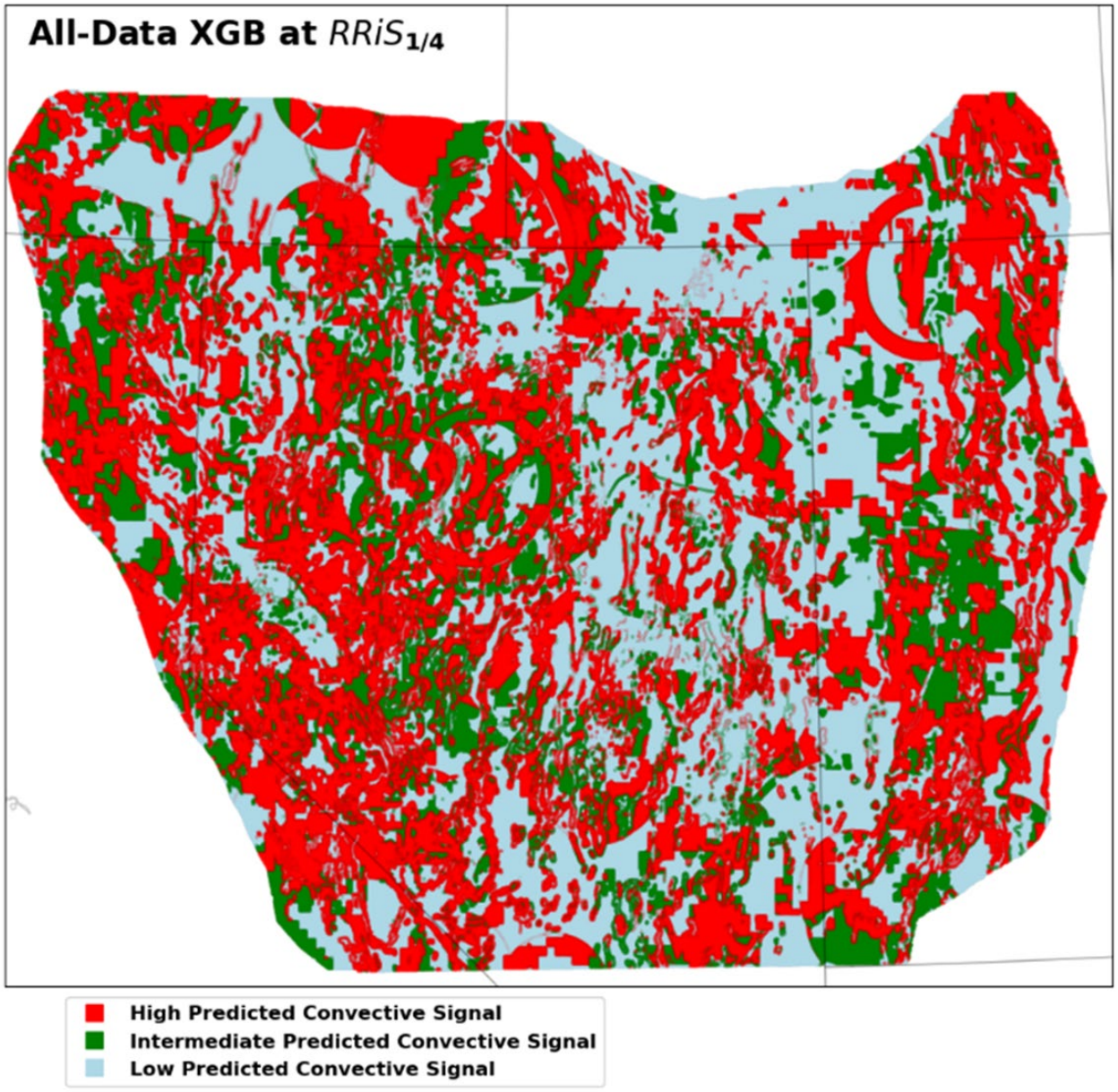

Figure A3: Predicted signals from conductive heat flow using XGBoost and the entire range of convective signals and $RRiS_{1/4}$ early stopping. The base map has been made using data from Natural Earth.

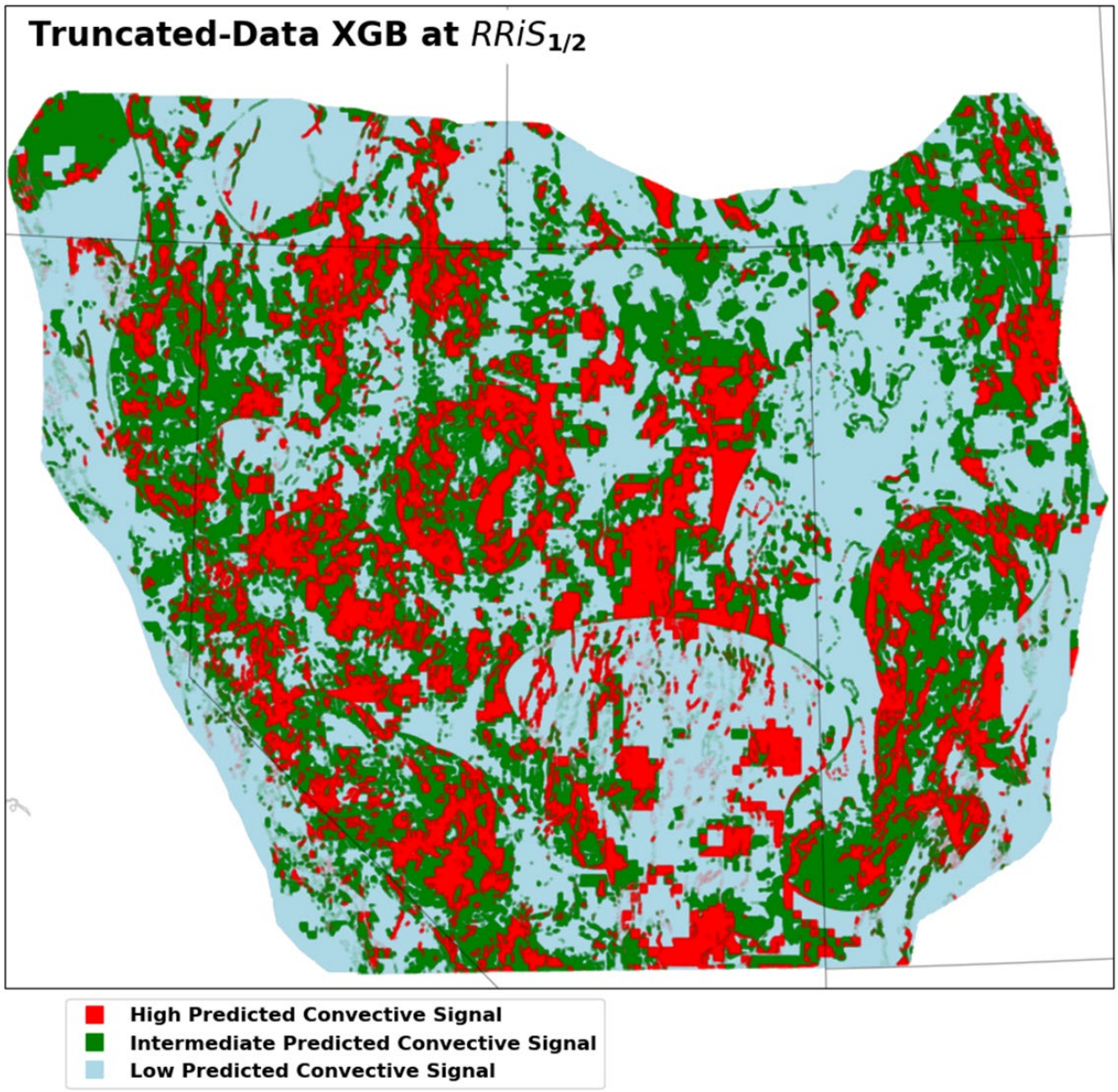

Figure A4: Predicted signals from conductive heat flow using XGBoost and convective signals with label values ranging from -25 to 200 mW/m^2 and RRiS$_{1/2}$ early stopping. The base map has been made using data from Natural Earth.

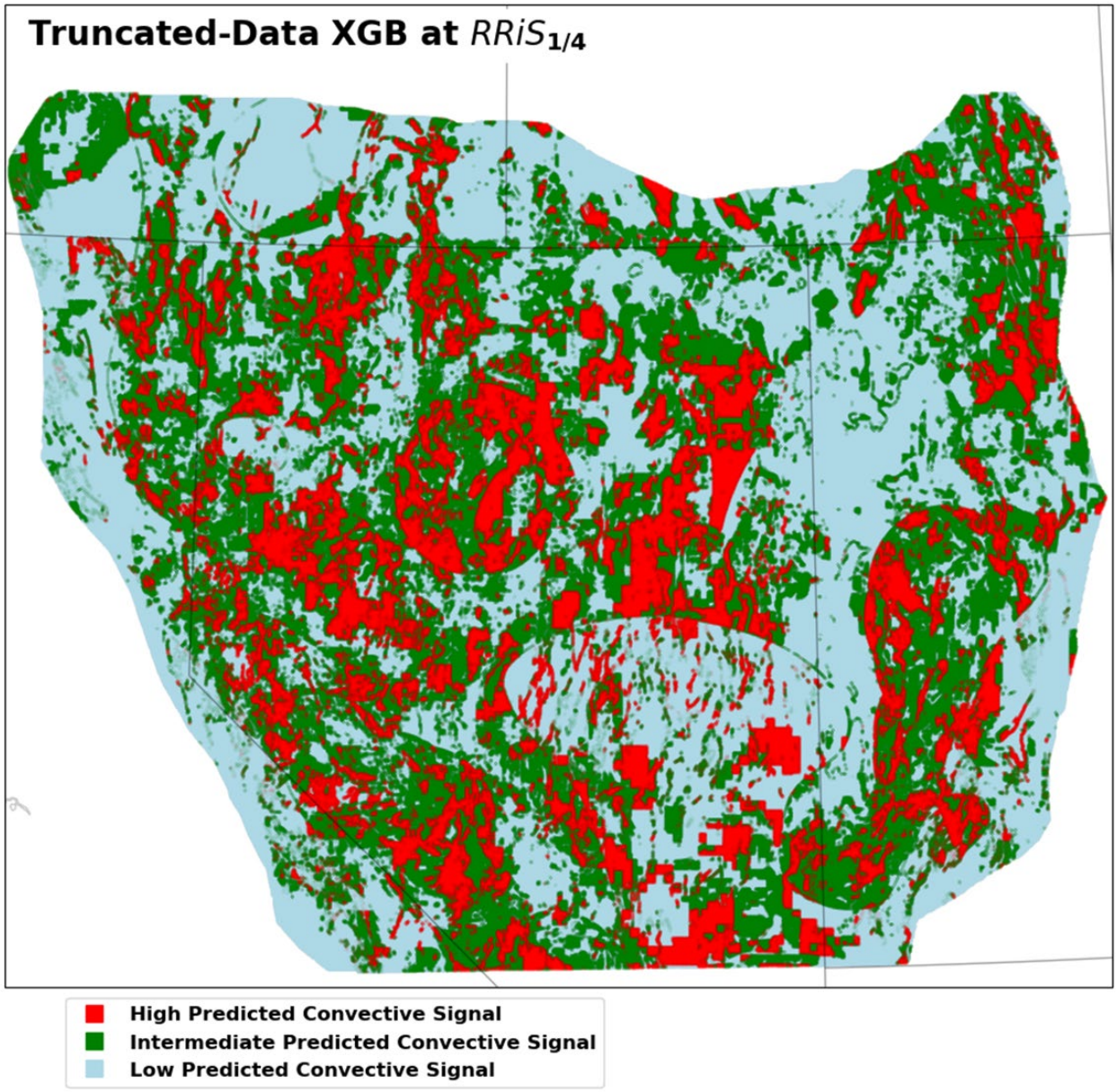

Figure A5: Predicted signals from conductive heat flow using XGBoost and convective signals with label values ranging from -25 to 200 mW/m^2 and RRiS$_{1/4}$ early stopping. The base map has been made using data from Natural Earth.

REFERENCES

Ayling, B., Faulds, J. E., Rivera, A., Koehler, R., Kreemer, C., Mlawsky, E., . . . Kleber, E. (2022a). Quaternary Faults (Publication no. 10.15121/1881483). Retrieved 29 March 2023, from Geothermal Data Repository https://gdr.openei.org/submissions/1391

Ayling, B., Faulds, J. E., Rivera, A., Koehler, R., Kreemer, C., Mlawsky, E., . . . Kleber, E. (2022b). Quaternary Volcanics (Publication no. 10.15121/1881483). Retrieved 29 March 2023, from Geothermal Data Repository https://gdr.openei.org/submissions/1391

Ayling, B., Faulds, J. E., Rivera, A., Koehler, R., Kreemer, C., Mlawsky, E., . . . Kleber, E. (2022c). Study Area Boundary (Publication no. 10.15121/1881483). Retrieved 23 August 2023, from Geothermal Data Repository https://gdr.openei.org/submissions/1391

Barbour, A. J. (2015). Pore pressure sensitivities to dynamic strains: Observations in active tectonic regions. *Journal of Geophysical Research: Solid Earth, 120*, 5863–5883. doi:10.1002/2015JB012201

Burns, E. R., Williams, C. F., Ingebritsen, S. E., Voss, C. I., Spane, F. A., & DeAngelo, J. (2015). Understanding heat and groundwater flow through continental flood basalt provinces: Insights gained from alternative models of permeability/depth relationships for the Columbia Plateau, USA. *Geofluids, 15*, 120-138. doi:10.1111/gfl.12095

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA.

DeAngelo, J., Burns, E. R., Gentry, E., Batir, J. F., Lindsey, C. R., & Mordensky, S. P. (2022). Heat flow maps and supporting data for the Great Basin, USA (Publication no.). U.S. Geological Survey data release, https://doi.org/10.5066/P9BZPVUChttps://www.sciencebase.gov/catalog/item/6297d2fad34ec53d276c5b28

DeAngelo, J., Burns, E. R., Gentry, E., Batir, J. F., Lindsey, C. R., & Mordensky, S. P. (2023). *New Maps of Conductive Heat Flow in the Great Basin, USA: Separating Conductive and Convective Influences.* Paper presented at the 48th Workshop on Geothermal Reservoir Engineering, Stanford, CA, USA.

Falgout, J. T., Gordon, J., Williams, B., & Davis, M. J. (2021). *SGS Advanced Research Computing, USGS Denali Supercomputer*. U.S. Geological Survey.

Faulds, J. E., Coolbaugh, M. F., & Hinz, N. (2021). *Inventory of structural settings for active geothermal systems and late Miocene (~8 Ma) to Quaternary epithermal mineral deposits in the Basin and Range province of Nevada*. Retrieved from Reno, NV:

Glen, J. M., Earney, T. E., Zielinski, L. A., Schermerhorn, W. D., Dean, B. J., & Hardwick, C. (2022). Regional geophysical maps of the Great Basin, USA. U.S. Geological Survey data release, http://doi.org/10.5066/P9Z6SA1Z

Huber, P. J. (1964). Robust Estimation of a Location Parameter. *Annals of Statistics, 53*(1), 73-101. doi:10.1214/aoms/1177703732

Kreemer, C., & Young, Z. M. (2023). Crustal Strain Rates in the Western United States and Their Relationship with Earthquake Rates. *Seismological Research Letters, 93*(6), 2990-3008. doi:10.1785/0220220153

Lee Rodgers, J., & Nicewander, W. A. (1988). Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician, 42*(1), 59-66. doi:10.1080/00031305.1988.10475524

Lundberg, S. M., & Lee, S.-I. (2017). *A Unified Approach to Interpreting Model Predictions*. Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

Mlawsky, E., & Ayling, B. F. (2021). The GBCGE Subsurface Database Explorer. from Nevada Bureau of Mines and Geology https://gdr.openei.org/submissions/1486

Mordensky, S. P., Burns, E. R., Lipor, J. J., & DeAngelo, J. (2023a). *Cursed? Why one does not simply add new data sets to supervised geothermal machine learning models.* Paper presented at the Geothermal Rising Conference, Reno, NV.

Mordensky, S. P., Lipor, J. J., DeAngelo, J., Burns, E. R., & Lindsey, C. R. (2023b). When Less Is More: How Increasing the Complexity of Machine Learning Strategies for Geothermal Energy Assessments May Not Lead toward Better Estimates. *Geothermics, 110*, 102662. doi:10.1016/j.geothermics.2023.102662

Natural Earth. (2023). August, 24, 2023. www.naturalearthdata.com

Peacock, J. R., & Bedrosian, P. A. (2022). Electrical Conductance Maps of the Great Basin, USA). U.S. Geological Survey data release, http://doi.org/10.5066/P9TWT2LU.

Tian, Y., & Zhang, Y. (2022). A comprehensive survey on regularization strategies in machine learning. *Information Fusion, 80*, 146-166. doi:10.1016/j.inffus.2021.11.005

Williams, C. F., & DeAngelo, J. (2008). Mapping Geothermal Potential in the Western United States. *GRC Transactions, 32*, 181-188.

Williams, C. F., Reed, M. J., Mariner, R. H., DeAngelo, J., & Galanis, S. P. (2008). Assessment of Moderate-and High-Temperature Geothermal Resources of the United States. *U.S. Geological Survey Fact Sheet 2008-3082*, 1-4.

Xue, L., Brodsky, E. E., Erskine, J., Fulton, P. M., & Carter, R. (2016). A permeability and compliance contrast measured hydrogeologically on the San Andreas Fault. *Geochemistry, Geophysics, Geosystems, 47*, 858-871. doi:10.1002/2015GC006167