PROCEEDINGS, 47th Workshop on Geothermal Reservoir Engineering
Stanford University, Stanford, California, February 7-9, 2022
SGP-TR-223

Multimodal Machine Learning for 3D Characterization of Hidden Groundwater and Geothermal Resources: Case Study, Lāna‘i Hawaii

Michael J. Friedel[1], Nicole Lautze[1], Erin Wallin[1], Aaron Rothfolk[1]

[1] Hawaii Groundwater & Geothermal Resource Center, University of Hawaii-Manoa

mfriedel@hawaii.edu

Keywords: Multimodal machine learning, multiphysics-Informed machine learning, 3D geothermal stratigraphic units, 3D hidden groundwater resources, 3D hidden geothermal resources 3D temperature prediction, 3D chloride prediction, 3D specific capacity prediction, 3D geology predictions, dike swarms, pluton, sill batholith, Moho, Lāna‘i Hawaii, transdisciplinary data product

ABSTRACT

The availability of freshwater and low-cost electricity are limiting factors for sustainable living in Hawaii. This raises the question: Can technology be developed to locate and characterize freshwater and geothermal resources simultaneously? We present a multimodal machine learning (MML) workflow to characterize the 3D distribution of features (physical and geochemical properties and state variables) along a groundwater-geothermal continuum. The success of this MML workflow is based on the availability and assimilation of mutual information to: (1) identify latent features using neural network proposals in a genetic algorithm with feature constraints, (2) organize related information across a hypersurface by competitive learning, (3) predict regionally continuous features by minimizing the quantization and topological errors with competitive learning, and (4) group statistically meaningful features based on the mode of stochastic k-means clusters. The proposed MML workflow is applied to a subset of the Hawaii Play Fairway data for Lāna‘i, Hawaii. These data include direct point measurements, such as borehole water level, geology, temperature, chloride concentration, isotopes, and specific conductivity; and derived 3D volume properties (numerically inverted), such as specific capacity, density, and electrical resistivity. Despite field data characteristics (disparate, scale dependent, spatially limited, sparse, and uncertain), the MML workflow yields a single 3D transdisciplinary data product whose voxels each contain statistical summaries of model features. Five-fold cross-validation (e.g., five randomly shuffled stratified split sets, each split with 80% training and 20% testing) reveals moderate generalizability of the MML model to independent data. Preliminary interpretation of continuous features reveals hidden freshwater (low temperature and low chloride concentration water) and geothermal (high temperature and brackish water) resources. In addition to integrating machine learning expertise and practices into the Play Fairway project outcomes, this study provides new capabilities for characterizing continuous subsurface hydrogeologic and geothermal features in the Hawaiian Islands for sustainable living (including Hawai‘i Island as part of the new Island Heat project sponsored by the Department of Energy) and at other geothermal sites worldwide.

1. INTRODUCTION

The availability of freshwater and low-cost electricity are limiting factors for sustainable living in Hawaii. This raises the question: *Can technology be developed to locate and characterize freshwater and geothermal resources simultaneously*? The Hawaii Play Fairway project, funded by the Geothermal Technologies Office (GTO), U.S. Department of Energy (DOE), produced a statistical methodology to integrate existing geologic, groundwater, and geophysical datasets relevant to subsurface heat, fluid, and permeability into a statewide resource probability map (Lautze et al., 2020, Lautze et al., 2022). The combined probabilities for heat, fluid, and permeability indicated the likelihood of a geothermal resource in the caldera region of Lāna‘i Island. This area, known as the Pālāwai Basin, exhibits the co-location of a high gravity signature and reduced resistivities at depth with elevated groundwater temperature. In Phase 3 of the Play Fairway project, drilling just outside the Pālāwai Basin confirmed increasing water temperature to a depth >1 km (Lautze and Thomas, 2021). Current trends in geothermal energy focus on the use of Machine/Deep Learning (ML) for industry problem-solution and decision-making (Smith et al., 2022).

In general, the ML applications fall into three classes: data-driven ML, physics-informed ML, and multimodal ML. According to Qin et al. (2022), data-driven ML models suffer from three limitations: 1) extensive data requirements, 2) lack of physical plausibility and interpretability, and 3) poor generalizability beyond training data. Physics-informed ML can overcome these shortcomings by either 1) embedding mass and/or energy balance equations into the ML algorithm, or 2) informing the ML algorithm using properties derived following the inversion of (multi)physics-based equations. In the (multi)physics-informed ML approach, the incorporation of physics into the ML algorithm typically relies on using simplified set of mass and/or energy balance equations (He et al., 2020; Qin et al., 2022) along with the addition of a physics-guided loss function to regularize training of the associated parameter types. In an alternative approach, a completed set of explicitly coupled water-heat-solute transport equations were used together with supervised ML algorithms that stochastically sampled different types of field measurements to reduce the computational burden and improve the parameter estimates (Friedel, 2020). In this study, we present the multimodal ML (MML) approach, which affords the possibility to incorporate any number of estimated property types (using traditional (multi)physics-based numerical model mass and/or energy calibrations and geophysical inversions, data-driven ML, and (multi)physics-informed ML) along with any number of point field measurements to inform the process.

The aim of this study is to demonstrate the efficacy in using an unsupervised MML workflow for constructing continuous transdisciplinary set of subsurface Geothermal Stratigraphic Units (GSUs). Conceptually, the GSUs reflect a single data product that is spatially characterized by a collection of continuous multiparameter voxels with individual parameters and their univariate statistics. We

hypothesize that the hydrogeologic, geothermal, and geophysical data collected across monitoring networks provide mutual information (measure of entropy describing mutual dependence among random variables) suitable for their assimilation, discovery, and prediction of continuous features when using the MML workflow. To test this hypothesis, we evaluate the ability of an unsupervised MML to achieve the following objectives: 1) develop open community datasets, 2) identify data acquisition targets with high value for future work, 3) identify new signatures to detect hidden groundwater and geothermal resources, and 4) foster new capabilities for characterizing subsurface temperature and permeability. This study extends the work of Friedel (2016) who applied an unsupervised MML workflow to construct a continuous set of hydrostratigraphic units (HSU) for characterizing shallow (<200m) groundwater systems, and Vesselinov et al. (2020) who used an unsupervised ML methodology to discover hidden signals in field data and extract their dominant attributes in a shallow (<200m) hydrothermal system.

2. METHODS

The MML workflow used to determine GSUs involves the following steps: (1) feature selection, (2) feature prediction, and (3) feature clustering.

2.1 Feature Selection

The identification of suitable features (variables) for building the MML model is undertaken using the nonlinear wrapper approach called learn heuristics with feature constraints. The wrapper model requires one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are selected (Yu and Liu, 2003). In this regard, the learn heuristics reflect the ML algorithm that is introduced into the Metaheuristics algorithm (Buscema et al., 2013). The ML algorithm is generalized for which multiple learning algorithms are evaluated, for example, Backpropagation, K-Nearest Neighbor, or Naïve Bayes, inside a genetic algorithm (GA) (metaheuristic). Constraints on the learn heuristic approach typically involve features such as one or more response variables. Given that the GA converges to a global minimum surface (not a single vector), the process is randomly reinitiated and run until the subset of optimal features can be identified based on their mode of values.

2.2 Feature Prediction

The Modified Self-Organizing Map (MSOM) procedure is used to predict feature vectors at unsampled locations. The prediction method is sufficiently robust to cope with feature vagaries due to sample size and extreme data insufficiency, even when >80% of the data are missing (Friedel and Daughney, 2016). The MSOM procedure involves the sequential application of competitive learning (self-organizing map) and estimation (minimization of objective function by competitive learning). The architecture of MSOM involves an input layer (signals from the environment) and an output layer (competitive feedback to the environment). The input layer comprises a set of nodes (neurons) that are connected one to another through a rectangular topology. The connections between inputs (data vectors) and nodes have weights, so a set of weights corresponds to each node.

In implementing the MSOM procedure, the competitive learning process iteratively modifies weights during the training phase so that the self-organized output pattern becomes consistent, meaning that the input pattern will produce the same self-organized output pattern. In doing so, the SOM iteratively maps each data sample as a vector (each variable is characterized as a cloud of data vectors) across a hypersurface on which data vectors closer to each other are more related (self-similar) than data vectors farther away. The learning algorithm may be summarized as follows (Kohonen, 2001):

1. Generate initial values including weights, radius, and learning parameter values
2. Select input vector from data set
3. Identify winning node, which is the closest node to the input vector using the Euclidean distance metric
4. Identify the neighborhood with the given radius using a Gaussian function
5. Update the weights for the winning node and all the nodes in the same neighborhood
6. Repeat steps 2 to 6 until the weight vectors reach a con-verged state
7. The estimation of missing values (sometimes referred to as imputation) is done simultaneously for all variables across the hypersurface (Kalteh et al., 2008).

Feature prediction is undertaken with the best-matching unit vectors supplied as initial values at locations of interest (Wang, 2003) that are iteratively perturbed while minimizing topographical and quantization error vectors during the competitive learning process (Kalteh et al. 2008).

Training and testing of the MSOM model are carried out using a stratified split-sample validation approach. In this approach, the data records are randomly shuffled and split with 80% set aside for training and 20% set aside for testing. In total, the original data set is split using this procedure N times. Each split characterizes a fold used in the N-fold cross-validation statistics to assess the generalizability of the unsupervised ML algorithm when presented independent field data. The metrics used to evaluate model performance include R-squared (a statistical measure representing the proportion of variance for the observed variable that is *explained* by the estimated variable) and mean squared error (predictive success) for continuous features; and accuracy (predictive success) and Cohen's Kappa

(comparison between the observed accuracy and the agreement expected due to chance) for binary (present or absent) features. Cohen 1960) suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as almost perfect agreement.

2.3 Feature Clustering

Grouping statistically meaningful features across the self-organized hypersurface is undertaken by k-means clustering (Vesanto and Alhoniemi 2000). The best partitioning for each number of clusters is determined based on the Euclidian distance criterion and interesting merges defined using the Davies-Bouldin index (Davies and Bouldin 1979). In this study, the optimal number of clusters is determined from repeating the k-means process to avoid convergence at local minima. The Davies-Bouldin validity index is selected as a matter of convenience, but other validity measures could be used.

3. CASE STUDY

3.1 Study Region

The MML study includes the Island of Lānaʻi (Lānaʻi) and adjacent Pacific Ocean located about 96 km southeast of Honolulu, Oahu (Figure 1). Lānaʻi is a cashew-shaped island with a maximum width of 29 km in the longest direction. The highest point on Lānaʻi is 1026 m with a land area of 364 km^2. The geology of Lānaʻi is described as an eroded extinct basaltic volcano that developed during one period of activity (Stearns, 1940). Structurally, the island has three primary rift zones (Northwest Rift zone, Southwest Rift zone, and South Rift zone) and caldera located in the Pālāwai Basin (Figure 2). The summit plateau resulted from collapse along the northwest rift zone. Lānaʻi island has fault breccias and dike complexes that lie in the rift zones radiating from the Pālāwai Basin. Relatively low permeability dikes often enclose porous lavas of relatively high permeability forming local compartmentalized reservoirs (Stearns,1940). Basaltic rocks exposed along the west coast are thinly bedded and considered the most permeable rocks due to cavities and fractures within and between lava flows and thought to permit the landward intrusion of sea water. Groundwater temperatures measured in a well on the rim of the Pālāwai Basin during Play Fairway Phase 3 vary from about 21C at the land surface to about 65C at a depth of 1km. Little is known about the subsurface groundwater-geothermal system below this depth, although temperatures at the Moho (about 13 km depth) are likely in the range of 444 C to 892 C (Schutt et al., 2018). Therefore, understanding the occurrence of groundwater and geothermal resources requires knowledge of the distribution and characteristics of these basalt-dike systems.

Figure1. Study map showing: (left) proximity of Lānaʻi to the Hawaiian Islands, and (right) Lānaʻi cross-section for which MML results are extracted (from the surface to 14.75 km) for evaluation and construction of the conceptual model.

3.2 Field Data

The field data used in this study reflects a subset of those data assembled and/or collected during The Hawaii Play Fairway project, funded by the DOE's GTO. This data set includes geologic, groundwater, and geophysical observations relevant to subsurface heat, fluid, and permeability for collected at the island of Lānaʻi (Lautze et al., 2020). For example, the geologic data includes information on basalt and dike rocks identified at the surface and in drill cores. Groundwater data collected in boreholes include physical properties (water level and specific capacity), aqueous chemistry (major ions, nutrients, and metals), environmental tracers (isotopes), and aqueous parameters (dissolved oxygen, specific conductivity, and temperature). These datasets reflect field sampling from the surface to a maximum borehole depth of about 1 km (Figure 2). To connect this sample region to greater depths, uncollated surface gravity and magnetotelluric measurements were collected and deterministically inverted to estimate 3D distribution of density and electrical resistivity values across Lānaʻi to depths of more than 14 km (Lautze et al., 2020).

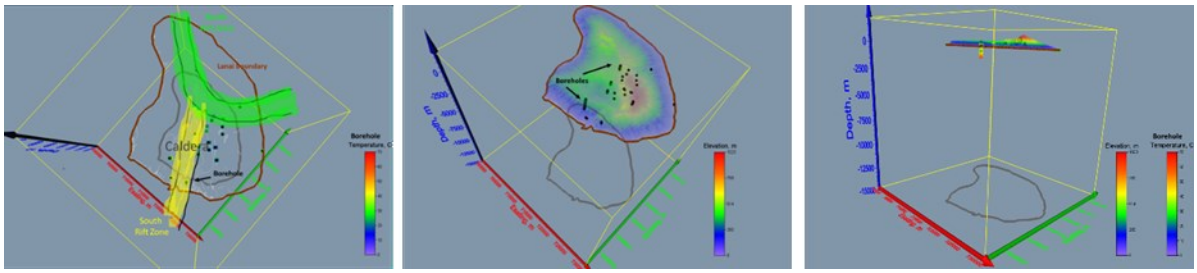

Figure 2. Study map showing Lāna'i coastal boundaries at sea level (brown) at 15km depth (gray); left panel includes primary rift zones: north (green) and south (yellow); caldera, borehole/well locations (shown as dots with color indicating temperature); middle and right panels show location of borehole/well locations with respect to elevation.

4. RESULTS AND DISCUSSION

4.1 Feature Selection

Of the original predictor variables (N = 54), the learn-heuristic based feature-selection process identified an optimal global set of variables (N = 27) suitable for predicting features beneath Lanai (Table 1). These variables are used in building of the unsupervised learning-based models. Important predictors that appear common to all three state variables include Water: Ocean, Geology: basalt and dike, Physical properties: density, resistivity, specific capacity; Aqueous properties: specific conductivity, oxygen reduction potential, and pH; Aqueous chemistry: HCO3, NO3, PO3, SO4, r, F, F, K, Na, Si, Sr, and Ca/Mg ratio; Stable isotopes: oxygen and hydrogen; and isotopic ratios including Ar, He, and Ne.

Constraints on learn heuristics

Informative features

		Type		Observation	
Category	Features	Discrete	Continuous	Measured	Derived
State variables	Head, m		X	X	
	Temperature, C		X	X	
	Chloride concentration, mg/l		X	X	
Water	Ocean	X		X	
Geology	Basalt	X		X	
	Dike	X		X	
Physical properties	Density, kg/m3		X		X
	Resistivity, ohm-m		X		X
	Specific capacity, m3/d/m		X		X
Aqueous properties	Specific conductivity		X	X	
	Oxygen reduction potential, mv		X	X	
	pH		X	X	
Aqueous chemistry	HCO3, mg/l		X	X	
	NO3, mg/l		X	X	
	PO4, mg/l		X	X	
	SO4, mg/l		X	X	
	Br, mg/l		X	X	
	F, mg/l		X	X	
	Fe, mg/l		X	X	
	K, mg/l		X	X	
	Na, mg/l		X	X	
	Si, mg/l		X	X	
	Sr, mg/l		X	X	
	Ca/Mg		X	X	
Stable isotopes	^{1}H and ^{2}H		X	X	
	^{16}O, ^{17}O and ^{18}O		X	X	
Isotopic ratios	4He/20Ne		X	X	
	20Ne/22Ne		X	X	
	40Ar/36Ar		X	X	
	38Ar/36Ar		X	X	

Table 1. Summary of features determined to be informative using the learn heuristics approach constrained to the state variables of head (m), temperature (C), and chloride concentration (mg/l). Physical properties: resistivity (ohm-m), specific capacity (m3/d/m); Aqueous properties: specific conductance, uM/s, oxygen reduction potential (mv); Aqueous chemistry: Br = bromide, (mg/l) F = Flouride (mg/l), Fe = iron (mg/l), HCO3 = bicarbonate (mg/l), K = potassium (mg/l), Na = sodium (mg/l), NO3 = nitrate (mg/l), PO4 = phosphate (mg/l), Si = silica (mg/l), SO4 = sulfate (mg/l), Sr = strontium (mg/l), Ca/Mg = calcium -magnesium ratio; Isotopes: abs 18O = absolute value of stable oxygen isotope, absolute value of Deuterium, absolute carbon isotope, Isotopic ratios: 4He20Ne = helium – neon ratio, 40Ar36Ar = agon ratio, 38Ar36Ar = Argon ratio.

4.2 Feature Prediction

Prior to training of the MSOM, features (field variables) were normalized by their data variance and randomly assigned (presenting the input vectors to the map sequentially using a randomly sorted database) as an initial set of map weight vectors. Application of the MSOM network to training data is done using a single fixed number of neurons and topological relations. The selected neural map shape (148 rows by 140 columns) is a toroid (wraps from top to bottom and side to side) with hexagonal neurons. Training of the map was conducted using both rough and fine phases. The rough training phase involved 20 iterations using a Gaussian neighborhood with an initial and final radius of 204 units and 51 units; and the fine training involved 400 iterations using a Gaussian neighborhood with an initial and final radius of 51 units and 1 unit. The initial and final learning rates of 0.5 and 0.05 decayed linearly down to 10^{-5}, and the Gaussian neighborhood function decreased exponentially from a decay rate of 10^{-1} iteration to 10^{-3}, providing reasonable convergence evidenced by similarity in their low quantization (q_e=0.073) and topographic (t_e=0.076) errors.

Testing and validation of the trained MSOM model are reflective of the five split sets previously discussed. The cross-validation statistics further reveal that the MSOM model has a moderate ability (kappa values > 0.53) to predict dikes and basalt with strong prediction accuracies (accuracy values > 0.99). For the sake of brevity, summary statistics and scatter plots are presented for independent observations and predictions of selected features that include temperature, chloride concentration, resistivity, and density. Note that the metrics for temperature predictions are reflective of those measured in the upper 1 km, because the magnitude of increasing temperatures at greater depths is unknown. In general, the cross-validation statistics (table 2) reveal similar observed to predicted values (minimum, average median and maximum) with reasonable prediction accuracies (standard deviation / square root of count): resistivity = 0.32 ohm-m, density = 1.2 kg/m3, temperature = 0.002 C, chloride concentration = 1.16 mg/. The associated scatter plots (Figure 3) of observed versus predicted values for these features have reasonable R-squared values (resistivity, ohm-m = 89.9%; density, kg/m3 = 99.9%, chloride, mg/l = 68.8%; temperature, C = 92.8%) albeit with some bias at smallest values. The bias at small values for density, conductivity, and chloride concentration is attributed to challenges in predicting these features at the ocean-basalt interface given the large voxels used during deterministic inversions and moderate ability to predict the presence or absence of basalt. Future studies may benefit by introducing higher frequency audiomagnetotelluric measurements to improve near surface estimates of resistivity at the coastal boundary and shallow depths. Likewise, the prediction of small density values may be reflective of the inability of MSOM to resolve density at the ocean-coastal boundary due to grid resolution.

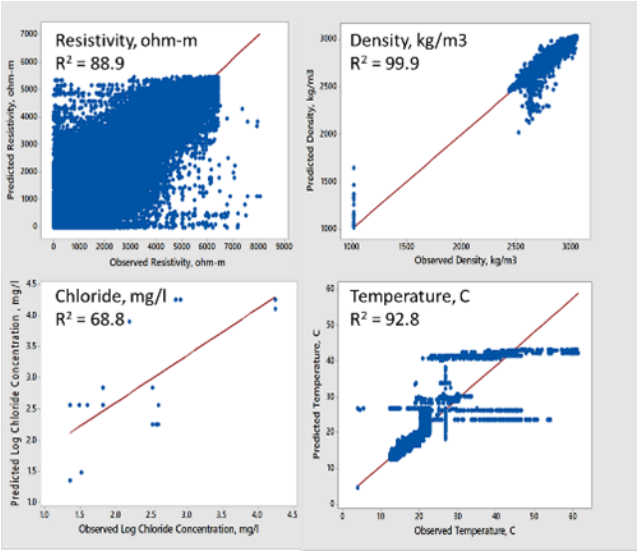

Figure. 3. Scatter plots showing observed versus predicted values for selected features with fitted line and R-squared values: resistivity, density, chloride concentration and temperature.

Table 2. Summary statistics for observed and predicted features: temperature (C), chloride concentration (mg/l), resistivity (ohm-m), and density (kg/m3). Note that statistics for the temperature are confined to those measured in the upper 1 km given that temperatures at the Moho are assumed to be about 758C (Schutt et al., 2018).

Temperature, C

Statistic	Observed, C	Predicted, C	Error, C
Minimum	4.00	4.00	0.00
Average	22.5	23.4	0.34
Median	22.4	23.4	0.03
Maximum	61.4	61.3	30.4
Standard deviation	4.69	8.19	1.22
Count	213824	1044532	213824

Chloride, mg/l

Statistic	Observed, mg/l	Predicted, mg/l	Error, mg/l
Minimum	23.0	23.0	0.0
Average	17999	12956	1.3
Median	18000	18000	0.0
Maximum	18000	18002	17670
Standard deviation	157.7	7974	533.3
Count	212919	1044530	212919

Resistivity, ohm-m

Statistic	Observed, ohm-m	Predicted, ohm-m	Error, ohm-m
Minimum	0.10	0.14	-86.5
Average	516	385	130
Median	100	35.1	25.4
Maximum	8066	5470	7181
Standard deviation	940	786	288
Count	799480	1044530	799480

Density, kg/m3

Statistic	Observed, kg/m3	Predicted, kg/m3	Error, kg/m3
Minimum	1020	1020	0.00
Average	1129	2016	21.9
Median	1020	2617	0.00
Maximum	3055	3037	2029
Standard deviation	420.6	862.4	182.6
Count	227176	1044530	227176

A natural outcome of using the trained MSOM model is the simultaneous prediction all model features. Given the large number of available features, a reduced number of features is selected for presentation that include state variables (temperature, chloride concentration), physical properties (specific capacity – a surrogate for permeability), geology (basalt and dike), and Geothermal Stratigraphic Units (GSU). Each of these features represent predicted values extracted from the 3D data cube along the cross-section A-B. The cross-section A-B trends SW-NE along a profile that begins at the Pacific Ocean on the west, crosses the caldera dike system, point of highest elevation, and ends at the Pacific Ocean on the east. Collectively, these figures reveal a heterogeneous groundwater –

geothermal system from which information can be useful in developing a conceptual model. Testing reveals that the prediction process preserves reasonable population statistics for the geologic, hydrogeologic and geophysical features, but the individual feature predictions reflect a random process (statistics change for different split sets). For this study, the cross-validation statistics demonstrate the ability of the MSOM to generalize when estimating continuous features from sparse field data characterizing the different support volumes.

This section presents a subset of features predicted while using the trained MSOM and 5 sets of independent observations. For example, the 3D density and resistivity predictions are presented in Figure 4. Some general spatial observations are that both density and resistivity are heterogeneous through the region. Some density features include minimum values associated with the ocean and greatest density inland at the caldera and to depth of about 15 km. The predicted resistivity also appears heterogenous through the region with features at sea level displaying more resistive character at and below the region of greatest elevation and conductive features similar in magnitude to the ocean but decreasing landward in areas located at the northern most point and southwestern point (toward point A on the cross section). Another slightly deeper region of high conductivity (low resistivity) appears at the eastern boundary of the island (toward point B on the cross section). These areas suggest regions landward intrusion of sea water with decreasing gradients over about 5 km. In the area of maximum elevation, the resistivities are greatest possibly due in part to groundwater recharge of freshwater. Below about 1 km, the anomalies appear to be vertically oriented as rectangular regions of alternating medium to high resistivity values.

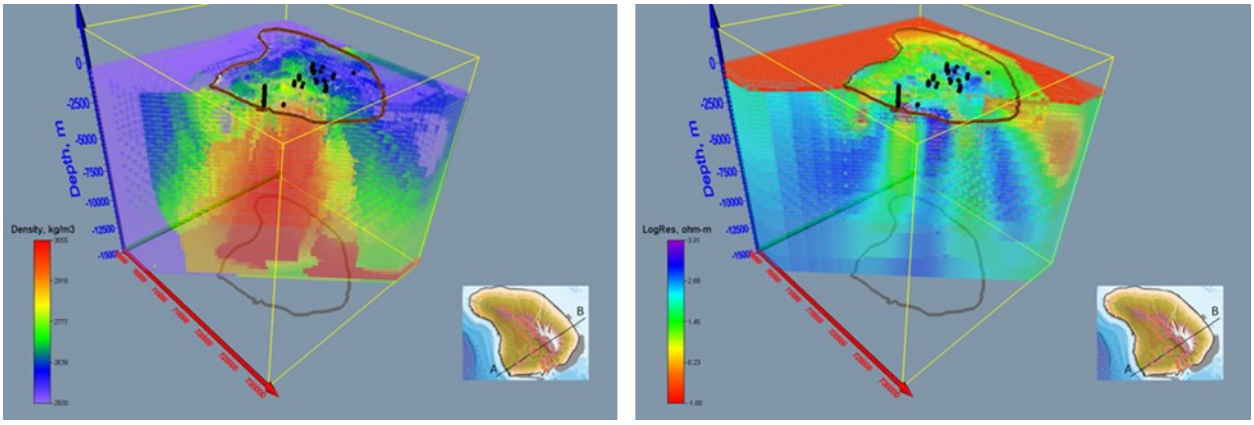

Figure. 4. Inverted 3D geophysical property distributions sliced horizontally at sea level and vertically along A-B: (left) density, kg/m3 (density > 2500 kg/m3), and (right) log resistivity, ohm-m.

The predicted spatial occurrence of basalt (left) and dike (right) as predicted using MSOM are shown in Figure 5. The general character of the predicted basalt is the prominence from surface to depths as great as about 10 km. One characteristic that the basalt appears to be draped over the region indicated as not basalt. The not basalt region appears as dike material with several interesting features. In fact, the dike material in the region of not basalt is interpreted as a domed shape pluton with roots below the Moho located at about 12.5 km depth. The change in character from the pluton at depth is interpreted as the possible location of the Moho. Other vertical features are interpreted as rising dike swarms to the west side of the pluton (between 5 and 8 km depth) and at the top of the pluton below the caldera region (between sea level to about 1 km depth). The later dike swarms correspond to the region denoted as a pipe zone by Malahoff and Woollard (1966) based on airborne total field magnetic anomalies. Lastly, the western dike swarm rose vertically to a depth of about 5 km below the surface where they terminate forming what appears to be a sill.

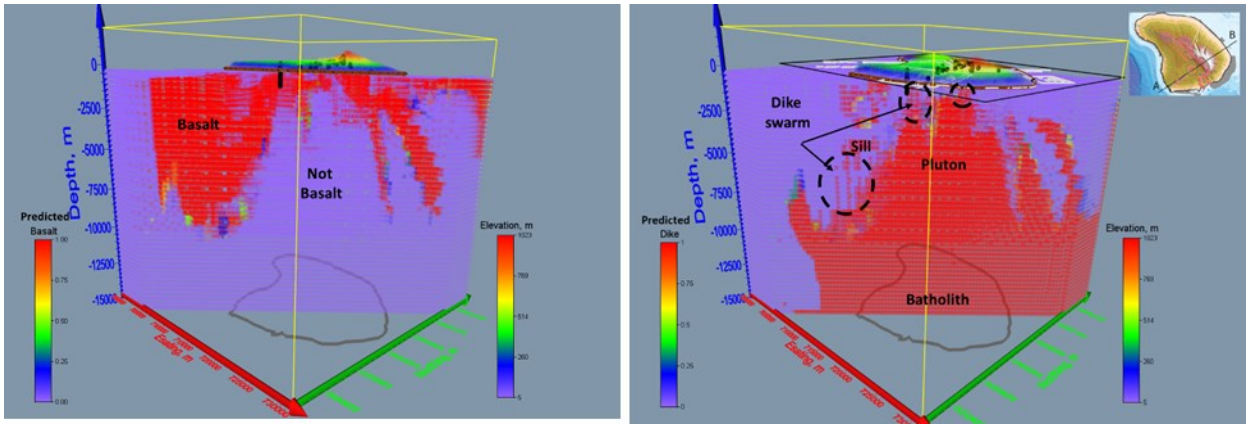

Figure 5. Predicted 3D geology distribution (probability) sliced horizontally at sea level and vertically along A-B: basalt (left) and dike (right). Dike image includes likely dike swarms, sill, pluton, and Moho (about 12.5 km). Location of shallow dike swarms correspond to airborne dipole magnetic anomalies (Malahoff and Woollard, 1966).

The predicted spatial occurrence of the ocean (left) and specific capacity (right) are shown in Figure 6. The character of the ocean appears correctly located adjacent to the island and increasing at depth to the western and eastern boundary. That said, the ocean at the western edge appears to extend to the bottom of the model domain. In fact, the actual ocean depth in this region is known to be about 6 km thereby revealing the apparent lack of resolution in resistivity and gravity information to inform the MSOM model. Improvements in predictions for this region of the model would likely require the addition of related properties derived from shipborne and/or ocean bottom geophysical measurements. The specific capacity predictions in this study represent the analog to permeability. In general, there is a trend characterized by low specific capacity at the eastern boundary increasing toward the west. This trend follows from an earlier discussion indicating that the most permeable basaltic rocks are located at the western edge of Lāna‘i. Also, the western dike swarm (between 5 and 8 km depth) is characterized as medium specific capacity, whereas the adjacent region to the east is characterized as low specific capacity.

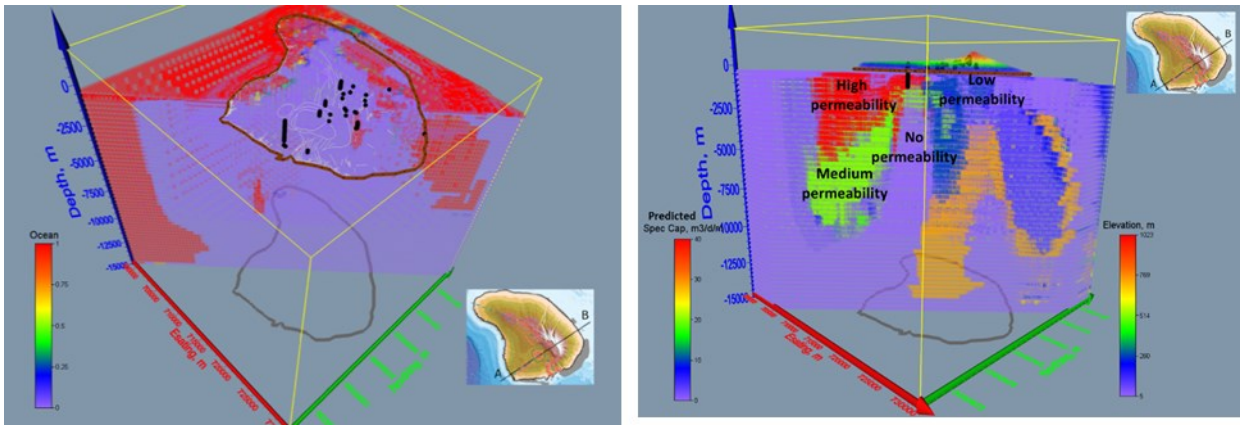

Figure 6. Predicted features sliced horizontally at sea level and vertically along A-B: (left) 3D ocean character (present//absent), (right) specific capacity (permeability analog) This figure reveals a trend where low specific capacity at the eastern boundary increases toward the west. This trend follows from an earlier discussion indicating that the most permeable basaltic rocks are located at the western edge of Lāna‘i . Well locations are black dots.

The predicted spatial distribution of temperature is over the range of 5°C to 758°C as shown in Figure 7. The upper left panel reveals the plan view temperature distribution at sea level with the digital elevation model overlying the island. The deep ocean at the model boundary reveals the coldest temperatures of about 5°C. The upper right panel reveals the plan view temperature distribution at sea level with no digital elevation model. Noteworthy are the cool temperatures near 20°C (blue) in the region of highest elevation and warm temperatures near 65°C (orange) in and around the caldera region. The lower left panel has interpreted regions of downward groundwater recharge from the region of highest elevation and cool temperatures and likely upward convective transport from the 758C source at the Moho through the interpreted dike swarm and sill region of medium specific capacity (surrogate for permeability) toward the warm 65C region of in the vicinity of the deepest well. The region directly below this well and adjacent to the interpreted region of convective transport is predicted to have the no specific capacity (no permeability) thereby supporting plausible convective transport through the more permeable region with conductive heat transfer away from the convective heat transport pathway. The lower right panel identified two geothermal prospects. The first geothermal prospect has predicted temperatures of about 65°C to 108°C and is located between 1 and 2 km depth in the vicinity of the caldera. The second and preferred geothermal prospect has predicted temperatures of about 149°C to 275°C located between 3 to 6 km depths in the vicinity of the western dike swarm and sill.

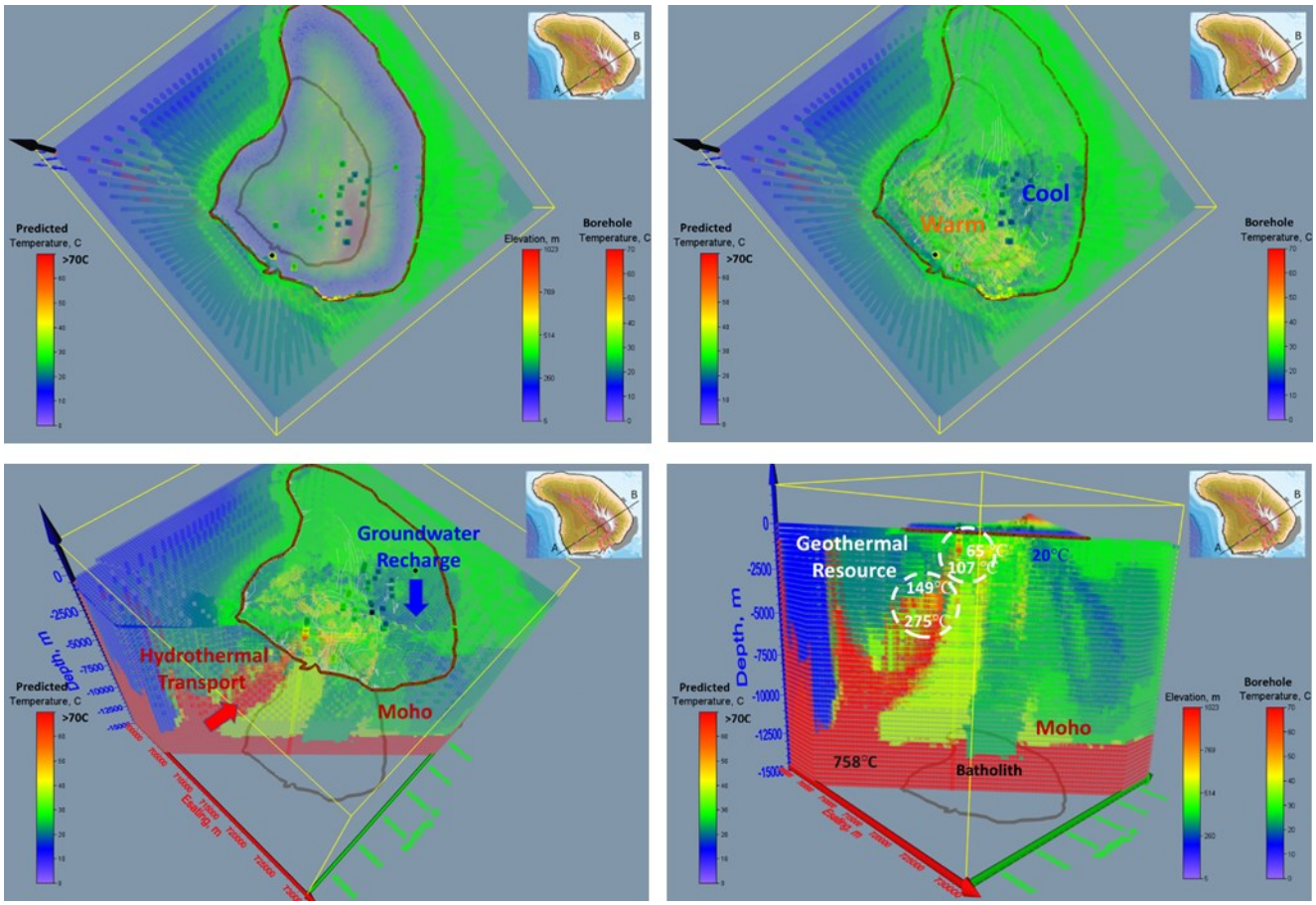

Figure 7. Predicted temperature over the range of 20C to 758C. Upper left: plan view temperature distribution at sea level with the digital elevation model. Upper right: plan view temperature distribution at sea level with no digital elevation model (lower left). Noteworthy are the cool temperatures (blue) attributed to recharge at high elevation and warm temperatures in and around the caldera region. Lower left: interpreted regions of groundwater recharge from the region of highest elevation and hydrothermal transport from the region below the Moho. Lower right: two geothermal prospects; first is about 65°C to 107°C between 1 and 2 km depth in the vicinity of the caldera, second is about 149°C to 275°C between 3 to 6km in the vicinity of the western dike swarm and sill.

The predicted spatial distribution of chloride concentration sliced horizontally at sea level and vertically along A-B is shown in Figure 8. The left linear concentration plot spans 23 mg/l (freshwater or no water) to 18000 mg/l (ocean water). The right log chloride concentration plot reveals heterogeneous water types that include freshwater, brackish water, and saline water. Two potential freshwater resource zones are indicated with circles. These freshwater resource zones have similar specific capacity but the zone to the west (closer to the caldera) has elevated temperatures, whereas the zone to the east has cooler temperatures. A conceptual groundwater system is constructed based on chloride concentration, geology, and temperature predictions (figure 9).

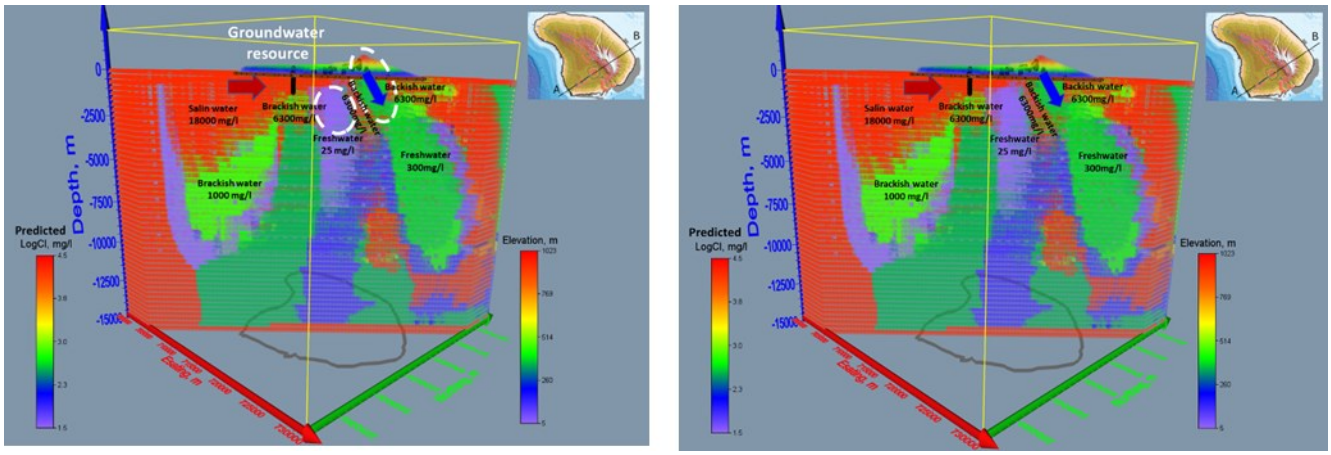

Figure 8. Predicted chloride concentration sliced horizontally at sea level and vertically along A-B. (left) linear concentration plot (23 mg/l (freshwater or no water) to 18000 mg/l (ocean water); (right) log chloride concentration plot showing heterogeneous water types that include freshwater, brackish water, and saline water. Two potential freshwater resource zones are present. The resources have similar permeability but zone to the west (closer to the caldera) has elevated temperatures, whereas the zone to the east has cooler temperatures.

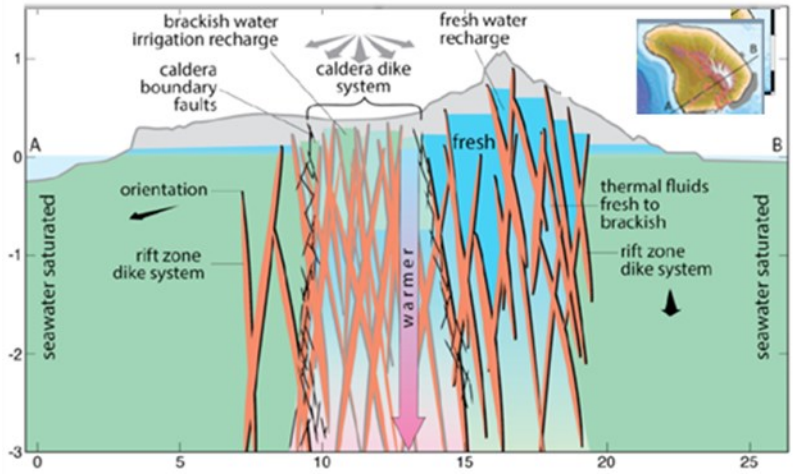

Figure 9. Conceptual model produced by Lautze and Thomas (unpublished) which agrees with MML predictions of continuous features along cross-section A-B

4.3 Feature Clustering

The joint interpretation of information across the MSOM network is undertaken by application of stochastic k-means clustering. In this case, the mode of k-means clusters is interpreted as the most likely and used to aggregate information. The clustering of similar information at various locations characterizes the variability of parameters associated with these Geothermal Stratigraphic Units (GSU). The distribution of GSUs mapped along the cross-section A-B on Lāna‘i reveal five basic groups shown in Figure 10: 1. Ocean. 2. Salinized basalt. 3. Basalt. 4. Basalt-dike transition. 5. Dike. Each GSU is characterized by univariate and spatial statistics that describe the physical and chemical properties comprising the geothermal model (Table 3). Other interpreted features include a pluton, batholith, depth to underplating/oceanic crust and Moho (Lahey et al., 2010).

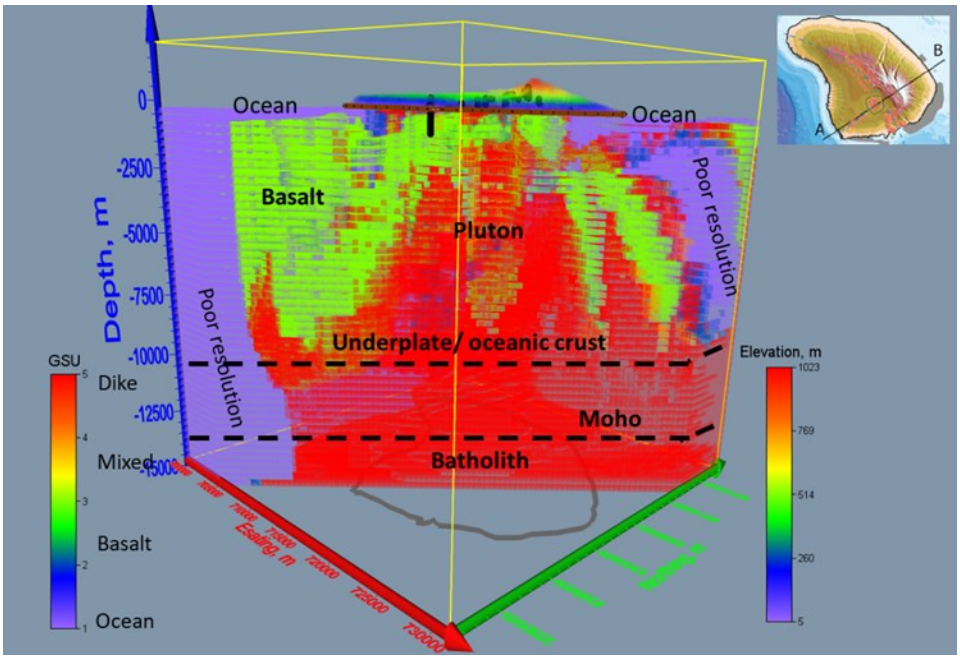

Figure 10. Predicted 3D Geothermal Stratigraphic Units sliced horizontally at sea level and vertically along A-B reveal five basic groups: 1. Ocean (purple), 2. salinized basalt (blue), 3. Basalt (green), 4. basalt-dike transition (orange), and 5. Dike (red). Each GSU is characterized by univariate and spatial statistics describing all the physical and chemical properties comprising the model (Table X). Well locations appear as black dots. Other interpreted features include a pluton, batholith, depth to underplating/oceanic crust and Moho (Lahey et al., 2010).

Table 3. Summary statistics for selected properties associated with Geothermal Stratigraphic Units. Location; Elev = elevation (m); State variables: Head = hydraulic head (m), cl = chloride concentration (mg/l), Temp = temperature (C); Physical properties: Res = electrical resistivity (ohm-m), SpecCap = specific capacity, (m3/d/m); Aqueous properties: Cond (specific conductance, uM/s, ORP = oxygen reduction potential (mv); Aqueous chemistry: Br = bromide, F = Flouride, Fe = iron, HCO3 = bicarbonate (mg/l), K = potassium, Na = sodium, NO3 = nitrate (mg/l), PO4 = phosphate (mg/l), Si = silica, SO4 = sulfate, Sr = strontium, Ca/Mg = calcium -magnesium ratio.

Group		Summary	Location			Water	Geology		State Variables			Physical properties			Aqueous properties		
GSU	Description	Statistic	Easting, m	Northing, m	Elev, m	Ocean	Dike	Basalt	Head, m	Temp, C	Cl, mg/l	SpeCap, m3/d/m	Density, kg/m3	Res, ohm-m	Cond, uM/s	ORP, mv	pH
1	Ocean	minimum	698886	2286203	-14745	1.00	0.00	0.00	0.0	4.0	16593	0.0	1020	0.1	0	51.9	8.00
		average	712723	2300699	-2042	1.00	0.00	0.00	0.0	58.0	17999	0.0	1325	119	47961	111.3	8.00
		median	709822	2299659	-319	1.00	0.00	0.00	0.0	24.4	18000	0.0	1020	0	48000	88.5	8.00
		maximum	730986	2318303	2055	1.00	0.01	0.01	0.1	758.0	18000	0.0	2960	3628	48000	283.1	8.00
		Standard deviation	10501	10412	3955	0.00	0.00	0.00	0.0	157.0	21	0.0	645	353	1267	55.0	0.00
2	Salanized basalt and dikes	minimum	700086	2286803	-14745	0.00	0.00	0.00	0.0	19.8	164	0.0	1020	0.1	0	52.1	7.95
		average	713052	2306872	-1822	0.45	0.29	0.25	10.6	49.9	17611	0.0	2062	60	46523	112.3	8.00
		median	711256	2309303	-945	0.36	0.03	0.00	0.0	26.8	18000	0.0	2515	2	48000	88.5	8.00
		maximum	730986	2318303	2055	0.98	1.00	1.00	465.1	758.0	18000	7.3	2948	3818	48000	283.1	8.00
		Standard deviation	8740	8012	3216	0.39	0.39	0.38	59.6	130.0	2102	0.3	716	319	7339	58.9	0.00
3	Basalt	minimum	700756	2287403	-11445	0.00	0.00	0.01	0.0	18.9	23	0.9	2478	4	0	52.1	7.14
		average	714600	2300283	-3573	0.00	0.01	0.99	42.9	29.5	5901	19.6	2717	445	1365	150.9	7.87
		median	713886	2298203	-2145	0.00	0.00	1.00	0.9	28.1	708	20.0	2703	233	166	196.2	7.90
		maximum	730986	2318303	855	0.00	0.99	1.00	484.2	73.0	18000	41.0	3011	4424	48000	283.1	8.00
		Standard deviation	7806	7622	3165	0.00	0.10	0.10	95.4	10.1	8072	13.3	96	516	6207	54.8	0.13
4	Dikes and Basalt	minimum	702786	2290103	-10245	0.00	0.00	0.00	0.0	18.9	31	1.4	2614	3	0	52.1	7.14
		average	723764	2298945	-3448	0.00	0.53	0.47	263.8	28.1	7623	25.1	2769	244	856	151.7	7.86
		median	724986	2298803	-2445	0.00	0.57	0.43	224.0	24.5	400	29.0	2747	34	167	196.2	7.91
		maximum	730986	2318303	306	0.25	0.99	0.99	484.3	46.0	18000	41.0	3048	4933	22370	227.4	8.00
		Standard deviation	5428	5197	2537	0.02	0.41	0.41	172.7	9.2	8823	14.9	92	603	1558	58.2	0.20
5	Dikes	minimum	698886	2286503	-14745	0.00	0.00	0.00	0.0	18.9	23	0.0	1020	0.1	0	52.1	7.14
		average	717025	2304238	-7022	0.00	0.83	0.17	229.3	182.3	6416	4.7	2768	748	14322	153.7	7.79
		median	717186	2304259	-7245	0.00	1.00	0.00	246.3	26.8	543	0.0	2807	237	2036	135.3	8.00
		maximum	730986	2318303	2055	0.80	1.00	1.00	485.6	758.0	18000	41.0	3055	8066	48000	283.1	8.00
		Standard deviation	7819	8007	4903	0.01	0.37	0.37	179.8	297.8	8376	8.9	262	1166	20843	58.2	0.34

Group		Summary	Location			Water Geology			Aqueous chemistry											
GSU	Description	Statistic	Easting, m	Northing, m	Elev, m	Ocean	Dike	Basalt	HCO3, mg/l	NO3, mg/l	PO4, mg/l	SO4, mg/l	Br, mg/l	F, mg/l	Fe, mg/l	K, mg/l	Na, mg/l	Si, mg/l	Sr, mg/l	Ca/Mg
1	Ocean	minimum	698886	2286203	-14745	1.00	0.00	0.00	139.0	0.70	0.10	2.65	60.97	1.38	0.02	357.1	9200	1.0	0.02	0.31
		average	712723	2300699	-2042	1.00	0.00	0.00	140.0	0.70	0.10	2.65	64.00	1.40	0.02	380.0	10000	1.0	1.04	0.31
		median	709822	2299659	-319	1.00	0.00	0.00	140.0	0.70	0.10	2.65	64.00	1.40	0.02	380.0	10000	1.0	0.38	0.31
		maximum	730986	2318303	2055	1.00	0.01	0.01	140.0	0.70	0.12	2.67	64.00	1.40	0.02	380.0	10000	1.1	2.66	0.32
		Standard deviation	10501	10412	3955	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.04	0.00	0.00	0.3	11	0.0	1.11	0.00
2	Salanized basalt and dikes	minimum	700086	2286803	-14745	0.00	0.00	0.00	103.6	0.62	0.10	2.65	7.51	0.59	0.01	27.4	311	1.0	0.04	0.31
		average	713052	2306872	-1822	0.45	0.29	0.25	139.8	0.70	0.10	2.65	63.19	1.40	0.02	376.8	9896	1.0	0.74	0.32
		median	711256	2309303	-945	0.36	0.03	0.00	140.0	0.70	0.10	2.65	64.00	1.40	0.02	380.0	10000	1.0	0.34	0.31
		maximum	730986	2318303	2055	0.98	1.00	1.00	140.0	0.72	1.51	3.81	64.00	1.40	0.02	380.0	10000	27.9	2.60	0.67
		Standard deviation	8740	8012	3216	0.39	0.39	0.38	1.7	0.00	0.05	0.05	4.70	0.04	0.00	22.8	713	0.9	0.84	0.01
3	Basalt	minimum	700756	2287403	-11445	0.00	0.00	0.01	52.5	0.48	0.00	2.65	0.06	0.00	0.00	1.9	16	1.0	0.04	0.31
		average	714600	2300283	-3573	0.00	0.01	0.99	84.9	1.69	2.25	27.11	3.78	0.11	0.17	22.6	526	1078	0.21	1.05
		median	713886	2298203	-2145	0.00	0.00	1.00	83.0	0.64	2.58	6.40	0.14	0.04	0.00	2.7	23	880	0.06	1.13
		maximum	730986	2318303	855	0.00	0.99	1.00	146.4	4.24	4.84	151.20	64.00	1.40	0.78	380.0	10000	1650	0.74	1.22
		Standard deviation	7806	7622	3165	0.00	0.10	0.10	25.9	1.63	2.12	33.57	13.69	0.30	0.31	80.8	2135	418	0.17	0.21
4	Dikes and Basalt	minimum	702786	2290103	-10245	0.00	0.00	0.00	52.5	0.48	0.00	2.65	0.06	0.00	0.00	1.9	16	1.0	0.04	0.31
		average	723764	2298945	-3448	0.00	0.53	0.47	71.5	0.90	2.05	11.49	0.65	0.04	0.06	4.7	74	1037	0.23	1.08
		median	724986	2298803	-2445	0.00	0.57	0.43	83.0	0.64	2.58	7.20	0.14	0.04	0.00	2.1	22	1130	0.38	1.13
		maximum	730986	2318303	306	0.25	0.99	0.99	142.8	4.26	3.87	76.00	64.00	1.40	0.78	380.0	10000	1650	0.74	1.22
		Standard deviation	5428	5197	2537	0.02	0.41	0.41	16.9	0.97	1.76	18.53	4.49	0.11	0.21	26.0	677	259.4	0.16	0.15
5	Dikes	minimum	698886	2286503	-14745	0.00	0.00	0.00	52.5	0.48	0.00	2.65	0.06	0.00	0.00	1.9	16	1.0	0.04	0.31
		average	717025	2304238	-7022	0.00	0.83	0.17	122.2	2.19	1.64	26.37	24.25	0.71	0.07	186.0	4842	645	0.57	0.66
		median	717186	2304259	-7245	0.00	1.00	0.00	140.0	0.70	0.10	6.40	1.88	0.13	0.02	12.7	127	842	0.38	0.84
		maximum	730986	2318303	2055	0.80	1.00	1.00	146.4	7.59	5.81	151.20	64.00	1.40	0.78	380.0	10000	1650	2.61	1.22
		Standard deviation	7819	8007	4903	0.01	0.37	0.37	26.5	2.56	2.31	38.14	29.35	0.67	0.20	186.8	4961	654	0.79	0.35

5. SUMMARY

In this study, a multimodal machine learning (MML) workflow was proposed, features selected, and modified self-organizing map (MSOM) model trained and tested using the Lanai subset of the Hawaii Play Fairway hydrogeologic, geothermal, and geophysical data. The Lanai MSOM model generalized to independent test data (five random 80/20 splits) despite data being characterized as sparse, spatially limited with different support volumes, and uncertain. The trained Lanai model was then used to predict continuous 3D features from which interesting merges of clusters characterized five 3D Geothermal Stratigraphic Units. The predicted features include state variables (such as hydraulic head, temperature, chloride concentration), physical properties (such as density, resistivity, specific capacity – permeability surrogate), geology (basalt and dike – generalized for plutonic and dike material), aqueous chemistry (major ions, metals, isotopes), and aqueous properties (pH, specific conductance, dissolved oxygen). Inspection of the predicted features revealed possible geologic attributes, such as basalt, dike swarms, sill, pluton, batholith, and Moho; possible groundwater attributes, such as downward groundwater recharge, saltwater intrusion, heterogenous chloride concentration, such as zones of freshwater, brackish water, and saline water; and geothermal attributes, such as upward convective hydrothermal transport along a gradient from the very hot Moho (758°C) to the near surface heterogeneous temperature distribution ranging from warm (65°C) to cool (20°C). Based on this information, two freshwater resources were identified from the surface to below 1 km depth: a warm water (50-65°C) region with intermediate specific capacity, low chloride concentration (<50 mg/l), and a cool water (about 20C) region with low specific capacity, low chloride concentration (<300 mg/l). Two geothermal resources were identified: one in the vicinity of the caldera where there is a warm (65°C) water at or near to the surface warmer down to 1 km depth, and a very hot hydrothermal plume being transported upward from the Moho (about 750°C) to commercially viable drilling depths of about 2.0 km to 6 km depth with temperatures of about 107°C to 275°C. The ability to determine Geothermal Stratigraphic Units (GSUs) from independent set of hydrogeologic, geothermal and geophysical measurements support their potential use in simultaneous characterization of combined groundwater and geothermal resources. Further, this novel approach affords the possibility for direct assignment of GSUs and their features to numerical model cells (or nodes), thereby minimizing ambiguity in the conceptualization to numerical modeling process including initial starting parameter values, boundary conditions, and geostatistical constraints in support of the calibration process. The MML workflow techniques used herein are recognized as novel in this application and their initial success warrants further research. The performance metrics in the Lānaʻi case study provide encouragement to continue this line of groundwater-geothermal research as part of the new Island Heat project funded by DOE's GTO.

ACKNOWLEDGEMENTS

This work was supported by a generous donation from the Pulama Lānaʻi to Lautze through the University of Hawaii Foundation, and the U.S. Department of Energy's Geothermal Technologies Office grant under award number DE-EE0006729.

REFERENCES

Buscema, M., Breda, M., Lodwick, W., 2013. Training with input selection and testing (TWIST) algorithm: a significant advance in pattern recognition performance of machine learning. J. Intell. Learn. Syst. 5, 29–38.

Cohen, J., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20, 37–46.

Connor, C.B., 2021, Magnetic anomaly over an inclined igneous dike. http://www.cas.usf.edu/~cconnor/magnetic_dike.html Last checked 12/5/2021

Davies, D.L., Bouldin, D.W., 1979, A cluster separation measure. IEEE Trans Patt Anal Mach Intel PAMI-1:224–227.

Friedel, M.J., 2016, Estimation and scaling of hydrostratigraphic units: application of unsupervised machine learning and multivariate statistical techniques to hydrogeophysical data, Hydrogeology Journal, 24, 2103-2122.

Friedel, M.J., Daughney, C., 2016, Statistical robustness of machine-learning estimates for characterizing a groundwater-surface water system, Southland, New Zealand, Achieving deep learning by systemizing machine learning with big data engines I, Poster IN11B-1619, Earth and Space Science Informatics, American Geophysical Union, San Francisco, CA, 2016.

He, Q.Z, Barajas-Solano, D., Tartakovsky, G., Tartakovsky, A.M., 2020, Physics-informed neural networks for multiphysics data assimilation with application to subsurface transport, Advances in Water Resources, 141, 2020, 103610, ISSN 0309-1708.

Hu, X., Steinsland, I., Simpson, D., Martino, S., & Rue, H., 2013. Spatial modelling of temperature and humidity using systems of stochastic partial differential equations arXiv, 1307.1402 v1.

Kalteh, A.M., Hjorth, P., Berndtsson, R., 2008, Review of the self-organizing map (SOM) approach in water resources: analysis, modeling and application. Environ Model Softw 23(7):835–845.

Leahy, G.M., Collins, J.A., Wolfe, C.J., Laske, G., Solomon, S.C., 2010, Underplating of the Hawaiian Swell - evidence from teleseismic receiver function, Geophys. J. Int., 183, 213-329.

Lautze, N.C., Ito, G., Thomas, D.M., Frazer, N., Martel, S., Hinz, N., Tachera, D.K., Hill, G., Pierce, H.A., Wannamaker, P.E., Martin, T., 2020. Play Fairway Analysis of geothermal resources across the State of Hawai'i: 4. Updates with new groundwater chemistry, subsurface stress analysis, and focused geophysical surveys. Geothermics 86,

Lindgren, F. Rue, H., 2011. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, J. R. Statist. Soc. B (2011)73, Part 4, 423–498.

Qin, Z., Jiang, A., Faulder, D., Cladouhos, T.T., and Jafarpour, B., A physics-guided deep learning model for prediction of geothermal reservoir performance, Proceedings, 47th Stanford Geothermal Workshop, Stanford University, Stanford, California.

Malahoff, A., Woollard, G.P., 1966, Magnetic surveys over the Hawaiian Islands and their geologic implications, Pac Sci 20(3): 265-311.

Schutt, D.L., Lowry, A.R., Buehler, J.S., 2018, Moho temperature and mobility of the lower crust in western United States. Geology 46(3), 219-222.

Smith, C.M., Nwosu, C., Aljubran, J., Ochie, K., Guidmundsdottir, H., and Okorafor, E.R., 2022, Recent trends in artificial intelligence for geothermal energy, Proceedings, 47th Stanford Geothermal Workshop, Stanford University, Stanford, California.

Stearns, H.T., 1940, Geology and ground-water resources of the islands of Lāna'i and Kahoolawe, Hawaii: Hawaii (Terr.) Division of Hydrography Bulletin 6, p. 177 p. (pt. 1, Geology and ground-water resources of Lāna'i, p. 1-115; pt. 2, Geology and ground-water resources of Kahoolawe, p. 117-173); 1 folded map in pocket (scale 1:62,500).

Vesanto, J., Alhoniemi, F., 2000, Clustering of the self-organizing map, IEEE Trans Neural Netw 11:586–600.

Vesselinov, V.V., Mudunuru M.K., Ahmmed, B., Karra, S., and Middleton, R.S., 2020, Discovering signatures of hidden geothermal resources based on unsupervised learning, Proceedings, 45th Stanford Geothermal Workshop, Stanford University, Stanford, California.

Wang, S., 2003, Application of self-organising maps for data mining with incomplete data sets. Neural Comput Applic 12:42–48

Yu, L, and Liu, H. Feature selection for high-dimensional data: A fast correlation-based filter solution, Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.