Predicting geothermal suitability for power production at the global scale through AI

Trumpy E.[1] and Coro G.[2]

[1] Institute of Geosciences and Earth Resources – National Research Council of Italy, via Moruzzi 1 – 56124 Pisa

[2] Institute of Information Science and Technologies "Alessandro Faedo" – National Research Council of Italy, via Moruzzi 1 – 56124 Pisa

eugenio.trumpy@igg.cnr.it

gianpaolo.coro@isti.cnr.it

Keywords: Geothermal energy, Artificial intelligence, Machine learning, Spatial probability distribution

ABSTRACT

Among renewables, geothermal energy has the highest baseload and sustainably allows natural electricity production, theoretically 24 hours a day, seven days a week, with low CO2 emission. Today, developing a geothermal project requires a preliminary stage of accurate, lengthy and expensive exploration activity to assess the viability of a study area for geothermal plant installation. In this study, we present the first global suitability map of geothermal sites as a reference to speed-up operations to locate more accurate studies and exploration campaigns. This map can act as valuable prior knowledge during assessments. It can help save time and money. Different environmental geospatial parameters potentially correlated to geothermal site suitability and geothermal power plant installations were collected from different sources, analysed, pre-processed and prepared as input data to a Maximum Entropy machine learning model. Our approach also complies with the Open Science principles because it allows different users and stakeholders to reproduce our results. Moreover, industry operators and policymakers can reuse our data, services, results, and web interfaces for other evaluations or to generate new maps at a regional scale.

1. INTRODUCTION

Many scientific disciplines, including Earth Science, increasingly use Artificial Intelligence (AI) methodologies, particularly those based on Machine Learning (ML) models. Geothermal exploration and resource assessment has recently started to use these methods too. The first stage of geothermal project development requires analysing the data produced by geophysical surveys and integrating heterogeneous datasets. Through AI, regional or small-scale assessments of geothermal resources can integrate and combine these data and identify suitable locations for siting geothermal wells and plants that would otherwise require time, invasive inspection, high costs and permissions from legislative authorities.

This work presents a worldwide map representing the places with the highest suitability for geothermal wells and plant installation. It summarises and complements the work of Coro and Trumpy (2020). A Maximum Entropy-based ML model (MaxEnt), mediated from ecological niche models, is used to build a spatial probability function based on real-valued vectors of environmental parameters. This function approximates the *prior* probability that a specific site is suitable for building a geothermal power plant. We trained our model on the environmental parameters of known efficient and operative geothermal power plants.

The data were retrieved from different repositories with different spatial distributions and resolutions. A pre-processing stage was necessary for noise and error removal, gap filling and alignment to a 0.5° global-scale resolution. This phase used algorithms to interpolate, fill data gaps, and produce homogeneous global-scale distributions. We used inverse distance weighting, kernel density, and Data-Interpolating Variational Analysis (DIVA) for this scope.

The environmental parameters used included: carbon dioxide, distance from several tectonic plates, historical earthquakes' density, depth and magnitude, elevation and depth, global heat flow, groundwater resources, precipitation, sediment thickness, and surface air temperature.

We applied MaxEnt to four different sets of these parameters: i) all parameters; ii) parameters with the highest contribution to the MaxEnt model; iii) parameters selected by an expert; iv) a combined set of expert- and model-selected parameters. The results showed similar patterns but several crucial discrepancies. We found that the optimal model used all parameters and correctly predicted the high suitability of operative and planned geothermal plants from the Global Geothermal Energy Database. Our model was projected at the global scale to produce a global suitability map of suitable sites for geothermal plants and wells installation (suitable geothermal sites). This map is a prior knowledge layer for site exploration, which can thus save time and money in the preliminary stages of geothermal site assessment, especially when few data are available. The map, and the way we produced it as a set of open-access Web services, can also support communication with citizens whose territories are involved in geothermal probing and power plant installation. Indeed, often these people are not clearly informed about the scientific reasons driving the selection of their territory and the potential benefits.

2. METHODOLOGY

The objective of this study was to produce a global-scale geothermal suitability map for installing geothermal power plants using the MaxEnt model. To this aim, we selected a set of global-scale datasets representing environmental parameters after considering their correlation to geothermal energy studies already highlighted by previous works. These parameters are thus associable with the suitability of an area for a geothermal plant installation. In particular, the selected parameters were (see Section 3 for their characterisation): i) carbon dioxide (Copernicus Atmosphere Monitoring Service - CAMS); ii) distance from convergence lines (from United States Geological Survey – USGS); iii) distance from diffuse lines (USGS); iv) distance from ridges lines (USGS); v) distance from transform lines (USGS); vi) earthquake density (Centennial Earthquake Catalog - CEC); vii) earthquake depth (CEC); viii) earthquake magnitudes (CEC); ix) elevation/depth (United States National Geophysical Data Center - NGDC); x) global heat flow (Davies 2013); xi) groundwater resources (World-wide Hydrogeological Mapping Assessment - WHYMAP); xii) precipitation (NASA Earth Exchange Platform); xiii) sediment thickness (Laske, 1997); xiv) surface air temperature (NASA Earth Exchange Platform); xv) current and planned geothermal plants (IGA – Global Geothermal Energy Database).

The data had heterogeneous resolutions and sometimes non-homogeneous distributions, which required a preparation phase for spatially aligning them and feeding the MaxEnt model. Resampling, temporal aggregation, and spatial alignment to a 0.5° resolution were performed through GDAL. Earthquake data had non-homogeneous distribution and were transformed into spatially continuous distributions through inverse weighted interpolation. Earthquake density was calculated through QGIS-kernel density estimation. In order to configure these processes, the spatial correlation between the observations was calculated through the Data-Interpolating Variational Analysis (DIVA) (Barth et al., 2010), available as-a-service on the D4Science e-Infrastructure (Assante et al., 2018). All data are openly available for consultation and reuse[1].

MaxEnt is a shallow ML model commonly used to forecast species distributions in ecological niche modelling. MaxEnt is applicable to general problems where a probability density function $\pi\ (\bar{x})$ should be approximated, based on real-valued vectors $(\bar{x})$. As a natural application, $\bar{x}$ is a set of environmental parameters related to a species' presence (Phillips et al., 2006a; Pearson, 2012; Coro et al., 2018). In these cases, $\pi\ (\bar{x})$ is the distribution of the species in the environmental parameter vector space, which translates to a spatial distribution when $\pi\ (\bar{x})$ is projected on the environmental parameters associated with the locations of the area under study. In this study *geothermal site suitability* is the phenomenon to be modelled, while the geothermal power plants currently in operation (from IGA) were the samples for training the model. MaxEnt internally estimates the coefficients of a linear combination of the input parameters, which is embedded in the Entropy maximisation function. These coefficients reflect the influence of each input parameter in the prediction of the training set locations (*percent* contributions) and thus approximate a variable predictive weight. These weights can be used to filter out the parameters that carry the lowest amount of information for the model. Only verified data of geothermal power plants were used for training, in order to improve model reliability through data quality enhancement. The MaxEnt model is openly available as an Open Science-oriented Web service[2] in the D4Science e-infrastructure, which allows to repeat the presented experiment with the attached data.

3. ENVIRONMENTAL PARAMETER CHARACTERISATION

The collected environmental parameters can be divided into homogeneously distributed data and sparse-point data. Here below, we characterise, per category, each environmental parameter used and the justification for including it in our experiment:

Homogeneously distributed data:

- Natural emissions of CO2 can be essential to assess the presence of hidden geothermal systems. Thus, it is informative for geothermal suitability assessment. For this reason, a global scale uniform distribution of carbon dioxide (CO2) flux at the soil was retrieved from CAMS. This dataset had a monthly time resolution and a 1° spatial resolution. For our scopes, it was averaged from January 1979 to December 2013 and re-sampled at a 0.5° spatial resolution (Fig. 1-a).

- The NGDC global datasets of elevation and depth were down-sampled from 0.33° to 0.5° resolution (Fig. 1-b and -c). This parameter is usually used to estimate thermal gradient during geothermal power plant siting and was thus selected for our experiment.

- The global heat flow distribution (Davies, 2013) (Fig. 1-d) represents the underground thermal state mainly affected by deep geological processes (i.e., radioactive decay of elements, tectonic setting, conduction, etc.). It includes the correlation between heat flow and underground geology; therefore, it was selected for our experiment. The original dataset was re-sampled from 2° to 0.5° resolution.

- A global sediment thickness map (Laske, 1997) was re-sampled from 1° to 0.5° (Fig. 1-e) and was used in our model as an indicator of the possible presence of geothermal reservoirs.

- The average surface air temperature and precipitation distributions from the NASA Earth Exchange Platform (Fig. 1-f and -g) were included because they correlated with heat flow, storage and transportation, and aquifer recharge. The original daily data of 2019 (the latest available) were annually averaged and re-sampled from 0.25° to 0.5° resolution.

- The WHYMAP vectorial map of groundwater resources (Fig. 1-h) includes large sedimentary basins suited for installing a geothermal power plant, especially for medium-to-high temperature scenarios. Therefore, it was selected and imported into our experiment after rasterization at 0.5° resolution.

[1] Downloadable at https://data.d4science.org/workspace-explorer-app/?folderId=Qk5YY3JMY0w1ODRqZmxtNkVQNXB6S3ROdWtkU1pYN1UrcXVSaVkyUW0vdmgvYXBYdFJ6Z0JkYWladVU0UWptVQ

[2] Usable after free registration at: https://services.d4science.org/group/biodiversitylab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.MAX_ENT_NICHE_MODELLING

Sparse-point data:

- The "Centennial Earthquake Catalog" includes instrumentally recorded earthquakes between 1900 and 2008. Information on location and magnitudes indicate a higher possibility of fractured rock presence and hence underground formation permeability correlated to a potential geothermal reservoir. Therefore, these data were retrieved, cleaned, averaged over time, and uniformed after removing earthquake events with no magnitude or temporal indication. A uniform distribution was calculated for magnitude, depth and density through inverse weighted interpolation and DIVA (Fig. 1-i-j) at 0.5° resolution. An earthquake density distribution was calculated through QGIS-kernel density estimation (Fig. 1-k).

- Geothermal systems are correlated with tectonic plate movements. Therefore, Earth-cells' distances from the USGS-data tectonic lineaments (i.e., convergent, transform, diffuse, ridge) were computed at 0.5° resolution (Fig. 1-l-o).

- We used the spots of current operative and future-planned geothermal power plants To assess model performance. The Global Geothermal Energy Database contains 133 expert-verified data (including currently operational plants only) and 60 future-planned or unverified operational plants. We used the 133 verified operational plants as the model training set and the other 60 points as the test set (Fig. 1-p).

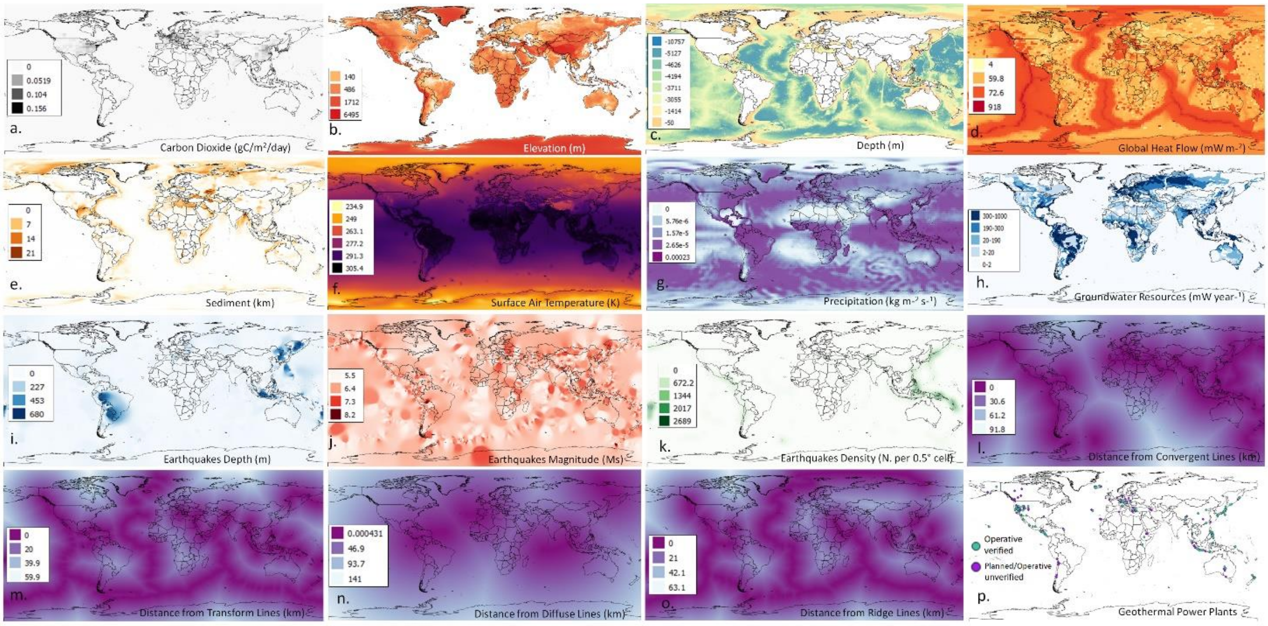

Figure 1: Maps of the environmental parameters used in our model: (a) carbon dioxide, (b) elevation, (c) depth, (d) global heat flow, (e) sediment thickness, (f) surface air temperature, (g) precipitation, (h) groundwater resources, (i) earthquake depth, (j) magnitude, and (k) density, distance from (i) convergent, (m) transform, and (n) diffuse, (o) ridge lines (p) operative and planned geothermal power plant.

3. RESULTS

To optimise model performance, MaxEnt was applied to four different sets of the selected environmental parameters: i) all parameters; ii) parameters carrying 95% of the total MaxEnt parameter percent contribution; iii) a sub-selection made by a geothermal energy expert based on his experience; iv) the union of the expert- and MaxEnt-selected parameters. MaxEnt was trained, for each parameter set, using active geothermal plants as the locations from which training vectors were extracted. Similar to applications in ecological niche modelling, its output was interpreted as a suitability score for geothermal plant installation. Values close to 1 indicated the highest suitability score. Based on the training set, the models and their decision thresholds were selected using the Area Under the Curve (AUC), i.e., the integral of the Receiver Operating Characteristic curve. AUC assesses a model's performance by simulating the balance between true positives and true negatives. We identified different possible decision thresholds based on AUC values to distinguish between suitable and unsuitable areas. These thresholds corresponded to i) the value maximising sensitivity (i.e., detected positives over real positives) (*sensitivity* threshold), ii) the value balancing omission rate (false negative rate) and sensitivity (*balanced* threshold), iii) the value that correctly predicted at least the 10-percentile of the training set (*10p* threshold). Table 1 reports the MaxEnt performance as the prediction accuracy (correct predictions over total) across the different parameter selections and thresholds. The related maps are reported in Fig. 2.

Table 1 - Performance of the four models built in our experiment. Accuracy is calculated for each decision threshold estimated by the Maximum Entropy models. The thresholds are reported in the rightmost columns.

Model	**AUC**	**Accuracy using sensitivity threshold**	**Accuracy using balanced threshold**	**Accuracy using 10p threshold**	**Sensitivity threshold**	**Balanced threshold**	**10p threshold**
All parameters	0.988	86.7%	76.2%	62.0%	0.02	0.136	0.319
MaxEnt-selected par.	0.980	92.4%	73.3%	67.6%	0.007	0.226	0.316
Expert-selected par.	0.985	90.5%	72.4%	67.6%	0.008	0.205	0.29
Expert and MaxEnt-selected par.	0.977	88.5%	72.4%	64.8%	0.024	0.242	0.308

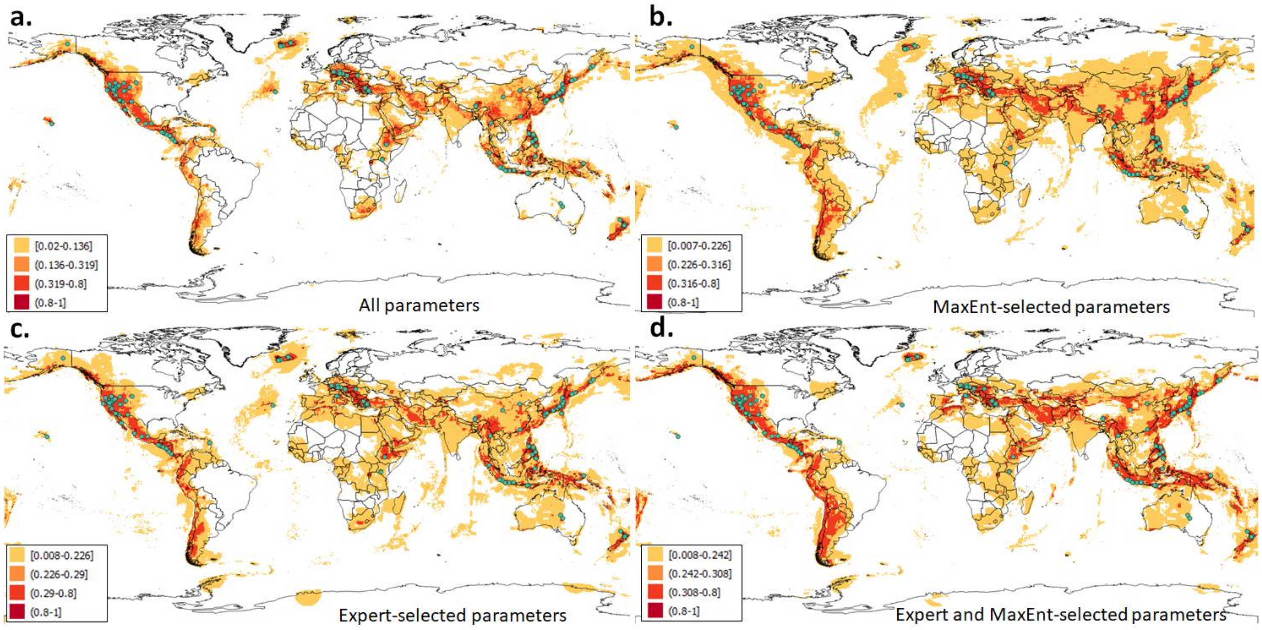

Figure 2: Global-scale geothermal suitability maps of four MaxEnt model using a) all environmental parameters, b) only the most important parameters as estimated by MaxEnt, c) the parameters selected by an expert, d) the intersection between expert- and MaxEnt-selected parameters.

Based on AUC, the optimal model was identified as the one using all parameters (0.988 AUC). For this model, the *sensitivity* threshold was the minimum to consider an area suitable. Figure 3 highlights the optimal model by assigning the minimal yellow categorisation to locations with a higher probability than the *sensitivity* threshold. The other ranges were used to determine medium-to-high locations. The MaxEnt model using MaxEnt-selected parameters reached the highest performance (92.4%), using the *sensitivity* threshold, but was less reliable on true-negative prediction given the lower AUC (0.980). Indeed, the visual comparison in Figure 2 shows that this model tends to overestimate the suitable locations. The other models did not reach the performance score of the one using all parameters. This indicates that all parameters carry essential information to predict suitable geothermal sites correctly.

The optimal model did not predict 8 geothermal power plants (out of 60) from the test set. These fall on locations with lower suitability scores than 0.02 (Fig. 3). However, they all represent geothermal plants with very low efficiency of production (e.g., in Canada) or co-production systems (e.g., in Florida) or with uncertain feasibility or still under evaluation (e.g., Canary Islands, Latvia, Australia).

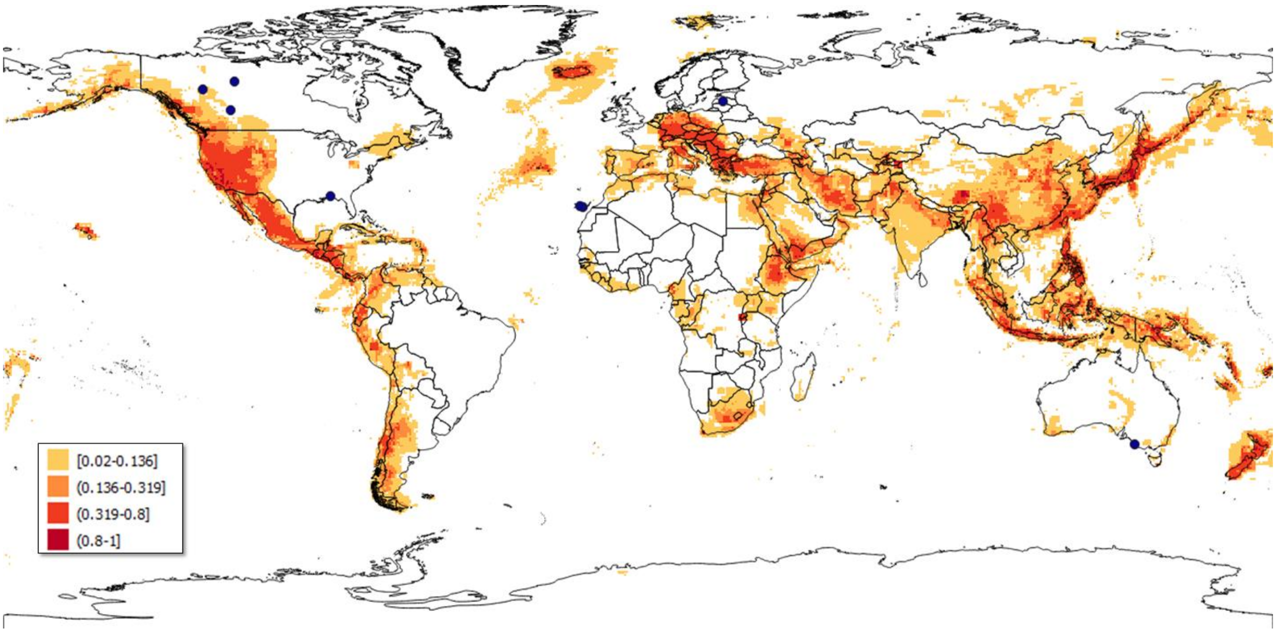

Figure 3: Optimal global geothermal suitability map produced by the MaxEnt model using all parameters. Warmer colours indicate higher suitability scores. Dots indicate geothermal power plants in the test set whose suitability was not predicted by the model.

4. CONCLUSIONS AND OUTLOOK

This study has presented the first global-scale map of the *prior* suitability of an area for geothermal power plant installation and geothermal energy production. A Machine Learning model was used to produce the area suitability distribution. Environmental parameters correlated with geothermal site suitability were pre-processed through inverse weighted interpolation, kernel density, DIVA, and Maximum Entropy to enhance result consistency and accuracy. The reliability of the map was tested against currently active and planned geothermal power plants and predicted 86.7% of the test data. In compliance with the Open Science paradigm directives, we made our model available as an open-access Web Service to allow experiment repetition and reproduction and model re-application to regional studies. Our future work will focus on increasing the map resolution by adding detailed regional-scale distributions. Moreover, we will test new parameters and combinations and compare the results with the indications of the global-scale map. Our work aims to lower mining risks, help policymakers build cost-effective energy management strategies, and communicate with citizens.

REFERENCES

Assante, M., Candela, L., Castelli, D., Cirillo, R., Coro, G., Frosini, L., Lelii, L., Mangiacrapa, F., Marioli, V., Pagano, P., *et al.* The Gcube System: Delivering Virtual Research Environments As-A-Service. Future Generation Computer Systems, NA (2018).

Barth, A., Alvera-Azcárate, A., Troupin, C., Ouberdous, M., Beckers, J.-M., A web interface for griding arbitrarily distributed in situ data based on data-interpolating variational analysis (diva). Advances in Geosciences 28 (2010), 29–37.

Coro, G., Vilas, L.G., Magliozzi, C., Ellenbroek, A., Scarponi, P., Pagano, P.: Forecasting the ongoing invasion of lagocephalus sceleratus in the mediterranean sea. Ecol. Model., 371 (2018), pp. 7-49

Coro, G. and Trumpy, E.: Predicting geographical suitability of geothermal power plants. Journal of Cleaner Production, 267, (2020), 121874.

Davis, J.H.: Global map of solid earth surface heat flow. Geochemistry, Geophysics, Geolsystems, 14 (10) (2013), pp. 4608-4622.

Laske, G.: A global digital map of sediment thickness. Eos Trans. AGU, 78(1997), p.F483.

Pearson, R.G., Species distribution modelling for conservation educators and practitioners. Synthesis. American Museum of Natural History. Available at: http://ncep.amnh.org (2012).

Phillipes, S.J., Andreson, R.P., Schapire, R.E.: Maximum entropy modelling of species geographic distributions. Ecol. Model., 190(3-4) (2006), pp. 231-259.