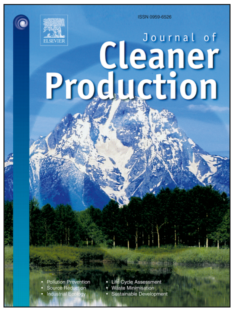
Journal Pre-proof



Predicting geographical suitability of geothermal power plants

Gianpaolo Coro, Eugenio Trumpy

PII: S0959-6526(20)31921-1

DOI: https://doi.org/10.1016/j.jclepro.2020.121874

Reference: JCLP 121874

To appear in: *Journal of Cleaner Production*

Received Date: 26 June 2019

Revised Date: 20 April 2020

Accepted Date: 22 April 2020

Please cite this article as: Coro G, Trumpy E, Predicting geographical suitability of geothermal power plants, *Journal of Cleaner Production* (2020), doi: https://doi.org/10.1016/j.jclepro.2020.121874.

Predicting Geographical Suitability of Geothermal Power Plants

Gianpaolo Coro[a], Eugenio Trumpy[b]

[a]Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo" – CNR, Pisa, Italy
[b]Istituto di Geoscienze e Georisorse – CNR, Pisa, Italy

Abstract

A large and increasing number of countries use geothermal energy as power source for domestic and industrial applications. Geothermal power plants produce energy out of this natural and renewable source in a sustainable way and contribute to reduce global warming. However, power plants effectiveness depends on the suitability of an area to geothermal energy production, which is a complex and unknown combination of many environmental factors. Nowadays, geothermal suitability assessments require invasive inspections, high costs, and legal permissions. Thus, having a global suitability map of geothermal sites as reference would be useful prior knowledge during assessments, and would help saving time and money. In this paper, the first suitability map of potential geothermal sites at global scale is presented. The map is the result of the application of data collection and preparation processes, and a Maximum Entropy model, to geospatial data potentially correlated with geothermal site suitability and geothermal plants operation. The reliability of our map is assessed against currently active and planned geothermal power plants. Our approach follows the Open Science paradigm that guarantees results reproduction and transparency, and allows stakeholders to reuse the produced standardised data, services, and Web interfaces in other experiments or to generate new maps at regional scale. Overall, our results can help scientists, industry operators, and policy makers in geothermal sites assessments. Also, our approach supports communication with citizens whose territories are involved in probing and assessments, in order to transparently inform them about the reasons driving the selection of their territory and the potential future benefits.

Keywords: Geothermal Energy, Artificial Intelligence, Renewable Energy, Environment, Machine Learning, Spatial Probability Distributions, Open Science

*Corresponding author
Email address: gianpaolo.coro@isti.cnr.it (Gianpaolo Coro)

1. Introduction

Geothermal energy is a natural, renewable, sustainable, and constant source of power that comes from the thermal energy stored in the Earth (Boyle, 2004). Geothermal power plants usually have low greenhouse gases emission and contribute to global warming reduction while supplying energy in a scalable way from rural villages to entire cities. Nowadays, 26 countries produce electricity from Earth's heat, for a total amount of 73.7 TWh/yr, and 83 countries use underground heat directly for many applications (e.g. heating, cooling, industrial processes, etc.) for a total of ~163 TWh/yr of thermal energy (Trumpy et al., 2015a; Bertani, 2016; Lund and Boyd, 2016; Breeze, 2019; IGA, 2020). The accessible electrical potential ranges between 35 and 200 GW, which however is 16 times the actual generation capacity (Roberts, 1978; Bertani, 2003; Stefansson, 2005; Tester et al., 2012; Manzella, 2017).

Unlocking the unused geothermal resource can help World nations (i) to be less dependent on energy imports, (ii) to foster the use of mixed energy in power plants, and (iii) to reduce greenhouse gases emission. Fostering geothermal energy usage and production through supply policies and geothermal projects requires to clearly identify and rank resources and opportunities. Usual sites exploration operations usually require time, invasive inspections with drilling probes, high costs, and permissions from legislative authorities (Barbier, 2002). In this context, having a global suitability map of suitable sites for geothermal plants installation (*suitable geothermal sites*) would be useful *a priori* knowledge to save time and money. At the same time, it would support communication with citizens whose territories are involved by geothermal probing and power plants installation, who are not usually transparently informed about the scientific reasons driving the selection of their territory and the potential benefits. A global suitability map would also support the understanding of the correlation between potential geothermal resources utilisation and the sustainability of geothermal life cycles within Life-Cycle Assessment (LCA) methodology (Tomasini-Montenegro et al., 2017; Parisi et al., 2019; Paulillo et al., 2019; Karlsdottir et al., 2020; Dumas et al., 2020).

Several examples exist of approaches that combine knowledge from different sources to build open-access overviews of geothermal resources at city or country scales (Coolbaugh et al., 2002; Noorollahi et al., 2007; Moghaddam et al., 2014; Limberger et al., 2014; Trumpy et al., 2015b). In the last decades, suitability maps have been produced for several types of renewable energy resources at regional scale, by mapping environmental resources onto operating plants through GIS tools (Seibt et al., 2005; Yousefi et al., 2007; Ramachandra and Shruthi, 2007; Angelis-Dimakis et al., 2011; Mourmouris and Potolias, 2013; Satkin et al., 2014; Bertermann et al., 2015; Sah and Wijayatunga, 2017). Also, many analytical approaches have assessed the suitability, sustainability, and continuous supply of renewable energy in certain countries or regions, using data envelop analysis and decision trees (e.g. multi-criteria analysis) (Atmaca and Basar, 2012; Troldborg et al., 2014; Martín-Gamboa et al., 2015; Ramazankhani et al., 2016; Siefi et al., 2017; Macharia et al., 2018; Karimi et al., 2019). Although being effective for their specific cases, these studies often take into account socio-economic factors - e.g. population density, industries, availability of skilled labour, etc. - together with environmental aspects that are very tied to the specific analysed region and are not always available for other regions. Thus, these methodologies are hardly reusable for other regional- or global-scale assessments. Further, they use models that tend to factorize parameters (by applying decision on one or few parameters at a time) or use linear combinations of parameters. Instead, the task of predicting suitable geothermal sites would demand a more complex combination of parameters within scalable decision systems. Fewer studies have used machine learning approaches (e.g. Artificial Neural Networks, Support Vector Machines, Random Forests, etc.) to model the correlation between specific parameters (e.g. natural emission baseline of CO_2) and the suitability of an area to geothermal energy exploitation (Santoyo et al., 2018; Shahab and Singh, 2019). However, also these studies focus on a few parameters that are available for the particular analysed region.

In this paper, the first suitability map of potential geothermal sites at global scale - with 0.5° resolution - is presented, produced through an Open Science approach that combines data collection and preparation tools with a Maximum Entropy machine learning model. The data used to train the models are geospatial data containing environmental information potentially correlated with geothermal site suitability and geothermal plants operation. Public data of both currently active and planned geothermal power plants are used to assess the performance and the reliability of the model.

The main potential stakeholders of this research are geologists, geothermal energy industry operators, territory citizens, and policy makers, who search for prior assessment of possible future-planned geothermal power plants. Our research also offers standardised data, services, and Web interfaces that allow these stakeholders to enhance our maps by producing higher resolution and regional-scale distributions. For these reasons, our approach follows the Open Science paradigm (Hey et al., 2009), because it aims at making all the research products re-usable, reproducible, and repeatable. All data are represented under standards of the Open Geospatial Consortium (OGC) in order to quickly combine them in computations. All processes and models are described through standards as well, in order to make them re-usable in other work flows and experiments.

In this paper, the used data, processing tools, and modelling techniques are first described (Section 2). Then, the maps produced by the models are reported and an estimate of the prediction performance of future planned geothermal sites is given (Section 3). Finally, the reliability of the optimal produced model is discussed with examples (Section 4) and final conclusions are drawn (Section 5).

2. Work Methodology

2.1. Data Collection and Preparation

The environmental parameters used for modelling were selected among a number of global scale datasets that have been correlated to geothermal energy studies by other works. Overall, these parameters can help assessing the *a priori* suitability of an area to a geothermal plant installation. The complete list of selected parameters is reported in Table 1. In the following sections, the used preparation tools, the primary data sources, and the data selection motivations are reported for all parameters.

2.1.1. Data Preparation Tools

The data used in our experiment came from heterogeneous sources and thus required preparation to be represented at the same resolution and used in our models. GDAL (OSGeo, 2019) was used to pre-process and re-sample raster files when needed. Point distributions were made spatially continuous (uniform distributions) through inverse weighted interpolation (IDW). Parameters values density was calculated through kernel density with QGIS (QGis, 2011). These algorithms require indicating the spatial extent of the correlation between the punctual data. To this aim, Data-Interpolating Variational Analysis (DIVA) was used (Barth et al., 2010), available as-a-Service through the D4Science e-Infrastructure (Coro et al., 2016; Assante et al., 2018). DIVA is largely used in oceanography to produce uniform distributions from punctual data and, as a first step, it estimates the spatial correlation between the data (Troupin et al., 2010; Schaap and Lowry, 2010; Troupin et al., 2012; Coro et al., 2018b). This correlation was calculated for all our punctual data and was used as input to IDW and kernel density.

DIVA and the other models used in our experiment require data to be represented under OGC standards for access and consumption (e.g. Web Coverage Service, OPeNDAP, Web Feature Service, etc.). To this aim, D4Science services were used to transform point and vector data into GIS layers, and raster data into NetCDF-CF files (Section 2.3). A 0.5° resolution was selected as common resolution for all data, because most data were available at a similar or higher resolution and 0.5° is fairly detailed for a global scale model.

2.1.2. Pre-existing Uniform Distributions

A global scale uniform distribution of carbon dioxide (CO_2) flux at soil is available on the Copernicus Atmosphere Monitoring Service with monthly estimates and a 1° spatial resolution (CAMS, 2019). This distribution is the result of a net flux inversion reanalysis of carbon dioxide between land, oceans, and atmosphere calculated through the PYVAR model. The CO_2 flux is one of the key drivers of the evolution of Earth's climate and has been largely involved in geothermal power plants studies, especially because of CO_2 emission, sequestration, and reuse in the plants (Pruess, 2006; Randolph and Saar, 2011b,a). Moreover, natural emission baseline of CO_2 has been already used to assess the presence of hidden geothermal systems and thus is a promising source of information for geothermal suitability assessment (Lewicki and Oldenburg, 2005; Santoyo et al., 2018). Several of these studies have been performed at local or regional scale and in proximity to volcanoes or grasslands, but no global assessment is available (Nykanen et al., 1995; Chiodini et al., 1999; Rodrigo-Naharro et al., 2013; Zhao et al., 2017; Aiuppa et al., 2019). For the scopes of our experiments, the monthly CO_2 raster data of Copernicus were averaged in time from January 1979 to December 2013 and re-projected at 0.5° spatial resolution (Figure 1-a). This long-term average is meant to combine, without prior weighting, CO_2 values preceding the higher industrialisation rate of the last years with natural presence of CO_2 in the soil, following a common approach used for other environmental parameters (NASA-NEX, 2015).

A global dataset of elevation and depth at 0.33° resolution (ETOPO2) is downloadable from the United Stated National Geophysical Data Center (NGDC) Web site (NOAA, 2001). The version used in our experiment (v2) includes localised corrections and integration of satellite, ocean sounding, and land data. For our scopes, this dataset was down-sampled to 0.5° resolution (Figures 1-b and -c). Elevation/depth information has been often involved in geothermal power plants projects, for example in the evaluation of connectivity and energy provisioning costs and to calculate thermal gradient (Prest et al., 2007; Limberger et al., 2018).

Based on more than 38,000 heat flow values, Davies (2013) built a 2° resolution global heat flow distribution map that represents the underground thermal state mainly affected by deep geological processes (i.e. radioactive decay of elements, tectonic setting, conduction etc.). This distribution (Figure 1-d) is reliable enough to be used in our model - after up-sampling to 0.5° resolution - in order to take into account the correlation between heat flow, geology, and geothermal power plants installations reported by other studies (DiPippo, 2012; Lu, 2018; Limberger et al., 2018).

A global sediment thickness map at 1° spatial resolution is available from Laske (1997). This map was obtained by combining high-resolution oceanic and tectonic maps with manually digitalised information. The map was conceived to model long-period global seismic data as dependent on crustal structure, which is related to the sediments layer. A 0.5° up-sampled version of this map (Figure 1-e) was used in our model to account for energy transportation issues in geothermal power plants. Further, sediments have been used by other studies as an indicator of possible presence of geothermal resources (Sharp Jr and Domenico, 1976; Sharp, 1978; Schellschmidt et al., 2010; Cacace et al., 2010; Nielsen et al., 2017).

Long-term daily forecasts from 1950 to 2100 are available at 0.25° resolution for minimum and maximum surface air temperature (SAT) and precipitation at surface on the NASA Earth Exchange platform (NASA-NEX, 2020). Forecasts are provided for 20 weather models based on the Coupled Model Intercomparison Project Phase 5 (CMIP5, 2019). In order to obtain representative information for SAT and precipitation, averaged data sets in time and space were produced at 0.5° resolution as in Coro et al. (2018a). The resulting average surface air temperature and precipitation distributions (Figures 1-f and -g) were involved in our model because of their correlation with heat flow, storage and transportation, and aquifer recharge (García-Gil et al., 2015; Bidarmaghz et al., 2016; Alhamid et al., 2016; Griebler et al., 2016; Nguyen et al., 2017).

The World-wide Hydrogeological Mapping and Assessment Programme (WHYMAP) publishes a vector map of groundwater and recharge that represents large sedimentary basins suited for groundwater exploitation (Richts et al., 2011). This map was rasterised to 0.5° resolution (Figure 1-h) and was involved in our model to take into account the use of groundwater by most geothermal power plants especially in the medium to high temperature scenarios (Armannsson and Kristmannsdottir, 1992; Barbier, 2002).

2.1.3. Distributions Estimated from Point Data

The "Centennial Earthquake Catalog" reports the locations and magnitudes of instrumentally recorded earthquakes between 1900 and 2008 (Engdahl et al., 1998; Engdahl, 2002). The Catalog reports large earthquakes generally over 5.5 *surface-wave magnitude scale* (Ms) in textual format, manually curated to correct and uniform magnitudes and locations. For our scopes, this catalogue was cleaned up of data in non-Ms units and averaged in time. Subsequently, a uniform 0.5° map was produced through DIVA and IDW. The result was a distribution map of earthquake magnitudes at global scale aiming at correlating geothermal power plants feasibility to pre-existing earthquakes' frequency and intensity, which may indicate higher permeability of rocks correlated to a possible geothermal reservoir (DiPippo, 2012; Kissling and Weir, 2005; Juncu et al., 2015; Clark et al., 2015). Overall, based on this information three 0.5° distributions were produced to represent earthquakes' depth (Figure 1-i), magnitude (Figure 1-j), and density (Figure 1-k).

Geothermal energy is overall correlated with the constant movement of Earth's crust plates (Barbier, 2002). When tectonic plates move against each other, there is increasing volcanic and earthquake activity. The areas where the plates collide, identify lines where different scenarios can occur, e.g. subduction (*con-*

vergent lines), mutual sliding in opposite direction (*transform* lines) or in the same relative motion (*diffuse* lines), and ridges formation (*ridge* lines). NASA and the United States Geological Survey publish vector files for these types of lines (Rybach and Muffler, 1981; Barbier, 2002; Glassley, 2018). In order to correlate geothermal sites to this information, the distances of 0.5° cell areas from the lines were calculated through a Java routine (in supplementary material) and four distributions were produced (Figures 1-l, -m,-n, -o).

In order to verify the performance of our model, locations of currently operating and future planned geothermal power plants were necessary. The Global Geothermal Energy Database of the International Geothermal Association (IGA) hosts these data (Trumpy et al., 2015a; IGA, 2020). This Web platform provides access to geothermal production data and promotes the development of geothermal energy through environmental and productivity information. A total of 133 expert-verified data of currently operating plants and 60 future planned and unverified operating plants were retrieved and used, in order to respectively produce train and test sets for our models (Figure 1-p and Section 3.2).

2.2. Modelling

2.2.1. Overview

Maximum Entropy (MaxEnt) is a machine learning model commonly used in ecological modelling, for example to forecast species distributions (Phillips et al., 2004, 2006a; Phillips and Dudik, 2008; Baldwin, 2009; Coro et al., 2015b, 2018c). MaxEnt is applicable to general problems where a probability density function $\pi(\bar{x})$ should be approximated, based on real-valued vectors. For example, the $\bar{x}$ vectors may refer to the environmental parameters correlated to a species presence or a phenomenon occurrence (Pearson, 2012; Coro et al., 2018c). The advantage of this model is that it can work with just positive examples (*presence-only* model). In this case, the vectors refer only to phenomenon occurrence locations, whereas non-occurrence is automatically estimated. Considering geothermal site suitability as the phenomenon to model, the geothermal power plants currently in operation were treated as positive examples to train the model. The model's input vectors were the environmental parameters associated with the areas where the geothermal plants operate. The MaxEnt implementation available as-a-Service on the D4Science e-Infrastructure (CNR, 2019; Phillips et al., 2019) was used to train the model, based on the environmental features reported in the previous section.

The training algorithm changes the model's parameters with the constraint that the density function is compliant with predefined mean values at the training-set locations and the entropy of the density function $H = -\sum \pi(\bar{x})\, ln(\pi(\bar{x}))$ is maximum on these locations (Elith et al., 2011). Verified geothermal power plants locations were used as training locations. MaxEnt performs a relative maximisation of the entropy function on these locations, with respect to the entropy values of the parameters of random points taken all over the world (*background points*, Phillips et al. (2006a)). The model uses a linear combination of the parameters, where the coefficients of the combination are changed to reflect the influence of each variable in predicting the training set locations (*percent contribution*). The training algorithm also records the dependency of the model's performance on the permutation of a variable values among the training vectors (*permutation importance*). After the training phase, percent contribution can be used to select the most influential environmental parameters, i.e. MaxEnt can be used as a filter to select the parameters that carry the highest quantity of information. For example, parameters with a contribution value higher than 5% of the maximum contribution value can be retained to possibly improve model's performance (Coro et al., 2015b), or generally to obtain a new projection based only on the most influential parameters (Coro et al., 2013, 2018c). A MaxEnt model trained on 0.5° resolution parameters can be used to produce probability values for the environmental parameters of 0.5° squared areas. Given the information represented in our training set, probability could be interpreted as suitability score for geothermal power plants.

One drawback of MaxEnt, is that it is very sensible to bias in the data, thus its performance increases if the training locations are reliable (Elith and Leathwick, 2009). Thus, in order to increase our model's reliability, only verified data of geothermal power plants were used for training. Our choice of MaxEnt was due to its comparable performance with respect to alternative presence-only models and to its availability as an Open Science-oriented service in D4Science, which allowed re-using it directly to our standardised data.

2.2.2. *Model Description*

In mathematical terms, MaxEnt uses the environmental vectors $\bar{x}$ of known geothermal power plant locations and of *background points* to estimate the probability $P(Suitability|\bar{x})$ that a site is suitable for a geothermal plant installation, conditioned on the environment. To this aim, the model estimates the ratio between the probability density $f(\bar{x})$ of the vectors across the area of interest and the probability density in the suitable locations $f_1(\bar{x})$. In fact, $P(Suitability|\bar{x})$ is related to $f(\bar{x})$ and $f_1(\bar{x})$ through the Bayes' rule:

$$P(Suitability|\bar{x}) = \frac{f_1(\bar{x})P(Suitability)}{f(\bar{x})}$$

where $P(Suitability)$ is the prior knowledge about the proportion of suitable sites in the analysed area (*prevalence*). The MaxEnt hypothesis is that the optimal $f_1(\bar{x})$ distribution is the closest distribution to $f(\bar{x})$. This is justified by the fact that $f(\bar{x})$ is a null model for $f_1(\bar{x})$, because without any training-set location there would be no expectation of preferred environmental conditions over the others. MaxEnt uses the Kullback-Leibler divergence (relative entropy) to measure the distance between the two functions:

$$d(f_1(\bar{x}), f(\bar{x})) = \sum_{\bar{x}} f_1(\bar{x}) log_2 \left(\frac{f_1(\bar{x})}{f(\bar{x})} \right)$$

and aims at minimising this distance. The model also imposes the constraint that $f_1(\bar{x})$ should reflect the observations at the training-set locations, i.e. $f_1(\bar{x})$ should report high-probability on parameters' values close to the parameters' means on the training set. It can be demonstrated that this characterization uniquely determines $f_1(\bar{x})$ as belonging to the following Gibbs distribution family (Phillips et al., 2006b):

$$f_1(\bar{x}) = f(\bar{x}) e^{\eta(\bar{x})}$$

where $\eta(\bar{x}) = \alpha + \beta\ h(\bar{x})$; α is a normalization constant that ensures that $f_1(\bar{x})$ sums to 1; h is an optional transformation of the vectors $\bar{x}$ that allows modelling complex relationships between the variables; β is a vector of coefficients to be estimated, which eventually allow calculating the *percent contribution* of each parameter. Thus, the target ratio of MaxEnt $f_1(\bar{x})/f(\bar{x})$ corresponds to $e^{\eta(\bar{x})}$. This means that MaxEnt needs to solve a log-linear model on the background and training vectors in order to estimate the α and β parameters, which is usually implemented through a penalised maximum likelihood algorithm (Phillips and Dudík, 2008).

2.3. Enabling Open Science

Making our methodology compliant with Open Science (OS) ensures longevity of data and processes through the compliance with the three "R"s of the scientific method: Reproducibility, Repeatability, and Re-usability (Hey et al., 2009). To this aim, we used the D4Science e-Infrastructure, a network of hardware and software resources that offers OS-compliant services for data publication and processing (Candela et al., 2016; Assante et al., 2018). D4Science allows re-using processes from several domains through Web interfaces and allows making them available to groups of scientists through Virtual Research Environments. Further, it offers a high-availability distributed storage system to host and publish data and a cloud computing system to process large amounts of data. In particular, D4Science allowed (i) importing raw data from the sources reported in the previous section, (ii) applying data preparation processes (e.g. DIVA), and (iii) publishing data under OGC representational standards through Web services and user interfaces. In particular, D4Science allowed transforming all environmental data into NetCDF-CF raster files and making them available through a Unidata Thredds service instance (John Caron and Davis, 2006), which enables access through OGC protocols and standards (e.g. OPeNDAP, Web Coverage Service, Web Feature Service, and Web Map Service). Point data of geothermal power plants were published as point maps through a GeoServer instance hosted by D4Science (Deoliveira, 2008). Standards are key to implement OS approaches and to reuse data in the processes. Indeed, having data represented through OGC standards allowed reusing the MaxEnt model provided by a biodiversity-oriented Virtual Research Environment that could work on OGC-compliant data. Further, each model training was executed on an OS-oriented cloud computing platform (DataMiner, Coro et al. (2017, 2015a)) that uses a network of 30 multi-core machines (Ubuntu 16 x86 64 with 16 virtual CPUs, 32 GB of random access memory, 100 GB of disk) and produces OGC-compliant maps out of the models results. DataMiner also records the complete set of experimental parameters (computational provenance) and publishes processes through the OGC Web Processing Service standard, which standardises input, output, and metadata and is compatible with several GIS tools (e.g. QGIS and ArcGIS). Overall, the OGC data and the processes reported in this paper can be openly revised, re-used in other experiments, and reproduced. This makes our approach compliant with Open Science.

3. Results

3.1. Parameters Selection

Four sets of environmental parameters were prepared for modelling. The first set contained all parameters (Table 2-a). The second set contained parameters selected using MaxEnt as a filter, i.e. by retaining only those with percent contribution higher than 5% of the maximum contribution (Table 2-b). This set helps evaluating the effect of excluding the less important contributors on the model's performance. The third set was a subset of parameters selected by an expert of geothermal energy based on his experience (Table 2-c). Notably, the expert excluded CO_2 because he considered only natural emission to be correlated with geothermal site suitability, whereas it is impossible to separate natural from artificial CO_2 emission in our dataset. The fourth set was the intersection between the MaxEnt-selected and the expert-selected parameters (Table 2-d). This last set aims at simulating an agreement on variables selection between MaxEnt and the expert.

3.2. Model Training

The cloud process reported in Section 2.2 was used to train four MaxEnt models focussing on the four groups of parameters reported in the previous section. The estimated MaxEnt π function was interpreted as a suitability score for geothermal power plant installation. The training set contained only verified power plants locations. Using this set, models' performance and optimal decision thresholds were estimated. In particular, the average Area Under the Curve (AUC) is the integral of the Receiver Operating Characteristic (ROC) curve that plots *sensitivity* ($\frac{True\ Positives}{True\ Positives+False\ Negatives}$) against 1-*specificity* ($1-\frac{True\ Negatives}{True\ Negative+False\ Positives}$). AUC values closer to 1 indicate better binary classification of sites as either suitable or unsuitable. Reference cut-off thresholds on the π function were calculated during the training phase (Phillips et al., 2019), which represent (i) the value maximising sensitivity (*optimal threshold*), (ii) the value balancing omission rate ($\frac{False\ Negatives}{True\ Positives+False\ Negatives}$) and sensitivity (*balanced threshold*), (iii) the value below which 10% of the training set classification falls (*10p threshold*). Using the test set of planned and operative sites reported in Section 2.1, accuracy was calculated for each threshold as $\frac{n.\ of\ geothermal\ sites\ correctly\ classified}{overall\ n.\ of\ geothermal\ sites}$.

3.3. Performance

Table 2 reports the *percent contribution* and the *permutation importance* of all environmental parameters in the four trained models. When a parameter has a high contribution and permutation importance the model strongly depends on it. In this view, carbon dioxide is a very important parameter although the expert excluded it from his choice. This high contribution is probably due to the inclusion of natural emission of CO_2 in our dataset, which unfortunately cannot be separated from artificial emission. A precise explanation would require an expert analysis focussing on this parameter only, in regions where these data are available. In the next section, a methodological approach to focus our model on this parameter at a regional scale is suggested.

Earthquake density had the highest permutation importance in all models, which indicates that this is another crucial parameter. Large amount of information is also carried by elevation/depth and global heat flow. Sediments have good importance and are also involved in the expert's choice but have less importance than the previously mentioned parameters. The permutation importance of surface air temperature is very similar to the sediments' one, but increases when information on carbon dioxide and precipitation is missing. This usually occurs in MaxEnt when parameters are correlated. Distances from tectonic lines have generally low percent contribution. Nevertheless, distance from convergent lines has high permutation importance, which means that this parameter carries valuable information.

Table 3 reports the effect of the different parameters selections on the models' performance. The related maps are reported in Figure 2. Generally, AUC is the most important measure because it accounts also for true negatives recognition (simulated through background points). According to AUC, the best model is the one using all variables (0.988 AUC). Instead, accuracy depends on the threshold used on the MaxEnt output distribution to distinguish between suitable and unsuitable areas. Accuracy is reported in Table 3 for each threshold. Coloured clusters in the images correspond to these thresholds, which change depending on the model. According to the *optimal* threshold, the model using the MaxEnt-selected parameters has the highest performance (92.4%). Using the *balanced* threshold, the model using all-parameters is the optimal one (76.2%), whereas using the *10p* threshold the models with MaxEnt-selected and expert-selected parameters have the highest accuracy. The model using the intersection between the expert- and MaxEnt-selected parameters has always the lowest performance. This indicates that the most important parameters do not carry alone all the information required to correctly predict suitable geothermal sites.

4. Discussion

Section 3.3 indicates that the parameters most correlated to suitable geothermal sites are carbon dioxide, earthquake density, elevation/depth, global heat flow, sediment thickness, and surface air temperature. Indeed, suitability is correlated with the complex combination of these parameters rather than with specific ranges of each parameter. Thus it not easy to identify high suitability ranges for each parameter separately. The MaxEnt response curves (in supplementary material) indicate that carbon dioxide is the only parameter where a high suitability range (between 0.05 and 0.156 $gCm^{-2}day^{-1}$) can be identified. Overall, models' performance indicates that all parameters carry useful information, although some of them are more correlated to geothermal site suitability. This agrees with other studies that have highlighted the correlation between each single parameter and geothermal energy exploitation (Section 2.1).

The maps in Figure 2 present similar patterns but several important discrepancies. Although the suitability score ranges in the maps are different, the coloured clusters have correspondent meaning: the yellow clusters can be interpreted as "low suitability" in all maps, the red clusters indicate "high suitability", and intermediate clusters indicate "medium suitability". Investigating the details of these maps requires inspecting the data (available in supplementary material), however the figures allow appreciating the general trends of the models, which were also numerically confirmed using a D4Science maps comparison tool (Coro et al., 2014). In particular, the models using MaxEnt-selected and expert-selected parameters tend to overestimate suitable locations with respect to the other models. The model using the intersection between expert- and MaxEnt-selected parameters tends to underestimate suitable locations, and thus has lower AUC and accuracy. The model using all parameters presents valid suitable locations and indicates realistic unsuitability locations. For example, unlike the other models it indicates general unsuitability in the South Italy peninsula, which is confirmed by other studies (Barbier, 2002).

Assuming the all-parameters model to be the optimal model, 8 geothermal plants from the test set have suitability lower than 0.02 and are thus "discarded" (Figure 3). Indeed, these points have different combinations of environmental characteristics with respect to close locations hosting operational power plants, and represent peculiar plants mostly having very low production or co-production system. In particular, compared with other correctly classified geothermal plants in the Gulf of Mexico, the discarded geothermal plant in Florida is characterised by higher sediment presence, lower heat flow and elevation, is closer to diffuse lines and more distant from transform lines, has more higher groundwater resources and higher surface air temperature, and much lower earthquake magnitudes and density. This combination makes this area less suitable to geothermal plants installation. Indeed, this is a planned power plant that is going to use also oil and gas wells and thus is not purely geothermal.

With respect to the other Australian geothermal power plants, the discarded plant in South Australia is more distant from diffuse lines, has more groundwater resources and precipitation, and lower surface air temperature and earthquake density. Indeed, the feasibility of this power plant is still under evaluation by the Australian government.

Compared with other suitable locations containing operative power plants in Canada, the three discarded Canadian power plants present lower elevation and CO_2 presence, lower earthquake magnitudes, much lower heat flow, higher groundwater resources level, and longer distance from transform and ridge lines. Indeed, these three plants have a very low energy production (about 0.2/0.3 MW), which is consistent with a low suitability indication by our model.

The areas of the two discarded geothermal plants in Canary islands differ from other suitable island locations (e.g. in the Caribbean Sea) especially for the lower earthquake depth, magnitude, and density, and for the lower precipitation quantity and surface air temperature. Finally, with respect to other power plants in central Europe, the discarded power plant in Latvia (with just 3MW planned production) is located in an area with lower heat flow and CO_2, longer distance from convergent lines, and higher earthquake magnitude. Indeed, locations in Canary islands and Latvia are planned power plants whose feasibility is still under evaluation, thus comparison with our map is not possible.

On the other hand, there are 52 planned geothermal power plants in areas that our model indicates as suitable. Among these, there are many examples of geothermal plants whose suitability was certified by expert-based assessment and whose locations fall in high probability ($> 60\%$) zones of our map. For example, the Imperial Valley in California is a region with a 2,950MW geothermal potential, where 403MW thermal energy is already generated by existing plants (Di Pippo and Lippmann, 2017). Our dataset of planned geothermal plants includes two ongoing projects in this region and our model indicates an 84% suitability score in the whole area, including the already operational plant sites. In western Iceland, our model agrees with four planned geothermal plants with a 73% score. These plants are very close to the Hellisheidi's 303MW geothermal power plant and are likely high-power plants (HGPS, 2020). The site of Latera in Italy already hosted an operational geothermal plant in the '90s and is identified with a 65% score. A new 14MW geothermal plant is planned for the next future in this site (Enel Green Power, 2020). The plant at Neuried (Germany) is a biomass/geothermal energy hybrid plant with a planned production of 4.2MW, which is under construction after a long assessment phase (Bertani, 2012). Our model identifies this site with a 70% suitability score. In Greece, the Lesvos island has been assessed as a promising geothermal area with medium enthalpy that could produce 8MW of thermal energy (Papachristou et al., 2016). Our model assigns a 64% suitability score to this site. Although these independent assessments used costly procedures, our model reliably agrees with them. This indicates that the model could have complemented expert opinion, or even reduced expensive procedures if used before the assessments.

Making both our model and data available as an Open-Science oriented tool, allows for quickly re-using them for regional or country-specific assessments. In fact, specific data available for a certain region can be used directly in our model, after standardization. These data can be combined with (or substituted to) the global-scale data presented in this paper. Also, our model can be executed with just one parameter available at a regional scale (e.g. natural emission baseline of CO_2) and can focus on a certain zone where this parameter has demonstrated high correlation with potential geothermal resources (Lewicki and Oldenburg, 2005). For example, the model could highlight the correlation between natural emission of CO_2 and geothermal suitability in a relatively small area where the baseline has been measured, in order to identify other potential geothermal sites. This projection could also verify the correspondence between the estimated geothermal suitability and anomaly in natural emission baseline of CO_2, as highlighted by other studies (Lewicki et al., 2008; Santoyo et al., 2018). Finally, our model can use - after standardization - the reconstructed information produced by other machine learning approaches (e.g. baseline CO_2, Santoyo et al. (2018)), which is useful especially when data are missing.

At a regional scale, there could be few reference geothermal plants to be used as a training set. In this case, also promissory geothermal zones already assessed through exploratory analysis should be included in the training set. Following a common practice in machine learning, a cross-validation strategy should be applied, which samples several times the training set and uses 80% of the data to train the model and 20% to test the model. Finally, either the average model or the model with the highest performance should be retained as the final model (Coro et al., 2018c). Regional maps produced in this way, can help decision-makers to support commercial projects, to revise exploratory analyses, and to obtain suitability assessments that overcome possible political or environmental hindrances.

5. Conclusions

In this paper, the first map of global suitability of an area to a geothermal power plant installation has been presented. The optimal MaxEnt model producing this map uses all proposed parameters (Table 1). The model's reliability has been demonstrated through real examples of known geothermal plant locations. Our results suggest that the produced map can be used as prior knowledge about the geothermal suitability of an area before standard exploration strategies, e.g. during the search for suitable drilling locations. The adopted standardization of data and processes has allowed to practically and quickly reuse models and processes from other domains (oceanography and ecological modelling). Further, the publication of results, data, maps, and models through an Open Science platform allows also citizens to be informed about the geothermal energy potential of their areas and increases the transparency of decisions and investments towards the large public.

Our future work will focus on practically applying the presented approach to help geothermal companies saving investment money and time, lower mining risks, and aid policy makers in energy management strategies. One specific region will be first selected together with experts, and our approach will be applied to evaluate the approximation made by the global scale map with respect to an expert-reviewed regional model. Further, our model will be used to (i) increase the map resolution at regional scale, (ii) test different parameter combinations and data available for that region (e.g. natural emission baseline of CO_2), (iii) investigate the correlation between specific regional parameter ranges and geothermal site suitability.

Disclosure statement

The authors declare no conflict of interest with the results produced for this paper.

References

Aiuppa, A., Fischer, T. P., Plank, T., Bani, P., 2019. Co 2 flux emissions from the earth's most actively degassing volcanoes, 2005–2015. Scientific reports 9 (1), 1–17.

Alhamid, M. I., Daud, Y., Surachman, A., Sugiyono, A., Aditya, H., Mahlia, T., et al., 2016. Potential of geothermal energy for electricity generation in indonesia: A review. Renewable and Sustainable Energy Reviews 53, 733–740.

Angelis-Dimakis, A., Biberacher, M., Dominguez, J., Fiorese, G., Gadocha, S., Gnansounou, E., Guariso, G., Kartalidis, A., Panichelli, L., Pinedo, I., et al., 2011. Methods and tools to evaluate the availability of renewable energy sources. Renewable and sustainable energy reviews 15 (2), 1182–1200.

Armannsson, H., Kristmannsdottir, H., 1992. Geothermal environmental impact. Geothermics 21 (5-6), 869–880.

Assante, M., Candela, L., Castelli, D., Cirillo, R., Coro, G., Frosini, L., Lelii, L., Mangiacrapa, F., Marioli, V., Pagano, P., et al., 2018. The gcube system: Delivering virtual research environments as-a-service. Future Generation Computer Systems, NA.

Atmaca, E., Basar, H. B., 2012. Evaluation of power plants in turkey using analytic network process (anp). Energy 44 (1), 555–563.

Baldwin, R. A., 2009. Use of maximum entropy modeling in wildlife research. Entropy 11 (4), 854–866.

Barbier, E., 2002. Geothermal energy technology and current status: an overview. Renewable and sustainable energy reviews 6 (1-2), 3–65.

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

Bertani, R., 2003. What is geothermal potential. IGA news 53, 1–3.

Bertani, R., 2012. Geothermal power generation in the world 2005–2010 update report. geothermics 41, 1–29.

Bertani, R., 2016. Geothermal power generation in the world 2010–2014 update report. Geothermics 60, 31–43.

Bertermann, D., Klug, H., Morper-Busch, L., 2015. A pan-european planning basis for estimating the very shallow geothermal energy potentials. Renewable energy 75, 335–347.

Bidarmaghz, A., Narsilio, G. A., Johnston, I. W., Colls, S., 2016. The importance of surface air temperature fluctuations on long-term performance of vertical ground heat exchangers. Geomechanics for Energy and the Environment 6, 35–44.

Boyle, G., 2004. Renewable energy. Renewable Energy, by Edited by Godfrey Boyle, pp. 456. Oxford University Press, May 2004. ISBN-10: 0199261784. ISBN-13: 9780199261789, 456.

Breeze, P., 2019. Power generation technologies. Newnes.

Cacace, M., Kaiser, B. O., Lewerenz, B., Scheck-Wenderoth, M., 2010. Geothermal energy in sedimentary basins: What we can learn from regional numerical models. Chemie der Erde-Geochemistry 70, 33–46.

CAMS, 2019. Flux inversion reanalysis of global carbon dioxide - fluxes and atmospheric concentrations. https://atmosphere.copernicus.eu/catalogue#/product/urn:x-wmo:md:int.ecmwf::copernicus:cams:prod:rean:co2:pid286.

Candela, L., Castelli, D., Coro, G., Pagano, P., Sinibaldi, F., 2016. Species distribution modeling in the cloud. Concurrency and Computation: Practice and Experience 28 (4), 1056–1079.

Chiodini, G., Frondini, F., Kerrick, D., Rogie, J., Parello, F., Peruzzi, L., Zanzari, A., 1999. Quantification of deep co2 fluxes from central italy. examples of carbon balance for regional aquifers and of soil diffuse degassing. Chemical Geology 159 (1-4), 205–222.

Clark, K., Lukovic, B., Villamor, P., Watson, M., McNamara, D. D., Milicich, S. D., Pummer, B., Ries, W., Rosenberg, M., Sepulveda, F., et al., 2015. New zealand geothermal power plants as critical facilities: an active fault avoidance study in the wairakei geothermal field, new zealand. In: World Geothermal Congress 2015. International Geothermal Association, pp. 19–25.

CMIP5, 2019. Coupled Model Intercomparison Project Phase 5. pcmdi.llnl.gov/mips/cmip5/.

CNR, 2019. Maximum Entropy Model Web Processing Service. https://services.d4science.org/group/biodiversitylab/data-miner?OperatorId=org.gcube.dataanalysis.wps.statisticalmanager.synchserver.mappedclasses.transducerers.MAX_ENT_NICHE_MODELLING.

Coolbaugh, M. F., Taranik, J. V., Rains, G., Shevenell, L. A., Sawatzky, D. L., Bedell, R., Minor, T. B., 2002. A geothermal gis for nevada: defining regional controls and favorable exploration terrains for extensional geothermal systems. Transactions-Geothermal Resources Council, 485–490.

Coro, G., Candela, L., Pagano, P., Italiano, A., Liccardo, L., 2015a. Parallelizing the execution of native data mining algorithms for computational biology. Concurrency and Computation: Practice and Experience 27 (17), 4630–4644.

Coro, G., Magliozzi, C., Ellenbroek, A., Pagano, P., 2015b. Improving data quality to build a robust distribution model for architeuthis dux. Ecological Modelling 305, 29–39.

Coro, G., Pagano, P., Ellenbroek, A., 2013. Combining simulated expert knowledge with neural networks to produce ecological niche models for latimeria chalumnae. Ecological modelling 268, 55–63.

Coro, G., Pagano, P., Ellenbroek, A., 2014. Comparing heterogeneous distribution maps for marine species. GIScience & remote sensing 51 (5), 593–611.

Coro, G., Pagano, P., Ellenbroek, A., 2018a. Detecting patterns of climate change in long-term forecasts of marine environmental parameters. International Journal of Digital Earth, 1–19.

Coro, G., Pagano, P., Napolitano, U., 2016. Bridging environmental data providers and seadatanet diva service within a collaborative and distributed e-infrastructure. Bollettino di Geofisica, 23–25.

Coro, G., Panichi, G., Scarponi, P., Pagano, P., 2017. Cloud computing in a distributed e-infrastructure using the web processing service standard. Concurrency and Computation: Practice and Experience 29 (18), e4219.

Coro, G., Scarponi, P., Pagano, P., 2018b. Enhancing argo floats datafire-usability. Bollettino di Geofisica, 53.

Coro, G., Vilas, L. G., Magliozzi, C., Ellenbroek, A., Scarponi, P., Pagano, P., 2018c. Forecasting the ongoing invasion of lagocephalus sceleratus in the mediterranean sea. Ecological Modelling 371, 37–49.

Davies, J. H., 2013. Global map of solid earth surface heat flow. Geochemistry, Geophysics, Geosystems 14 (10), 4608–4622.
URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/ggge.20271

Deoliveira, J., 2008. Geoserver: uniting the geoweb and spatial data infrastructures. In: Proceedings of the 10th International Conference for Spatial Data Infrastructure, St. Augustine, Trinidad. pp. 25–29.

Di Pippo, R., Lippmann, M., 2017. The shrinking salton sea and its impact on geothermal development. https://geothermal.org/PDFs/Articles/17JulyAug.pdf.

DiPippo, R., 2012. Geothermal power plants: principles, applications, case studies and environmental impact. Butterworth-Heinemann.

Dumas, P., Garabetian, T., Manfrida, G., Fiaschi, D., Parisi, M. L., Loschetter, A., Maury, J., Ravier, G., Blanc, I., Perez-Lopez, P., et al., 2020. Geoenvi project: Tackling the environmental concerns for deploying geothermal. In: Proceedings World Geothermal Congress 2020, Reykjavik, Iceland, April 26 – May 2, 2020. WGC, pp. 1–8.

Elith, J., Leathwick, J. R., 2009. Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. Annual Review of Ecology, Evolution, and Systematics 40 (1), 677–697.

Elith, J., Phillips, S. J., Hastie, T., Dudík, M., Chee, Y. E., Yates, C. J., Jan. 2011. A statistical explanation of MaxEnt for ecologists. Diversity and Distributions 17 (1), 43–57.

Enel Green Power, 2020. Technical Report on the New Latera Geothermal Power Plant. `https://www.dariotamburrano.it/wp-content/uploads/2017/10/RELAZIONE-TECNICA-LATERA.pdf`.

Engdahl, E. R., 2002. Global seismicity: 1900-1999. International handbook of earthquake and engineering seismology, 665–690.

Engdahl, E. R., van der Hilst, R., Buland, R., 1998. Global teleseismic earthquake relocation with improved travel times and procedures for depth determination. Bulletin of the Seismological Society of America 88 (3), 722–743.

García-Gil, A., Vázquez-Suñe, E., Alcaraz, M. M., Juan, A. S., Sánchez-Navarro, J. Á., Montlleó, M., Rodríguez, G., Lao, J., 2015. Gis-supported mapping of low-temperature geothermal potential taking groundwater flow into account. Renewable Energy 77, 268–278.

Glassley, W. E., 2018. Geology and hydrology of geothermal energy. Power Stations Using Locally Available Energy Sources: A Volume in the Encyclopedia of Sustainability Science and Technology Series, Second Edition, 23–34.

Griebler, C., Brielmann, H., Haberer, C. M., Kaschuba, S., Kellermann, C., Stumpp, C., Hegler, F., Kuntz, D., Walker-Hertkorn, S., Lueders, T., 2016. Potential impacts of geothermal energy use and storage of heat on groundwater quality, biodiversity, and ecosystem processes. Environmental Earth Sciences 75 (20), 1391.

Hey, T., Tansley, S., Tolle, K. M., 2009. The fourth paradigm: data-intensive scientific discovery. Vol. 1. Microsoft research Redmond, WA.

HGPS, 2020. Hellisheidi Geothermal Power Station - South Iceland. `https://www.extremeiceland.is/en/attractions/hellisheidi-geothermal-power-station`.

IGA, 2020. International Geothermal Association Database. `https://www.geothermal-energy.org/explore/our-databases/geothermal-power-database/`.

John Caron, U., Davis, E., 2006. Unidata's thredds data server. In: 22nd International Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology. pp. 1–3.

Juncu, D., Árnadóttir, T., Ali, T., Hooper, A., 2015. Numerical modelling of crustal deformation due to fluid extraction and re-injection in the hengill geothermal area in south iceland. In: EGU General Assembly Conference Abstracts. Vol. 17. p. 13455.

Karimi, H., Soffianian, A., Seifi, S., Pourmanafi, S., Ramin, H., 2019. Evaluating optimal sites for combined-cycle power plants using gis: comparison of two aggregation methods in iran. International Journal of Sustainable Energy, 1–12.

Karlsdottir, M. R., Heinonen, J., Palsson, H., Palsson, O. P., 2020. Life cycle assessment of a geothermal combined heat and power plant based on high temperature utilization. Geothermics 84, 101727.

Kissling, W., Weir, G., 2005. The spatial distribution of the geothermal fields in the taupo volcanic zone, new zealand. Journal of Volcanology and Geothermal Research 145 (1-2), 136–150.

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

Lewicki, J., Fischer, M., Hilley, G., 2008. Six-week time series of eddy covariance co2 flux at mammoth mountain, california: Performance evaluation and role of meteorological forcing. Journal of Volcanology and Geothermal Research 171 (3-4), 178–190.

Lewicki, J. L., Oldenburg, C. M., 2005. Near-surface co2 monitoring and analysis to detect hidden geothermal systems. Tech. rep., Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States).

Limberger, J., Boxem, T., Pluymaekers, M., Bruhn, D., Manzella, A., Calcagno, P., Beekman, F., Cloetingh, S., van Wees, J.-D., 2018. Geothermal energy in deep aquifers: A global assessment of the resource base for direct heat utilization. Renewable and Sustainable Energy Reviews 82, 961–975.

Limberger, J., Calcagno, P., Manzella, A., Trumpy, E., Boxem, T., Pluymaekers, M., van Wees, J.-D., 2014. Assessing the prospective resource base for enhanced geothermal systems in europe. Geothermal Energy Science 2 (1), 55–71.

Lu, S.-M., 2018. A global review of enhanced geothermal system (egs). Renewable and Sustainable Energy Reviews 81, 2902 – 2921.
URL http://www.sciencedirect.com/science/article/pii/S1364032117310341

Lund, J. W., Boyd, T. L., 2016. Direct utilization of geothermal energy 2015 worldwide review. Geothermics 60, 66–93.

Macharia, M. W., Gachari, M. K., Mundia, C. N., Kuria, D. N., et al., 2018. A gis-based approach for exploring geothermal resources along part of the kenyan rift. Journal of Applied Sciences, Engineering and Technology for Development. JASETD, 254–265.

Manzella, A., 2017. Geothermal energy. EPJ Web Conf. 148, 00012.
URL https://doi.org/10.1051/epjconf/201714800012

Martín-Gamboa, M., Iribarren, D., Dufour, J., 2015. On the environmental suitability of high-and low-enthalpy geothermal systems. Geothermics 53, 27–37.

Moghaddam, M. K., Samadzadegan, F., Noorollahi, Y., Sharifi, M. A., Itoi, R., 2014. Spatial analysis and multi-criteria decision making for regional-scale geothermal favorability map. Geothermics 50, 189–201.

Mourmouris, J., Potolias, C., 2013. A multi-criteria methodology for energy planning and developing renewable energy sources at a regional level: A case study thassos, greece. Energy Policy 52, 522–530.

NASA-NEX, 2015. Space apps challenge. https://web.archive.org/web/20150924122046/https://nex.nasa.gov/nex/projects/1348/wiki/general_data_access_and_apis/.

NASA-NEX, 2020. The NASA Earth Exchange Platform. nex.nasa.gov.

Nguyen, A., Pasquier, P., Marcotte, D., 2017. Borehole thermal energy storage systems under the influence of groundwater flow and time-varying surface temperature. Geothermics 66, 110–118.

Nielsen, L. H., Sparre Andersen, M., Balling, N., Boldreel, L. O., Fuchs, S., Leth Hjuler, M., Kristensen, L., Mathiesen, A., Olivarius, M., Weibel, R., 2017. The geothermal energy potential in denmark-updating the database and new structural and thermal models. In: EGU General Assembly Conference Abstracts. Vol. 19. p. 7296.

NOAA, 2001. ETOPO2 Global 2 Arc-minute Ocean Depth and Land Elevation from the US National Geophysical Data Center (NGDC).
URL \url{https://doi.org/10.5065/D6668B75}

Noorollahi, Y., Itoi, R., Fujii, H., Tanaka, T., 2007. Gis model for geothermal resource exploration in akita and iwate prefectures, northern japan. Computers & geosciences 33 (8), 1008–1021.

Nykanen, H., Alm, J., Lang, K., Silvola, J., Martikainen, P. J., 1995. Emissions of ch4, n2o and co2 from a virgin fen and a fen drained for grassland in finland. Journal of Biogeography, 351–357.

OSGeo, 2019. GDAL - Geospatial Data Abstraction Library. https://www.gdal.org/.

Papachristou, M., Mendrinos, D., Dalampakis, P., Arvanitis, A., Karytsas, C., Andritsos, N., 2016. Geothermal energy use, country update for greece. In: European geothermal congress, Strasbourg, France. pp. 19–24.

Parisi, M. L., Ferrara, N., Torsello, L., Basosi, R., 2019. Life cycle assessment of atmospheric emission profiles of the italian geothermal power plants. Journal of Cleaner Production 234, 881–894.

Paulillo, A., Striolo, A., Lettieri, P., 2019. The environmental impacts and the carbon intensity of geothermal energy: a case study on the hellisheiði plant. Environment international 133, 105226.

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

Phillips, S. J., Anderson, R. P., Schapire, R. E., 2006a. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190 (3-4), 231–259.

Phillips, S. J., Anderson, R. P., Schapire, R. E., 2006b. Maximum entropy modeling of species geographic distributions. Ecological Modelling 190 (3), 231 – 259.
URL http://www.sciencedirect.com/science/article/pii/S030438000500267X

Phillips, S. J., Dudik, M., 2008. Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation. Ecography 31, 161–175.

Phillips, S. J., Dudík, M., Schapire, R. E., 2004. A maximum entropy approach to species distribution modeling. In: Proceedings of the twenty-first international conference on Machine learning. ACM, p. 83.

Phillips, S. J., Dudík, M., 2008. Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. Ecography 31 (2), 161–175.
URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0906-7590.2008.5203.x

Phillips, S. J., Miroslav, D., E., S. R., 2019. Maxent software for modeling species niches and distributions (version 3.4.1). http://biodiversityinformatics.amnh.org/open_source/maxent/.

Prest, R., Daniell, T., Ostendorf, B., 2007. Using gis to evaluate the impact of exclusion zones on the connection cost of wave energy to the electricity grid. Energy Policy 35 (9), 4516–4528.

Pruess, K., 2006. Enhanced geothermal systems (egs) using co2 as working fluid—a novel approach for generating renewable energy with simultaneous sequestration of carbon. Geothermics 35 (4), 351–367.

QGis, D., 2011. Quantum gis geographic information system. Open Source Geospatial Foundation Project 45.

Ramachandra, T., Shruthi, B., 2007. Spatial mapping of renewable energy potential. Renewable and sustainable energy reviews 11 (7), 1460–1480.

Ramazankhani, M.-E., Mostafaeipour, A., Hosseininasab, H., Fakhrzad, M.-B., 2016. Feasibility of geothermal power assisted hydrogen production in iran. International Journal of Hydrogen Energy 41 (41), 18351–18369.

Randolph, J. B., Saar, M. O., 2011a. Combining geothermal energy capture with geologic carbon dioxide sequestration. Geophysical Research Letters 38 (10).

Randolph, J. B., Saar, M. O., 2011b. Coupling carbon dioxide sequestration with geothermal energy capture in naturally permeable, porous geologic formations: Implications for co2 sequestration. Energy Procedia 4, 2206–2213.

Richts, A., Struckmeier, W. F., Zaepke, M., 2011. Whymap and the groundwater resources map of the world 1: 25,000,000. In: Sustaining groundwater resources. Springer, pp. 159–173.

Roberts, V. W., 1978. Geothermal energy prospects for the next 50 years. Electric Power Research Institute (EPRI), Report ER-611-SR, 4–1.

Rodrigo-Naharro, J., Nisi, B., Vaselli, O., Lelli, M., Saldaña, R., Clemente-Jul, C., del Villar, L. P., 2013. Diffuse soil co2 flux to assess the reliability of co2 storage in the mazarrón–gañuelas tertiary basin (spain). Fuel 114, 162–171.

Rybach, L., Muffler, L. J. P., 1981. Geothermal systems: principles and case histories. Chichester, Sussex, England and New York, Wiley-Interscience, 1981. 371 p.

Sah, B. P., Wijayatunga, P., 2017. Geographic information system-based decision support system for renewable energy development: An indonesian case study. Think-Asia.

Santoyo, E., Acevedo-Anicasio, A., Pérez-Zarate, D., Guevara, M., 2018. Evaluation of artificial neural networks and eddy covariance measurements for modelling the co2 flux dynamics in the acoculco geothermal caldera (mexico). International Journal of Environmental Science and Development 9 (10).

Satkin, M., Noorollahi, Y., Abbaspour, M., Yousefi, H., 2014. Multi criteria site selection model for wind-compressed air energy storage power plants in iran. Renewable and Sustainable Energy Reviews 32, 579–590.

Schaap, D. M., Lowry, R. K., 2010. Seadatanet–pan-european infrastructure for marine and ocean data management: unified access to distributed data sets. International Journal of Digital Earth 3 (S1), 50–69.

Schellschmidt, R., Sanner, B., Pester, S., Schulz, R., 2010. Geothermal energy use in germany. In: Proceedings World geothermal congress. Vol. 152. p. 19.

Seibt, P., Kabus, F., Hoth, P., 2005. The neustadt-glewe geothermal power plant–practical experience in the reinjection of cooled thermal waters into sandstone aquifers. In: Proceedings Word Geothermal Congress, Antalya (Turkey). pp. 1–4.

Shahab, A., Singh, M., 2019. Comparative analysis of different machine learning algorithms in classification of suitability of renewable energy resource. In: 2019 International Conference on Communication and Signal Processing (ICCSP). IEEE, pp. 0360–0364.

Sharp, J., 1978. Energy and momentum transport model of the ouachita basin and its possible impact on formation of economic mineral deposits. Economic Geology 73 (6), 1057–1068.

Sharp Jr, J., Domenico, P., 1976. Energy transport in thick sequences of compacting sediment. Geological Society of America Bulletin 87 (3), 390–400.

Siefi, S., Karimi, H., Soffianian, A., Pourmanafi, S., 2017. Gis-based multi criteria evaluation for thermal power plant site selection in kahnuj county, se iran. Civil Engineering Infrastructures Journal 50 (1), 179–189.

Stefansson, V., 2005. World geothermal assessment. In: Proceedings of the world geothermal congress. pp. 24–29.

Tester, J. W., Drake, E. M., Driscoll, M. J., Golay, M. W., Peters, W. A., 2012. Sustainable energy: choosing among options. MIT press.

Tomasini-Montenegro, C., Santoyo-Castelazo, E., Gujba, H., Romero, R., Santoyo, E., 2017. Life cycle assessment of geothermal power generation technologies: An updated review. Applied Thermal Engineering 114, 1119–1136.

Troldborg, M., Heslop, S., Hough, R. L., 2014. Assessing the sustainability of renewable energy technologies using multi-criteria analysis: Suitability of approach for national-scale assessments and associated uncertainties. Renewable and Sustainable Energy Reviews 39, 1173–1184.

Troupin, C., Barth, A., Sirjacobs, D., Ouberdous, M., Brankart, J.-M., Brasseur, P., Rixen, M., Alvera-Azcárate, A., Belounis, M., Capet, A., et al., 2012. Generation of analysis and consistent error fields using the data interpolating variational analysis (diva). Ocean Modelling 52, 90–101.

Troupin, C., Machin, F., Ouberdous, M., Sirjacobs, D., Barth, A., Beckers, J.-M., 2010. High-resolution climatology of the northeast atlantic using data-interpolating variational analysis (diva). Journal of Geophysical Research: Oceans 115 (C8).

Trumpy, E., Bertani, R., Manzella, A., Sander, M., 2015a. The web-oriented framework of the world geothermal production database: a business intelligence platform for wide data distribution and analysis. Renewable Energy 74, 379–389.

Trumpy, E., Donato, A., Gianelli, G., Gola, G., Minissale, A., Montanari, D., Santilano, A., Manzella, A., 2015b. Data integration and favourability maps for exploring geothermal systems in sicily, southern italy. Geothermics 56, 1–16.

Yousefi, H., Ehara, S., Noorollahi, Y., 2007. Geothermal potential site selection using gis in iran. In: Proceedings of the 32nd workshop on geothermal reservoir engineering, Stanford University, Stanford, California. pp. 174–82.

Zhao, Z., Dong, S., Jiang, X., Liu, S., Ji, H., Li, Y., Han, Y., Sha, W., 2017. Effects of warming and nitrogen deposition on ch4, co2 and n2o emissions in alpine grassland ecosystems of the qinghai-tibetan plateau. Science of the Total Environment 592, 565–572.

Data	Primary Source
Carbon Dioxide	Copernicus Atmosphere Monitoring Service
Distance from Convergent Lines	United States Geological Survey
Distance from Diffuse Lines	United States Geological Survey
Distance from Ridge Lines	United States Geological Survey
Distance from Transform Lines	United States Geological Survey
Earthquake Density	Centennial Earthquake Catalog
Earthquake Depths	Centennial Earthquake Catalog
Earthquake Magnitudes	Centennial Earthquake Catalog
Elevation/Depth	United Stated National Geophysical Data Center
Global Heat Flow	Davies (2013)
Groundwater Resources	World-wide Hydrogeological Mapping and Assessment Programme
Precipitation	NASA Earth Exchange Platform
Sediments	Laske (1997)
Surface Air Temperature	NASA Earth Exchange Platform
Current and Planned Geothermal Plants	IGA - Global Geothermal Energy Database

Table 1: Summary of all used parameters along with their primary sources. Details about how these data were accessed and post-processed are given in the article.

Parameter name	Percent contribution	Permutation importance
a. All parameters		
Carbon Dioxide	38.7	14.9
Earthquake Density	30.5	29.4
Elevation/Depth	10.1	21.4
Global Heat Flow	6.3	2.7
Sediments	3.8	1.5
Surface Air Temperature	2.3	1.3
Earthquake Depths	1.7	0.8
Distance from Convergent Lines	1.6	19.7
Earthquake Magnitudes	1.5	0.6
Distance from Transform Lines	1.1	0.4
Distance from Diffuse Lines	0.8	2.2
Precipitation	0.7	1.5
Distance from Ridge Lines	0.6	3
Groundwater Resources	0.3	0.6
b. MaxEnt-selected parameters		
Carbon Dioxide	41.4	18
Earthquake Density	33.4	41.9
Elevation/Depth	10.9	28.9
Global Heat Flow	7.1	5.2
Sediments	4.3	3.3
Surface Air Temperature	2.9	2.7
c. Expert-selected parameters		
Earthquake Density	41.6	32.3
Elevation/Depth	27.5	33.5
Surface Air Temperature	10.5	2.8
Global Heat Flow	6.7	2.3
Sediments	3.7	1.3
Earthquake Depths	2.5	2.5
Earthquake Magnitudes	1.8	0.5
Distance from Converging Lines	1.8	19.5
Distance from Transform Lines	1.2	0.5
Groundwater Resources	1.1	1.1
Distance from Diffuse Lines	1	1.5
Distance from Ridge Lines	0.5	2.4
d. Expert and MaxEnt-selected parameters		
Earthquake Density	45.5	50
Elevation/Depth	30.7	36.9
Surface Air Temperature	12	5.6
Global Heat Flow	7.6	3.5
Sediments	4.3	4.1

Table 2: Percent contribution and permutation importance of the environmental parameters involved in our experiment, calculated by a Maximum Entropy model: (a) all parameters (the most important ones are highlighted in bold), (b) parameters having the highest importance for Maximum Entropy, (c) parameters selected by an expert (the most important ones are highlighted in bold), (d) intersection between expert- and Maximum Entropy-selected parameters.

Model	**AUC**	**Accuracy using optimal threshold**	**Accuracy using balanced threshold**	**Accuracy using 10p threshold**	**Optimal threshold**	**Balanced threshold**	**10p threshold**
All parameter	**0.988**	86.7%	**76.2%**	62.0%	0.02	0.136	0.319
MaxEnt-selected par.	0.98	**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

Table 3: Performance of the four models built in our experiment in terms of Area Under the Curve (AUC) and accuracy. Accuracy is calculated for each decision threshold estimated by the Maximum Entropy models. These thresholds are reported in the rightmost columns.

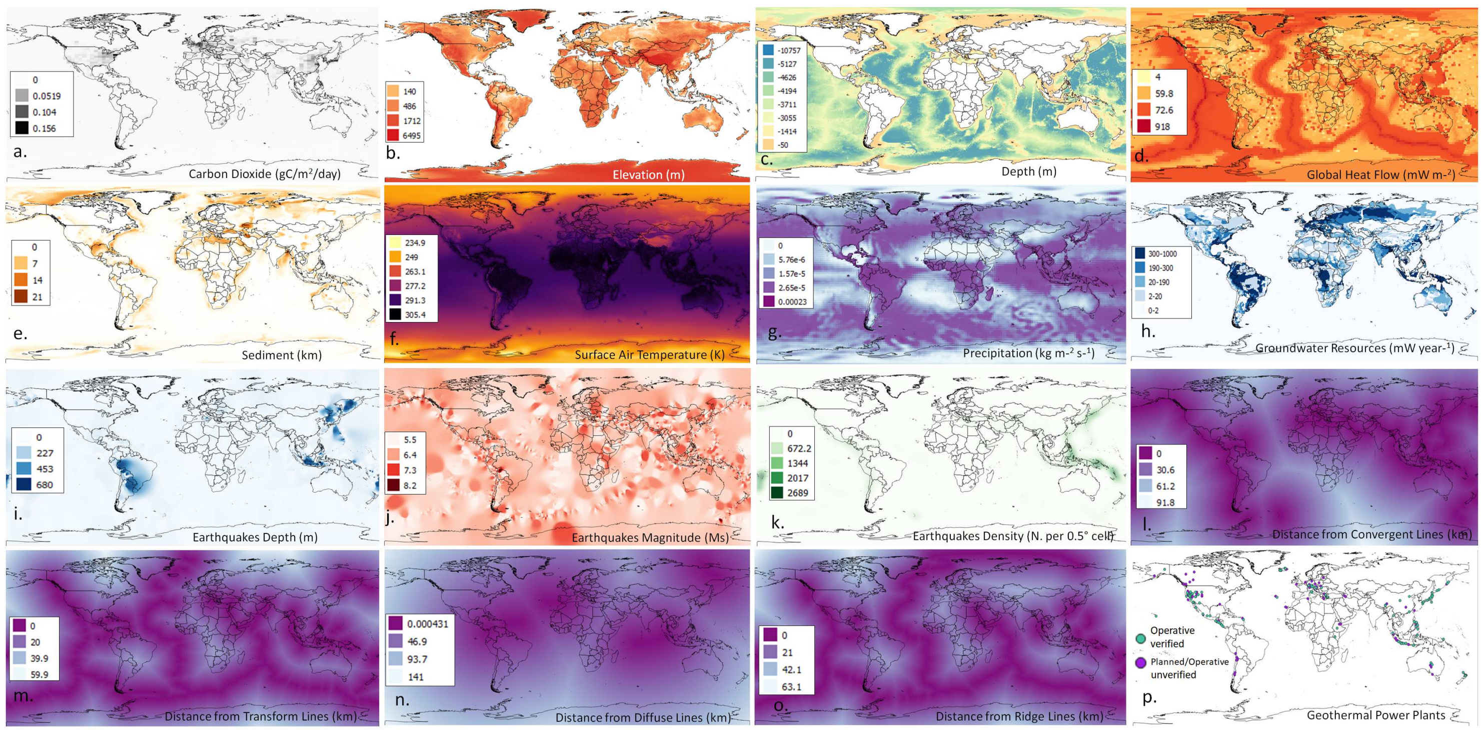


Figure 1: Visual comparison of the global data used in our model: (a) carbon dioxide, (b) elevation, (c) depth, (d) global heat flow, (e) sediment thickness, (f) surface air temperature, (g) precipitation, (h) groundwater resources, (i) earthquake depth, (j) magnitude, and (k) density, distance from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

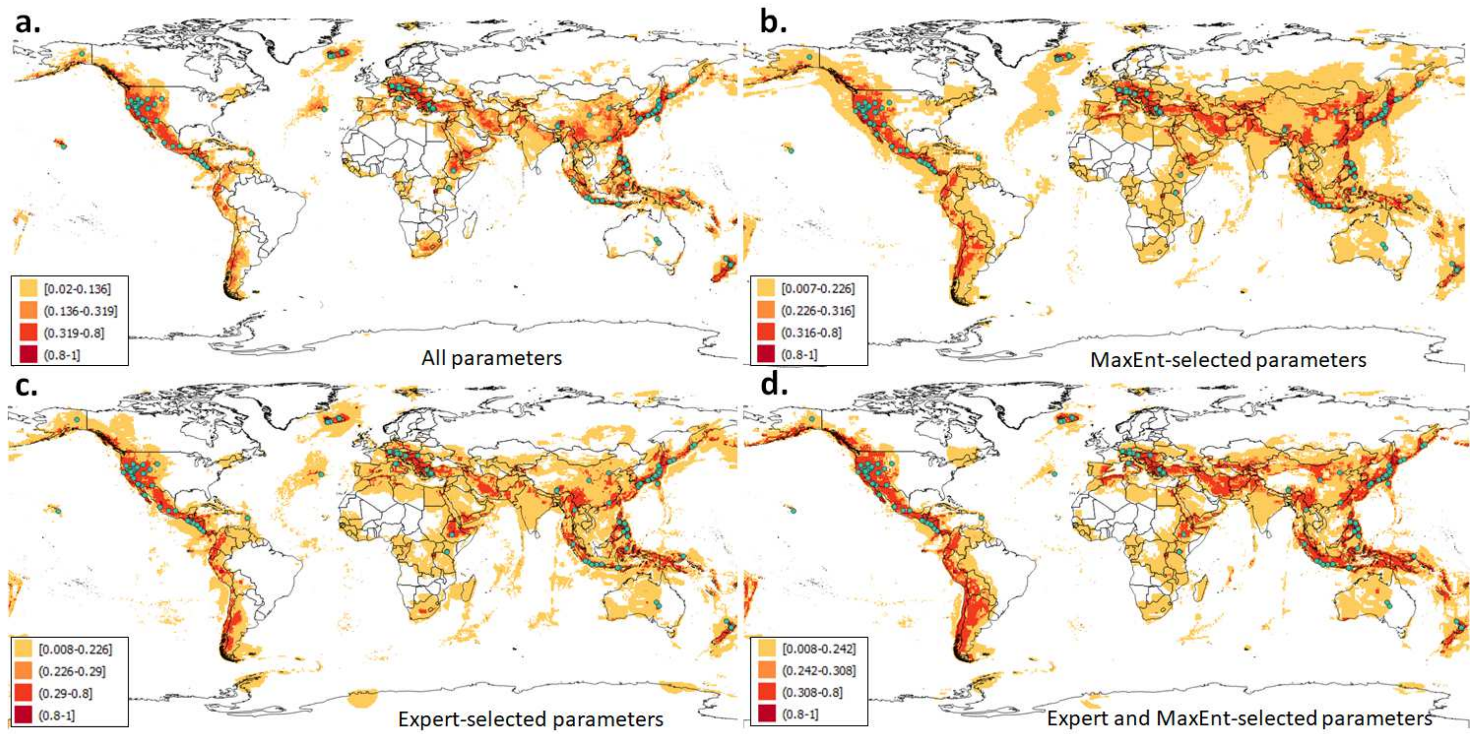


Figure 2: Global geothermal suitability distribution of four Maximum Entropy models using (a) all environmental parameters, (b) only the most important parameters according to the Maximum Entropy model, (c) parameters selected by an expert, (d) the intersection between expert- Maximum Entropy-selected parameters. Warmer colours indicate higher suitability scores. Dots indicate geothermal power plants locations used to train the models. The colours clusters identify suitability score ranges that are built over the decision thresholds calculated by the models.

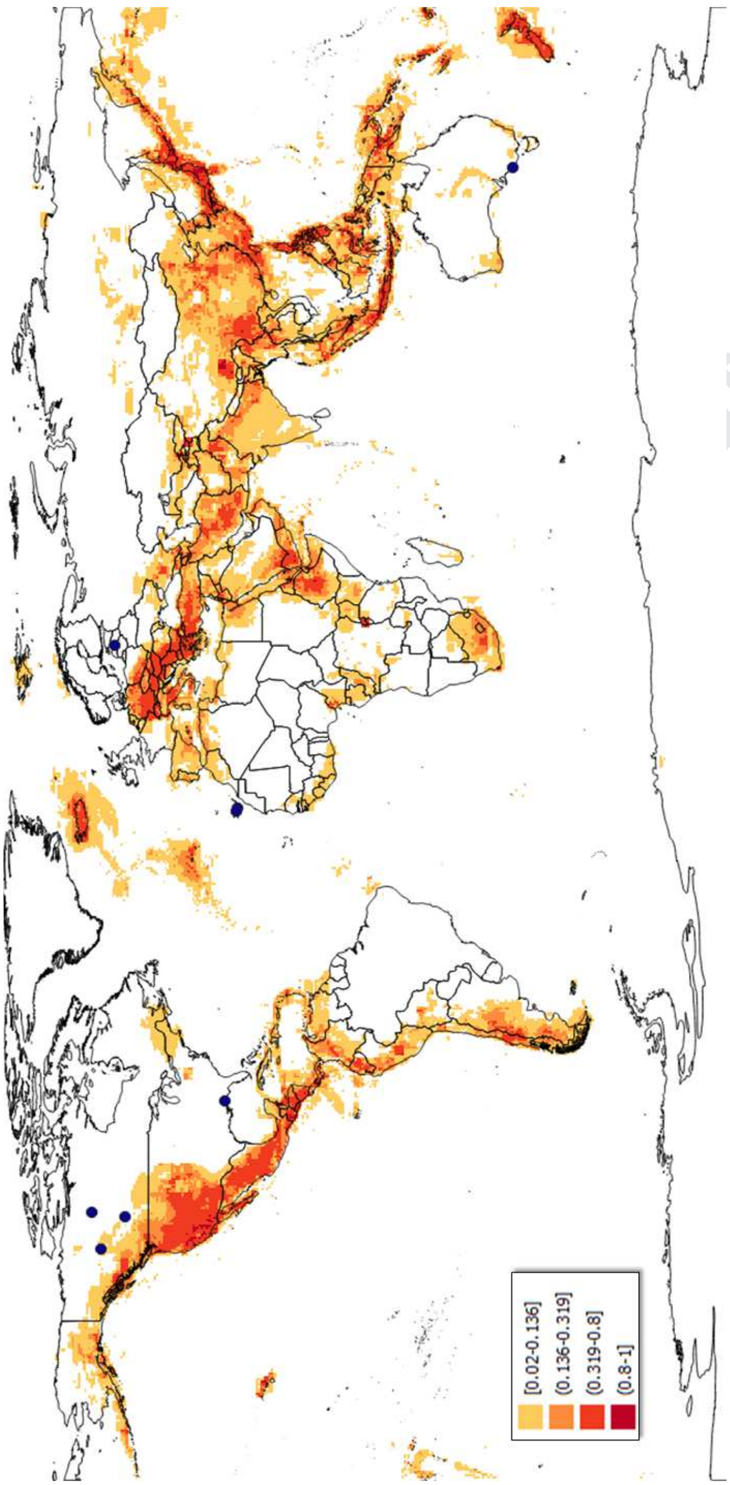


Figure 3: Optimal geothermal suitability distribution produced by the Maximum Entropy model using all parameters. Warmer colours indicate higher suitability scores. Dots indicate geothermal power plants in the test set whose suitability is not predicted by the model.

CRediT author statement

Gianpaolo Coro: Conceptualization, Methodology, Writing, Software

Eugenio Trumpy: Conceptualization, Methodology, Writing, Data curation

Highlights

- The first global-scale map of the suitability of an area to geothermal energy production and plant installation is presented
- Geospatial data correlated with geothermal site suitability are processed through Data-Interpolating Variational Analysis and Maximum Entropy modelling
- The reliability of the map is tested against currently active and planned geothermal power plants
- An Open Science-compliant tool is proposed to allow stakeholders increase the map resolution and revise parameters
- Target stakeholders are scientists, industry operators, and policy makers, for transparently assessing geothermal sites and improving communication with citizens

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Gianpaolo Coro (on behalf of the authors)