Preprocessing approaches in machine learning-based groundwater potential mapping: an application to the Koulikoro and Bamako regions, Mali

Víctor Gómez-Escalonilla [1*], Pedro Martínez-Santos[1], Miguel Martín-Loeches[2]

5   [1]UNESCO/UNITWIN Chair Appropriate Technologies for Human Development. Department of Geodynamic, Stratigraphy and Paleontology, Faculty of Geology, Complutense University of Madrid, C/José Antonio Novais 12, 28040 Madrid, Spain.
[2]Department of Geology, Geography and Environment. Geology UD, University of Alcalá, Alcalá de Henares, Madrid, Spain.

Correspondence to: V. Gómez-Escalonilla (vigome01@ucm.es)

Abstract. Groundwater is crucial for domestic supplies in the Sahel, where the strategic importance of aquifers can only be
10 expected to increase in the coming years due to climate change. Groundwater potential mapping is gaining recognition as a valuable tool to underpin water management practices in the region, and hence, to improve water access. This paper presents a machine learning method to map groundwater potential and illustrates it through an application to two regions of Mali. A set of explanatory variables for the presence of groundwater is developed first. Several scaling methods (standardization, normalization, maximum absolute value and min-max scaling) are used to avoid the pitfalls associated with the reclassification
15 of explanatory variables. A number of supervised learning classifiers is then trained and tested on a large borehole database (n=3,345) in order to find meaningful correlations between the presence or absence of groundwater and the explanatory variables. This process identifies noisy, collinear and counterproductive variables and excludes them from the input dataset. Tree-based algorithms, including the AdaBoost, Gradient Boosting, Random Forest, Decision Tree and Extra Trees classifiers were found to outperform other algorithms on a consistent basis (accuracy >0.85), whereas maximum absolute value and
20 standardization proved the most efficient methods to scale explanatory variables. Borehole flow rate data is used to calibrate the results beyond standard machine learning metrics, thus adding robustness to the predictions. The southern part of the study area was identified as the better groundwater prospect, which is consistent with the geological and climatic setting. From a methodological standpoint, the outcomes lead to three major conclusions: (1) because there is a priori no way to know which algorithm will work better on a given dataset, we advocate the use of a large number of machine learning classifiers, out of
25 which the best are subsequently picked for ensembling; (2) standard machine learning metrics may be of limited value when appraising map outcomes, and should be complemented with hydrogeological indicators whenever possible; and (3) the scaling of the variables helps to minimize bias arising from expert judgement and maintains robust predictive capabilities.

Keywords: big data, climate change, water access, random forest, decision trees, supervised classification

1 Introduction

Water is crucial for human beings. Water provides food security, cleanliness and hydration, which translates into health, economic activity and arguably, better education opportunities (United Nations, 2002, 2010). Today, 2.5 billion people depend exclusively on groundwater for their domestic supply (Grönwall and Danert, 2020). Groundwater is particularly crucial in most of the Sahel, where rainfall and surface water are absent for several months (Llamas and Martínez-Santos, 2005; Díaz-

35 Alcaide et al., 2017). In a context of climate change, where rainfall will decrease in most arid and semiarid regions and drought episodes will intensify (Arneth et al., 2019), groundwater resources are expected to be increasingly relied upon.

Groundwater potential mapping (GPM) is recognized as a valuable tool to underpin planning and development of groundwater resources (Elbeih, 2015). GPM may be understood as a means to estimate groundwater storage in a given region, as a measure

40 of the probability of finding groundwater, or as a prediction as to where the highest borehole yields may occur (Díaz-Alcaide and Martínez-Santos, 2019). In practice, however, it consists in computing spatially-distributed estimates for a target variable (groundwater potential) based a set of explanatory variables (Díaz-Alcaide and Martínez-Santos, 2019). GPM typically relies on existing cartography, digital elevation models obtained from satellite, aerial photographs, satellite imagery and geophysical information (Schetselaar et al., 2007).

45

There are two main approaches to GPM: expert-based decision systems and machine learning methods. Expert-based systems have existed for a long time (DEP, 1993), and include multi-influence factor techniques (Magesh et al., 2012; Nasir et al., 2018; Martín-Loeches et al 2018) and analytical hierarchy processes (Mohammadi-Behzad et al., 2019; Al-Djazouli et al., 2020), among others. Machine learning is comparatively newer. In the current context of computer science development,

50 advances in machine learning techniques present enormous upside in terms of predicting the outcome of a target variable based on complex associations of explanatory variables. Hence, machine learning is well suited to map groundwater potential. Algorithms used in the GPM literature include Mixture Discriminant Analysis (Al-Fugara et al., 2020), Random Forest (Kalantar et al., 2019; Moghaddam et al., 2020), Boosted Regression Tree (Naghibi et al., 2016), Logistic Regression (Ozdemir, 2011; Chen et al., 2018; Nhu et al., 2020), Support Vector Machines (Naghibi et al., 2017b), Neural Networks (Lee

55 et al., 2012; Panahi et al., 2020) and Ensemble methods (Naghibi et al., 2017a; Martínez-Santos and Renard, 2020; Nguyen et al., 2020b).

GPM works under the assumption that the presence of groundwater can be partially inferred from surface features. Some of the most frequently used explanatory variables are lithology, geological lineaments, landforms, topography, soil, land use/land

60 cover, drainage and slope-related variables, rainfall, and vegetation indices (Jha et al., 2007). Supervised classification algorithms are trained to find the associations between these variables and known groundwater data. Once the algorithms yield accurate predictions, their findings are extrapolated to estimate groundwater potential in a given study area.

Because the number of available boreholes to train and test the algorithms is usually "small", and because the number of explanatory variables can be relatively high, a crucial issue in machine-learning studies is how explanatory variables should
65   be reclassified in order to minimize noise. Some of the explanatory variables are categorical, and can be represented as integers. Others are however continuous, and are best reclassified in ranges prior to being assigned an integer number. This is typically performed by means of classic methods, such as quantiles, standard deviation intervals or natural breaks (Jenks, 1967). Sometimes the intervals are based directly on expert criteria, which means that a bias may be incorporated from the beginning of the process.

70

The outcomes of machine learning GPM studies are almost invariably assessed by means of standard big data metrics such as precision, recall, and area under the receiver operating characteristic curve. While useful, these are of limited value in cases where the input dataset consists solely of unambiguous examples. Furthermore, there are question marks as to whether these metrics are truly representative for the development of spatially-distributed estimates (Martínez-Santos et al., 2021). In those
75   instances, using ad hoc calibration elements, such as complementary field information, can contribute to a better interpretation of the outcomes. Within this context, this research presents two main additions to the literature. In the first place, it explores different scaling methods. The goal is to avoid the pitfalls associated with the reclassification of explanatory variables. Scaling is thus advocated as an essential part of algorithm training, since each subsequent procedure depends on the choice of unit for each feature (Huang et al., 2015). Furthermore, scaling is expected to transform feature values based on a defined rule, so that
80   all scaled features have the same degree of influence (Angelis and Stamelos, 2000). The second novelty has to do with the way the outcomes are evaluated. Borehole flow rates are used as a means to complement standard machine learning metrics, thus providing additional robustness to predictions. This is all demonstrated through the application of machine learning techniques to map groundwater potential across two regions of Mali.

2 Material and methods

85   ### 2.1 Study area

This work pertains to the Koulikoro and Bamako regions of Mali (Fig. 1), which span a joint surface in excess of 90,000 km^2. The study area features three distinct climate zones, including tropical savannah in the south, hot arid steppe in the central and northern parts, and hot arid desert in the north (Traore et al., 2018). Mean yearly temperatures are relatively uniform from a spatial perspective (27 °C), although seasonal oscillations are observed. The coolest temperatures take place in the south in the
90   month of January (19 °C) and the hottest in the north in May (41 °C). There is a clear rainfall gradient, from 300 mm/year in the north to 1200 mm/year in the south.

From a hydrogeological perspective, four major aquifer units are distinguished (Traore et al., 2018). These include basement aquifers, aquifers linked to fractures and intergranular porosity of consolidated sedimentary rocks (Precambrian and 95 Paleozoic), aquifers formed in intrusive volcanic rocks, and aquifers in unconsolidated sedimentary materials.

Basement aquifers are mostly located towards the south of the Koulikoro region. These are characterized by a thick weathered mantle. The average thickness of the weathered formation over the basement in this region is between 10 and 50 meters. Water in these aquifers is preferentially located in the weathered mantle, and, within this, the lower part is generally more transmissive 100 due to lower clay content. The upper part is less permeable to flow, but can still be important as a groundwater reservoir. Fractures can produce significant quantities of water, although their storage capacity is typically low (Martín-Loeches et al., 2018). Borehole yields range from 4 to 6 m^3/hour (Traore et al., 2018).

The Precambrian metasedimentary materials are located in the central part of the Koulikoro region. Metasediments are 105 considered a mixed permeability aquifer: low permeability layers provide higher storage, while more fractured layers present higher permeability and lower storage. Mean aquifer thickness ranges from 30 to 50 meters and the average yield varies from 5 to 10 m^3/hour. Some boreholes however exceed 100 m^3/hour. The Paleozoic rocks located in the north are determined by fractures that allow water to flow through the sandstone and limestone layers. Average borehole yields are around 6 m³/hour and the fractured horizons are about 40-45 m thick. Finally, unconsolidated sedimentary materials are composed by shales and 110 argillaceous sandstone interbedded with limestone. The average borehole yields around 7 m^3/hour. The thickness of the saturated zone ranges from less than 100 m to over 400 m (Traore et al., 2018).
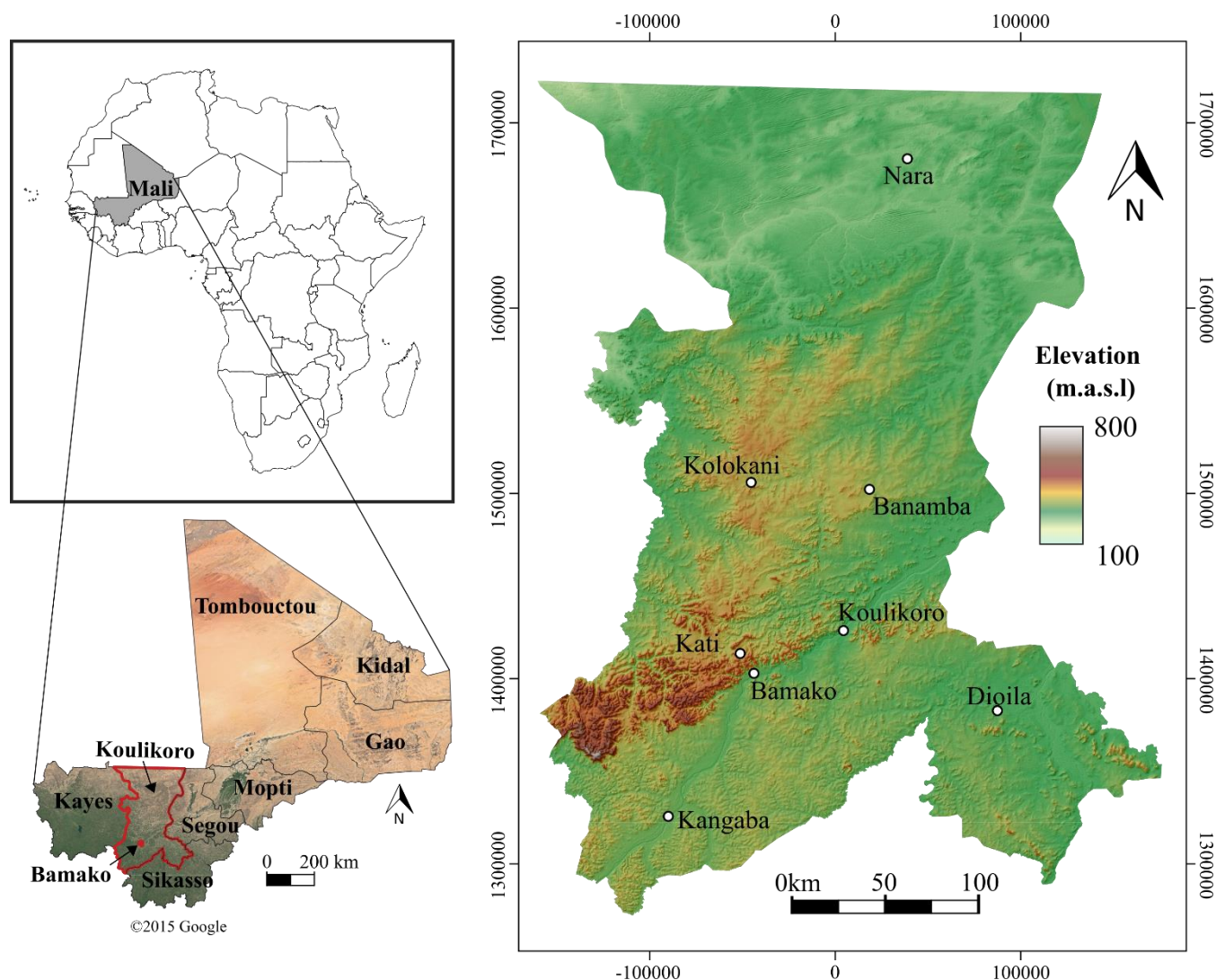
Figure 1. Study area. Koulikoro and Bamako regions located in southwestern Mali (from Map data ©2015 Google).

2.2 Borehole database

Borehole data were provided by Direction Nationale de l'Hydraulique (2010). The database contains information on 5,387 boreholes (3,772 successful and 1,615 unsuccessful), distributed across 1,605 human settlements. The smallest unit of data aggregation is the village level. Available information includes the total number of boreholes per settlement, the borehole success rate, the mean flow rate (m^3/h), and the number of boreholes within a given flow rate range (less than 5 m^3/h, between 5-10 m^3/h and over 10 m^3/h). In most cases, there is also information on the mean borehole depth, mean water table depth and mean electric conductivity (μS/cm). Figure 2 shows that the most common yields range between 1.5 and 4.5 m^3/h, while the mean borehole depth ranges between 50-80 metres. This can be assumed to be the thickness of the weathered formation over

the fresh granite and is roughly consistent with the studies conducted in Burkina Faso (Courtois et al., 2010). Water table depth ranges mostly between 5 and 15 meters.

125 There is a considerable number of villages with a 100% success rate (530), many of which have a single borehole (452). For algorithm training purposes, this raises the question as to whether villages with a small number of boreholes are statistically representative, particularly in cases where the mean yield is low. Besides, it creates an imbalance in the input dataset. Villages with fewer than five boreholes were therefore omitted from the sample.

130
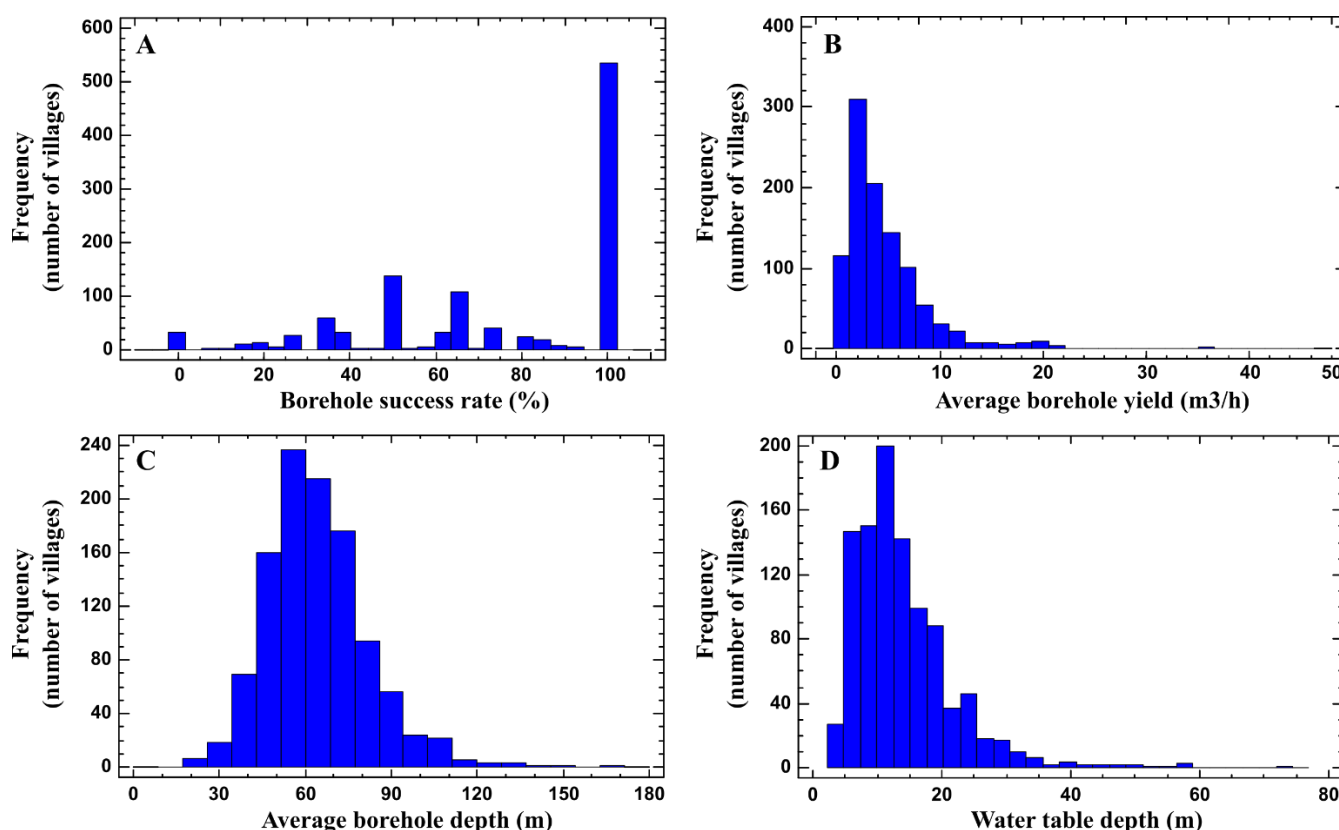
Figure 2. Frequency analysis of the four borehole-related variables. (A) Borehole success rate (B) Average borehole yield (C) Average borehole depth and (D) Water table depth

2.3 Predictive mapping software

135 MLMapper v2.0 is an upgrade of the software developed by Martínez-Santos and Renard (2019). MLMapper is a QGIS3 plugin that allows for the development of predictive maps of a given target variable based on a series of explanatory variables (Fig. 3). MLMapper relies on a variety of supervised learning classifiers from the SciKit-Learn 0.24.1 toolbox (Pedregosa et

al., 2011). Version 2.0 uses 19 different algorithms, including support vector machines (SVC), linear support vector machines (LVC), logistic regression (LRG), decision tree classifier (CRT), random forest classifier (RFC), K-neighbor classification (KNN), linear discriminant analysis (LDA), Gaussian naïve Bayes classification (NBA), multilayer perceptron neural network (MLP), Ada-boost classifier (ABC), quadratic discriminant analysis (QDA), gradient boosting classification (GBC), Gaussian process classifier (GPC), ridge classifier (RID), stochastic gradient descent linear classifier (SGD), perceptron (PER), passive aggressive classifier (PASSI), nu-support vector classifier (nuSVC), and extra-trees classifier (ETC).
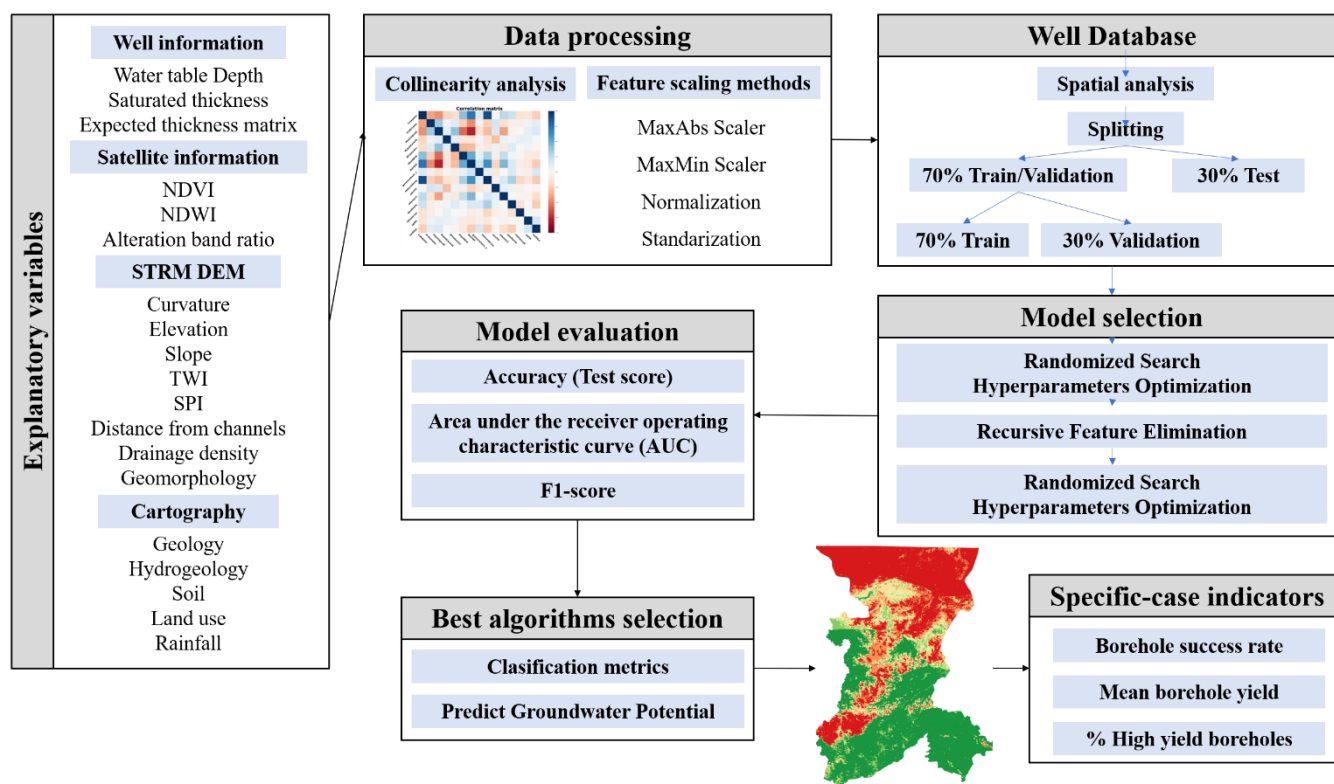
Figure 3. Conceptual model of the predictive mapping procedure with MLMapper v2.0.

2.4 Definition of target and explanatory variables

For the purpose of this study, groundwater potential is defined as the likelihood of a drilled borehole being successful. Successful boreholes are those that yield sufficient water to justify the installation of a hand pump (>0.5 m^3/h) (Foster et al., 2006). The target variable is therefore binary, and can be interpreted as the presence/absence of groundwater.

For algorithm training, villages where more than 50% of wells were known to be successful were labelled "positive". The positive classification also applies to those villages with more than one high yield borehole (>10m^3/h). The others were labelled "negative". The resulting input dataset consisted of 650 villages, out of which 390 were labelled positive and 260 were negative. This comprises information from 3,345 boreholes, out of which 2,101 were successful and 1,244 were unsuccessful.

155

Sixteen explanatory variables were selected based on an extensive review of the GPM literature (Díaz-Alcaide and Martínez-Santos 2019). Explanatory variables include lithology (Fig. 4), landforms, land use, soil, expected thickness matrix (Fig. 5), vegetation-related indices (NDWI, NDVI), clay mineral alteration, drainage density, stream power index, topographic wetness index, clay content, rainfall, expected saturated thickness, water table depth, slope, and slope curvature (Fig. 6). An additional

160   layer with mean borehole flow rates per village was developed for the purpose of calibrating the results.

Geology constrains the presence of groundwater to an important extent. Figure 4 shows the major geological domains of the study area (BGS 2021). The rocks that make up the Precambrian craton (south) are composed mainly of gneiss, schist and quartzite, representing metamorphosed volcanic-sedimentary sequences. The original sedimentary layers, which include shale,

165   arkose, gravel and conglomerate, were intercalated with volcanic rocks, such as basalt, gabbro, dolerite, rhyolite and tuff. This domain is in the south of Koulikoro region. Further north, metasedimentary rocks of Proterozoic age, predominantly low-medium grade metamorphosed sandstones, with varying amounts of mudstone and limestone, take up over 50% of the study area. Volcanic outcrops (basalts and gabbros) are located in the central sector and in the northern end. Sedimentary rocks (sandstone, limestone and shale) of Cambrian-Carboniferous age and Cretaceous-Tertiary age occur in the northern third of

170   the study area. Quaternary fluvial deposits associated with the Niger River are observed along the riverbed (Traore et al., 2018).

Soils are important in GPM because soil characteristics such as permeability, grain shape, grain size, and void ratio control recharge (Fig. 5c). Soil descriptions of the study area were obtained from the European Soil Data Centre (Dewitte et al., 2013).

175   Integration of land use and land cover is often used in groundwater potential mapping studies because human activities alter hydrological dynamics (Díaz-Alcaide and Martínez-Santos, 2019). For instance, croplands and forests, located in the southern part of the study area, are associated with high groundwater potential because ploughing, root development and biological activity favour infiltration (Fig. 5b). Areas close to permanent water bodies also tend to correlate with a higher groundwater potential (Naghibi et al., 2017a). In contrast, urban settlements and wastelands are assumed to have low groundwater potential

180   due to the presence of impervious surfaces and the absence of moisture, respectively (Magesh et al., 2012). Another important variable in terms of aquifer recharge is precipitation, as both can be assumed to be correlated to some extent. Rainfall data in this case represents the mean annual precipitation for the 1950-2009 interval (Fig. 6l).
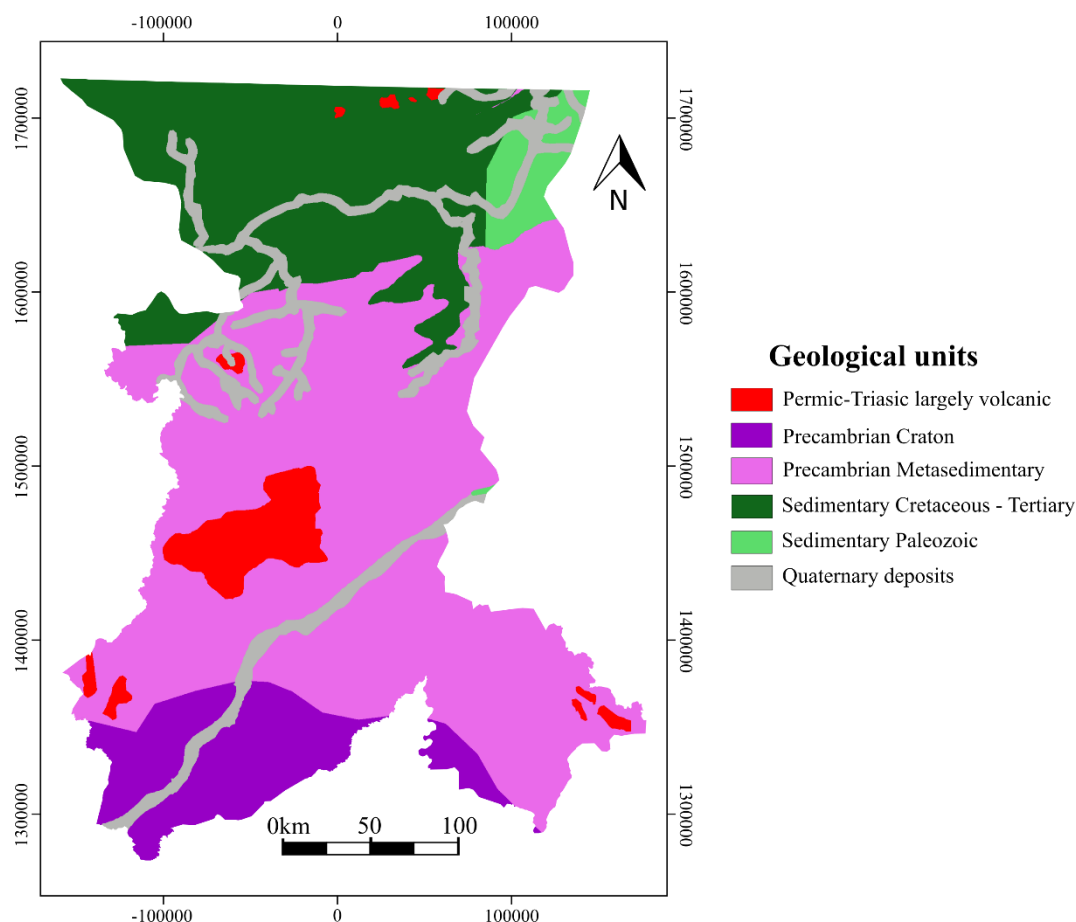
Figure 4. Geological map with the main units that outcrops in the study area (adapted from BGS 2021)

185

Boreholes in the study area are often drilled until the unaltered bedrock is reached. As a result, borehole depth can be a suitable proxy for aquifer thickness (Fig. 5d). Because the borehole database includes static level measurements, an expected saturated thickness layer was computed by subtracting one from the other (Fig. 5d).

190 The digital elevation model (DEM) was obtained from the radar-based Shuttle Radar Topography Mission (SRTM), with a resolution of 1 arcsecond (30 m). DEMs are relevant because shallow groundwater flow and infiltration are partially conditioned by surface features and parameterized by properties that can be extracted from the surface data (Elbeih, 2015). In this case, the DEM was used to develop the curvature, slope, topographic wetness index (TWI), stream power index (SPI) and geomorphology layers (Fig. 6b,c,i,j; Fig. 5a). It also was used to obtain the channel network, which is used in turn to elaborate 195 a drainage density map (Fig. 6h).

The topographic wetness index represents the ease with which water may accumulate at the surface. This is computed for each pixel as per Eq. 1, where α is the local upslope area draining through a certain point per unit contour length and tanβ is the local slope (Beven and Kirkby, 1979).

200    $TWI = \ln(\alpha / \tan\beta)$ (1)

Similarly, the Stream Power Index provides a measure of the erosive power of flowing water. It is calculated as per Eq. 2, where SCA shows the specific catchment area (m^2) and δ represents the local slope gradient calculated (Moore et al., 1991).

$SPI = SCA \times \tan\delta$ (2)

205

The channels extracted from the DEM are used to develop the drainage density and distance from channels maps. Drainage density is computed as the total length of the streams per catchment unit area. Distance from channels was developed by extracting all major channels into a separate layer and developing 500, 1500 and 2500 meter buffers.

210   QGIS 3.0's Geomorphon plugin was used to prepare the landform map (Jasiewicz and Stepinski, 2013). This approach uses DEM for the classification and mapping of landform features based on the principle of pattern recognition, rather than differential geometry. By default the Geomorphon plugin classifies landforms in ten different categories. Because some of them can be expected to play a similar role in the context of GPM, these were subsequently regrouped in four (Fig. 5a).

215   Satellite monitoring does not penetrate deep into the ground, but provides information about features that may be associated with shallow groundwater (Díaz-Alcaide and Martínez-Santos, 2019). This can be important in the case at hand, where the borehole database shows the static level to remain around 5-15 m below the surface throughout most of the study area (Fig. 6a). Vegetation-related indices can be useful in this context, particularly when computed at the end of the dry season (Fig. 6e,f). Take for instance the normalized difference vegetation index (NDVI), which is an estimate of vegetation vigour and is
220   derived from the response of vegetation to red and visible infrared wavelengths (Xie et al., 2008). Similarly, the normalized difference water index (NDWI) is used as a measure of the amount of water in the vegetation or soil moisture (Xu, 2006). Based on Landsat 8 products, the NDVI and the NDWI are computed as per Eq. 3 and Eq. 4, respectively, where B3 represents the green band (0.53 - 0.59 μm), B4 is the red band (0.64 - 0.67 μm) and B5 is the near infrared band (0.85 - 0.88 μm).

$NDVI\ (Landsat\ 8)\ =\ (B5 - B4)\ /\ (B5 + B4)$ (3)

225    $NDWI\ (Landsat\ 8)\ =\ (B3 - B5)\ /\ (B3 + B5)$ (4)

Clay content in the first few meters of the surface largely determines the percolation of water into the aquifer. Therefore, an additional clay content layer (g/kg) in the top two meters of the terrain was considered (Poggio and de Sousa, 2020). This layer

is obtained by state-of-the-art machine learning methods that use global soil profile information and covariate data to model
the spatial distribution of soil properties around the world. (Fig. 6k). To complement the information on clay content on the

230   subsurface, an additional layer has been developed by combining bands 6 and 7 of Landsat 8 (Ourhzif et al., 2019). This layer
provides information on clay content on the surface and the relationship with infiltration. Clay content on the surface is
calculated as per Eq. 5, where B6 is the short-wave infrared 1 and B7 the short-wave infrared 2.

$$Alteration\ (clay\ minerals\ reflectance)\ =\ B6\ /\ B7 \tag{5}$$

235
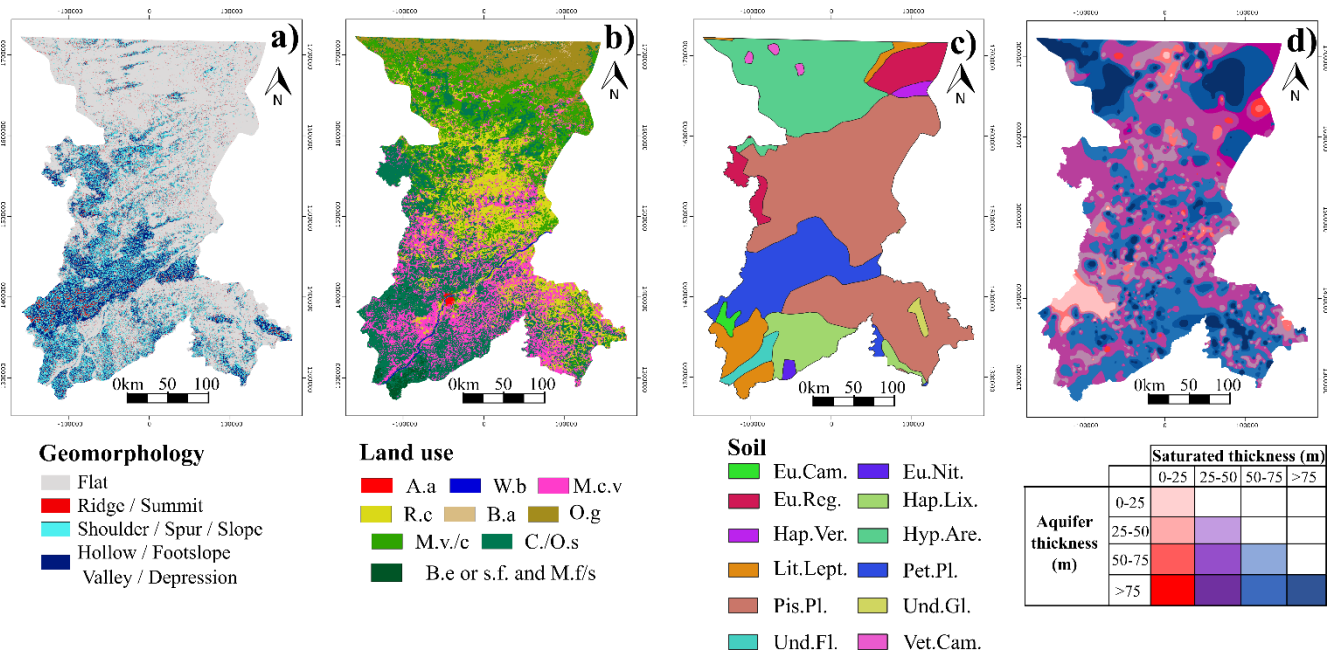
Figure 5. Explanatory variables used to predict the GPM: a) geomorphology b) Land use (A.a = artificial areas; W.b = water bodies; M.c.v
= Mosaic cropland vegetation; R,c = Rainfed cropland; B.a = Bare areas; O.g = Open grassland; M.v./c = Mosaic vegetation/cropland; C./O.s
= Close to open shrubland; B.e or s.f. and M.f/s = Broadleaved evergreen or semidecidous forest and Mosaic forest / shrubland) c) Soil
(Eu.Cam. = Eutric Cambisols; Eu.Nit. = Eutric Nitrisols; Eu.Reg. = Eutric Regosols; Hap.Lix. = Haplic Lixisols; Hap.Ver. = Haplic

240   Vertisols; Hyp.Are. = Hypoluvic Arenosols; Lit.Lept. = Lithic Leptosols; Pet.Pl. =Petric Plinthosols; Pis.Pl. = Pisoplinthic Plithosols;
Und.Gl. = Undifferentiated Gleysols; Und.Fl. = Undifferentiated Fluvisols; Vet.Cam = Vetric Cambisols) d) expected thickness matrix.
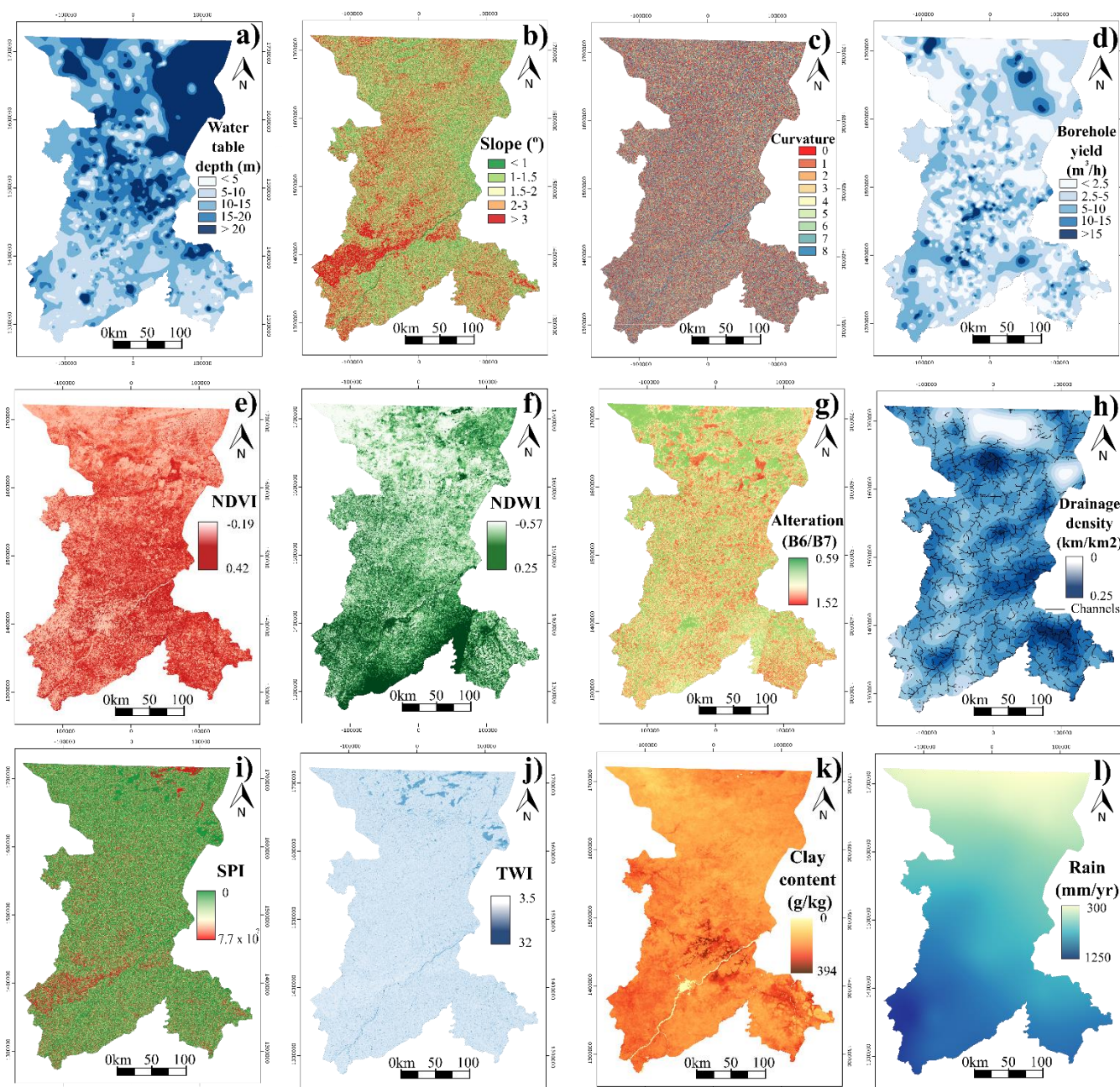
Figure 6. Explanatory variables used to predict the GPM: a) water table depth (meters) b) slope (degree) c) curvature d) borehole yield (m3/h) e) normalized difference vegetation index (NDVI) f) normalized difference water index (NDWI) g) alteration band ratio (B6/B7) h) Drainage density i) Stream power index (SPI) j) topographic wetness index (TWI) k) Clay content (g/kg) l) rainfall (mm/year)

2.5 Preprocessing of explanatory variables

Machine learning algorithms are designed to find meaningful associations between the explanatory variables and a given target variable. Preprocessing of explanatory variables is important to avoid bias. Take for instance the issue of ranges. Rainfall in the study area oscillates between 300 and 1200 mm/year, while NDVI and NDWI values only range between -1 and +1. Since
250   some algorithms attribute weight to explanatory variables based on sheer magnitude, scaling is advisable to prevent rainfall from having an undue bearing on the results. Furthermore, those algorithms that use gradient descent as an optimization technique (logistic regression, neural networks), require data to be scaled because range differences lead to different step sizes in the gradient descent formula for each feature, and computational time increases as a result. Algorithms such as K-neighbors and support vector machines are affected by different ranges because these classifiers rely on distances between data points to
255   determine their similarity.

A major goal of this study is to analyse and compare different methods to preprocess explanatory variables. The idea is to rely on raw data as much as possible, instead of reclassifying it into intervals generated statistically or by expert criteria. Four scaling methods were therefore used: standardization, maximal absolute scaler (MaxAbs), maximal-minimal scaler (MaxMin)
260   and normalization (Pedregosa et al., 2011).

The standardized explanatory variables are obtained by removing the mean and scaling to unit variance (Zheng and Casari, 2018). Centering and scaling happen independently on each explanatory variable by computing the relevant statistics on the samples in the training set. Mean and standard deviation are then stored to be used on later data using transform. This is
265   obtained by means of Eq. 6, where x is the value of the sample, u is the mean of the training samples, and s is the standard deviation of the training samples.

$$\tilde{x} = (x - u) / s \tag{6}$$

The MinMax method transforms features by scaling each explanatory variable to a given range. This estimator scales and
270   translates each feature individually so that it lies within the range given in the training set (between zero and one in the case at hand). Thus, this scaling method is robust to small standard deviations in feature data (Pedregosa et al., 2011). MinMax scaling is performed as per Eq.7.

$$\tilde{x} = (x - min(x))/(max(x) - min(x)) \tag{7}$$

275   The MaxAbsScaler works in a similar way, but the training data is scaled to fall within the [-1, 1] range by dividing by the largest maximum value of each feature. Finally, under the normalizer procedure, each sample is rescaled regardless of other

samples so that its norm ('L1', 'L2' or 'max') equals one (Pedregosa et al., 2011). The L1 norm attempts to minimize the sum of the absolute differences between the target value and the estimated values. In turn, the L2 norm uses least squares, while the L-Max norm rescales by the maximum of the absolute values. The L2 norm is by far the most common and was adopted as the procedure of choice.

2.5 Supervised classification routine

MLMapper incorporates a series of routines to enhance the conventional train/test process. These include collinearity checks, random-search parameter fitting, and recursive feature elimination. A selection of the best algorithms and a basic arithmetic ensemble is also performed at the end in order to appraise the degree of agreement among classifiers.

Collinearity occurs when two or more variables are highly correlated. This can affect the performance of the classifiers by attributing extra weight to an input variable or by adding noise to the final outcomes. Interpretability can also be impaired because the regression coefficients of certain algorithms are not uniquely determined (Martínez-Santos et al., 2021). MLMapper incorporates a collinearity analysis function to prevent collinearity from adversely affecting the results. The collinearity analysis is performed before running the algorithms.

Random-search parameter fitting increases the accuracy of the predictions by identifying the best combination of those parameters that govern each algorithm. The random search cross-validation function needs an algorithm, a scoring metric to evaluate the performance of the different hyperparameters and a dictionary with the hyperparameter names and values. In regard to the previous version of MLMapper, which used grid-search cross validation, this provides additional flexibility and reduces computational time. The number of iterations was established at 500 in the case of this study.

Once the optimal parameters distributions have been selected, recursive feature elimination cross validation (RFECV) is performed. RFECV is only available for those algorithms that use feature importance or coefficient weight attributes. The goal of RFECV is to identify the most important explanatory variables for each algorithm, as well as to eliminate those features that either incorporate noise or detract from the accuracy of predictions.

Classifiers are trained first on the initial set of explanatory variables. The weight of each feature is obtained, either through a coefficient attribute or through a feature importance attribute. The least important explanatory variables are then pruned from the model. This procedure is recursively repeated on the pruned set until the optimal number of features to select is eventually reached (Pedregosa et al., 2011). Once RFECV is complete, algorithms are run again on the input dataset (excluding noise and counterproductive variables) and hyperparameters are re-optimized. This step yields important information for the purpose of interpreting the maps, as it provides the user with the relative weight of explanatory variables.

310

Each algorithm predicts groundwater potential throughout the study area, assigning a value of 1 to pixels with positive groundwater potential and 0 to pixels where groundwater is expected to be absent. Each algorithm operates differently and relies on a different combination of explanatory variables, which inevitably leads to discrepancies in the predictions. In order to analyse the degree of agreement between the classifiers, an ensemble map is developed by computing the arithmetic mean

315   at the pixel scale of those algorithms exceeding 0.85 predictive accuracy. Green pixels mean that all the best-performing algorithms agreed on a positive groundwater potential outcome (arithmetic mean = 1). Conversely, red zones represent those pixels where all the best-performing algorithms agreed on a negative groundwater potential (arithmetic mean = 0). Intermediate colours represent various degrees of agreement among the algorithms.

320   **2.6 Machine learning metrics for algorithm evaluation**

Outcomes of machine learning studies are often evaluated on the grounds of accuracy (test score) and the area under the receiver operating characteristic curve (AUC). Test score is calculated as the number of successful predictions over the total number of attempts in the test dataset, thus providing a direct measure of predictive accuracy. In turn, the receiver operating characteristic curve shows the performance of a classification model at all classification thresholds. The AUC is a probabilistic

325   metric that assesses the degree to which algorithms can distinguish between classes. Test score and AUC rank algorithms on a scale of 0 to 1, a higher score implying better performance. If the classes are balanced, a score of 0.5 on either metric suggests that the prediction is no better than a random estimate.

Because the ratio between positive and negative classes in the input dataset is approximately 3:2, the F-1 metric was also used

330   to analyse the quality of the predictions in each of the classes separately. F1-score is the harmonic mean of precision and recall. Precision is calculated like the ratio of true positives over the sum of true and false positives. Precision therefore shows the classifier's ability not to label as positive a sample that is negative. In turn, recall is the ratio of true positives over the sum of true positives and false negatives. In other words, recall represents the ability of the classifier to find all positive samples.

3 Results and discussion

335   **3.1 Collinearity analysis**

Figure 7 presents the results of the collinearity analysis. Pair-wise correlation coefficients are expressed as a colour palette ranging from -1.0 (inverse correlation) to 1.0 (direct correlation). There is no hard threshold as to what is an acceptable level of correlation between two variables, though the literature shows that values between 0.4 and 0.85 may be acceptable (Dormann et al., 2013). In this case, no strong correlations were found among explanatory variables except when appraising the expected

340   saturated thickness versus the expected thickness matrix. This pair renders a correlation coefficient of 0.85, which is close to

the upper acceptability threshold. Because the number of non-surface features in the explanatory dataset is limited, a closer look was taken at the importance of these two variables. RFECV reveals that both are relied upon by the best performing algorithms. In addition, the thickness matrix has seldom been used in the literature. These reasons led to the decision of keeping both variables.
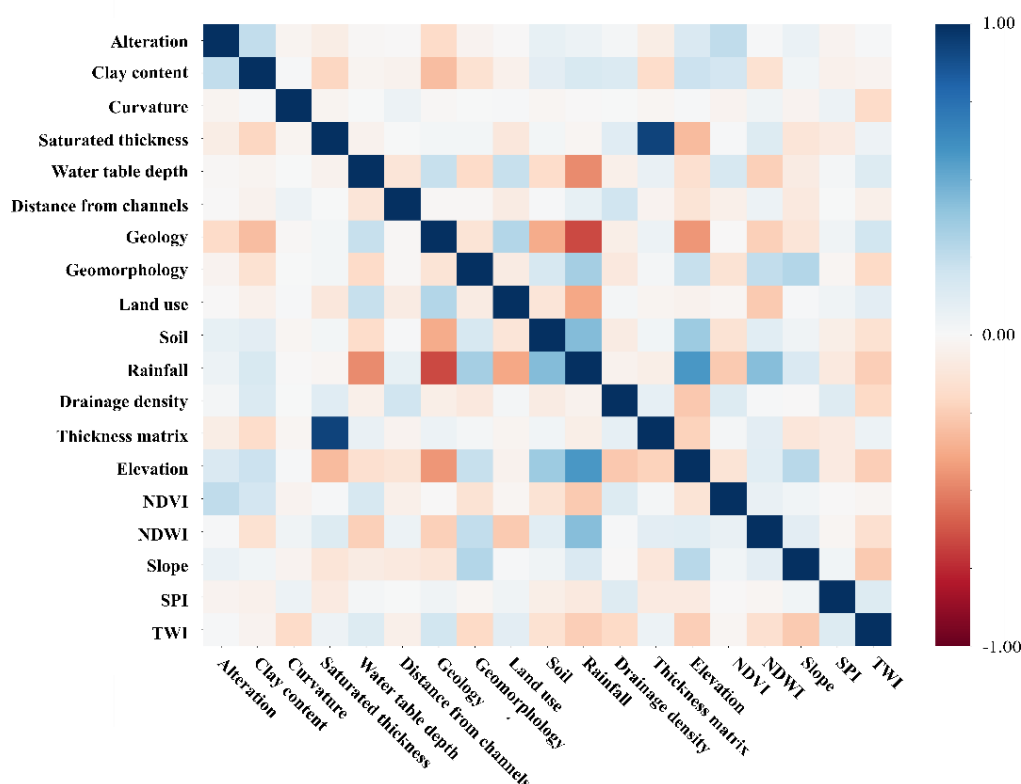
Figure 7. Pair-wise collinearity analysis for all explanatory variables.

3.2 Model evaluation and scaling method selection

Table 1 shows the results of the best algorithms for each scaling method. The five best algorithms for each scaling method were selected and the mean of each metric was computed to analyse performance. All scaling methods render a mean test score for the top five algorithms in excess 0.85, which implies that the predictions are accurate in all cases. The standardization scaling method presents the highest value (0.867), followed by the MaxAbs method (0.866), the normalization method (0.860) and the MaxMin method (0.857). All algorithms render a higher F1 score for the positive outcome, which is the most common one in the input database. However, the F1 score for the negative class is always higher than 0.81. This indicates that the algorithms are capable of distinguishing positive and negative classes accurately.

The AUC exceeds 0.90 in all cases. This implies that all the best-performing algorithms have a high probability of adequately predicting the target variable. When comparing the scaling methods, the standardized presents the highest scores for test score and F1 score for the negative class while the MaxAbs method present the second-best performance in terms of test scoring, F1

360   score for the negative class and AUC metrics. Hyperparameter optimization by the random search procedure led to the scores in Table 2.

Table 1. Results of the top five algorithms for each scaling method. Refer to section 2.6 for the definition of each metric (Test score = optimized test score; F1 score 0 = f-1 score false; F1 score 1 = f-1 score true; AUC = area under curve). The mean of each metric has been

365   calculated for the different reclassification methods to compare the results.

Scaling method	Better algorihms	Test Score	Test score mean	F1 score 0	F1 score 0 mean	F1 score 1	F1 score 1 mean	AUC	AUC mean
Standarize	Decision Tree	0.872		0.850		0.890		0.893	
	Random Forest	0.852		0.820		0.870		0.911	
	AdaBoost	0.862	0.867	0.830	0.836	0.880	0.886	0.912	0.909
	Gradient Boosting	0.872		0.840		0.890		0.914	
	Extra Trees	0.878		0.840		0.900		0.914	
Normalization	RandomForestClassifier	0.847		0.810		0.870		0.901	
	GradientBoostingClassifier	0.857		0.830		0.880		0.905	
	SGDClassifier	0.857	0.860	0.830	0.828	0.880	0.882	-	0.907
	PassiveAggressiveClassifier	0.862		0.830		0.880		-	
	ExtraTreesClassifier	0.878		0.840		0.900		0.916	
MaxMin	Decision Tree	0.857		0.820		0.880		0.913	
	Random Forest	0.857		0.830		0.880		0.912	
	Gradient Boosting	0.847	0.857	0.820	0.824	0.870	0.880	0.903	0.911
	Ridge Classifier	0.842		0.800		0.870		-	
	Extra Trees	0.883		0.850		0.900		0.918	
MaxAbs	Decision Tree	0.867		0.840		0.890		0.912	
	Random Forest	0.852		0.820		0.870		0.911	
	AdaBoost	0.867	0.866	0.830	0.834	0.890	0.886	0.898	0.910
	Gradient Boosting	0.872		0.840		0.890		0.916	
	Extra Trees	0.872		0.840		0.890		0.914	

The results in Table 1 show that the best scaling methods based on the score obtained in machine learning metrics are standardized and MaxAbs. Therefore, the focus of ensuing analyses will be placed on these two.

370

Noticeably, the best performing classifiers all belong in the family of tree-based algorithms. Decision Tree (DTC), Random Forest (RFC), AdaBoost Classifier (ABC), Gradient Boosting (GBC) and ExtraTrees (ETC) consistently outperformed other

classification methods throughout the process. While tree-based algorithms do not necessarily require feature scaling (Li et al. (2017)), the outcomes of this experience show that different scaling methods lead to slightly different performances. The best

375 test scores were obtained by the ETC algorithm (0.878 and 0.872, for standardize and MaxAbs methods respectively), while the test scores of the other algorithms remained above 0.85. The F1-score for the positive groundwater potential is above 0.87 in all cases, while the F1-score for the negative potential decreases slightly (0.82 to 0.85). AUC scores oscillate between 0.89 and 0.91, which reinforces the idea that the best-performing classifiers are able to accurately predict the target.

380 By their own nature, tree-based algorithms can be expected to yield good results in groundwater potential studies. DTC is popular technique, largely due to the ease with which the internal logic of the algorithm can be interpreted (Naghibi et al., 2016; Moghaddam et al., 2020). In turn, RFC is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). RFC is one of the most powerful machine learning classifiers in general, and typically renders useful outcomes in predictive mapping

385 (Naghibi et al., 2017b; Martínez-Santos and Renard, 2020). The GBC algorithm is an ensemble method in which multiple weak classifier trees are used to build a strong classifier (Friedman, 2001). In the same way, the basic principle of ABC (Freund and Schapire, 1997) is to fit a sequence of weak learners into repeatedly modified versions of the data. The predictions of all of them are combined by a weighted majority vote to produce the final prediction. The data modifications at each iteration of the so-called boosting consist of applying weights to each of the training samples (Pedregosa et al., 2011). The literature shows

390 that GBC and ABC have been successfully used as ensemble methods in the development of GPM maps (Nguyen et al., 2020a; Martínez-Santos and Renard, 2020). Finally, for the ETC algorithm (Geurts et al., 2006), randomization goes a step further in the way splits are computed. As in random forests, a random subset of candidate features is used, but instead of looking for the most discriminative thresholds, random thresholds are extracted for each candidate feature and the best of these randomly generated thresholds is chosen as the splitting rule. This usually allows the variance of the model to be reduced somewhat

395 further (Pedregosa et al., 2011).

Other algorithm families have been successfully applied in GPM research. Support vector machines (Martínez-Santos and Renard, 2020), logistic regression (Ozdemir, 2011; Chen et al., 2018), neural networks (Moghaddam et al., 2020; Panahi et al., 2020) have all been demonstrated to be efficient predictors of groundwater potential, even when pitched against tree-based

400 algorithms. This suggests that differences in input databases and the choice of explanatory variables may constrain the applicability of each algorithm in each given context. Furthermore, the inherent complexity of supervised classification makes it difficult for users to understand the internal logic of the algorithms, which means it is frequently impossible to predict which one will perform best in a given situation. Incorporating various algorithms families in GPM is therefore perceived as a suitable course of action in GPM studies.

405

Table 2. Algorithm hyperparameters (description after Pedregosa et al., 2011) and adjusted outcomes. Max_depth = The maximum depth of the tree; max_features = The number of features to consider when looking for the best split.; min_samples_leaf = The minimum number of samples required to be at a leaf node; min_samples_split = The minimum number of samples required to split an internal node; random_state = Controls both the randomness of the bootstrapping of the samples used when building trees and the sampling of the features to consider when looking for the best split at each node.; n_estimators = Number of trees in the ensemble methods; algorithm = method used by Ada Boost (discrete SAMME boosting algorithm or real SAMME.R boosting algorithm); learning_rate = shrinks the contribution of each classifier.

410

Classifier	Hyperparameter	Optimized parameter value	
		Standardize method	MaxAbs method
Decision Tree	max_depth	3	6
	max_features	0.9	0.6
	min_samples_leaf	8	5
	min_samples_split	0.1	0.1
	random_state	0	0
Random Forest	max_depth	6	5
	max_features	1.0	1.0
	min_samples_leaf	8	8
	n_estimators	141	196
	random_state	0	0
Ada Boost	Algorithm	'SAMME'	SAMME.R
	learning_rate	0.795	0.109
	n_estimators	340	170
	random_state	0	0
Gradient Boosting	max_depth	2	2
	max_features	0.6	0.2
	min_samples_leaf	20	20
	n_estimators	100	160
	random_state	0	0
Extra Trees	max_depth	9	8
	max_features	0.8	0.8
	min_samples_leaf	1	5
	n_estimators	520	350
	random_state	0	0

3.3 Importance of explanatory variables

415 Naghibi and Pourghasemi (2015) indicate that the importance of explanatory variables in groundwater potential mapping is considerably influenced by the approach used in an investigation and the study area properties. Outcomes shows that elevation, rainfall, geology and drainage density, among others, are the most important factors conditioning the groundwater potential, which is in agreement with results obtained by other authors in different geographical contexts (Ozdemir, 2011; Naghibi and Pourghasemi, 2015; Nguyen et al., 2020b).

420

A major advantage of incorporating recursive feature elimination is that it eliminates part of the expert bias associated with the choice of explanatory variables. In this case, the fact that all variables are used by at least two of the best-performing algorithms suggests that the initial choice of explanatory variables was appropriate (Fig. 8). However, feature selection reveals clear differences among the classifiers. Under standardized scaling, RFC only required three explanatory variables to predict

425 groundwater potential (precipitation, expected saturated thickness and elevation). ABC and GBC used eight each, while ETC and DTC used eleven and fourteen, respectively. Under MaxAbs scaling, the RFC and ABC used three and four variables, respectively. GBC algorithm worked with six, DTC used sixteen and ETC seventeen.

All algorithms agree on the importance of elevation and precipitation. The combined importance of these two variables for the

430 RFC, GBC, ABC, DTC and ETC algorithms is 0.90, 0.71, 0.47, 0.90, and 0.42, respectively, under the standardize scaling method. For the MaxAbs method, the combined weights amount to 0.90, 0.51, 0.64, 0.87 and 0.47, respectively. Other commonly used features include the expected saturated thickness (eight out of ten cases), drainage density (six), NDWI (six), geology (five), clay content (five) and NDVI (five). The relative weight of other variables varies depending on the algorithm and scaling method, but most agree on the importance of expected saturated thickness, slope, geology and drainage density.

435 The least used variables are the alteration layer, curvature, soil and SPI.

3.4 Groundwater potential maps

Classifier outcomes were extrapolated to produce groundwater potential maps. Figure 9 shows the groundwater potential predictions rendered by each of the five best-performing algorithms under the two most effective scaling methods. Red areas are those in which the algorithms have found a combination of explanatory variables leading to a negative potential. In turn,

440 green zones represent a positive groundwater potential. All maps show a gradient characterized by the predominance of positive areas in the south to a greater proportion of negative areas in the north. This appears to be related to rainfall patterns. Other major negative zones occur around mountain outcrops of the southwest.
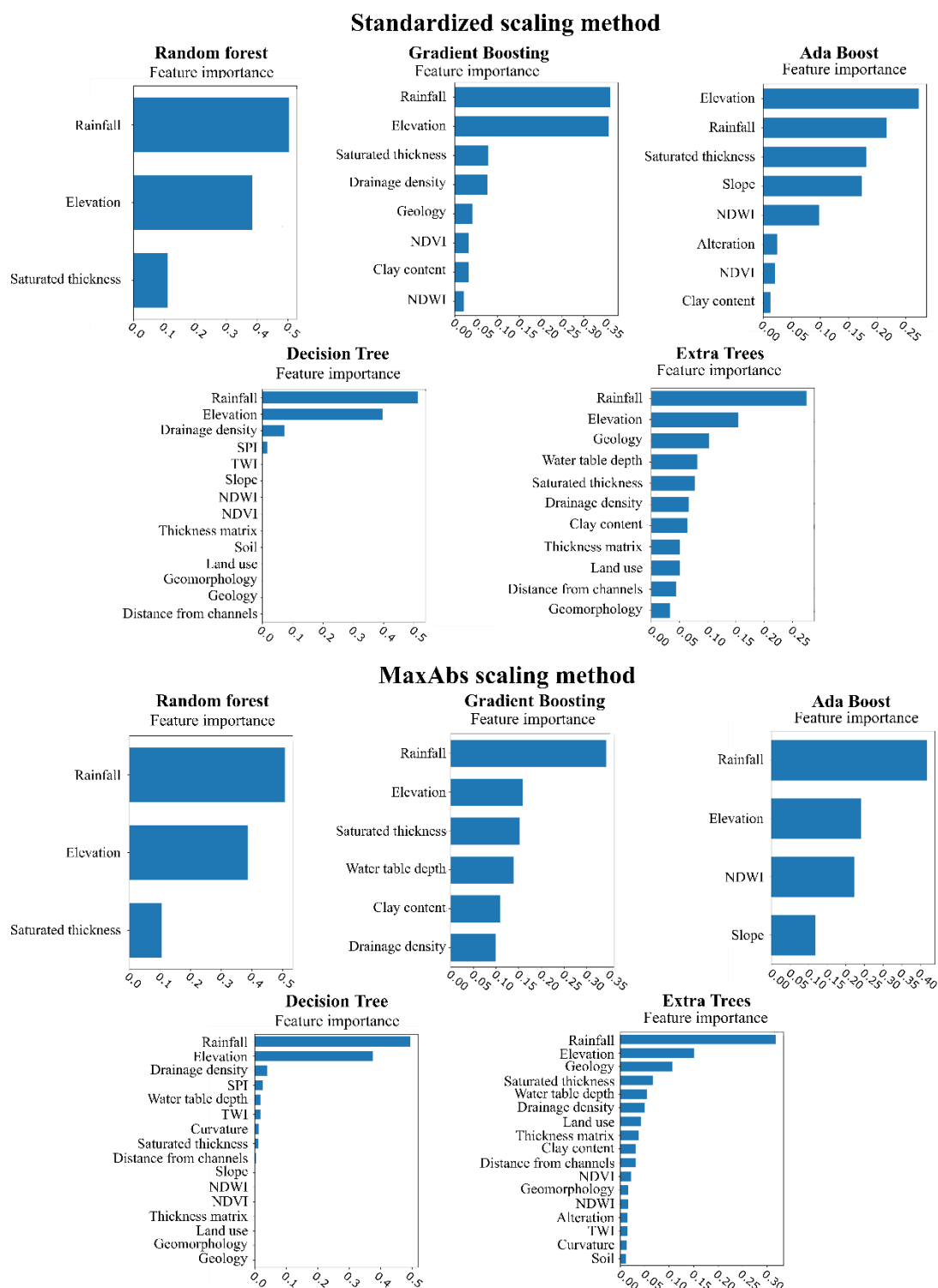
Figure 8. Feature importance calculated for the best performing tree-based algorithms using the standardized and MaxAbs scaling methods.
445 The sum of all variable weights equals one.

Overall, maps are remarkably similar, although the MaxAbs scaling method renders a slightly greater proportion of negative potential areas. The agreement map (Fig. 10) allows for an analysis of discrepancies among the best performing algorithms. The MaxAbs scaling method predicts a low groundwater potential (pixel value 0-0.2) for 41% of the study area, while about 40% is identified as a high groundwater potential (>0.8) and the remaining 19% consists of zones with moderate potential (0.2 450 - 0.8). The standardized scaling method renders a high groundwater potential for 42%, low groundwater potential for 38%, and various degrees of moderate potential for 20% of the region.
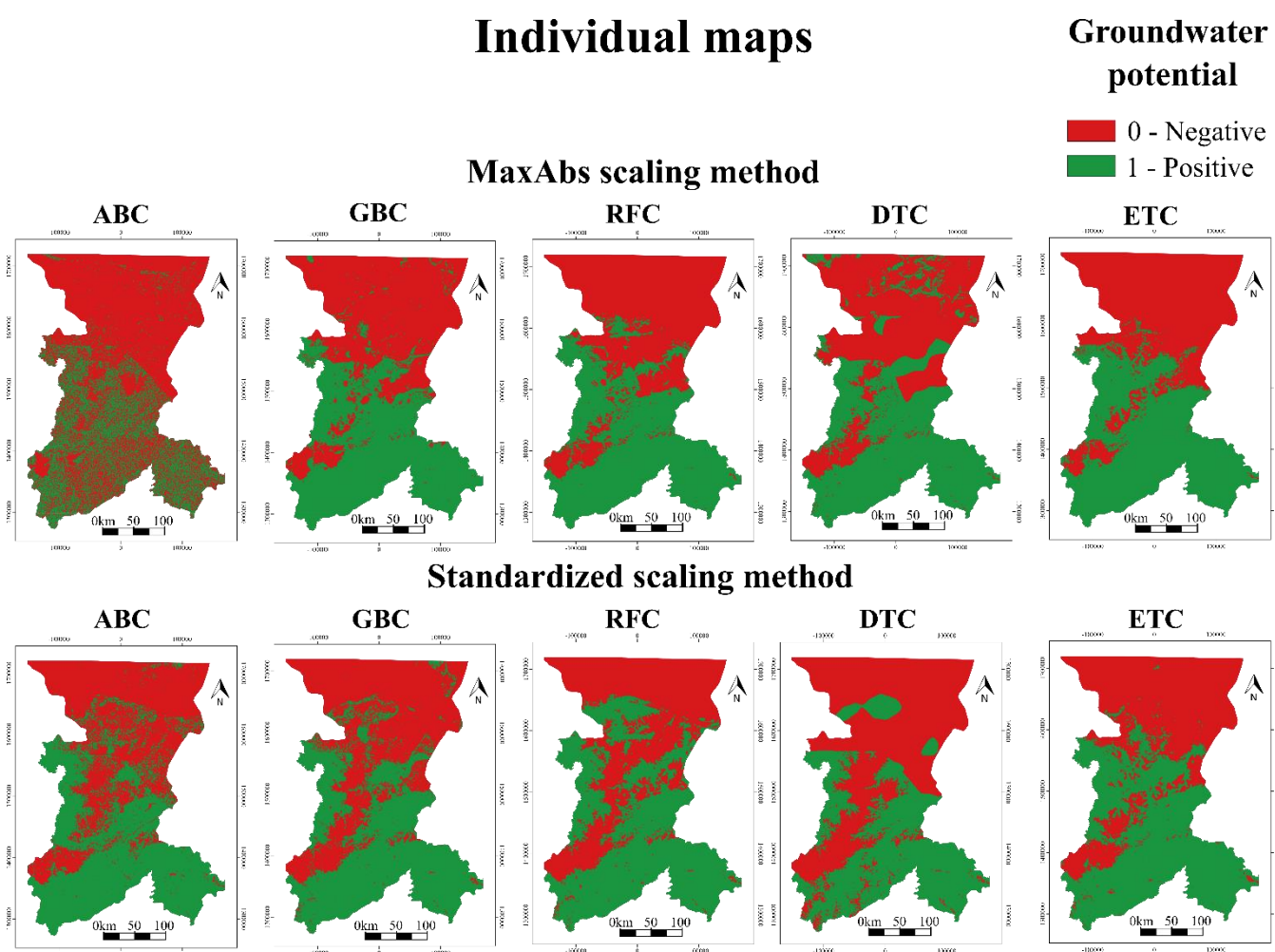
Figure 9. Mapping outcomes of the top five supervised classification algorithms for the two best performing scaling methods. At the top the 455 MaxAbs scaling method, below it the standardized scaling method. From left to right: AdaBoost classifier, Gradient Boosting classifier, Random Forest classifier, Decision Tree classifier and Extra Trees classifier.

Martínez-Santos et al. (2021) show that standard machine learning metrics do not necessarily provide an unambiguous measure of classifier performance when predicting spatially-distributed outcomes. These authors argue that case-specific indicators

460   could provide additional insights on the results. In this case, an independent dataset of groundwater data (borehole yield) was
       used to cross-check machine-learning predictions. Villages featuring more than five boreholes (n=334) were relied upon for
       this purpose. These villages were classified in three groups based on the predicted groundwater potential ("low" <0.2;
       "moderate" 0.2 to 0.8; and "high" >0.8). Table 3 shows the mean borehole success rate, the mean borehole yield, and the
       percentage of boreholes with a flow rate in excess of 10 m^3/h in each category. The average water table depth is also provided
465   for reference.

Agreement maps

A) MaxAbs scaling method   **B) Standardized scaling method**
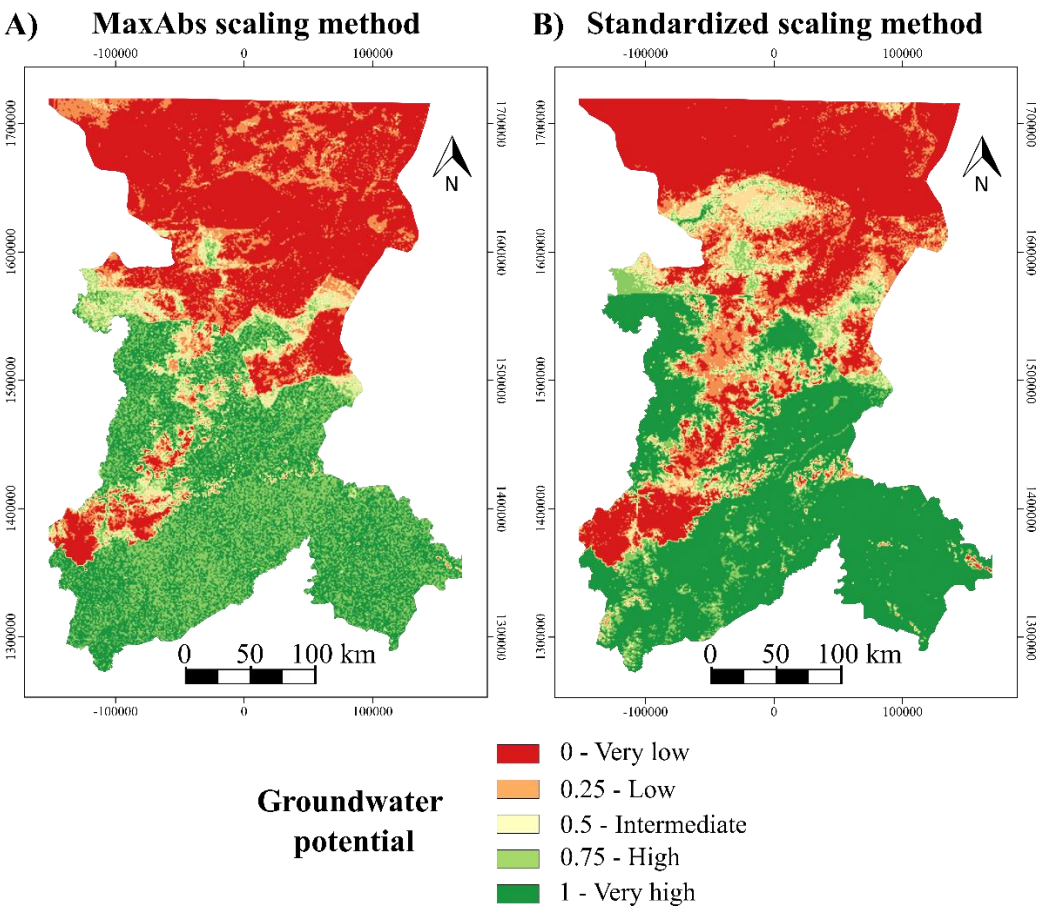
Figure 10. Mapping outcomes of the agreement map for (A) MaxAbs scaling method (B) Standardized scaling method.

470   A first conclusion that springs to mind is that about 60% of the villages with more than five boreholes fall within the high
       groundwater potential description, while a further 20% present a moderate groundwater potential. To an extent, this was to be

expected, as the larger populations develop in areas where groundwater is available. Table 3 also shows groundwater potential outcomes to match borehole flow rates reasonably.

475 The mean borehole success rate in areas labelled as high groundwater potential is 75.9% for the MaxAbs scaling method and 73.7% for the Standardized method. In contrast, the mean success rate for villages in the low groundwater potential category is 33.6% and 32.1%, respectively. These outcomes suggest that both scaling methods and, in general, the machine learning approach, are adept at identifying areas where boreholes are more likely to be successful.

480 Both the mean borehole yield and the percentage of high yield boreholes ($>10 m^3/h$) provide interesting benchmarks for comparison. The mean borehole yield is consistent with map outcomes. Under the MaxAbs scaling method, the mean borehole yield is 1.39 m^3/h in the villages located within low groundwater potential areas and 3.92 m^3/h in areas where groundwater potential is predicted to be high. Furthermore, 76.1% of the high yield boreholes fall within high groundwater potential areas, while just 6.6% are located in low groundwater potential areas. Results are similar for the standardized scaling method. Mean 485 borehole yield is 1.44 m^3/h and 3.63 m^3/h for the low and high potential categories, respectively, while 74.6% of the high yield boreholes are located in villages within high groundwater potential zones.

Table 3. Results of the analysis based on the agreement map categories for the MaxAbs and Standardized scaling methods. The analysis was performed for villages with more than 5 wells. The average success rate, average flow rate, average flow rate with the negative boreholes 490 and water table depth as well as the percentage of wells with a yield higher than 10m³/h are shown.

Groundwater potential	Number of villages	Borehole success rate (%)	Mean borehole yield (m³/h)	Mean borehole yield - only positive boreholes (m³/h)	% High yield boreholes (>10 m³/h)	Average water table depth (m)
MaxAbs scaling method						
Low	70	33.6	1.39	3.58	6.6%	18.6
Moderate	69	50.5	2.23	4.36	17.3%	15.1
High	195	75.9	3.92	5.10	76.1%	12.1
Standardized scaling method						
Low	65	32.1	1.44	3.96	6.6%	19.3
Moderate	70	55.5	2.78	4.57	18.8%	16.1
High	199	73.7	3.63	4.87	74.6%	11.8

On a final note, the literature features few examples of groundwater potential studies in the study area. Perhaps the only systematic precedent is the one carried out by Díaz-Alcaide et al. (2017). These authors performed a national-scale assessment of groundwater potential for the Republic of Mali based on the same borehole database that has been used in this research.

495   They however aggregated their data at the commune scale, rather than at the village scale, and relied on different classification principles. Their outcomes are thus to be interpreted at a different spatial resolution. This notwithstanding, their groundwater potential assessment for the Bamako and Koulikoro regions is similar to the one obtained by means of machine learning classification. The lower groundwater potential areas were identified in the north and the higher groundwater potential ones were found mostly in the south. This attests to the capacity of machine learning approaches to realistically depict field

500   conditions, and thus, to underpin groundwater development.

4 Conclusions

Machine learning applications are gaining recognition as valuable tools to underpin groundwater resources management. The ease with which machine learning algorithms find complex associations within large datasets, together with their ability to develop accurate predictions, open up a whole new dimension to the analysis of groundwater data. Artificial intelligence

505   approaches may improve water access in areas such as the Sahel by providing additional guidance to borehole drilling initiatives. Within this context, our research combined state-of-the-art machine learning tools, including collinearity checks, recursive feature selection, feature importance computation and cross-validation mechanisms to map groundwater potential in two regions of Mali. Tree-based algorithms consistently outperformed other algorithm families such as support vector machines and neural networks. Since this cannot be expected to occur in every case, we advocate the use of a large number of

510   machine learning classifiers and the subsequent selection of the best performers as a sensible course of action. The same logic entails that a range of scaling methods should be used to prevent expert-based bias associated with the reclassification of explanatory variables.

A crucial finding of this research is that conventional machine learning metrics (test score, area under the receiver operating

515   characteristic curve), can be more representative of algorithm performance than of the actual field conditions. This is particularly relevant when attempting to develop spatially-distributed predictions. Double-checking algorithm results with an independent groundwater dataset (borehole flow rates in this case) is recommended to ensure that map outcomes are accurate. Machine learning approaches are thus seen as a means to underpin borehole siting initiatives at the regional scale, although it is recognized that local-scale fieldwork is needed for optimal outcomes.

520   *Author contribution*

VGE: Conceptualization, Formal analysis, Methodology, Software, Writing – original draft preparation. PMS: Conceptualization, Methodology, Software, Funding acquisition, Writing – review & editing. MML: Conceptualization, Writing – review & editing

Competing interests

525   The authors declare that they have no conflict of interest.

Acknowledgements

This work has been funded under research grant RTI2018-099394-B-I00 of Spain's Ministerio de Ciencia, Innovación y Universidades. The first author received an FPI grant from the Ministerio de Ciencia e Innovación to develop his PhD within this project PRE2019-090026. The second author received a Salvador de Madariaga grant (PRX18/00235) from Spain's

530   Ministerio de Educación, Cultura y Deporte to carry out a 3-month research stay at the Université de Neuchâtel, Switzerland, where the original version of the software used in this paper was developed. The authors thank the Direction Generale de l'Hydraulique of Mali for making available its borehole database.

References

Al-Djazouli, M. O., Elmorabiti, K., Rahimi, A., Amellah, O., and Fadil, O. A. M.: Delineating of groundwater potential zones
535   based on remote sensing, GIS and analytical hierarchical process: a case of Waddai, eastern Chad, GeoJournal, https://doi.org/10.1007/s10708-020-10160-0, 2020.

Al-Fugara, A., Pourghasemi, H. R., Al-Shabeeb, A. R., Habib, M., Al-Adamat, R., Al-Amoush, H., and Collins, A. L.: A comparison of machine learning models for the mapping of groundwater spring potential, Environ. Earth Sci., 79, 206, https://doi.org/10.1007/s12665-020-08944-1, 2020.

540   Angelis, L. and Stamelos, I.: A simulation tool for efficient analogy based cost estimation, Empir. Softw. Eng., 5, 35–68, https://doi.org/10.1023/A:1009897800559, 2000.

Arneth, A., Barbosa, H., Benton, T., Calvin, K., Calvo, E., Connors, S., Cowie, A., Davin, E., Denton, F., and van Diemen, R.: Summary for policymakers, 2019.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology / Un modèle à base
545   physique de zone d'appel variable de l'hydrologie du bassin versant, Hydrol. Sci. Bull., 24, 43–69, https://doi.org/10.1080/02626667909491834, 1979.

Breiman, L.: Random Forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

British Geological Survey (BGS). Africa Groundwater Atlas. Geology. http://earthwise.bgs.ac.uk/index.php/Geology, last access: 2 February 2021.

550   Chen, W., Li, H., Hou, E., Wang, S., Wang, G., Panahi, M., Li, T., Peng, T., Guo, C., Niu, C., Xiao, L., Wang, J., Xie, X., and Ahmad, B. B.: GIS-based groundwater potential analysis using novel ensemble weights-of-evidence with logistic regression and functional tree models, Sci. Total Environ., 634, 853–867, https://doi.org/10.1016/j.scitotenv.2018.04.055, 2018.

Courtois, N., Lachassagne, P., Wyns, R., Blanchin, R., Bougaïré, F. D., Somé, S., and Tapsoba, A.: Large-Scale Mapping of Hard-Rock Aquifer Properties Applied to Burkina Faso, Groundwater, 48, 269–283, https://doi.org/10.1111/j.1745-6584.2009.00620.x, 2010.

DEP: Carte hydrogeologique du Burkina Faso. Feuille Ouagadougou. Echelle 1:50 000., Ministère de l'Eau and Directorat Général de la Coopération au Développement Pays Bas, Ouagadougou, Burkina Faso, 1993.

Dewitte, O., Jones, A., Spaargaren, O., Breuning-Madsen, H., Brossard, M., Dampha, A., Deckers, J., Gallali, T., Hallett, S., Jones, R., Kilasara, M., Le Roux, P., Michéli, E., Montanarella, L., Thiombiano, L., Van Ranst, E., Yemefack, M., and Zougmore, R.: Harmonisation of the soil map of Africa at the continental scale, Geoderma, 211–212, 138–153, https://doi.org/10.1016/j.geoderma.2013.07.007, 2013.

Díaz-Alcaide, S. and Martínez-Santos, P.: Review: Advances in groundwater potential mapping, Hydrogeol. J., 27, 2307–2324, https://doi.org/10.1007/s10040-019-02001-3, 2019.

Díaz-Alcaide, S., Martínez-Santos, P., and Villarroya, F.: A Commune-Level Groundwater Potential Map for the Republic of Mali, Water, 9, 839, https://doi.org/10.3390/w9110839, 2017.

Direction Nationale de l'Hydraulique (Ed.): Données Hydrogeologiques et des Forages. Direction Nationale de l'Hydraulique, 2010.

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S.: Collinearity: a review of methods to deal with it and a simulation study evaluating their performance, Ecography, 36, 27–46, https://doi.org/10.1111/j.1600-0587.2012.07348.x, 2013.

Elbeih, S. F.: An overview of integrated remote sensing and GIS for groundwater mapping in Egypt, Ain Shams Eng. J., 6, 1–15, https://doi.org/10.1016/j.asej.2014.08.008, 2015.

Foster, S., Tuinhof, A., and Garduño, H.: Sustainable Groundwater Management. Lessons from Practice, Case profile collection Groundwater Development in Sub-saharan Africa. A Strategic Overview of Key Issues and Major Needs, vol. 15, 2006.

Freund, Y. and Schapire, R. E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, J. Comput. Syst. Sci., 55, 119–139, https://doi.org/10.1006/jcss.1997.1504, 1997.

Friedman, J. H.: Greedy Function Approximation: A Gradient Boosting Machine, Ann. Stat., 29, 1189–1232, 2001.

Geurts, P., Ernst, D., and Wehenkel, L.: Extremely randomized trees, Mach. Learn., 63, 3–42, https://doi.org/10.1007/s10994-006-6226-1, 2006.

Grönwall, J. and Danert, K.: Regarding Groundwater and Drinking Water Access through A Human Rights Lens: Self-Supply as A Norm, Water, 12, 419, https://doi.org/10.3390/w12020419, 2020.

Huang, J., Li, Y.-F., and Xie, M.: An empirical analysis of data preprocessing for machine learning-based software cost estimation, Inf. Softw. Technol., 67, 108–127, https://doi.org/10.1016/j.infsof.2015.07.004, 2015.

Jasiewicz, J. and Stepinski, T. F.: Geomorphons — a pattern recognition approach to classification and mapping of landforms, Geomorphology, 182, 147–156, https://doi.org/10.1016/j.geomorph.2012.11.005, 2013.

Jenks, G. F.: The Data Model Concept in Statistical Mapping, Int. Yearb. Cartogr., 7, 186–190, 1967.

Jha, M. K., Chowdhury, A., Chowdary, V. M., and Peiffer, S.: Groundwater management and development by integrated
590 remote sensing and geographic information systems: prospects and constraints, Water Resour. Manag., 21, 427–467,
https://doi.org/10.1007/s11269-006-9024-4, 2007.

Kalantar, Al-Najjar, Pradhan, Saeidi, Halin, Ueda, and Naghibi: Optimized Conditioning Factors Using Machine Learning
Techniques for Groundwater Potential Mapping, Water, 11, 1909, https://doi.org/10.3390/w11091909, 2019.

Lee, S., Song, K.-Y., Kim, Y., and Park, I.: Regional groundwater productivity potential mapping using a geographic
595 information system (GIS) based artificial neural network model, Hydrogeol. J., 20, 1511–1527,
https://doi.org/10.1007/s10040-012-0894-7, 2012.

Li, T., Jing, B., Ying, N., and Yu, X.: Adaptive Scaling, ArXiv170900566 Stat, 2017.

Llamas, M. R. and Martínez-Santos, P.: Intensive Groundwater Use: Silent Revolution and Potential Source of Social
Conflicts, J. Water Resour. Plan. Manag., 131, 337–341, https://doi.org/10.1061/(ASCE)0733-9496(2005)131:5(337), 2005.

600 Magesh, N. S., Chandrasekar, N., and Soundranayagam, J. P.: Delineation of groundwater potential zones in Theni district,
Tamil Nadu, using remote sensing, GIS and MIF techniques, Geosci. Front., 3, 189–196,
https://doi.org/10.1016/j.gsf.2011.10.007, 2012.

Martínez-Santos, P. and Renard, P.: Mapping Groundwater Potential Through an Ensemble of Big Data Methods,
Groundwater, 58, 583–597, https://doi.org/10.1111/gwat.12939, 2020.

605 Martínez-Santos, P., Aristizábal, H. F., Díaz-Alcaide, S., and Gómez-Escalonilla, V.: Predictive mapping of aquatic
ecosystems by means of support vector machines and random forests, J. Hydrol., 126026,
https://doi.org/10.1016/j.jhydrol.2021.126026, 2021.

Martín-Loeches, M., Reyes-López, J., Ramírez-Hernández, J., Temiño-Vela, J., and Martínez-Santos, P.: Comparison of
RS/GIS analysis with classic mapping approaches for siting low-yield boreholes for hand pumps in crystalline terrains. An
610 application to rural communities of the Caimbambo province, Angola, J. Afr. Earth Sci., 138, 22–31,
https://doi.org/10.1016/j.jafrearsci.2017.10.025, 2018.

Moghaddam, D. D., Rahmati, O., Panahi, M., Tiefenbacher, J., Darabi, H., Haghizadeh, A., Haghighi, A. T., Nalivan, O. A.,
and Tien Bui, D.: The effect of sample size on different machine learning models for groundwater potential mapping in
mountain bedrock aquifers, CATENA, 187, 104421, https://doi.org/10.1016/j.catena.2019.104421, 2020.

615 Mohammadi-Behzad, H. R., Charchi, A., Kalantari, N., Nejad, A. M., and Vardanjani, H. K.: Delineation of groundwater
potential zones using remote sensing (RS), geographical information system (GIS) and analytic hierarchy process (AHP)
techniques: a case study in the Leylia–Keynow watershed, southwest of Iran, Carbonates Evaporites, 34, 1307–1319,
https://doi.org/10.1007/s13146-018-0420-7, 2019.

Moore, I. D., Grayson, R. B., and Ladson, A. R.: Digital terrain modelling: A review of hydrological, geomorphological, and
620 biological applications, Hydrol. Process., 5, 3–30, https://doi.org/10.1002/hyp.3360050103, 1991.

Naghibi, S. A. and Pourghasemi, H. R.: A Comparative Assessment Between Three Machine Learning Models and Their
Performance Comparison by Bivariate and Multivariate Statistical Methods in Groundwater Potential Mapping, Water Resour.
Manag., 29, 5217–5236, https://doi.org/10.1007/s11269-015-1114-8, 2015.

Naghibi, S. A., Pourghasemi, H. R., and Dixon, B.: GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran, Environ. Monit. Assess., 188, 44, https://doi.org/10.1007/s10661-015-5049-6, 2016.

Naghibi, S. A., Moghaddam, D. D., Kalantar, B., Pradhan, B., and Kisi, O.: A comparative assessment of GIS-based data mining models and a novel ensemble model in groundwater well potential mapping, J. Hydrol., 548, 471–483, https://doi.org/10.1016/j.jhydrol.2017.03.020, 2017a.

Naghibi, S. A., Ahmadi, K., and Daneshi, A.: Application of Support Vector Machine, Random Forest, and Genetic Algorithm Optimized Random Forest Models in Groundwater Potential Mapping, Water Resour. Manag., 31, 2761–2775, https://doi.org/10.1007/s11269-017-1660-3, 2017b.

Nasir, M. J., Khan, S., Zahid, H., and Khan, A.: Delineation of groundwater potential zones using GIS and multi influence factor (MIF) techniques: a study of district Swat, Khyber Pakhtunkhwa, Pakistan, Environ. Earth Sci., 77, 367, https://doi.org/10.1007/s12665-018-7522-3, 2018.

Nguyen, P. T., Ha, D. H., Jaafari, A., Nguyen, H. D., Van Phong, T., Al-Ansari, N., Prakash, I., Le, H. V., and Pham, B. T.: Groundwater Potential Mapping Combining Artificial Neural Network and Real AdaBoost Ensemble Technique: The DakNong Province Case-study, Vietnam, Int. J. Environ. Res. Public. Health, 17, 2473, https://doi.org/10.3390/ijerph17072473, 2020a.

Nguyen, P. T., Ha, D. H., Avand, M., Jaafari, A., Nguyen, H. D., Al-Ansari, N., Van Phong, T., Sharma, R., Kumar, R., Le, H. V., Ho, L. S., Prakash, I., and Pham, B. T.: Soft Computing Ensemble Models Based on Logistic Regression for Groundwater Potential Mapping, Appl. Sci., 10, 2469, https://doi.org/10.3390/app10072469, 2020b.

Nhu, V.-H., Rahmati, O., Falah, F., Shojaei, S., Al-Ansari, N., Shahabi, H., Shirzadi, A., Gorski, K., Nguyen, H., and Bin Ahmad, B.: Mapping of Groundwater Spring Potential in Karst Aquifer System Using Novel Ensemble Bivariate and Multivariate Models, Water, 12, 985, https://doi.org/10.3390/w12040985, 2020.

Ourhzif, Z., Algouti, A., Algouti, A., and Hadach, F.: Lithological mapping using Landsat 8 OLI and ASTER multispectral data in Imini-Ounilla district south High Atlas of Marrakech, ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci., XLII-2/W13, 1255–1262, https://doi.org/10.5194/isprs-archives-XLII-2-W13-1255-2019, 2019.

Ozdemir, A.: Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey), J. Hydrol., 405, 123–136, https://doi.org/10.1016/j.jhydrol.2011.05.015, 2011.

Panahi, M., Sadhasivam, N., Pourghasemi, H. R., Rezaie, F., and Lee, S.: Spatial prediction of groundwater potential mapping based on convolutional neural network (CNN) and support vector regression (SVR), J. Hydrol., 588, 125033, https://doi.org/10.1016/j.jhydrol.2020.125033, 2020.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., and Cournapeau, D.: Scikit-learn: Machine Learning in Python, Mach. Learn. PYTHON, 2825–2830, 2011.

Poggio, L. and de Sousa, L.: SoilGrids250m 2.0 - Clay content, 2020.

Schetselaar, E. M., Harris, J. R., Lynds, T., and de Kemp, E. A.: Remote Predictive Mapping 1. Remote Predictive Mapping (RPM): A Strategy for Geological Mapping of Canada's North, Geosci. Can., 34, 2007.

660   Traore, A. Z., Bokar, H., Sidibe, A., Upton, K., Ó Dochartaigh, B., and Bellwood-Howard, I.: Africa Groundwater Atlas: Hydrogeology of Mali, http://earthwise.bgs.ac.uk/index.php/Hydrogeology_of_Mali, 2018.

United Nations: General comment no. 15. The right to water. UN Committee on Economic, Social and Cultural Rights, United Nations, 2002.

United Nations: Resolution A/RES/64/292. United Nations General Assembly, United Nations, 2010.

665   Xie, Y., Sha, Z., and Yu, M.: Remote sensing imagery in vegetation mapping: a review, J. Plant Ecol., 1, 9–23, https://doi.org/10.1093/jpe/rtm005, 2008.

Xu, H.: Modification of Normalized Difference Water Index (NDWI) to Enhance Open Water Features in Remotely Sensed Imagery, Int. J. Remote Sens., 27, 3025–3033, https://doi.org/10.1080/01431160600589179, 2006.

Zheng, A. and Casari, A.: Feature Engineering for Machine Learning, O'Reilly Media, Inc, 218 pp., 2018.

670