Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-261-RC2, 2021

Comment on hess-2021-261

Anonymous Referee #2

Referee comment on "Preprocessing approaches in machine-learning-based groundwater potential mapping: an application to the Koulikoro and Bamako regions, Mali" by Víctor Gómez-Escalonilla et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-261-RC2, 2021

The manuscript **"Preprocessing approaches in machine learning-based groundwater potential mapping: an application to the Koulikoro and Bamako regions, Mali "** represents an important contribution aligned with the objective of the HESS journal and can interest the scientific community working on machine learning applied in water management.

Concerning the scientific quality, I think that the used scientific approach and applied methods are interesting but the sections of the manuscript have unbalanced structure and some sections are inappropriate and need in-depth analysis with improving the used English language. For that, I think this paper needs major modification and resubmission

- **General Comments**

The introduction :

-the section dedicated to the Reviews of literature concerning Groundwater potential mapping studies should be more developed with the presentation of the brief results of the pertinent studies.

- the introduction missed the presentation of the water resources problems in the study area and the need to elaborate the Groundwater potential map

Then the results discussed must be more in-depth, especially by explaining the results of the GPM obtained in connection with the hydrogeological context of the study area and the used explanatory parameters.

The methodology

The hydrogeological context of the studies area is unfairly presented; then the explanatory parameters used are unclearly presented. It is important to explain in-depth these used data to enrich the explanation of the results of GPM.

Revision suggestions:

ABSTRACT:

line9: "Groundwater is crucial for domestic supplies in the Sahel"

it is necessary to precise the location. which Sahel?

Line11 & 12: "This paper presents a machine learning method to map groundwater potential and illustrates it through an application to two regions of Mali".

It is poorly structured sentences!

Line 13: “A set of explanatory variables for the presence of groundwater is developed first”

I suggest to replacing the presence of groundwater by **groundwater occurrence**

Line17: “This process identifies noisy, collinear and counterproductive variables and excludes them from the input dataset”:

It is a result details, I suggest deleting this sentence.

Line 18, 19 & 20: “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 standardization proved the most efficient methods to scale explanatory variables”.

I suggest replacing by:

The results shows that the 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 standardization proved the most efficient methods to scale explanatory variables.

Line 22 & 23: "From a methodological standpoint, the outcomes lead to three major conclusions":

I suggest replacing by: The outcomes of this study lead to three major conclusions

Introduction

Line 38 & 39: "Groundwater potential mapping (GPM) is recognized as a valuable tool to underpin planning and development of groundwater resources (Elbeih, 2015)".

I suggest replacing by

Groundwater potential mapping (GPM) is recognized as a valuable tool to underpin planning and exploration of groundwater resources (Elbeih, 2015).

Line 41 & 42: "In practice, however, it consists in computing spatially-distributed estimates for a target variable (groundwater potential) based a set of explanatory variables"

What are the explanatory variables, you should explain them, I suggest to replace these sentences by:

However, it consists of computing spatially distributed estimates for a target variable (groundwater potential) based a set of dependent variables such as soil, lineaments, slope, geology, landforms, lithology, and drainage density (Díaz-Alcaide and Martínez-Santos 2019a)

Line 42, 43 & 44: "GPM typically relies on existing cartography, digital elevation models obtained from satellite, aerial photographs, satellite imagery and geophysical information (Schetselaar et al., 2007)".

The GPM based on the assembling of data from different sources. I suggest replacing by:

GPM typically relies on the compilation of data derived from existing maps, aerial photographs, satellite imagery, and airborne geophysical information (Schetselaar et al. 2008).

Line 46: "There are two main approaches to GPM: expert-based decision systems and machine learning methods".

I suggest replacing by:

Recently, expert-based decision systems and machine learning methods have been implemented in many groundwater studies.

Line 46 &47: "**Expert-based systems have existed for a long time (DEP, 1993)"**

I suggest replacing by: Expert-based system methods have been used for a long time (DEP, 1993)

Line 52 & 53: **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),**

I suggest replacing by: In literature, The Machine Learning Algorithms used in the GPM studies include Mixture Discriminant Analysis (Al-Fugara et al., 2020), Random Forest (Kalantar et al., 2019; Moghaddam et al., 2020),

Line 58: "GPM works under the assumption that the presence of groundwater can be partially inferred from surface features"

I suggest replacing by:

The GPM is based on a common assumption is that the groundwater occurrence can be partially inferred from surface features.

Line 60 & 61: Supervised classification algorithms are trained to find the associations between these variables and known groundwater data.

The data are trained using the algorithm not the algorithms are trained: I suggest replacing by:

These explanatory variables are trained using Supervised classification algorithms to find the associations between them and known groundwater data.

Line 64 & 65: add reference.

Line 68: add reference

Line 71 & 72: 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.

I suggest replacing by:

The outcomes of GPM studies using machine learning algorithms are almost invariably assessed by means of standard big data metrics such as…. And add reference to this observation

Line 76 to Line 80: 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 all scaled features have the same degree of influence (Angelis and Stamelos, 2000).

I suggest replacing by:

Within this context, this research presents two main additions to the literature. In the first place, it explores different scaling methods 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

all scaled features have the same degree of influence (Angelis and Stamelos, 2000). (This is d detail of methodology I propose to add to the methodology section)

2 Material and methods

2.1 Study area

Line 93 to 111: I suggest to add a hydrogeological section or a geologic map to highlight the aquifers units of the study area

Line 89 to 101: "Water in these aquifers is preferentially located in the weathered mantle, and, within this, the lower part is generally more transmissive 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)"

I suggest replacing by:

In these aquifers, groundwater flows preferentially in the weathered mantle, and, within this, the lower part is generally more transmissive due to lower clay content. The upper part is less permeable to flow but can still be important as a groundwater reservoir where the fractures can raise the reservoir permeability although their storage capacity is typically low (Martín-Loeches et al.,2018).

Line 107: "**Some boreholes however exceed 100 m3/hour"**

I suggest replacing by:

However, some boreholes yield exceeds 100 m3/hour

Line 107 & 108: "The Paleozoic rocks located in the north are determined by fractures that allow water to flow through the sandstone and limestone layers".

I suggest replacing by:

In the north, the fractured Paleozoic rocks allow water to flow through the sandstone and limestone layers.

2.2 Borehole database

Line 115: Borehole data were provided by Direction Nationale de l'Hydraulique (2010)

I suggest replacing by:

Borehole data were provided by the National Water Directorate (DNH, 2010)

Line 115 to 116: "**The database contains 115 information on 5,387 boreholes (3,772 successful and 1,615 unsuccessful), distributed across 1,605 human settlements".**

I suggest replacing by:

The database contains information of 5,387 boreholes (3,772 successful and 1,615 unsuccessful), distributed across 1,605 fields.

Line 121 to 123: "This can be assumed to be the thickness of the (Courtois et al., 2010). Water table depth

I suggest replacing by:

There is a considerable number of boreholes with a 100% success rate (530), many villages are supplied by a single borehole

Line 126 to 127: 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

I suggest replacing by:

This raises the question in the application of algorithm in the choice of training datasets, especially to whether villages with a small number of boreholes are statistically representative, particularly in cases where the mean yield is low

Line 145: Figure 3: correct the word classification metrics

Line 156 to 157: Sixteen explanatory variables were selected based on an extensive review of the GPM literature (Díaz-Alcaide and Martínez-Santos 2019).

I think to explain in detail this extensive review in the Introduction part

Line 161: you should add a description of the main factors that can influence the groundwater recharge before explaining the description of each used variables or factors in the groundwater potential mapping

Line 162: Geology constrains the presence of groundwater to an important

extent

I think to delete this sentence

Line 173: Soils are important in GPM because soil characteristics such as permeability…

I suggest replacing by:

Soil is important factor to determine the groundwater occurrence …….

Line 174: Soil descriptions of the study area were obtained from the European Soil Data Centre

You should describe the main soils of the study area types and their characteristics

Line 175 and 176: 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).

I suggest replacing by:

Integration of land use and land cover is often used in groundwater potential mapping studies because Land use changes, which are mostly induced by human activities, affect hydrological dynamics (Díaz-Alcaide and Martínez-Santos, 2019).

Line 175 to 180: you should describe the land use of your study area and the data used for the elaboration of this map

Line 182: You should add the reference of used rainfall data

Line 184: Figure 4: you should add the lineaments and faults in the geological map

Line 191 & 192: 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)

I suggest replacing by:

The topography is a relevant factor in groundwater distribution, storage, and flow, as well as surface runoff and infiltration are partially conditioned by surface features and parameterized by properties that can be extracted from the surface data (Elbeih, 2015)

Line 197: The topographic wetness index

I suggest replacing by:

The Topographic Wetness Index (TWI)

Line 243: 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 245 (g/kg) l) rainfall (mm/year)

What is the difference of the figure 6g (alteration band ratio (B6/B7)) and the figure 6k (Clay content); in the text it means the same information line 230 to 233: **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.**

ð□□´ð□□□ð□□i
ð□□□ð□□□ð□□□ð□□ið□□□ð□□
□ð□□□ (ð□□□ð□□□ð□□□ð□□¦
ð□□□ð□□□ð□□□ð□□□ð□
□□ð□□□ð□□□ð□□
ð□□□ð□□□ð□□□ð□□□ð□□□ð□□□ð□□ið□□□ð□□□ð□□□ð□□□)
= ð□□µ6 / ð□□µ7 (5)

Line 267: reference of equation 6

Line 273: reference of equation 7

Line 380 to 400: I find this paragraph should be added to the introduction section to explain the use of used algorithms in literature

Line 437: Classifier outcomes were extrapolated to produce groundwater potential maps

What you want to say it is not clear!

Line 437 to 438: Figure 9 shows the groundwater potential predictions rendered by each of the five best-performing algorithms under the two most effective scaling methods

I suggest adding the abbreviations of used algorithms and scaling methods between parentheses

Line 447: The agreement map (Fig. 10) allows for an analysis of discrepancies among the best performing algorithms.

What you want to say about the agreement map!

Line 455: **Figure 9.** Mapping outcomes of the top five supervised classification algorithms for the two best performing scaling methods. At the top the 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.

I suggest to add number or letter for each map like:

- AdaBoost classifier, (b) Gradient Boosting classifier, (c) Random Forest classifier, (d) Decision Tree classifier and (e) Extra Trees classifier.

Line 492 to 494: "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".

This is a literature review about similar studies in pilot area, I suggest to add in the Introduction section