Referee comment to:

HESS-2021-261 Revision report
"Preprocessing approaches in machine learning-based groundwater potential mapping: an application to the Koulikoro and Bamako regions, Mali" by Gómez-Escalonilla et al. (2021).

[1] General Comments
Gómez-Escalonilla et al. (2021) provide an interesting machine learning learning-based groundwater potential mapping in the Koulikoro and Bamako regions of Mali. The authors have used machine learning models for groundwater potential mapping (GPM) in two regions in Mali and evaluated their models. Also, their models explore the potential factors affecting groundwater potential. The paper is interesting and within the scope of the HESS journal. In general, machine learning is well-placed in HESS. The authors have done very diligent work by summarizing many publications applyingmachine learning. The manuscript can be interesting to the scientific community working on machinelearning applied in hydrology. The work is of importance; but at the present state, I would not recommend it for publication because certain comments need to be addressed with major revisions.

We thank Reviewer #1 for a series of insightful comments. We have strived to incorporate them all to the manuscript. We believe this has helped us improve the quality of our work.

[2] The introduction is very general. It should be worked out why this study with machine learning is necessary, knowing that machine learning is a "Blackbox model" and what its benefit is with other methods such as fuzzy logic, the frequency ratio, weight of evidence, or multi-criteria decision analysis (MCDA).

Agreed. A key point is that methods such as Weight of evidence, Frequency Ratio and Multi-criteria decision analysis require a grouping of the variables in intervals. A bias is generated from the outset, since these intervals rely almost exclusively on expert criteria. We attempt to show that, if pre-processing is involved, machine learning algorithms can work directly with raw data, thus discarding a potentially biased clustering of explanatory variables. Another important advantage of supervised classification is that the power of artificial intelligence can be harnessed to find complex associations among explanatory variables that might otherwise pass unnoticed

In order to incorporate this comment to the manuscript, the third paragraph of the introduction has been rewritten as:

"The literature shows that there are two main approaches to GPM, namely, expert-based decision systems and machine learning methods. Expert-based system methods have been used for a long time (DEP, 1993). These include multi-influence factor techniques (Magesh et al., 2012; Nasir et al., 2018; Martín-Loeches et al 2018), analytical hierarchy processes (Mohammadi-Behzad et al., 2019; Al-Djazouli et al., 2021), and Dempster-Shafer models (Mogaji and Lim 2018, Obeidavi et al 2021). Other frequently used methods are weight of evidence and frequency ratio analysis (Falah and Zeinivand, 2019; Boughariou et al., 2021). Machine learning is comparatively newer. A major difference between machine learning and expert approaches is that supervised classification uses the advantages of artificial intelligence to find complex associations among explanatory variables that might otherwise pass unnoticed. Hence, machine learning is well suited to map complex spatially-distributed variables such as groundwater occurrence. The GPM literature showcases a wide variety of supervised classification approaches. Thus, Al-Fugara et al. (2020) used mixed discriminant analysis to map spring potential in a watershed of Jordan; much like Odzemir (2011) mapped spring potential in a Turkish basin by means of a logistic regression method. Random forests have proved adept at mapping groundwater potential, both in mountain bedrock aquifer (Moghaddam et al., 2020), as well as in large metasedimentary basins (Martínez-Santos and Renard 2019). Other supervised classification methods include boosted regression trees (Naghibi et al., 2016), support vector machines (Naghibi et al., 2017b), neural networks (Lee et al., 2012; Panahi et al., 2020) and Ensemble methods (Naghibi et al., 2017a; Martínez-Santos and Renard, 2019; Nguyen et al., 2020b)."

[3] In addition, the objectives are not clear and are included at various locations in the manuscript (see Page 13, Line 255- 260, with the sentence "A major goal of this study…). You need to improve the manuscript correctly.

Agreed. In order to clarify the goals we have rewritten some sentences of the last paragraph of the introduction, which now reads:

"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 instances, using ad hoc calibration elements, such as complementary field information, can contribute to a better interpretation of the outcomes. The objective of this research is to build on the existing literature by presenting two main methodological additions. In the first place, we explore different scaling methods in order to avoid the pitfalls associated with the reclassification of explanatory variables. 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 method is demonstrated through the application of machine learning techniques to map groundwater potential across two regions of Mali. The geographical setting also represents an added value to the literature. Indeed, while there are numerous examples of GPM studies based on artificial intelligence in other continents (Naghibi et al., 2017a; Chen et al., 2018; Panahi et al., 2020), these approaches remain uncommon across Africa.

We found another instance in the manuscript that could be misleading in terms of the objectives (section 2.5, second paragraph). We have rewritten it as: *"The rationale behind using preprocessing approaches 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) and normalization (Pedregosa et al., 2011)".*

[4] In the Introduction section only studies for the other continent are presented. It would be interesting to see how studies in other regions in Africa deal the groundwater potential mapping with machine learning or others methods?

Agreed. We thank Reviewer #1 for a suggestion that allows us to highlight another important novelty of our work. GPM studies based on artificial intelligence are uncommon in Africa. To our knowledge, the only known precedent is a paper by Martinez-Santos and Renard (2019), in which the authors mapped groundwater potential in the Baoulé basin, Mali. We now note this in the last three sentences of the introduction: *"This method is demonstrated through the application of machine learning techniques to map groundwater potential across two regions of Mali. The geographical setting also represents an added value to the literature. Indeed, while there are numerous examples of GPM studies based on artificial intelligence in other continents (Naghibi et al., 2017a; Chen et al., 2018; Panahi et al., 2020, these approaches remain uncommon across Africa".*

We also agree that there are many non-machine learning GPM studies in Africa. To acknowledge this, the second paragraph of the introduction now reads:

"Groundwater potential mapping (GPM) is recognized as a valuable tool to underpin planning and exploration of groundwater resources (Elbeih, 2015). GPM may be understood as a means to estimate groundwater storage in a given region, as a measure 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). 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. GPM often uses existing cartography, digital elevation models, aerial photographs, satellite imagery and geophysical information (Díaz-Alcaide and Martínez-Santos, 2019). Recent years have witnessed a growing interest in groundwater potential studies in Africa, largely as a result of the need to achieve the Sustainable Development Goal #6. The majority of these work with a combination of remote sensing, geographic information systems and geophysics (Delgado 2018, Adeyeye et al., 2019, Magaia et al 2018, Mpofu et al 2020, Owolabi et al 2020, Saadi et al 2021, Al-Djazouli et al. 2021), while others rely directly on the interpretation of information from borehole databases (Díaz-Alcaide et al 2017)."

[5] Also, the authors should be linking the issues of groundwater resources in the context of The Sustainable Development Goals (SDGs) to motive thereader at the end of the introduction.

Agreed. We concur on the importance of groundwater to achieve the Sustainable Development Goals (SDGs). We now provide a mention to SDG 6 in the second paragraph of the introduction: "*Recent years have witnessed a growing interest in groundwater potential studies in Africa, largely as a result of the need to achieve the Sustainable Development Goal #6. The majority of these efforts use a combination of remote sensing, geographic information systems and geophysics (Delgado 2018, Adeyeye et al., 2019, Magaia et al 2018, Mpofu et al 2020, Owolabi et al 2020, Saadi et al 2021, Al-Djazouli et al. 2021), while others rely directly on the interpretation of information from borehole databases (Díaz-Alcaide et al 2017)".*

Also, we now mention the relevance of our work to SDG 6 in the second paragraph of the conclusions:

"A crucial finding of this research is that conventional machine learning metrics (test score, area under the receiver operating 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. In the context of Sustainable Development Goal #6, predictive groundwater maps such as those developed in the course of this research may be of use to private investors willing to participate in improving water access in remote regions, as well as to government officers, cooperation funds and international donors".

[6] **You could add this reference in the introduction section: A new method to map groundwater potential at a village scale, based on a comprehensive borehole database. An application to Sikasso, Republic of Mali by Ana Carolina Gonçalves Delgado, 2018.**

Agreed. The suggested reference is a product of our own research group. It has been added to the second paragraph of the introduction section.

[7] Add a new section (or put some sentences) about the limitations of machine learning techniques to study groundwater potential zones.

Agreed. We have added a limitations section (section 3.5) where we present the main limitations of our approach. This section reads:

3.5 Limitations

Predictive groundwater mapping presents to key uncertainties, namely, deficiencies in data quality (e.g., small sample sizes, missing covariates, biased and missing data), as well as errors in the structural nature and specifications of the model (Rahmati et al. 2015). In spatial modeling studies the sample size has proven to be a significant factor affecting the predictive abilities of the models (Guisan et al., 2007). Moghaddam et al. (2020) analyzed the influence of sample size on GPM, concluding that there is a significant decrease in AUC values when the sample size is 25% of the input dataset. In the present case,

with 650 ground-truth points, the effect of sample size could be ruled out as per this standard.

Input data was made available by government officers after careful field evaluation. From this perspective, the database is considered of high quality. However, it is also true that key hydrogeological variables are missing. Groundwater potential was evaluated in binary form (positive/negative) because the borehole database provides little information in terms of borehole productivity. Spatially distributed estimates of aquifer transmissivity, storage coefficients and yield would allow for a multi-class assessment, which in turn would provide a more realistic zoning of groundwater potential. Besides, the village scale resolution presents some shortcomings. As per the database, all boreholes within a given village have the same coordinates. This makes it difficult to train the algorithms when, for instance, the village overlies a non-homogeneous hydrogeological setting. A more detailed knowledge of how the boreholes are distributed in space would be expected to enhance the outcomes.

Along the same lines, the resolution of certain explanatory variables is potentially problematic. Take for instance soil and lithology, which are only available at the regional scale, and which might constrain groundwater potential to an important extent. Gómez-Escalonilla et al. (under review) explains that inroads can be made in the use of dynamic explanatory variable. Take for instance evapotranspiration or seasonal fluctuations (VV- and VH-polarization intensity (backscattering coefficient) and VV- polarization coherence (interferometric correlation) from Sentinel-1 time-series from which temporal descriptors are derived. Furthermore, Worthington (2015) shows that groundwater modeling in bedrock aquifers is complex because there is often substantial flow through fractures, and the interconnectivity and magnitude of these fractures are usually uncertain. In this context, geophysical techniques can provide useful information that improves the ability to predict GPM. However, the absence of geophysical information also limits our knowledge about subsurface structures.

The lack of interpretability is a major drawback of machine learning algorithms. Except for the simple decision tree, whose internal logic is straightforward, all classifiers work according to a complex internal architecture. Therefore it is nearly impossible to understand the reasoning that leads most algorithms to a given conclusion beyond computing feature importance. This in turn results in extensive trial and error before optimal outcomes are achieved.

Overfitting can lead to spurious results in machine learning models. Overfitting occurs when the model does not generalize well with the training data (i.e. it tries to fit the training data perfectly at the risk of missing the underlying associations between the explanatory and target variables). To address this, techniques such as splitting the initial data set into separate training and test subsets, cross-validation, regularization and ensembling may be used (Dietterich, 1995; Yeom et al., 2018). Case-specific indicators may also be used to appraise the results beyond standard machine learning metrics, thus adding to the robustness of predictions (Martinez-Santos et al 2021a and 2021b). In the case at hand, this is achieved by comparing map outcomes with the limited available data on borehole yield.

[8] The overfitting problem is one of the drawbacks that affect the accuracy of models in machine learning. Take into account this issue in the introduction.

Agreed. To maintain the flow of the text, we have placed our comment on overfitting in the limitations section, rather than in the intro. Please see also our answer to [7].

[9] **The topic of validation of Groundwater Potential Map by is not mentioned. In my opinion, it is oneof the major limitations of the study. If so, this topic should be discussed in more detail.**

We agree that the original manuscript does not speak of validation explicitly. However, sections 3.2 and 3.4 present very detailed validation procedures and scores. Most of section 3.2, including Table 1, is devoted to outlining the results in terms of standard machine learning metrics (f-1, AUC, test score). These represent different takes on how well each of the models is able to "guess" an unknown outcome based on what it learned during the training process. This is the routine validation procedure of any machine learning study.

Then, in section 3.4, we go beyond standard machine learning validation metrics to compare the maps we produced with actual borehole yield data (see also Table 3). This is an additional validation procedure that is seldom carried out in groundwater potential studies. It adds robustness to our results. Please note this is also a chief conclusion of our work. The conclusions section (second paragraph) reads:

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

[10] Did you limit the validation of your model with cross-validation? Or do you have the intention to integrate the external validation?

Please see [9] above. External validation is provided by comparing machine learning validated outcomes with a different dataset (borehole flow rates).

[11] The authors need to describe in the methodology section a sub-section of "Multicollinearity analysis" before presenting the results in section 3.1. Did you use the variance inflation factor (VIF) and tolerance (TOL) indices as are customarily used to estimate multicollinearity of predictive factors in machine learning modelling? If so, explain in the manuscript.

Agreed. We now describe this in the second paragraph of section 2.6, which now reads:

"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 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. Collinearity analysis is performed before running the algorithms. Pairwise correlation among explanatory variables is computed and correlation coefficients are expressed in a range between -1.0 (inverse correlation) and 1.0 (direct correlation). Highly-correlated explanatory variables may be excluded from the algorithm training procedure."

As explained in the text above we used a different procedure (not VIF/TOL). Our approach is based on pairwise correlation analyses (see also Fig 8).

[12] What is the effect of sample size on the different machine learning models for groundwater potential mapping in your research?

We carried out many test runs. The first one included all villages (1605 human settlements). We observed that the machine learning models could not find clear associations between explanatory variables and groundwater potential. We concluded that human settlements with a single borehole might not be statistically representative, especially in cases where the average yield is low (these could represent minimum extraction flow rates determined by the type of pump, rather than by aquifer parameters). Different sample sizes were defined as per the number of boreholes in the villages. The analysis led to select villages with five or more boreholes as our optimal sample.

[13] Did you try to make a sensitivity analysis of the effect of each factor (explanatory variables) on thegroundwater potential map, i.e., when you decide to eliminate one or more factors??

Yes. The RFECV procedure described in section 2.6 (third and fourth paragraphs) includes a sensitivity analysis. This procedure analyzes the impact of eliminating each of the explanatory variables. Once all variables are appraised, the number of features leading to the highest test score is selected. This automatic

procedure is responsible for discarding or keeping each variable.

[14] What is the resolution chosen to develop the thematic layers? Because the various GIS layers come with different spatial resolutions. You need to clarify this aspect.

Agreed. As suggested, we now include the explanatory variables used in groundwater potential mapping in a Table. The information about the resolution of each explanatory variable were included in the Table below:

Table 1. Explanatory variables used in GPM. The scale/resolution, acquisition time and source of data for each factor are provided.

Explanatory variables	Scale/resolution	Time (dd/mm/yyyy)	Source of data
Alteration Band Ratio	30 meters	07-16/03/2020	Own elaboration from Landsat 8
Clay content	250 meters	N/A	SoilGrids250m 2.0
Curvature	30.53 meters	N/A	Own elaboration from DEM
Saturated thickness	30.53 meters	N/A	Own elaboration from DEM and borehole database
Water table Depth	30 meters	2010	Own elaboration from DNH (2010)
Distance from channels	30.53 meters	N/A	Own elaboration from DEM
Geology	1:5 million	N/A	British Geologycal Survey
Geomorphology	30.53 meters	N/A	Own elaboration from DEM
Land use	300 meters	2009	ESA Climate Change Initiative
Soil	1:3M	N/A	European Soil Data Centre
Rainfall	0.5°	1950-2009	CRU TS 3.21 dataset (Climatic Research Unit at the University of East Anglia)
Drainaige density	30.53 meters	N/A	Own elaboration from DEM
Thickness matrix	30.53 meters	N/A	Derived from DEM and borehole database
Elevation (DEM)	30.53 meters	23/09/2014	Shuttle Radar Topography Mission (SRTM)
NDVI	30 meters	07-16/03/2020	Own elaboration from Landsat 8
NDWI	30 meters	07-16/03/2020	Own elaboration from Landsat 8
Slope	30.53 meters	N/A	Own elaboration from DEM
SPI	30.53 meters	N/A	Own elaboration from DEM
TWI	30.53 meters	N/A	Own elaboration from DEM

[15] The sub-sections in Section 2 on "Material and methods" need to be reorganised for better reading. For example, you could define the title "Materials and methods": Study area; Data used (Borehole database, Explanatory variables/Thematic layers, etc); and Methods (Random Forest, AdaBoost, Gradient Boosting, Decision Tree, and Extra Trees classifiers); Tools used to process the data, etc.

Agreed. We have rearranged this section so that:
2.1 Study area.
2.2 Data (instead of "borehole database"), including the old section 2.4 (now 2.2.1 and 2.2.2, please see [31] below)
2.3 Predictive mapping software
2.4 Preprocessing of explanatory variables

2.5 Supervised classification routine
2.6 Machine learning metrics for algorithm evaluation

[16] A table of data sources must be put to increase the clarity and to ease readers' understanding.

Agreed. Please see [14].

[17] Do you have performed quality control of datasets before the modelling?

Yes. The borehole database was verified on site by government officers (Direction Nationale de l'Hydraulique). Then they made it available for us.

[18] Did you use all explanatory variables to map groundwater potential in Figure 10? If so, could you specify the variables used to develop the final products? Because, on Page 20 of MS, you mentioned that the outcomes show that elevation, rainfall, geology, and drainage density, among others, are the most important factors conditioning the groundwater potential. Did you use thesefour (04) explanatory variables in reality? Specify exactly the variables used in the final models.

Agreed, the way we presented this in the manuscript could lead to confusion. Each model uses a different set of explanatory variables Each model picks its own by means of the Recursive Feature Elimination Cross Validation procedure. The second paragraph of section 3.3 now reads:

"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. However, feature selection reveals clear differences among the classifiers. Under standardized scaling, RFC only required three explanatory variables to predict 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. The variables used by each of the best-performing algorithms are presented in Figure 8".

[19] **The conclusion is very general. To be check according to the revision of MS.**

Agreed. We have added the importance of this research to SDG 6 and to stakeholders in the second paragraph of the conclusions:

"A crucial finding of this research is that conventional machine learning metrics (test score, area under the receiver operating 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. In the context of Sustainable Development Goal #6, predictive groundwater maps such as those developed in the course of this research may be of use to private investors willing to participate in improving water access in remote regions, as well as to government officers, cooperation funds and international donors."

We have also added a paragraph where we deal with case-specific outcomes to make our conclusions more concrete (last paragraph of the conclusions).

From a regional perspective, groundwater potential closely resembles rainfall. Despite the predominance of low-permeability basement rocks, medium to high groundwater potential observed in southern areas seems to be associated with high rainfall and well-developed weathering mantles. In contrast, good conditions for groundwater storage in the north present a low potential due to limited precipitation. The

central part is characterized by a medium groundwater potential in the plains, as well as by a high potential in alluvial sediments of the major river systems and low potential in the highlands .

[20] Limitations of the research should be addressed.

Agreed. We have added a new section on limitations. Please see [7].

[21]I suggested the authors separate the results and discussion. I suggest developing in the new discussion section "the model validation/performance and comparison"; "assessment of variable importance"; limitations of the research", etc.

We thank Reviewer #1 for this suggestion. We observe that many papers in this journal present results and discussion together. Given the complexity of our manuscript, we fear that changing the structure of this section now could potentially lead us to the loss of important information. Therefore, we prefer to keep it as it is unless strictly necessary.

[22] **Furthermore,try to compare the outcomes of your research with other studies available in the literature on themapping of groundwater potential such as the GIS-based Dempster–Shafer model, etc.**

We agree that a reference to the Dempster-Schafer model was missing. It is however difficult to provide a meaningful comparison with it because we would need to apply it specifically to our study region to draw conclusions. Besides, we lack the experience with that particular model to make comparisons on a methodological level. Instead, we now mention it in the literature review and provide references. The third paragraph of the introduction now reads:

"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). These include multi-influence factor techniques (Magesh et al., 2012; Nasir et al., 2018; Martín-Loeches et al 2018), analytical hierarchy processes (Mohammadi-Behzad et al., 2019; Al-Djazouli et al., 2020), and Dempster-Shafer models (Mogaji and Lim 2018, Obeidavi et al 2021). Other frequently used methods are weight of evidence and frequency ratio analysis (Falah and Zeinivand, 2019; Boughariou et al., 2021). Machine learning is comparatively newer. A major difference between machine learning and these approaches is that supervised classification uses the advantages of artificial intelligence to find complex associations among explanatory variables that might otherwise pass unnoticed. Hence, machine learning is well suited to map complex spatially-distributed variables such as groundwater occurrence. 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 et al., 2012; Panahi et al., 2020) and Ensemble methods (Naghibi et al., 2017a; Martínez-Santos and Renard, 2019; Nguyen et al., 2020b)."

[23] The discussion is incomplete, authors must address the uncertainty in groundwater potential mapping (deficiencies of data quality; biased and absent data, sample sizes, missing covariates, and also errors in the structural and nature of the model, etc). Add some references.

Agreed. We have added a new section on limitations. Please see [7].

[24] In the Conclusion section, specify the utility of this research and the potential users. For example,who could use the prepared maps?

Agreed. We have improved the second paragraph of the conclusions. It now reads:

A crucial finding of this research is that conventional machine learning metrics (test score, area under the receiver operating 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. In the context of Sustainable Development Goal #6, predictive groundwater maps such as those developed in the course of this research may be of use to private investors willing to participate in improving water access in remote regions, as well as to government officers, cooperation funds and international donors.

[25] Is it possible to improve the performance of the best machine learning models in your study? Which additional predicting variable (s) (even if such information is scarce) could be added to improve the results?

Yes. It is definitely possible, but we would need better input data (which is currently unavailable). We explain this in the newly added section 3.5). Please see [7].

[26] I think that Table 2 on Page 19 could be placed in a new section in Supporting Information.

Agreed. Fixed.

[27] Abstract section: Page 1. The abstract is very long. It should be shortened and focused. The abstract should be thoroughly revised according to the revision of the manuscript.

Agreed. The abstract is now down from 350 to 250 words. We believe it is more focused now. It reads:

"Groundwater is crucial for domestic supplies in the Sahel, where the strategic importance of aquifers will increase in the coming years due to climate change. Groundwater potential mapping is a valuable tool to underpin water management in the region, and hence, to improve drinking water access. This paper presents a machine learning method to map groundwater potential in two regions of Mali. A set of explanatory variables for the presence of groundwater is developed first. Scaling methods (standardization, normalization, maximum absolute value and min-max scaling) are used to avoid the pitfalls associated with the reclassification of explanatory variables. Noisy, collinear and counterproductive variables are identified and excluded from the input dataset. Twenty machine learning classifiers are 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. Tree-based algorithms (accuracy >0.85) consistently outperformed other classifiers. Maximum absolute value and standardization proved the most efficient scaling techniques. 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. Outcomes lead to three major conclusions: (1) picking the best performers out of a large number of machine learning classifiers is recommended as a good methodological practice; (2) standard machine learning metrics should be complemented with additional hydrogeological indicators whenever possible; and (3) variable scaling helps minimize expert bias".

[28] <u>Keywords</u>: I propose to delete "big data; climate change and water access"; and add "Groundwaterpotentiality, and GIS".
Agreed. Fixed

[29] Specific comments: Page 2. Paragraph 1. Line 2, rewrite the sentence "Today, 2.5 billion people…." by "Today, 2.5 billionpeople around the World…".
Agreed. Fixed.

[30] Page 2. Paragraph 3. Line 1. Introduce a sentence before this: "There are two main approaches toGPM: expert-based decision systems and machine learning methods".
Agreed. Fixed.

[31] Page 3. For section 2.4, Material and methods, I suggest separating them into two sections.
2.4.1 Definition of target
2.4.2.2 Explanatory variables/Thematic layers
In this new section, I propose to describe the explanatory variables by order according to Figure 6. Also, I propose to prepare explanatory variables used in groundwater potential mapping in a Table. Inyour Table, you could, for example, put in 4 columns (Type of data layers/ Explanatory variables;scale/resolution, time, available format, and source of data).

Agreed. The first two paragraphs are now section 2.2.1 (Target variable). The remainder of this section is now 2.2.2 Explanatory variables.

We have prepared the table as suggested. Please see our answer to [14].

We have reworked Figure 6 so that it matches the explanation in the text in order.

[32] Page 6. Line 3. Specify correctly if the numbers 530 and 452 are the number of villages?

Agreed. The text is correct. Both numbers refer to the number of villages. 530 is the number of villages with a 100% success rate, of which 452 have only one borehole.

[33] Page 9: Put in order Figure 5a; before Figure 6.

Agreed. Fixed.

[34] Page 9: Rewrite the last sentence by also, it was used to.

Agreed. Fixed. It now reads *"It was also used to obtain…"*

[35] Page 9. Figure 4 must be centered.

Agreed. Fixed.

[36] Page 12: Fig.12: order the number of figures following the description found on page 9, i.e.: curvature,slope, topographic wetness index (TWI),

Agreed. Fixed.

[37] Page 13: Number 2.5 was repeated on page 14. Check it. I have the impression that the authors did not take the time to proofread the document.

Fixed. We did take the time to proofread the document. Minor mistakes happen.

[38] Page 13: 2nd paragraph. The objectives of the study mentioned at various locations in the manuscript should be summarized at the end of the introduction (see comments above). Why did you put the maingoal of the study here? I think that the objective must be found in the introduction section.

Agreed. Fixed. This was just a reminder of the goal. We have deleted it as explained above, so that there

is no confusion. Please see our answer to [3].

[39] Page 14. 3rd paragraph. Did you fix the number of iterations at 500 in this study? It is the default valueof model? Justify how this number was established.

Agreed. The third paragraph in section 2.6 now reads:

"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. A sensitivity analysis of the number of iterations was performed. The best compromise between results and computational cost were obtained by fixing the number of iterations at 500. No significant improvement in scoring metrics was observed for higher values, whereas running times increased considerably."

[40] Page 17: you mentioned in the first line that "The AUC exceeds 0.90 in all cases". I'm not sure about this affirmation because, if you analyse Table 1, you observe that in the MaxAbs scaling method AdaBoost shows an AUC value of 0.898. Could you rewrite your sentence to take into account this case? Or maybe use AUC mean because this value exceeds in all cases.

Agreed. The sentence now reads:
"AUC exceeds 0.90 in all cases except for the AdaBoost algorithm (0.898) and Decision Tree algorithm (0.893)."

[41] Page 20. First-line (Line 415). You mentioned Naghibi and Pourghasemi (2015), the citation is incorrectbecause you have three authors: Seyed Amir Naghibi & Hamid Reza Pourghasemi & Barnali Dixon. At the end of this paragraph again, you mentioned (Naghibi and Pourghasemi, 2015; Nguyen et al. 2020b). Due to this error for the citation in two places, I propose to check all references.

We believe our referencing is correct. Naghibi and Pourghasemi (2015) is included in the reference list (page 28, last-line). This reference only has two authors:

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.

The other paper by Naghibi, Pourghasemi and Dixon (Naghibi et al 2016) is quoted in the introduction.

We have double-checked all references anyway.

[42] Page 21: Move Figure 8 on page 21 under the section of "3.3 Importance of explanatory variables"

Agreed. Fixed.

[43] Page 23: I propose to add the well locations/boreholes on the two maps in Figure 8. If possible, put on these two maps: well training and well validation with different colours of points.Also, make clear your legend of Fig.8 with the classes well-defined.
For example:
(0- 0.2) Very low;
(0.2- 45) Low;
Etc.
Change the term "Intermediate" in the legend by "Moderate". It is most appropriate. Change

“Groundwater potential” to "Groundwater potential classes”.

Agreed. Fixed. We assume the reviewer refers to Figure 10 below.

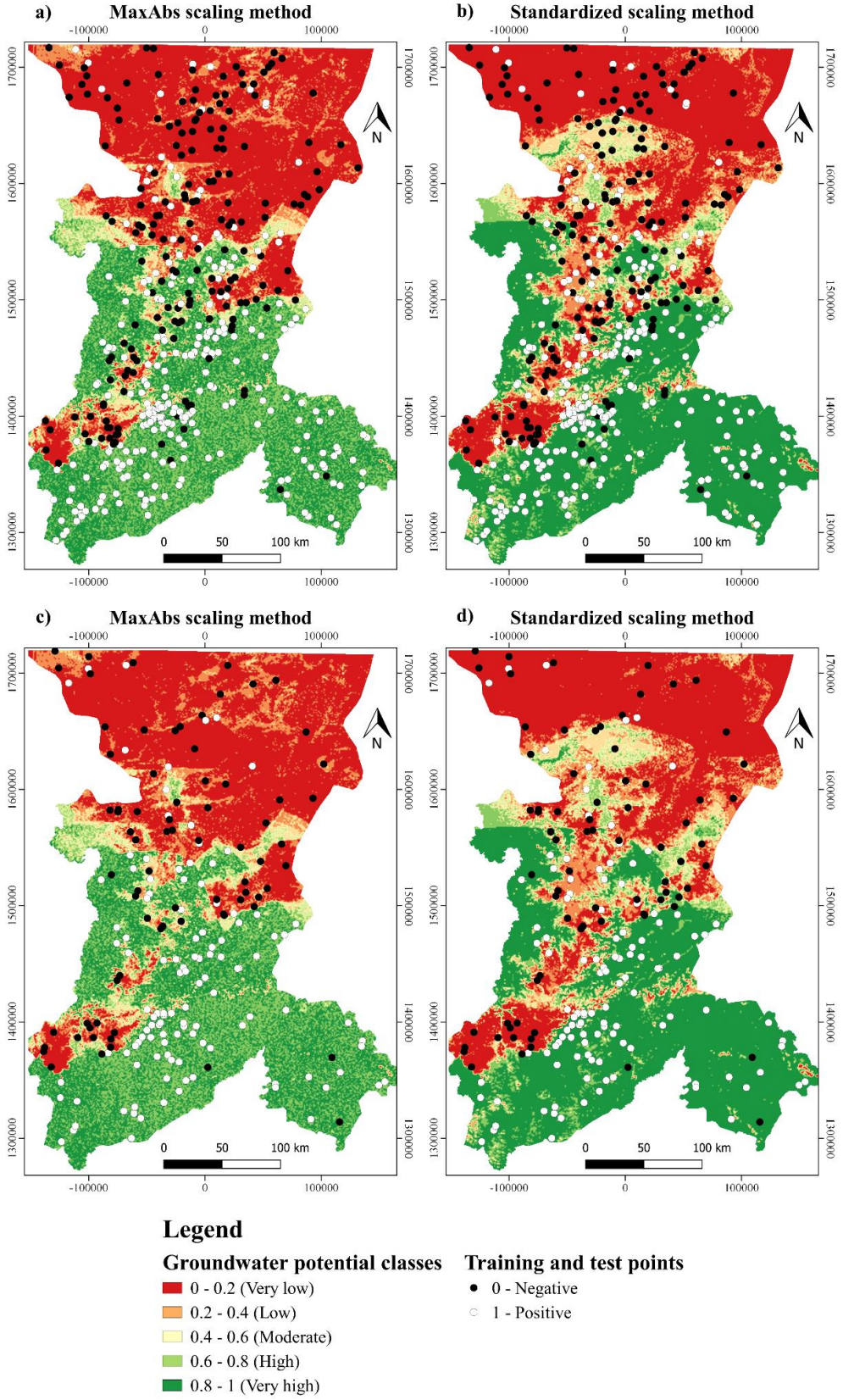

Figure 11. Mapping outcomes of the agreement map for MaxAbs scaling method and Standardized scaling method. a) Training points on the MaxAbs scaling method agreement map. b) Training points on the Standardized scaling method agreement map. c)

Testing points on the MaxAbs scaling method agreement map. d) Testing points on the Standardized scaling method agreement map.

[44] Page 21. I repeat the need to clarify my request mentioned above (see general comments). When you analyse feature importance calculated in Figure 8, you observe that some explanatory variables are not important in the models. Could you explain more how many variables did you select to produce the outcomes of Figure 10?

Agreed. Please see our answer to [18].

[45] Page 23. Why did you choose to classify villages in three classes based on groundwater potential, andyou show the outcomes of Groundwater potential in five classes?

Agreed. To clarify: classes are potential outcomes. There are only two (positive and negative), and groundwater potential is classified in positive and negative for each algorithm (Figure 9). However, ensembling allows for a finer classification. In Figure 10 the arithmetic mean of all five best classifiers is computed at the pixel level. This renders six possible values (0, 0.2, 0.4, 0.6, 0.8 and 1), which represent the agreement among the best classifiers for each pixel. This in turn allows for five intervals (0-0.2 Very low; 0.2-0.4 Low; 0.4 - 0.6 Moderate; 0.6 - 0.8 High; 0.8 - 1 Very high). We represent each of these agreement level as a groundwater potential outcome.

Furthermore, by dividing the villages into five classes, the sample of villages falling into high, moderate and low potentials is very small (approximately 20 points per class compared to 170 for high potentials and 70 for low potentials). It is therefore difficult to draw large conclusions from such comparatively small samples, and so we have chosen to group them in such a way that the number of villages in each category is more balanced (nearly 70).

[46] Page 24. Could explain more why Groundwater potential classes are three in Table 3 compared to Figure 10, where we found five classes?

Agreed. Please see [45] above

[47] Page 15. Section 3 on Results and discussion. Please add a new sub-section on "Validation on machinelearning models.

Please see [9].

[48] Reference section Page 27. Rewrite this reference: Direction Nationale de l'Hydraulique (Ed.): Données Hydrogeologiques et des Forages. Direction Nationale de l'Hydraulique du Mali, 2010. to the precise country name.
Agreed. The reference now reads:

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

[49] Page 29. Precise the link and access date of this reference: Poggio, L. and de Sousa, L.: SoilGrids250m 2.0 - Clay content, 2020.
Agreed. The reference now reads:

Poggio, L. and de Sousa, L.: SoilGrids250m 2.0 - Clay content, https://soilgrids.org/, Access date: 15/02/2020, 2020.

[50] Page 30.
Add the access date of the reference Traore, A, Z., et al.
Agreed. The reference now reads:

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, Access date: 27/10/2020, 2018.

[51] Add the link and access date of this reference: United Nations: Resolution A/RES/64/292. UnitedNations General Assembly, United Nations, 2010.
Agreed. The reference now reads:

United Nations: Resolution A/RES/64/292. United Nations General Assembly, United Nations, https://undocs.org/pdf?symbol=en/a/res/64/292, Access date: 10/03/2021, 2010.

[52] Technical corrections. Page 5. Line 115-120. Add the unity of mean water depth to be coherent in the sentence, because youhave put the unity of mean electric conductivity.
Agreed. Fixed.

[53] Page 6: in Figure 2 B, write correctly m^3/h
Agreed. Fixed.

[54] Page 8. Line 160-165, add the comma in (BGS, 2021). Also, on Page 9 and the title of Figure 4, add acomma to the same reference.
Agreed. Fixed.

[55] Page 11, line 5. "semiarid" "semi-arid"
Agreed. Fixed.

[56] Page 13: Equation 6; define $\tilde{x}$
Agreed. Fixed.

[57] Page 25. Delete "s" in Conclusion.
Agreed. Fixed.

On a final note, we would like to thank Reviewer #1 again for a thorough review of our manuscript. We hope our answers will be enough to merit publication in HESS.

References:

Adeyeye, O.A., Ikpokonte E.A., Arabi, S.A. GIS-based groundwater potential mapping within Dengi area, North Central Nigeria, The Egyptian Journal of Remote Sensing and Space Science, Vol 22, 2, 175-181, ISSN 1110-9823, https://doi.org/10.1016/j.ejrs.2018.04.003, 2019.

Al-Djazouli, M. O., Elmorabiti, K., Rahimi, A., Amellah, O., and Fadil, O. A. M.: Delineating of groundwater potential zones 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, 2021.

Boughariou, E., Allouche, N., Ben Brahim, F. et al. Delineation of groundwater potentials of Sfax region, Tunisia, using fuzzy analytical hierarchy process, frequency ratio, and weights of evidence models. Environ Dev Sustain 23, 14749–14774. https://doi.org/10.1007/s10668-021-01270-x. 2021.

Delgado, A. C. G. A new method to map groundwater potential at a village scale, based on a comprehensive borehole database. An application to Sikasso, Republic of Mali. 2018.

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.

Dietterich, T. Overfitting and undercomputing in machine learning. ACM computing surveys (CSUR), 27(3), 326-327. 1995

Falah, F., and Zeinivand, H. GIS-Based Groundwater Potential Mapping in Khorramabad in Lorestan, Iran, using Frequency Ratio (FR) and Weights of Evidence (WoE) Models. Water Resour 46, 679–692 https://doi.org/10.1134/S0097807819050051. 2019.

Gómez-Escalonilla, V., Vogt, M.L., Destro, E., Isseini, M., Origgi, G., Djoret, D., Martínez-Santos, P. and Holecz F. Delineation of groundwater potential zones by means of ensemble tree supervised classification methods in the eastern Lake Chad basin. Geocarto International. Under Review.

Guisan, A., Graham, C.H., Elith, J., Huettmann, F., the NCEAS Species Distribution Modelling Group. Sensitivity of predictive species distribution models to change in grain size. Divers. Distrib. 13, 332–340. 2007.

Gumma, M.K., Pavelic, P. Mapping of groundwater potential zones across Ghana using remote sensing, geographic information systems, and spatial modeling. Environ Monit Assess 185, 3561–3579. https://doi.org/10.1007/s10661-012-2810-y, 2013.

Magaia, L.A., Goto, Tn., Masoud, A.A., and Koike, K. Identifying Groundwater Potential in Crystalline Basement Rocks Using Remote Sensing and Electromagnetic Sounding Techniques in Central Western Mozambique. Nat Resour Res 27, 275–298. https://doi.org/10.1007/s11053-017-9360-5, 2018.

Mall, I., Diaw, M., Madioune, H. D., Ngom, P. M., and Faye, S.. Use of Remote Sensing and GIS for Groundwater Potential Mapping in Crystalline Basement Rock (Sabodala Mining Region, Senegal). GIS. Nova Science Publishers, 317, 2014.

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.

Martínez-Santos, P., Díaz-Alcaide, S., De la Hera, A., Gomez-Escalonilla, V. A multi-parametric supervised classification algorithm to map groundwater-dependent wetlands. Journal of Hydrology 603 (2021) 126873.2021a.

Martínez-Santos, P., Aristizábal, H.F., Díaz-Alcaide, S., Gomez-Escalonilla, V. Predictive mapping of aquatic ecosystems by means of support vector machines and random forests: an application to the Valle del Cauca region, Colombia. Journal of Hydrology 595 (2021) 126026 DOI: 10.1016/j.jhydrol.2021.126026. 2021b.

Mogaji, K.A., Lim H.S. (2018) Application of Dempster-Shafer theory of evidence model to geoelectric and hydraulic parameters for groundwater potential zonation, NRIAG Journal of Astronomy and Geophysics, 7:1, 134-148, DOI: 10.1016/j.nrjag.2017.12.008. 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.

Mpofu, M., Madi, K., and Gwavava, O. Remote sensing, geological, and geophysical investigation in the area of Ndlambe Municipality, Eastern Cape Province, South Africa: Implications for groundwater potential, Groundwater for Sustainable Development, Vol 11, 100431, ISSN 2352-801X, https://doi.org/10.1016/j.gsd.2020.100431, 2020.

Obeidavi, S., Gandomkar, M., Akbarizadeh, G., Delfan, H. (2021). Evaluation of Groundwater Potential using Dempster-Shafer Model and Sensitivity Analysis of Effective Factors: A case study of North Khuzestan Province. Remote Sensing Applications: Society and Environment 22 (2021) 100475. 2021.

Owolabi, S.T., Madi, K., Kalumba, A.M. and Orimoloye, I.R. A groundwater potential zone mapping approach for semi-arid environments using remote sensing (RS), geographic information system (GIS), and analytical hierarchical process (AHP) techniques: a case study of Buffalo catchment, Eastern Cape, South Africa. Arab J Geosci 13, 1184. https://doi.org/10.1007/s12517-020-06166-0, 2020.

Rahmati, O., Pourghasemi, H. R., & Melesse, A. M. Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: A case study at Mehran Region, Iran. Catena, 137, 360–372. doi:10.1016/j.catena.2015.10.010. 2015.

Saadi, O., Nouayti, N., Nouayti, A., Dimane, F., and Elhairechi, K. Application of remote sensing data and geographic information system for identifying potential areas of groundwater storage in middle Moulouya Basin of Morocco, Groundwater for Sustainable Development, Vol 14, 100639,ISSN 2352-801X, https://doi.org/10.1016/j.gsd.2021.100639, 2021

Worthington, S.R. Diagnostic tests for conceptualizing transport in bedrock aquifers. J. Hydrol. 529, 365–372. 2021.

Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. Privacy risk in machine learning: Analyzing the connection to overfitting. In 2018 IEEE 31st Computer Security Foundations Symposium (CSF) (pp. 268-282). IEEE. 2018.