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).

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 applying machine learning. The manuscript can be interesting to the scientific community working on machine learning 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.

- 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). 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.

- 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? Also, the authors should be linking the issues of groundwater resources in the context of The Sustainable Development Goals (SDGs) to motive the reader at the end of the introduction.

- 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.

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

- 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.

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

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

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

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

- 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.

- 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.

- A table of data sources must be put to increase the clarity and to ease readers’ understanding.

- Do you have performed quality control of datasets before the modelling?

- 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 these four (04) explanatory variables in reality? Specify exactly the variables used in the final models.

- The conclusion is very general. To be check according to the revision of MS.

- Limitations of the research should be addressed.

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

- 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.

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

- 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?

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

- **Abstract section:**

- Page 1. The abstract is very long. It should be shortened and focused.

Keywords: I propose to delete “big data; climate change and water access”; and add “Groundwater potentiality, and GIS”.

- The abstract should be thoroughly revised according to the revision of the manuscript.

Specific comments:

Page 2. Paragraph 1. Line 2, rewrite the sentence “Today, 2.5 billion people….” by “Today, 2.5 billion people around the World…”.

Page 2. Paragraph 3. Line 1. Introduce a sentence before this: “There are two main approaches to GPM: expert-based decision systems and machine learning methods”.

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. In your Table, you could, for example, put in 4 columns (Type of data layers/ Explanatory variables; scale/resolution, time, available format, and source of data).

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

Page 9: Put in order Figure 5a; before Figure 6.
Page 9: Rewrite the last sentence by also, it was used to.

Page 9. Figure 4 must be centered.
Page 12: Fig.12: order the number of figures following the description found on page 9, i.e.: curvature, slope, topographic wetness index (TWI),

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.

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 main goal of the study here? I think that the objective must be found in the introduction section.

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

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.

Page 20. First-line (Line 415). You mentioned Naghibi and Pourghasemi (2015), the citation is incorrect because 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.

Page 21: Move Figure 8 on page 21 under the section of “3.3 Importance of explanatory variables” on
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”.

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?

Page 23. Why did you choose to classify villages in three classes based on groundwater potential, and you show the outcomes of Groundwater potential in five classes?
Page 24. Could explain more why Groundwater potential classes are three in Table 3 compared to Figure 10, where we found five classes?

Page 15. Section 3 on **Results and discussion**. Please add a new sub-section on “Validation on machine learning models.

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.
Page 29. Precise the link and access date of this reference: Poggio, L. and de Sousa, L.: SoilGrids250m 2.0 - Clay content, 2020.

Page 30.

- Add the access date of the reference **Traore, A, Z., et al.**
- Add the link and access date of this reference: United Nations: Resolution A/RES/64/292. United Nations General Assembly, United Nations, 2010.

Technical corrections

Page 5. Line 115-120. Add the unity of mean water depth to be coherent in the sentence, because you have put the unity of mean electric conductivity.
Page 6: in Figure 2 B, write correctly m^3/h
Page 8. Line 160-165, add the comma in (BGS, 2021). Also, on Page 9 and the title of Figure 4, add a comma to the same reference.
Page 11, line 5. “semiarid” “semi-arid”

Page 13: Equation 6; define $\tilde{x}$
Page 25. Delete “s” in Conclusion.