Referee comment to:

EGUsphere-2023-1898 Revision report

"High-resolution long-term average groundwater recharge in Africa estimated using random forest regression and residual interpolation" by Pazola et al. (2023).

General Comments

Pazola et al. (2023) provide an interesting machine learning and residual interpolation for groundwater recharge mapping at the continental scale of Africa.

The authors have used machine-learning called random forest model to estimate groundwater recharge across Africa. In addition, their models explore the potential factors affecting groundwater potential.

The paper is interesting and within the scope of the EGUsphere journal. In general, machine learning is well-placed in EGUsphere. The authors have done very diligent work by summarizing many publications applying machine learning and linear mixed models. The manuscript can be interesting to the scientific community working on machine learning applied in hydrology. The manuscript is very well written and we thank the authors for adding the codes, however at the present state; I would not recommend it for publication because certain comments need to be addressed again.

- The introduction is well. 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). The overfitting problem is one of the drawbacks that affect the accuracy of models in machine learning. Why did decide to choose the Random Forest compared to LLM models? It would be interesting if you compare machine-learning models and physics-based models to estimate groundwater recharge.

- Can you explain why the choice of the period of modelling 1981-2010? Because the input data in Table S1 has multiple Periods.

- Did you limit the validation of the random forest model with cross-validation? Alternatively, do you have the intention to integrate the external validation by compiling local raw data?
- The authors need to highlight deep the uncertainty in GIS data resampling. According to the authors what was the influence of the data resampling (0.5° spatial resolution and 0.1° spatial resolution) in the different models (LLM and RF models)?
- We know that RF is robust against the multicollinearity of features. Did you try to test the multicollinearity of predictive factors? If not, please can you use the variance inflation factor (VIF) and tolerance (TOL) indices as are customarily used to estimate the multicollinearity of all predictive factors in machine learning modelling? For example, we think that Precipitation and ET are not a problem for parameter estimation because Aridity is based on P and ET. Can you give more explanations?

- Can you explain to us the difference between the final variables in your random forest model compared to the variables selected in the study of Moeck et al. (2020)? *A global-scale dataset of direct natural groundwater recharge rates: A review of variables, processes and relationships*. https://doi.org/10.1016/j.scitotenv.2020.137042. Please cite this reference in your study.

- What is the effect of training dataset sample size on the performance/quality during the implementation of the RF model?

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

- We know that the various GIS layers come with different spatial resolutions. Why did you choose to develop the final map at 0.1° spatial resolution? Can you explain the choice of this type of resolution?

- Do you have performed/checked quality of GeoTIFF datasets before the modelling?

- In the discussion, the authors must address the uncertainty in the GIS explanatory dataset used to estimate groundwater recharge (deficiencies of data quality; biased and absent data, sample sizes, missing covariates, etc.).

- Is it possible to improve the performance of the random forest model developed in your study? Which additional predicting variable (s) (even if such information is scarce) could be added to improve the results?

- Why you did not test the continental scale model at the country level/scale by using the best variables retained in your final model? In others words, Can you validate your machine learning model at the local scale?

- **Abstract section:**

Line 10: Put semicolon ";" between 0.83 and 0.88

Specific comments:

Page 2:
Line 23, replace ~ by the word approximatively.
Line 28. Add, "s" in the word "contributes".

Page 3:
Line 76. Add the article "a" in this sentence "A recent study by Huang et al. (2019) employed a multi-layer perception network"……
Line 88. In this sentence, "In the field of groundwater modelling. The RF technique has… please check the *dot* between modelling and The RF.

Page 4.

Line 92. It may be interesting to show the equivalent of the spatial resolution like 0.5° in terms of distance (km) for more appreciation.

Line 107. Add the term 'the two" before "different models.

Line 108. We think that this paragraph "*Section 2 summarises the study area and the spatial characteristics of its groundwater resources, and outlines the data sources and the model development process. Section 3 presents the results of the modelling experiments. Section 4 discusses these results in the wider context and critically evaluates the developed model*" is not very important here and can be removed and keep just the sentence starting by "this study is accompanied by a Supporting Material that provides extensive information on the predictors used and additional analyses that extend the investigation presented in this paper.

Page 5. The study area section is not clearly presented. For example when the authors say that: "*These provide a basis for the division of the continent into 8 climatic regions, most of which experience high interannual rainfall seasonality".* We need to present clearly with a little section these 8 climatic regions. Please improve this section.

Page 6.

Make sure all your Figures are correctly inserted. Because, for example, the map of Figure 1 cuts the sentence in Line 151.

Line 160. Add "The" before number of wet days.

Page 7. Line 169. Just say: To create the groundwater recharge map…

Page 12 and Page 13. Add some reference to justify your finding in semi-arid and arid context results such as Burkina Faso, Ethiopia, etc. Please Line 327 to Line 356.

Page 12. Insert the Table 1. Optimal random forest hyperparameters found through random search with cross-validation for different random forest model variants used in this study at the end of this sentence: " *The model underestimates these samples (136 obs/38 pred, 221 obs/64 pred, 266,…"*

Page 16. Again, a map of Figure 3 divides the sentence in Line 380. Need to be arranged.

Technical correction

Page 6. Line 160. Please put a space between 300 and meter.