RESPONSES TO REVIEWER 2:

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.

R9 => We thank reviewer 2 for their positive comments above and are of the view that the responses and revisions to the manuscript now warrant publication of the paper in a revised form.

General Comments

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.

R10 => We worked with a relatively small dataset (134 points in total) and a set of explanatory variables that are correlated with each other. These were the most important factors that influenced the algorithm choice. There are multiple algorithms that could be applied to this problem (e.g. support vector machine, Gaussian process regression, random forest, gradient boosting decision tree, XGBoost, symbolic regression) and we chose RF as it was previously successfully applied to large-scale groundwater studies (Podgorski and Berg, 2020) and to studies with datasets of a similar size (Pham et al. 2022). RF is robust to overfitting, as the final prediction is an average of predictions from multiple decision trees, each trained with a different subset of data. In addition to RF, we applied residual kriging to explicitly account for variability in LTA recharge observations around the fitted values. We note that there are published studies applying machine learning with MCDA in hydrology, but it is unclear how it could be applicable to this regression problem. Weight of evidence could be an interesting addition to assess feature importance; this is examined in this work using other methods. Frequency ratio and fuzzy logic could introduce additional value to our analysis their inclusion is beyond the scope of the current analysis.

Comparing ML models with physical models would likely be a valuable exercise but it constitutes a separate study beyond the scope of the current analysis as would comparisons from a set of different algorithms (listed above).

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

R11 => The choice of the modelling period is dictated by the original input data, namely determinations of mean annual recharge from a variety of methods covering the period of 1981 to 2010. We used secondary publicly available data and we have not produced any data on our own. CGIAR-CSI data (Aridity, PET) are only available for 1970-2000. It is not possible to get average values from that dataset for 1981-2010. CGIAR-CSI is widely used and is representing long term climatic average. We decided that it can be representative for the period our analysis. Land cover is a categorical variable. The revised manuscript uses a different dataset (Historical Land-Cover Change and Land-Use Conversions Global Dataset - NOAA dataset) with a mode value for 1981-2010. When revising the input sources, we noticed an error in Table S1. Landcover data was not used to create models that generated maps at 0.5° and 0.1° spatial resolutions. Also, the number of wet days is missing (Harris et al. 2020), as it was used in the variable selection process. Table S4 is updated accordingly.

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?

R12 => An external validation is preferable, but the analysis is rooted in limited historical data compiled from local raw data from different regions.

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)?

R13 => Most predictive factors were available at a higher resolution than 0.5° or 0.1° so each input raster was upscaled using bilinear (continuous data) or mode (categorical data) resampling methods to an appropriate resolution (0.5° or 0.1°). Bilinear interpolation is a good standard technique but loss of ultra small-scale details is inevitable. The revised manuscript contains a note of uncertainty in GIS data resampling.

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?

R14 => By design random forest should not be affected by correlated features. We focused on prediction, not on interpretability. A detailed analysis of feature importance was outside of the scope of this analysis but could be extended using these suggestions.

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.*

R15 => Moeck et al. (2020) point out that recharge estimates based solely on climatic variables can be misleading and that vegetation and soil structure have an explanatory power too. It's a reasonable assumption and this could be addressed in a separate study that looks more carefully into variable importance and focus on interpretability of machine learning models.

In our study, variables were selected for the model to match the data used previously in the Linear Mixed Model by MacDonald et al. (2021); most of the data are the same datasets. The MacDonald data set underwent a more thorough and transparent QA to give a curated dataset using techniques only appropriate to the African environment. Interestingly although local factors in soil and geology are important in controlling local recharge as shown by the residuals in the model, they do not improve large scale continental model - as discussed in MacDonald et al. 2021. In the follow up paper from the Moeck 2020 paper, only climatic factors are used for global modelling (Berghuijs et al. 2022, https://doi.org/10.1029/2022GL099010)

We add these points to the discussion section in the revised manuscript, citing Moeck et al. (2020) and Berghuijs et al. (2022).

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

R16 => It has a marginal effect.

```
train/test ratio 70% to 30%
Mean R2 train (log): 0.94
Mean R2 test (log): 0.63
Mean out-of-bag error: 0.63

train/test ratio 75% to 25%
Mean R2 train (log): 0.94
Mean R2 test (log): 0.60
Mean out-of-bag error: 0.65

train/test ratio 80% to 20%
Mean R2 train (log): 0.94
Mean R2 test (log): 0.61
Mean out-of-bag error: 0.65
```

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?

R17 => We relied on built-in feature importances of RF algorithm and investigated changes to R^2 metric when gradually adding factors to identify which factors have very little or no explanatory power.

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?

R18 => It's the highest possible resolution that we could obtain raster data for the FLDAS soil moisture dataset (McNally et al. 2018) (at the time of the study). Global hydrological models typically work at 0.5° spatial resolution, and large-scale prediction maps (e.g. MacDonald et al., 2021) are produced at that resolution too. We sought to demonstrate that data-driven modelling can create prediction maps at a higher resolution, given input data of good quality. We are aware that there is a high uncertainty in the predictions and do not claim that a higher resolution is better (see R1); prediction maps at a higher resolution can, however, be obtained from a similar effort as lower resolution maps.

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

R19 => Yes, we checked missing values and values range.

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.).

R20 => We recognise that considerable uncertainty will exist in gridded datasets (their representativity over 100 km^2 to 2500 km^2) used to estimate recharge. We opted for high-quality, published, and peer-reviewed datasets, and their origins are outlined in Table S1.

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?

R21 => We employed the most appropriate gridded datasets available based on the necessity of inclusion. It may be possible that better, detailed representation of vegetation and soil structure variables may improve results but this constitutes a separate study beyond the aims of this one.

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?

R22 => We welcome the suggestion of conducting basin-level analyses, in an area with dense observational network but are unaware of the availability of datasets currently permitting such an analysis.

Minor comments

R23 => All following suggestions were considered and implemented, unless stated otherwise.

Abstract section:

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

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. 3

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.

We respectfully decline this stylistic request to remove this paragraph since these sentences helpfully map out the order and logic of the manuscript.

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.

REFERENCES

Berghuijs, W. R., Luijendijk, E., Moeck, C., van der Velde, Y., & Allen, S. T. (2022). Global recharge data set indicates strengthened groundwater connection to surface fluxes. *Geophysical Research Letters*, 49, e2022GL099010.

Cuthbert, M. O., Taylor, R. G., Favreau, G., Todd, M. C., Shamsudduha, M., Villholth, K. G., ... & Kukuric, N. (2019). Observed controls on resilience of groundwater to climate variability in sub-Saharan Africa. *Nature*, *572*(7768), 230-234.

Fox, E. W., Ver Hoef, J. M., & Olsen, A. R. (2020). Comparing spatial regression to random forests for large environmental data sets. *PloS one*, *15*(3), e0229509.

Harris, I., Osborn, T. J., Jones, P., & Lister, D. (2020). Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Scientific data*, *7*(1), 109.

MacDonald, A. M., Lark, R. M., Taylor, R. G., Abiye, T., Fallas, H. C., Favreau, G., ... & West, C. (2021). Mapping groundwater recharge in Africa from ground observations and implications for water security. *Environmental Research Letters*, *16*(3), 034012.

McNally, A., Arsenault, K., Kumar, S., Shukla, S., Peterson, P., Wang, S., ... & Verdin, J. P. (2017). A land data assimilation system for sub-Saharan Africa food and water security applications. *Scientific data*, *4*(1), 1-19.

Meinshausen, N., & Ridgeway, G. (2006). Quantile regression forests. *Journal of machine learning research*, *7*(6).

Moeck, C., Grech-Cumbo, N., Podgorski, J., Bretzler, A., Gurdak, J. J., Berg, M., & Schirmer, M. (2020). A global-scale dataset of direct natural groundwater recharge rates: A review of variables, processes and relationships. *Science of the total environment*, *717*, 137042.

Pham, Q. B., Tran, D. A., Ha, N. T., Islam, A. R. M. T., & Salam, R. (2022). Random forest and nature-inspired algorithms for mapping groundwater nitrate concentration in a coastal multi-layer aquifer system. *Journal of Cleaner Production*, *343*, 130900.

Podgorski, J., & Berg, M. (2020). Global threat of arsenic in groundwater. *Science*, *368*(6493), 845-850.