RESPONSES TO REVIEWER 1:

The authors use an RF approach to generate spatial long-term average groundwater recharge for Africa based on 134 recharge values from the literature and compare their results with the field observations and a previous publication using an LMM (linear mixing model). The results are generated and compared for two spatial resolutions. The RF approach is very similar to LMM but offers a higher spatial variability than LMM and therefore also shows small-scale trends.

Even though the approach is generally ok, the manuscript is very well written and the workflow and code(s) is available through github (which I really appreciate), I still have some critical points that should be considered and discussed in detail in a revised version.

I'm somewhat unsure about the better spatial resolution of the results. Just because the resolution is better doesn't mean the results are more reliable. There is a very large uncertainty due to the few observations and their distribution but the maps suggest a much better and more robust result and this is dangerous. What would be the next step with the results or what can the better spatial resolution be used for? If the data is extracted directly from the maps (for water budget calculations, for example) this can lead to very distorted results, as the simulated recharge values are very uncertain for many areas. I believe the whole uncertain should be better discussed and the maps must better highlight the uncertainties (maybe with transparent colors, see my comment below)

R1 => We thank Reviewer 1 for their positive comments on our manuscript and note their concerns over the possible inference that high-resolution predictions are more robust. We do not claim that higher resolution data are more robust yet understand that this potentially could be implied, especially given comparison with predictions of the linear mixed model that were considerably smoothed out. We address the issue of uncertainty by constructing prediction intervals for each grid cell using Quantile Random Forest (Meinshausen, 2006; Fox et al. 2020). Although RF provides information on the conditional mean of the output variable, QRF instead provides information on the conditional distribution function of the response. By providing the prediction intervals, the reader and potential user are informed of the underlying prediction uncertainty.

Current global hydrological models typically operate at 0.5° spatial resolution, and large-scale prediction maps like MacDonald et al. (2021) have similarly been produced at this resolution. There is, nevertheless, an on-going trend toward hyper-resolution models (e.g. 0.1°) at continental to global scales. There is thus a need for robust approaches to the development of empirically derived datasets at higher spatial resolutions to test large-scale recharge models and support recharge mapping.

I wonder why, for example, seasonality in precipitation is not present in the climatic input data. In some regions, precipitation only falls in a few months and therefore the processes for recharge are significantly different for conditions when precipitation is distributed *throughout the year. Yes, LMM or RF show a good fit /regression, but certain parameters may compensate for the missing input. Also, of course, the relative importance does not show the importance of seasonality but only because this has not been tested in the RF (although it was in the previous work using LMM, but this is not transferable directly to the RF approach).*

R2 => Seasonality in precipitation dominates the hydrology of all modelled areas on continental Africa whether in the equatorial humid tropics, tropical drylands or sub-tropical locations. This analysis estimates recharge at annual timescales and thus does not specifically capture seasonal variability in precipitation. We thank Reviewer 1 for bringing to our attention the fact that the number of wet days is not mentioned in the manuscript. It was originally considered as an input variable but it was not selected for the final model due to its weak influence. This point is included in the revised paper (e.g. Tables S1 and S4 in the Supplementary Material). The data source for the number of wet days is Harris et al. (2020), which was used in the LMM study by MacDonald et al. (2021). After rerunning the analysis, we confirm that the number of wet days was not included in the final models due to its weak explanatory power. It showed a strong correlation with NDVI and its inclusion in the predictor set did not improve the model fit in terms of R^2 for training and testing datasets respectively: (1) model with # of wet days 0.93/0.79; and (2) model without # of wet days 0.93/0.81.

Similar for depth to groundwater table (or call it unsaturated zone thickness) which is important for recharge processes, rate and timing. How important is this input for the RF algorithm and for the process description. I also wonder why distance to rivers is not included as an (raster)input, perhaps paired with discharge rates. This would help to better capture the important process of groundwater-surface water interaction and bank filtration, which many of the authors know better than I do.

R3 => The observational dataset on groundwater recharge, compiled by MacDonald et al. (2021) only includes diffuse recharge points. Focussed recharge is an important recharge regime, especially in drylands (Cuthbert et al., 2019), with strong seasonality in precipitation but is not specifically reported in the dataset. Consequently, we did not include predictors related to surface water-groundwater interactions as the objective of our analysis was to compare directly the RF model to another data-driven model (LMM) by MacDonald et al. (2021). There are other possible explanatory factors that we could have been considered besides the groundwater table depth such as soil structure and vegetation but this would render differences between the RF and LM models when our aim was to compare these modelling methods.

Of course there is a large uncertainty in the precipitation data sets and in the timing of recharge, but wouldn't it be possible to minimize these uncertainties and also the scaling (regression is dominated by the high recharge values) significantly by using the recharge / precipitation ratio and obtain more robust results? It would be nice if this can be discussed and tested more.

R4 => We welcome this suggestion from Reviewer 1 to minimize uncertainties associated with precipitation datasets using a recharge/precipitation ratio (i.e. the proportion of precipitation that is converted to recharge). However, given the established non-linear (power law) relationship between recharge and precipitation (see LMM – MacDonald et al. (2021) and RF models), we see no computational advantages to employing such an approach.

How does the spatially uneven distribution of the observations affect the results? Wouldn't it make more sense to show only the more robust areas and show the very uncertain ones transparently? Since not all climatic conditions have been covered, would clustering be useful to minimize the spatial discrepancy and influence?

R5 => We demonstrated that some data points have impact on recharge predictions in different regions (e.g. inclusion of zero-recharge points located in Sahara amplifies the predicted high-recharge values in the humid regions). Therefore, such simple uncertainty indicator could be misleading as well. We cannot exclude that the opposite can be true too, namely inclusion of more high-recharge observation might have an impact on predictions in more arid regions. We also showed that the model is biased towards dry regions, as historically these areas were of interest for groundwater studies. Data scarcity in humid regions leads to high residual in predicted vs observed values. In the revised manuscript, we use Quantile Random Forest to construct prediction intervals and based on the results and provide maps visualising the prediction uncertainty.

Is the correlation of the aridity index with precipitation and ET not a problem for parameter estimation and generally with all estimation methods? Aridity is based on P and ET, and I wonder what is the advantage of using all three parameters? Looking at the SI, precipitation and aridity are the most important parameters, and I wonder what the results would look like if only aridity was used. When I see table S4, I wonder why the results look almost the same for training and test, even if only P us used.

R6 => Correlation of precipitation, ET (evapotranspiration) and AI (Aridity Index) is not an issue for the algorithm itself but it’s true that these variables might altogether represent redundant input. From the point of view of the model, any of these correlated features can be used as the predictor, with no concrete preference of one

over the others. We decided to keep all these variables as, when used together, the fit of the model was marginally improved.

Regarding the data in Table S4, please see our rationale above. Precipitation explains most of the variability in GW recharge, better than Aridity Index. We checked model performance with aridity alone and the model fit in this case wasn't as good as with precipitation as the only input. There is a small improvement in the model fit when all three variables P + PET + AI are used, compared with P alone or P + PET.

Predictor set	**R^2 train (log)**	**R^2 test (log)**
Precip	**0.90**	**0.74**
Aridity	**0.90**	**0.61**

I'm not an expert on RF, but aren't the results validated using the ROC curve and sensitivity, specificity and accuracy rather than just the regression? That would be more informative about the model results and robustness instead of using only a regression, or?

R7 => All these concepts are reserved for classification problems. The model performance in a classification problem is assessed through a confusion matrix from which accuracy, sensitivity, and specificity are obtained from. For regression problems, different metrics are computed such as mean square error or coefficient of determination, which can show how accurately predicted values match known values; they were used in this study.

Line 451: Also process based models require careful input selection and quantification of uncertainties in the input dataset.

R8 => We agree.

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.