Response to Reviewer Comments – Reviewer 3

We thank the editor and the three reviewers for their constructive comments and suggestions. We thank all three reviewers' comments that the text is well written and for their recognition of our study as a valuable resource for groundwater practitioners. We believe that in addressing their comments, the manuscript will be considerably improved and be ready for publication.

Most questions were about minor text updates and queries. Two reviewers asked for further comparisons between our outputs and previous investigations. We present a suggested approach to address these comments, including new figures for both the manuscript and the supporting information.

We believe that these additions directly address reviewer concerns, clearly showing the impact of the distribution of our estimates as a primary control on the differences in recharge estimates between our study and previous studies.

Our responses to the reviewer's comments (**RC**) are provided below as author's comments (**AC**). To help with the assessment of our responses, we colour coded our responses into agreement (green), partial agreement (yellow) and disagreement (red). When referring to text excerpts in our manuscript, we have provided the line number and whether ~~text has been removed~~, or if new text is added.

RC1: In this study, the authors used chloride mass balance (CMB) to derive long-term groundwater recharge rate estimates for Australia. A random forest model was built and tested using 17 relevant climatological, geologic, hydrologic, and static soil/vegetation variables as the predictors. The random forest model was validated using the point-scale CMB recharge rate estimates and the best-performing model was based on 8 of the 17 variables. The 8-variable model was used to generate the median, 5th, and 95th percentiles of recharge rate for the entire Australia. Overall, the manuscript is very well written. The experiments are set up in an organized and thoughtful way. I enjoy reading the discussion section where the authors provide guidance for practitioners to use the dataset. I only have some major and minor comments as outlined below.

> **AC:** Thanks for your interest, positive feedback, attention to detail and helpful comments on our manuscript that intend to improve our work.

RC2: Major comment:

Table 1: I have questions on the temporal evolution of these factors and the importance of the temporal component of the model. Depth to water table is a time-varying variable. Specify what value of depth to water table is used in this study.

> **AC:** **We agree (minor change to the manuscript suggested).**
>
> **Suggested addition to Table 1 of manuscript**
>
> Suggested addition to the "Description" column of the "Depth to water table" row in Table 1: Output of global numerical groundwater model. Mean simulated water table depth.

RC3: CMB is a method that measures long term (hundreds to thousands of years) groundwater recharge rate. I notice the authors use different time periods for different input features. My two questions are 1) Are those periods the longest time periods with data availability? 2) If yes to question 1), the time periods of data availability are still time periods that cannot match up the residence time of chloride which is on the order of hundreds to thousands of years). How did the authors go about that? The authors can do a sensitivity analysis on using different time periods of input variable to test the sensitivity of their model results to the choice of input time periods.

> **AC:** **We partially agree (no change suggested).** Yes, we have chosen the input datasets with the longest time periods available. The reviewer correctly highlights the residence time of chloride and that the data available and used in our study still would not perfectly match up the residence time of chloride. However, for the large scale of our study, the datasets used (i.e., rainfall, runoff, evapotranspiration, etc.) are the best available datasets to use for the CMB method. We believe that a sensitivity analysis on using different time periods would not be required for this study as: (1) The residence time of chloride is on the order of hundreds to thousands of years, (2) we utilise the datasets with the longest time period available, (3) testing the sensitivity of shorter time periods (i.e., decadal time period) would not be appropriate for the long residence time of chloride (i.e., minimum of hundreds of years).

RC4: Minor comments:

Line 105: "…we identified 17 different gridded datasets (Table 1)." Is the distance to coast a gridded dataset? If it is a gridded dataset, specify the spatial resolution in Table 1. Otherwise, change the wording on Line 105.

> **AC:** **We agree (minor changes to the manuscript suggested).**
>
> **Suggested text changes in manuscript**
>
> Suggested revision at line 105: To investigate factors that influence groundwater recharge, we identified 17 different spatial datasets – 16 of which are available as gridded maps ~~gridded datasets~~ (Table 1).

RC5: Table 1: The categories do not make total sense to me. Geology seems to belong to "Surface processes and hydrogeological" category. "geomorphological" can be changed to "soil properties". For sand, silt, and clay fractions, the description states they are 100 to 200 cm interval. Does this mean the input features are for the 100 -200 cm vertical layer? If yes, justify why choosing a deeper soil layer instead of values for the entire soil column.

> **AC:** **We agree (minor changes to the manuscript suggested) with the changes to Table 1 and Table 2. We partially agree with the latter part of the comment (no change made).**
>
> **Suggested changes to table in manuscript**
>
> Suggested revision in Table 1: Move Geology under the "Surface processes and hydrogeological" category. Change "geomorphological" category to "soil properties".
>
> Suggested revision in Table 2: Move Geology under the "Surface processes and hydrogeological" category. Change "geomorphological" category to "soil properties". Change all mentions of "geomorphological" or "geomorphology" to "soil properties".
>
> **Author comments**
>
> Yes, the input datasets for sand, silt and clay fractions are for the 100 to 200 cm vertical layer. This layer was chosen as it was the largest interval available out of a range of different intervals (i.e., 0 to 5 cm, 5 to 15 cm, 15 to 30 cm, 30 to 60 cm, 60 to 100 cm, and 100 to 200 cm), and therefore, would be the most effective interval at controlling recharge.

RC6: Line 205: Step (6) removes cases where estimated recharge equals or exceeds mean annual rainfall. Explain why and how did that happen. Could this be related to the errors underlying the estimation of recharge rates?

> **AC: We partially agree (no change suggested).** Step 6 in our data filtering process was implemented to remove recharge estimates that are unusually higher than the mean annual rainfall available at the location. There are three plausible reasons how such recharge estimates could occur: (1) under-estimation of groundwater chloride concentration used in our study due to limited measurements not being representative of actual long-term average, (2) error in the chloride deposition map used in our study (i.e., over-estimate of chloride deposition at recharge site), (3) the recharge estimate is in an area that receives additional sources of water that has significantly lowered the concentration of groundwater chloride (e.g., in an irrigation area). More detailed information was provided in the supporting information and referred to in line 206 of the manuscript.

RC7: Line 220: What is the "typical practice"? Specify the name of the method.

> **AC: We agree (minor change to the manuscript suggested).**
>
> **Suggested text changes to manuscript**
>
> Suggested revision at line 219: The dataset was split into a randomly selected training subset (70 %) and validation subset (remaining 30 %) following the train test split procedure ~~typical practice~~ (e.g., West et al., (2023); Sihag et al., 2020; Rahmati et al., 2016).

RC8: Line 221: Each tree in the random forest model (the model) was trained on n randomly selected observations, with replacement (i.e., bootstrapping) from the training subset, where n is equal to the total number of observations in the training subset.

> **AC: We disagree (no change suggested).** We are unsure what the reviewer was implying with this comment as the reviewer has presented text verbatim from the manuscript. We note that we apply the random forest analyses in a routine way.

RC9: Line 306: "The recharge area of these deep systems is likely to be hundreds of kilometres away from the bore location, whereas our analyses assume recharge occurs within the 0.05° × 0.05° pixel from the chloride deposition map that contains the bore." How does this influence the results or how do the authors address this question.

> **AC: We disagree (no change suggested).** We believe this has been addressed in line 304 to line 306. As our study aims to estimate recharge for the shallow water table aquifer system which we assume is receiving recharge at or close to the groundwater bore location, we filter out groundwater chloride samples with sample depths more than 150 metres below ground surface, following previous published examples of this threshold. The filtered out deep samples are the ones that we believe recharge likely occurs "hundreds of kilometres away from the bore location…", and hence are omitted from the analyses.

RC10: Line 314: "The mean recharge rate…", do the authors mean "spatial mean recharge rate"?

> **AC:** We p**artially agree (minor changes to the manuscript suggested).** We have addressed this with a suggested revision.
>
> **Suggested text changes in manuscript**

Suggested revision at line 314: The mean values of recharge rates for R_{50}, R_{95}, and R_5 (i.e., the point datasets) are 43.5 mm y^{-1}, 113.4 mm y^{-1}, and 25.8 mm y^{-1}, respectively.

RC11: Figure 3 and Line 320: Add the map of precipitation to Figure 3 to assist the comparison.

AC: We p**artially agree (minor changes to the manuscript suggested).** We believe referring to a map of precipitation would assist in the comparison; however, rather than adding it to Figure 3, we suggest the following revision to the text.

Suggested text changes in manuscript

Suggested revision at line 319: As expected, high recharge rates are mostly located in areas with high precipitation, i.e., in the tropical north, along the east coast, and in north-western Tasmania (see Figure 3 and rainfall map in Figure S1a of the supporting information), while low recharge rates are mostly located inland from the coast.

RC12: Table 2 and Line 344: The best-performing 7-variable grouping has a performance as good as the 8-variable grouping. The less the number of variables, the lower the computation cost and potential of over-fitting. Why not choose the 7-variable grouping?

AC: **We partially agree (no change suggested).** The reviewer is correct in highlighting that the less the number of variables, the lower the computation cost; however, this is not true for the potential of over-fitting. We favoured the model not over-fitting the training data and the model's performance on unseen data rather than lower computational cost (which is not an issue for 8 variables at 200 to 250 trees).

RC13: Line 369: typo, "200 trees" should be "250 trees"

AC: **We agree (minor changes to the manuscript suggested).** We thank the reviewer for their attention to detail.

Suggested text changes in manuscript

Suggested revision at line 368: The R_{50} random forest model achieved a training score of R^2: 0.772, 'out-of-bag' score of R^2: 0.716, external validation test score of R^2: 0.732 and 10-fold cross validation R^2: 0.715, with ~~200~~ 250 trees in the random forest (Figure 5).

RC14: Line 445-459: Could the covariance/correlation between variables influence the feature importance of a specific variable? For example, precipitation, distance to coast, and elevation are correlated. Will using all three variables in the model introduce redundant information and potentially increase their explanatory power?

AC: We p**artially agree (no changes suggested).** The reviewer is correct in highlighting that the covariance/correlation between variables could influence the feature importance of a specific variable. To use the reviewer's example, if precipitation is the strongest variable out of the group of correlated variables, then precipitation is likely to be chosen more often in a scenario where all three correlated variables are randomly chosen at a split in the random forest tree. If this was the case, then it would affect the ranking of these variables, i.e., precipitation may rank higher than the others. However, the random selection of variables at each split in the random forest tree (which is part of the random forest algorithm/method) is in place to limit these kinds of biases. Using all three variables may increase their explanatory power (i.e., higher ranking in the feature importance); however, it does not introduce redundant information. Rather, it produces a better performing model as the algorithm always chooses the variable that splits the data the best.

RC15: Line 450: The sentence in the parenthesis reads weird. Rephrase.

AC: **We agree (minor changes to the manuscript suggested).**

Suggested text changes in manuscript

Suggested revision at line 449: Feature importance may be influenced by factors such as variable cardinality (i.e., tendency to ~~score~~ give higher importance to variables with many unique levels ~~higher importance~~ as they offer more opportunities for splitting the data; Strobl et al., 2007).

RC16: Line 515: Geology seems to be an important factor to explain the overestimation in the model. However, on line 450, the authors state that geology was not included in the highest-performing model because it cannot split the data due to low cardinality. It reads conflict to me. The authors should explain more.

AC: We p**artially agree (minor changes to the manuscript suggested).** In line 515 we are suggesting that geology may be important to limit recharge where low permeability bedrock outcrops at or sub-crops close to the ground surface. However, the geology spatial dataset does not provide sufficient detail to differentiate between low permeability bedrock and more permeable, highly fractured bedrock. Therefore, we believe a revision to the text will remove this misunderstanding.

Suggested text changes in manuscript

Suggested revision at line 515: ~~We do not account for geology in our model~~ No geological dataset is available that provides detailed spatial information on the permeability of bedrock; therefore, modelled recharge rates can be significantly overestimated in areas such as where low permeability bedrock outcrops at the surface and underestimated in areas where highly fractured bedrock exists.