Reviewer 1:

The authors have performed several new analyses in order to respond to my admittedly lengthy list of requests, and I appreciate their efforts. While I think another round of major revisions is required, I do think this manuscript is on its way to being acceptable for publication. In this review, line numbers refer to the 'track changes' document.

We thank Reviewer 1 for their continued review of our manuscript. We appreciate the assessment that the manuscript is moving in the right direction and have taken the remaining points seriously. The line numbers refer to the track-changes document.

First, I will respond to a few points in the authors' response:

"Spatial overlap of known limitations and observed performance: The reviewer's interpretation rests on the assumption that poor model performance in Figure 5 should spatially coincide with the regions of known GLOBGM limitations flagged in Figure 4. We respectfully challenge the validity of this assumption...". My interpretation was actually somewhat more general: if a paper states that results are less reliable in some regions (and hence more reliable in some others) I expect that the high-reliability regions display better performance. If the results are in fact quite comparable between the regions, why delineate regions at all? Or what limitations have not been considered that are reducing performance in the high-reliability regions?

We thank the Reviewer for this clarification. We acknowledge that our original wording, "GLOBGM has been shown to provide less accurate results", seemed to suggest that we had formally evaluated GLOBGM against observations and found its accuracy in these regions to be lower than elsewhere. That was not our intention. Figure 4 and accompanying text are meant to indicate where specific process representations in GLOBGM are simplified or absent (Karst, permafrost, mountainous areas). Its purpose is to give an honest and upfront statement of the model's limitations. We have rephrased the relevant text to make this clear, please see lines 281-282, 452-453, Figure 4.

We wish to clarify the intention of our response we presented in the previous round of reviews to the Reviewer's comment:

"***Figure 5a, and the analyses shown in Figure D3 show that the regions of poor performance do not line up well with the regions in Figure 4 where GLOBGM is expected to provide less accurate results due to limitations in process representation. In addition, there are no significant differences in performance between locations outside or inside major abstraction areas (Fig. D3)'***

We presented the boxplots in our previous response precisely to show that the point-based observations give no indication that performance differs between the regions depicted in Figure 4 and those outside them. We hope that the statement above and the inclusion of these boxplots in the Appendix makes this point more clearly now.

"Then regarding the suggestion to focus only on regions where the model matches historical observations we respectfully disagree with this approach. Groundwater monitoring networks are heavily concentrated in the Global North (Bäthge et al. 2026) and in addition groundwater models are overwhelmingly concentrated in high-GDP countries with sparse

coverage over Africa, South Asia, and South America (Zamrsky et al. 2025). Restricting our model output to regions with good simulation and observation agreement would systematically exclude the Global South. We argue that providing physically-based model estimates, while acknowledging uncertainty, is more informative than providing nothing for these regions."

I appreciate the point, but there is a fine line between reporting highly uncertain results and reporting results that we know are probably wrong, without clearly stating this. The regions the authors point out that have sparse coverage (Africa, South America, South Asia) contain substantial areas where the model disagrees strongly with GRACE observations. If the model is wrong in these locations, releasing these results is probably worse for decision-makers than releasing no results at all. The manuscript is on its way to a better discussion of uncertainty, and I think that I will be comfortable with publication with more clear and upfront statements about the model failures.

GRACE comparison: Thanks for including this. However, more needs to be said about the comparison.

- For what percentage of the global land surface does GLOBGM show the same direction of change as GRACE?

- For what percentage of the land surface do the two estimates agree within some margin of error?

- Is poor meteorological forcing really the only cause of this poor performance? There's a wealth of research on groundwater and recharge in sub saharan Africa, and a couple of papers are cited, but surely some more of it could be useful? What about shifts in precipitation intensity? eg. Taylor et al (2013) https://doi.org/10.1038/nclimate1731, Zhang et al (2016) https://doi.org/10.1002/hyp.10809, Jasechko and Taylor (2015) https://doi.org/10.1088/1748-9326/10/12/124015. Is heavy-rainfall-driven recharge well-represented in GLOBGM? Perhaps there are other important processes that are also missing from the model. I don't think these issues need to be corrected in the model for this publication, but I do think that they need to be pointed out so readers can judge for themselves whether they trust the results for a particular region.

In addition to sub-saharan Africa and Siberia, I think it's worth pointing out other areas where the model does not match GRACE: Peninsular India, the Alps, the Andes, - the Caspian Sea and south of the Black Sea, the North American Great Lakes, North-western North America, Mexico, Northern South America, southeast US, northeast US and eastern Canada, Japan, Iceland, New Zealand, Australia is also quite dissimilar if you look at it region-by-region. I don't agree that the simulations for northern Africa agree with GRACE. Perhaps for the Sahara desert, but not for the populated coastal regions.

The authors have improved the discussion of the model performance, but the model limitations (for example, for the above regions) should also be made clear in the abstract.

We thank the Reviewer for this constructive framing of this issue. We agree with the reviewer that it is important for modeling studies to be upfront about limitations and uncertainty and provide enough information on limitations and uncertainty so that readers can judge the reliability of the results for a particular region. This is why we planned on releasing the spinup information, quality assurance mask once the manuscript is accepted, we will now also include the GRACE/GRACE-FO agreement information (see line 794).

The Reviewer is requesting that we include "more clear and upfront statements about model failures". In the table below we summarize where in the manuscript we discuss model limitations and uncertainty; the rows in bold are new additions in response to the Reviewer's current comments.

RC Table 1: Summary table highlighting where in the manuscript limitations are acknowledged and discussed, new additions to the manuscript are marked in bold.

Section / Line #	**Limitation**	**Connection to Reviewer comment**
Abstract (20)	**Statement of limitation acknowledgement in abstract.**	**Relates to "*The authors have improved the discussion of the model performance, but the model limitations (for example, for the above regions) should also be made clear in the abstract.*"**
2.1.2 (147 – 156)	Recharge downscaling limitation	-
4.1 (611-640)	Uncertainty in recharge estimation, calibration challenges, initial condition estimation.	-
2.4.2 (313-314)	Observation data sparsity	
4.2 (648-649)	Delays in deep unsaturated zones.	-
2.4.1 (280-289) **2.4.2 (290-335)** 3.2.1 (451-463) **3.2.2 (490-515)** 3.4 (567-570) **4.2 (650-656)**	**Process limitation: karst, permafrost, mountainous terrain**	**Encapsulates the following regions highlighted by the Reviewer: t*he Alps, the Andes, south-eastern US, eastern Canada, north-eastern US, Mexico, North-western North America, Northern Africa.***
4.2 (663-667)	**GRACE disagreement: islands, small archipelagos; ocean-land leakage**	**Encapsulates the following regions highlighted by the Reviewer: Japan, New Zealand, Iceland.**
4.2 (661-663)	**GRACE disagreement: large water bodies**	**Encapsulates the following regions highlighted by the Reviewer: *North American Great Lakes, Caspian Sea and south of the Black Sea.***
4.2 (656-664)	**GRACE disagreement: due to process representation.**	***Relates to "Is poor meteorological forcing really the only cause of this poor performance?..." and " I don't agree that the simulations for northern Africa agree with GRACE. Perhaps for the Sahara desert, but not for the populated coastal regions"***

In the revised manuscript we have improved the analysis which compares GLOBGM TWSA and GRACE/GRACE-FO TWSA. We have included the percentage of area which agrees with GRACE/GRACE-FO and what percentage agrees within some degree of error margin with GRACE/GRACE-FO (methods: lines 313-335, results: 500-515, discussion: 650-667, Figure E1). This has also allowed us to benchmark these values against previous studies and report that the results of GLOBGM are comparable (see lines 652-654). In addition, we have elaborated on the potential mechanism that may be the cause (see RC Table 1). The reviewer is also requesting that we include the list of regions they have identified as disagreeing with GRACE/GRACE-FO in the abstract. We thank the reviewer for their thorough evaluation of the spatial discrepancies between GRACE/GRACE-FO and GLOBGM. We have expanded our discussion on limitations (see RC Table 1) and linked them to the regions presented in Figure 4 which largely encapsulates these regions listed. From the updated analysis, we find that Australia and Peninsular India largely agree with GLOBGM. Where they do not agree, we have included discussion on why this may be (see RC Table 1). It is not feasible to list every region in the abstract. We have, however, expanded the sentence which was already meant to indicate that we consider and discuss limitations in the manuscript so that readers are made aware of this section from the onset (see lines 20).

L497: For these updated performance scores, am I correct that the authors only screened out locations where the following criteria are met? ((the observed and simulated trends are <0.1 m) AND (rs<0)) OR ((the observed and simulated heads are <1 m) AND the KGENPskill<0)? Obviously, if you discard only the poor simulations (rs<0 or KGENPskill<0) then the overall skill improves! If the authors want to set some criteria (based on heads/trends) to filter results before evaluating performance, then they also must discard skilful predictions that meet those criteria. This is a basic principle of model evaluation.

The Reviewer is correct that this introduces selection bias. We have reworked this section and have only dropped the cases where observed heads are < 1 m, when referring to score improvements. Since the denominator of the bias ratio is the mean observed heads, it is the main cause for excessively bad KGE-NP_{skill} scores as the values approach 0 (lines 476 – 490, Fig. D1, D2). In addition, we have used the Theil-Sen slope to split the validation results into trending vs stable. This approach allows us to draw inference on the performance of different systems without selecting based on the metric being scored.

What is the justification for using three different thresholds to mask out non-significant or small trends? The authors use 0.1 m/year for observed well-based trends, 0.01 m/year for raster trends, and 0.001 m/year for spatially aggregated trends. I would argue 0.1 m/year is not a negligible trend. Over 75 years a change of 7.5 m in a region with a shallow water table could either cause widespread groundwater flooding or cause traditional wells to dry up. Regardless, the authors should pick a threshold, justify it, and apply it consistently.

We agree with the Reviewer that a common threshold is a better choice. We have opted for 0.001 $m.yr^{-1}$ for the spatially aggregated trends and raster maps (see Figures 9, 10, G1, G2 and line 419). For the post-hoc analysis, for splitting the dataset into trending vs stable we now use The rationale behind why we use a different threshold for the observation wells and the maps is because for a 1km raster cell is a spatial average, which means that the variance is suppressed. Since we separate stable from trending based on the observation well time series which have higher variance compared to our 1km grid cell a higher threshold is required. We initially chose 0.1 $m.yr^{-1}$ following Jasechko et al. 2024. However, we agree with the reviewer that it this may be too high for this application and that it is important to be conservative. We now use 0.01 $m.year^{-1}$ and have added justification in the legend of Figure D1.

L 153: This assumption [] in many settings

Should have been “This assumption holds in many climatic settings.” (line 152).

L 382: Here you should state if the comparison is to the bias-corrected results or the raw GLOBGM predictions

This is a valid point. We have made this clearer in the text. We included both raw and ML-corrected products in the comparison. However, the primary comparison refers to the uncorrected GLOBGM v1.1 values in relation to the other models. For transparency we also show the ML-corrected product (ML-GLOBGM v1.1) so readers can see the effect of the optional bias correction (lines 385-386).

L 413: If I understand correctly, the authors remove grid cells with negligible slope before taking the spatial average? What is the justification for this?

We thank the reviewer for flagging this; the wording was ambiguous. No grid cells were masked before the spatial aggregation. The gridded and aggregated trends were computed independently. For the gridded representation the Theil-Sen slope and Mann-Kendall test were computed independently. The thresholds mentioned in the text were applied to plotting. For the spatially aggregated trends: For each spatial unit and each year we first compute the area-weighted mean water table depth over all cells in that unit and only then estimate a Theil-Sen slope and perform a Mann-Kendall on the aggregated time series. The masking is then applied to the trend of the aggregate. We have updated the text to make this clear (lines 405-413).

Based on Figure E1 it seems that trend magnitudes greater than 0.02 occur almost nowhere. Is this true? Or is the legend on Figure E1 incorrect? It seems inconsistent with Figure 9.

In the updated manuscript, we have used a different GRACE/GRACE-FO product, so the underlying data are now different. The legend in Figure E1 is correct. Figure E1 and Figure 9 are not comparable. Figure E1 shows total water storage anomaly, whereas Figure 9 shows water table depth trend.

L507: I think Fig. E1 is the wrong cross-reference.

This text has now been replaced.

Figure D1: This bar graph is quite confusing. At first I read it as '47.1% of locations with |slope|<0.1 have rs<0'. I think what it is trying to say is 'of the locations with rs<0, 47.1% have |slope|<0.1.

This figure has been removed.

Fig 11: I think these are anomalies from the 1970 starting point? A negative WTD does not make sense.

Correct, we have updated the legend for clarity and made it more clear in the methods (line 412, Fig. 11).

L625: typo

Thank you. We have removed the double "using" (line 602).

L716: I do not see agreement in the Gulf of Guinea or in Niger. GLOBGM predict negative changes in these regions, GRACE is strongly positive.

We agree with the reviewer; in the updated analysis we have refrained from naming this as a region of agreement and marked it as a region with disagreement due to the presence of karst (line 511).

L721: "Despite these structural differences..." The authors have only identified one structural difference.

Agreed, we have changed this sentence (line 697).

L752: It's a bit disappointing that the authors chose to compare their results only to the two example studies from the US that I pointed out. A quick google search shows that there are many more (large-scale) regional modelling studies that could be compared, eg. for Brazil, the UK, Germany, and China. Also, Meixner predicted decreased recharge across most of the southwestern US, and no change/slight increases in northern systems. Is this really in line with the GLOBGM v1.1 predictions of "rising water tables across most of the western contiguous United States"?

The reviewer is correct to question our framing of the Meixner et al. (2016) reference. In the previously revised manuscript, the Meixner study referred to specific aquifer systems and in the revised manuscript we have been more accurate with its attribution (line 732). In the revised manuscript, we strived to find and include more regional studies in this section.

We agree that regional studies provide useful comparisons to GLOBGM results. To allow for a like-for-like comparison, it is important that these studies provide information on groundwater heads or water table depth. The majority of studies we have been able to find focus on recharge or groundwater storage and not heads or water table depth. We have been able to identify some studies that cover the regions the Reviewer refers to but without being provided references to these it is challenge. We have reworked Section 4.3 so that our primary comparison relies on the only (almost) global counterpart (Costantini et al. 2024) and we have included the below references where appropriate, (Section 4.3).

Meixner, T., Manning, A. H., Stonestrom, D. A., Allen, D. M., Ajami, H., Blasch, K. W., Brookfield, A. E., Castro, C. L., Clark, J. F., Gochis, D. J., Flint, A. L., Neff, K. L., Niraula, R., Rodell, M., Scanlon, B. R., Singha, K., and Walvoord, M. A.: Implications of Projected Climate Change for Groundwater Recharge in the Western United States, Journal of Hydrology, 534, 124–138, https://doi.org/10.1016/j.jhydrol.2015.12.027, 2016.

Kumar, R., Samaniego, L., Thober, S., Rakovec, O., Marx, A., Wanders, N., Pan, M., Hesse, F., and Attinger, S.: Multi-Model Assessment of Groundwater Recharge Across Europe Under Warming Climate, Earth's Future, 13, e2024EF005 020, https://doi.org/10.1029/2024EF0050.

Jackson, C. R., Bloomfield, J. P., and Mackay, J. D.: Evidence for Changes in Historic and Future Groundwater Levels in the UK, Progress in Physical Geography: Earth and Environment, 39, 49–67, https://doi.org/10.1177/0309133314550668, 2015.

Parry, S., Mackay, J. D., Chitson, T., Hannaford, J., Magee, E., Tanguy, M., Bell, V. A., Facer-Childs, K., Kay, A., Lane, R., Moore, R. J., Turner, S., and Wallbank, J.: Divergent Future Drought Projections in UK River Flows and Groundwater Levels, Hydrology and Earth System Sciences, 28, 417–440, https://doi.org/10.5194/hess-28-417-2024, 2024.

Wu, W.-Y., Lo, M.-H., Wada, Y., Famiglietti, J. S., Reager, J. T., Yeh, P. J.-F., Ducharne, A., and Yang, Z.-L.: Divergent Effects of Climate Change on Future Groundwater Availability in Key Mid-Latitude Aquifers, Nature Communications, 11, 3710, https://doi.org/10.1038/s41467-020-17581-y, 2020.

Wunsch, A., Liesch, T., and Broda, S.: Deep Learning Shows Declining Groundwater Levels in Germany until 2100 Due to Climate Change, Nature Communications, 13, 1221, https://doi.org/10.1038/s41467-022-28770-2, 2022.

Condon, L. E., Atchley, A. L., and Maxwell, R. M.: Evapotranspiration Depletes Groundwater under Warming over the Contiguous United States, Nature Communications, 11, 873, https://doi.org/10.1038/s41467-020-14688-0, 2020.

Lastly, there are a number of typos throughout the revised manuscript, and I would encourage the authors to proofread carefully.

We thank the Reviewer for highlighting this. We have gone through the full manuscript carefully and corrected the remaining typos and grammar issues, the edits are visible in the track change document.

References:

Jasechko, S., Seybold, H., Perrone, D. et al. Rapid groundwater decline and some cases of recovery in aquifers globally. Nature 625, 715–721 (2024). https://doi.org/10.1038/s41586-023-06879-8

Yang, C., Jia, Z., Xu, W., Wei, Z., Zhang, X., Zou, Y., McDonnell, J., Condon, L., Dai, Y., and Maxwell, R.: CONCN: a high-resolution, integrated surface water–groundwater ParFlow modeling platform of continental China, Hydrol. Earth Syst. Sci., 29, 2201–2218, https://doi.org/10.5194/hess-29-2201-2025, 2025.