This study reports the results of a global groundwater model (GLOBGM v1.1) run at 1 km resolution under 3 CMIP6 scenarios and 5 climate models. Van Jaarsveld et al predict that groundwater levels will rise, on average, on most continents but that regions of groundwater depletion will persist. The study is ambitious in scope, and I applaud the authors for an attempt to make climate change predictions for an important component of the earth system. Unfortunately, the study suffers from several issues in methodology, validation, and interpretation, and because of this I believe the results are not reliable. Given the potential for the results to inform groundwater management policy, I think it would be inappropriate and potentially harmful to publish the manuscript. I unfortunately must recommend rejection.

We thank Referee #1 for the thorough review and the numerous constructive comments and suggestions, which will help to improve the manuscript. However, we respectfully but firmly disagree with the conclusion that the model results are too unreliable to assess groundwater dynamics and trends. We are a bit taken aback by the use of the terms “inappropriate and potentially harmful”, which we feel is an overly strong dismissal that is not conducive to open scientific discourse. We would like to highlight that we have been deliberately transparent about the accuracy of our results (more so than in some previous global groundwater studies) and we maintain that our model is sufficiently accurate to assess regional trends in groundwater resources. In the following response, we provide additional analyses and context to support our case that the conclusions of this study are robust.

We also note that the accuracy of these results should not be compared to the accuracy reached by regional to local groundwater models that are built for water management at that scale using local data on hydrogeology and forcing. Global models are for global scale assessments of past and future trends, which can be identified even if locally deviations from observations exist. This is evident from global climate models (CMIP) or global hydrological models (ISIMIP) that can show large deviations from observations locally (e.g., Krysanova et al. 2020), but nevertheless provide valuable tools to robustly project changes in the future (Schewe et al. 2014).

Referee #1 makes the point below that lack of the accuracy should lead to reducing the resolution to 5 arcminutes as before, suggesting that this better reflects the accuracy. However, we note that we still compare 30 arc-second resolution(~ 1 km^2) grid cell against point measurements and therefore still face a large scale gap. And the results shown here do a better job than the models with a 5 arcminute resolution. The main reason is that topography and surface-groundwater interactions are better resolved at higher resolutions. There is therefore no reason to decrease the resolution, as long as the accuracy of the results is made clear.

That being said, we agree with Referee #1 that further improvements in the accuracy assessment and the discussion of the results is possible. We will address to the specific concerns below but start by discussing in general the accuracy of the results in comparison to other global models. In addition, the additional analysis presented here will be incorporated into the revised manuscript.

My primary concern is the accuracy of the model and how it is evaluated. Overall, at approximately one third of wells the correlation to observations is negative: the model predicts the wrong direction of change. In my opinion, this is not good enough to make predictions about future trends. Based on Figure 5, panel a, the regions that show poor performance do not line up particularly well with the regions in Figure 4 where "GLOBGM has been shown to provide less accurate results". It seems the poor performance is primarily in places with high GW abstraction. Also, at 10%-20% of the wells the bias ratio is less than 0.1, meaning the simulated water table depth is more than 10 times too low? Again, this seems quite poor. Given that the model struggles to predict groundwater levels in regions where groundwater data exist and the subsurface is relatively well-understood, I am sceptical of the predictions for the rest of the world.

In our initial validation we did not account for the fact that multiple observation wells were situated within a single grid cell. We have now included a filtering step where if multiple observation wells occur within a single grid cell we chose the one with the longest temporal coverage. The response below is based on the updated validation procedure and will be incorporated into the revised manuscript. In the revised manuscript we will update the results and retrain the ML algorithm on this filtered dataset.

<u>GLOBGM_v1.1 in comparison to other global groundwater modelling efforts:</u>

We have compared our results with previous global groundwater modelling efforts. We compared the steady state results with those of Fan et al. (2017), De Graaf et al. (2017), Reinecke et al. (2019) and Verkaik et al. (2024). Note that Fan et al. (2017) and Reinecke et al. (2019) provide steady state results only. This shows that results are better than all of these models except Fan et al. (2017) when looking at steady state estimates (Figure RC1 1 & Table RC1 1). The table also compares the results of this version of the transient model with those of Verkaik et al. (2024) and De Graaf et al. (2017), showing again that the results are better (Figure RC1 2 & Table RC1 1). The conclusion is that our results are on par or better than previous modelling results.

Also, we compared the simulated trends with those of GRACE/GRACE-FO (Chandanpurkar et al. 2025). For fair comparison we performed a spatial moving averaging approach of the yearly groundwater levels with a Gaussian spatial filter with a standard deviation of 75 km resembling approximately the GRACE/GRACE-FO footprint. We compare the 5 arc-minute PCR-GLOBWB TWS trends by which the groundwater component is replaced with GLOBGM estimates and the groundwater storage trends from GLOBGM only separately. The results show that similar trends can be observed, especially around the groundwater and abstraction hotspots (Figure RC1 3). The largest differences between GACE/GRACE-FO and the simulated TWS can be seen in Sub-Saharan Africa and north-eastern Siberia, most likely caused by the poor meteorological forcing over these regions (Liu et al. 2024).

Table RC1 1: Performance comparison of global groundwater head models across steady-state and transient metrics.

Model	Resolution	Steady State MAE (m)	Median Monthly KGE-NP	Median Annual KGE-NP
Fan et al. (2013)	1km	19.3	-	-
ML_globgm_v1.1	1km	19.9	0.296	0.26
globgm_v1.1	1km	22.8	0.297	0.24
globgm_v1.0 (Verkaik v1.0)	1km	25.5	0.15	0.23
Reinecke et al (2019)	10 km	27.2	-	-
De Graaf et al (2017)	10 km	45.9	0.089	0.11

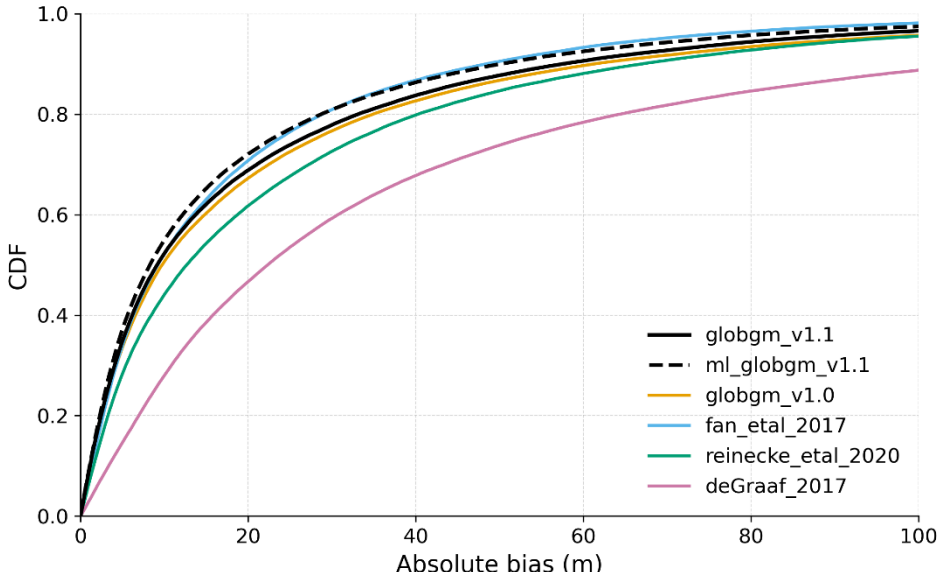

Figure RC1 1: Cumulative distribution functions (CDFs) of absolute bias in simulated water steady state water table depth (WTD) for six global groundwater models evaluated against observed well data.

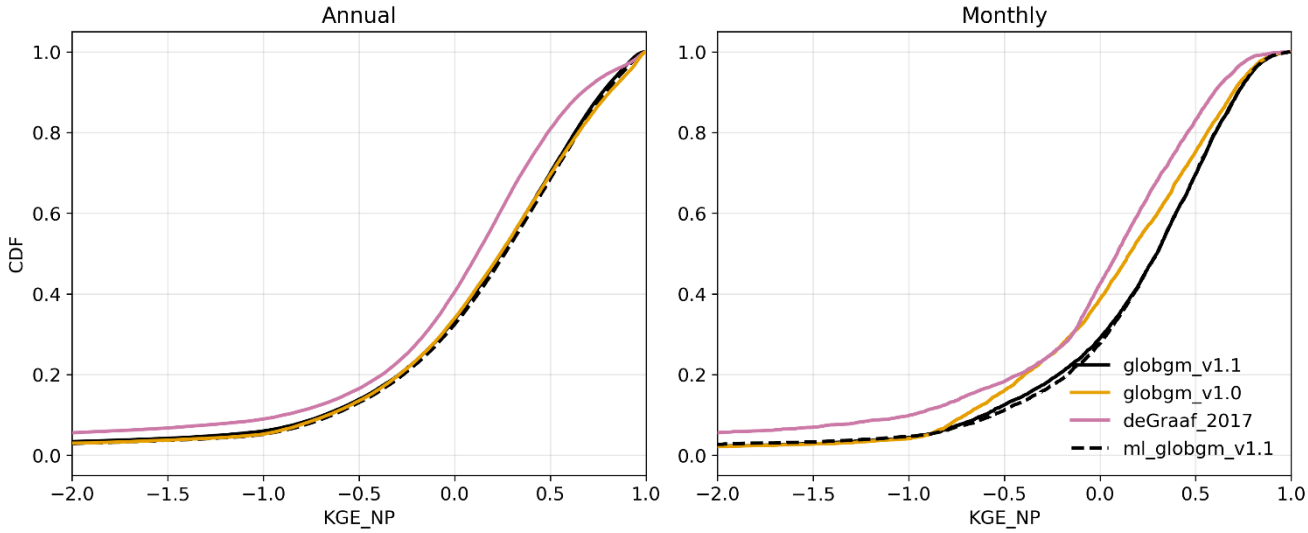

Figure RC1 2: Cumulative distribution functions (CDFs) of model performance metrics for four global groundwater models, De Graaf (2017), GLOBGM v1.1, ML_GLOBGM_v1.1, GLOBGM_v1.0, evaluated against observed groundwater head time series at annual (top row) and monthly (bottom row) timescales.

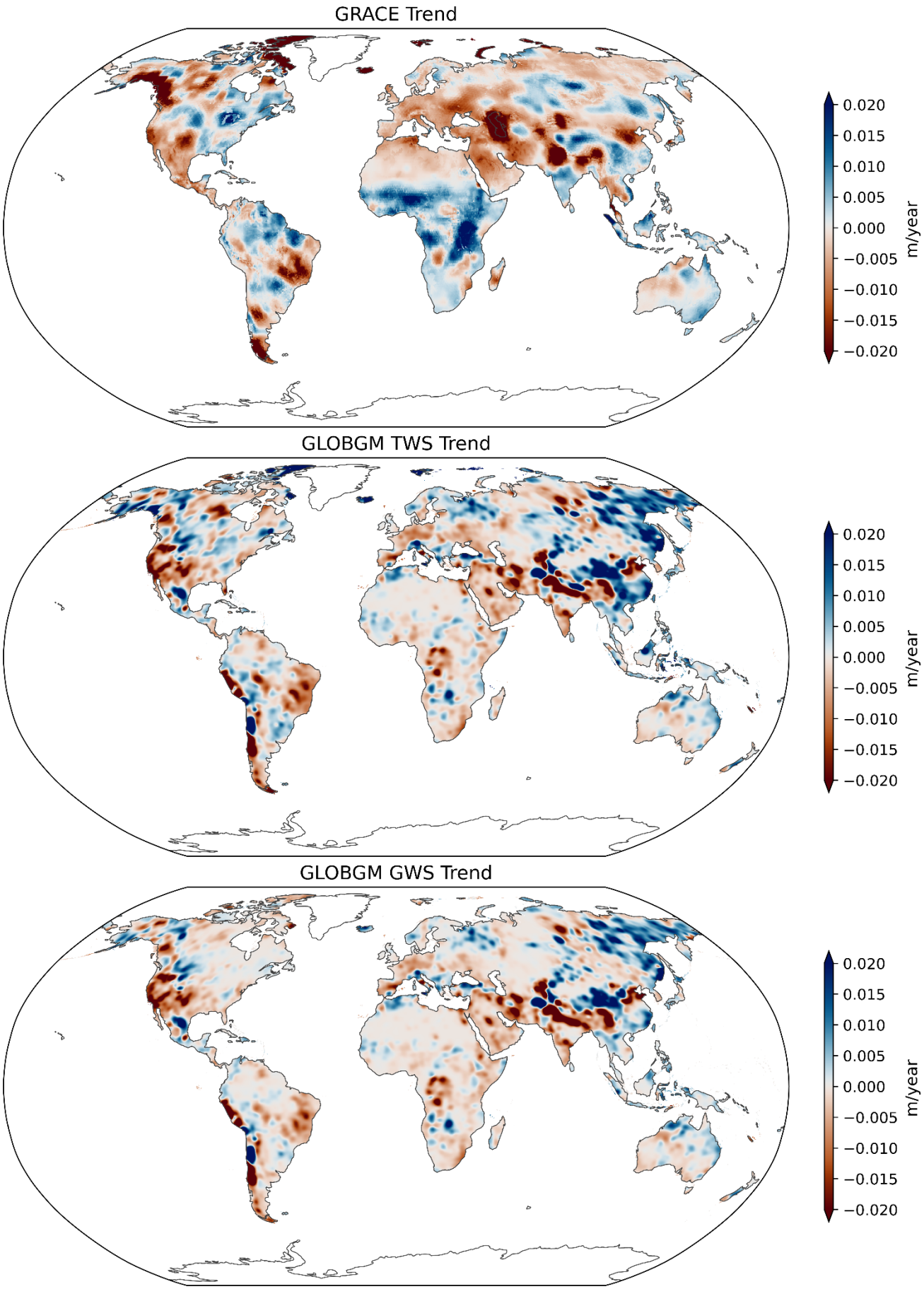

Figure RC1 3: (Top) Observed TWS trends derived from GRACE/GRACE-FO satellite gravimetry data. (Middle) Simulated TWS trends from the PCR-GLOBWB / GLOBGM. (Bottom) Simulated GWS trends from GLOBGM.

GLOBGM_v1.1 in comparison to other regional groundwater modelling efforts:

Although we caution against comparing global models to regional models, we can rely on a previously published manuscript which compared the performance of GLOBGM v1.0 to ParFlow over China (Yang et al. 2025), which shows that GLOBGM v1.0 performs similarly to the regional model CONCN 1.0 which is based on ParFlow (Figure RC1 4).

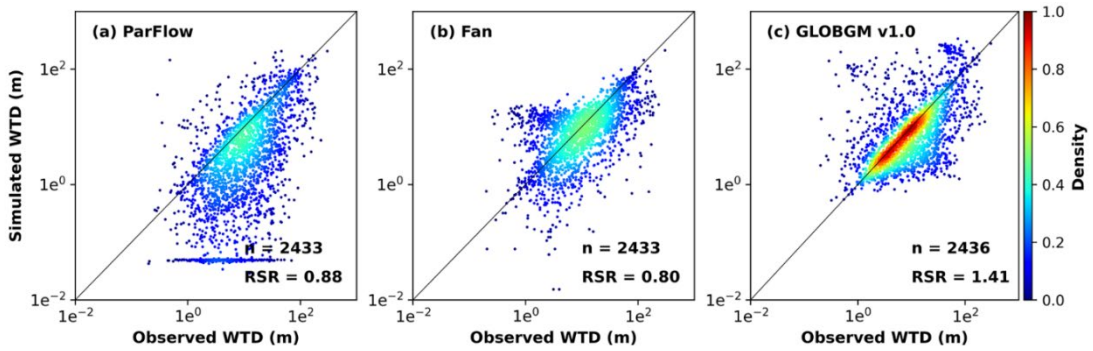

Figure RC1 4: Sourced from Yang et al. 2025: Scatterplots of simulations vs. observations of water table depth. The simulations are long-term average water table depths of 1981-2010 in the current work, of 2004-2014 in Fan's model, and of 1958–2015 in GLOBGM v1.0. The observations are the averages of 2018.

The concerns regarding negative correlations:

In the original manuscript negative correlations between simulated and observed groundwater heads, occur in approximately *one fifth* of locations for the monthly data and a quarter when considering the annual data - *not one third* as suggested by the reviewer. From the updated validation procedure this pattern holds.

We agree with the reviewer that the correlations in Figure 5 relay a pessimistic view of models ability to reproduce trends in groundwater heads. However, these figures mask an important nuance, which is that for the majority of observation wells the trends in simulated and observed groundwater heads are very small such that very small differences in trends get penalized disproportionately in the KGE-NP function when these display opposing trends. To better investigate the reproduction of trends we compared the Theil-Sen slopes of Figure 8 with the Theil-Sen slopes fitted to the heads at observation wells.

We first show that for the locations with a negative spearman correlation, 47 % of both simulated and observed have negligible slopes (Figure RC1 5). We used a threshold of 0.1 m year^{-1} as indicative of stable groundwater trends (Jasechko et al. 2024). In addition, when excluding these points from the overall distribution, we show that only 15% of locations a meaningful opposing direction correlation (Figure RC1 5).

We also analysed simulated vs observed Theil-Sen slopes. Results show that simulated and observed trends agree in direction for 61% of the wells and if the problematic wells (with negligible and opposing directions) are excluded this rises to 74%.

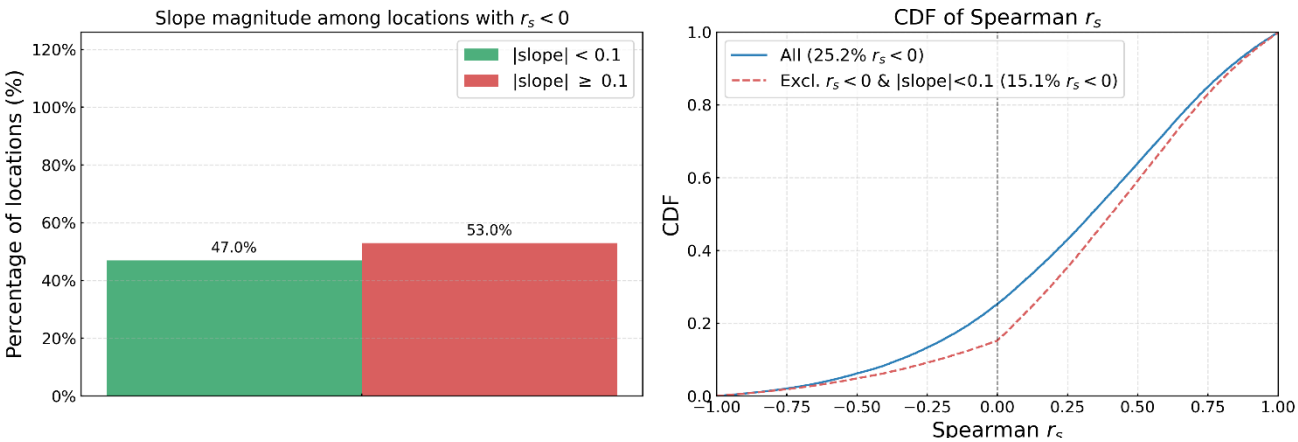

Figure RC1 5: Proportion of locations with negative Spearman rank correlations (rs < 0) partitioned by slope magnitude. (Right) Cumulative distribution function (CDF) of Spearman rs values for all locations (solid blue) and for the subset excluding locations where rs < 0 and |slope| < 0.1 (dashed red; 15.1% with rs < 0). The vertical dashed line marks rs = 0

In addition, we would like to note that from Figure 5 it is clear to see that the annual and monthly data correlation is worse for the deeper wells compared to the shallower wells. This shows that a lack of correlation can be attributed to the delays that occur in deeper unsaturated zones that are not modelled in our approach. We will add this point to the discussion.

<u>Spatial overlap of known limitations and observed performance:</u>

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 on both logical and empirical grounds. The spatial distribution of observation data in Figure 5 is not random; it is strongly constrained by data availability. With the majority of observation wells occurring in concentrated regions with high groundwater abstraction (e.g., North America, Europe, South Asia which are largely outside of the regions highlighted in Figure 4). Critically, if approximately 85% of all observations (both good and poor, Figure RC1 6) fall outside of Figure 4 regions, then any spatial pattern of performance will predominately appear outside of these regions. The concentration of poor performing locations outside of Figure 4 is therefore a direct consequence of where the data exists. The spatial mismatch the reviewer identifies is not diagnostic of model behavior but of data locality.

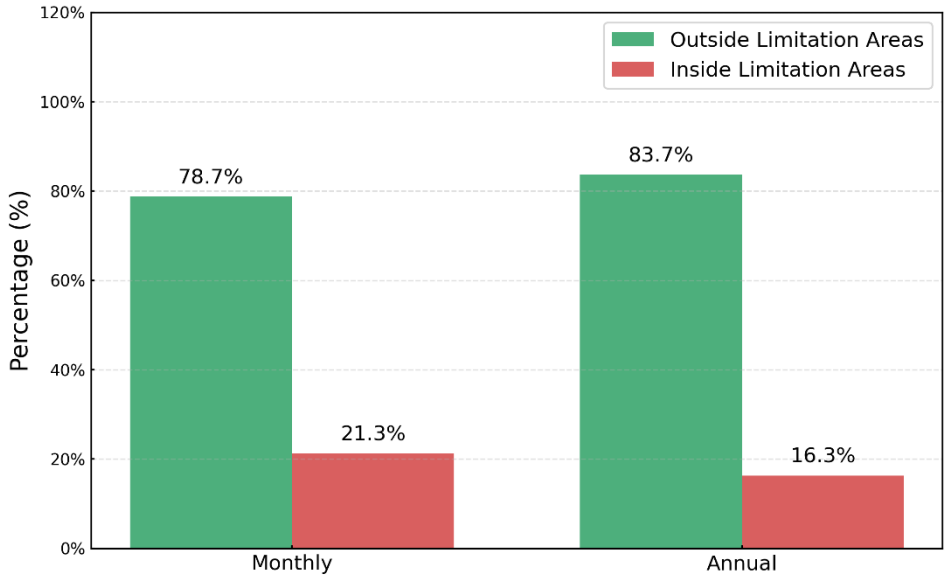

Figure RC1 6: Proportion of groundwater head observations located outside and inside the regions of known GLOBGM limitations (Figure 4), for both monthly and annual time series datasets.

Nonetheless, we have calculated box plots of KGE-NP, r_spearman, alpha_fdc and beta separately for the regions marked in Figure 4 the remaining land area and separately (Figure RC1 7). This shows that the distributions of KGE-NP and its components (r_spearman, alpha_fdc, and beta) are broadly comparable between stations inside and outside these regions, for both monthly and annual datasets. The absence of a meaningful difference in performance between the two groups demonstrates that the spatial pattern identified by the reviewer is not diagnostic of the model accuracy but is instead a straightforward artifact of where observations exist.

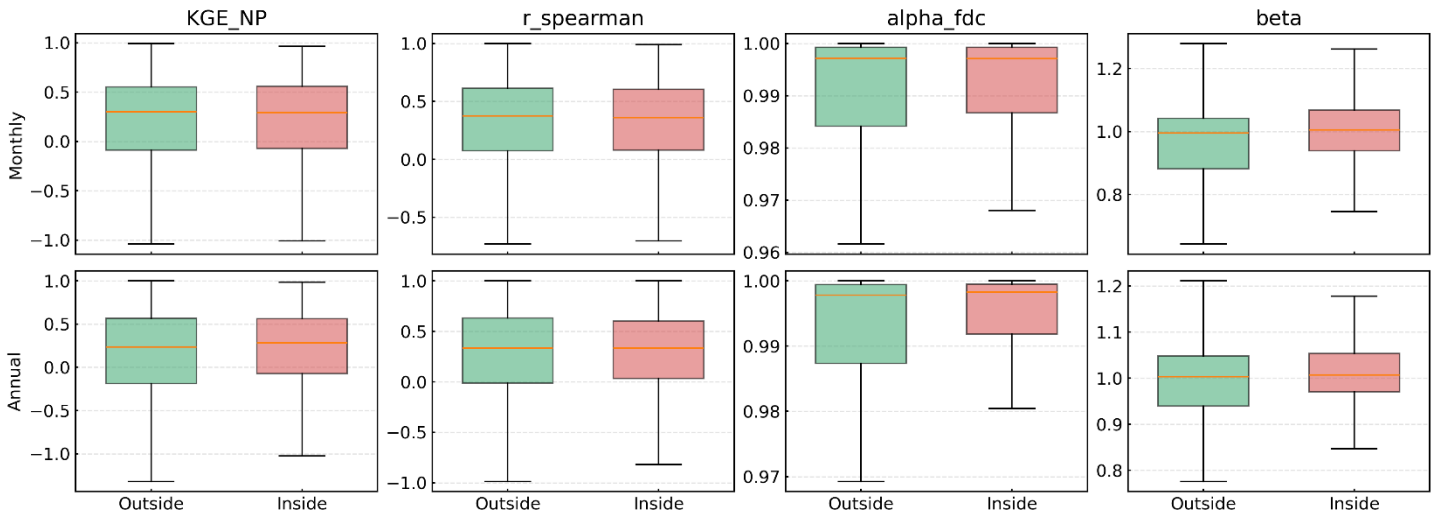

Figure RC1 7: Distribution of KGE-NP performance metric components, Spearman correlation (r_spearman), flow duration curve slope ratio (alpha_fdc), and bias ratio (beta), for observations located outside and inside the GLOBGM limitation regions (Figure 4), shown for monthly (top row) and annual (bottom row) datasets.

We also examined whether model performance differs systematically between high-abstraction and low-abstraction regions (Figure RC1 8). The distributions of KGE-NP and its components are broadly comparable (if anything, marginally better) for stations located within high-abstraction areas. This suggests that GLOBGM does not perform systematically worse in regions of intensive groundwater use. We therefore conclude that model performance is reasonably consistent across contrasting abstraction regimes, further supporting the robustness of our results.

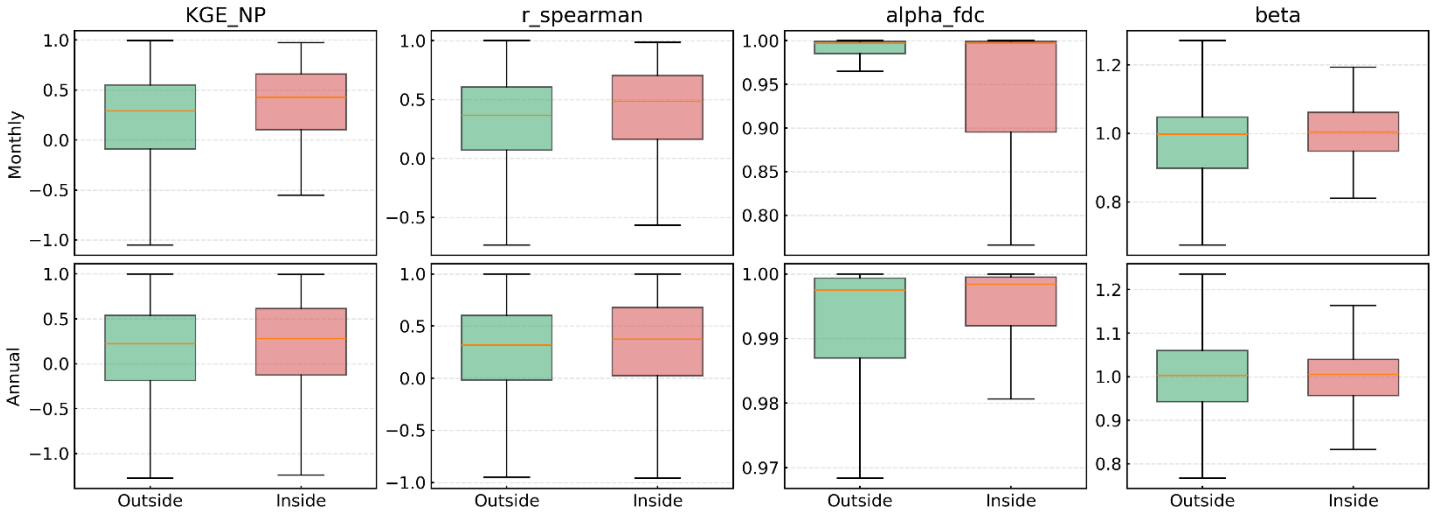

Figure RC1 8: Distribution of KGE-NP performance metric components, Spearman correlation (r_spearman), flow duration curve slope ratio (alpha_fdc), and bias ratio (beta), for observations located outside and inside abstraction hotspots (Figure 2), shown for monthly (top row) and annual (bottom row) datasets

<u>On the bias ratio:</u>

The low bias ratios are an artefact of the results of locations with hydraulic heads close to zero (sea-level), resulting in small numbers in the denominator. The same issue would have occurred when calculating bias-ratios for the water table depth in areas with shallow water tables. Therefore, it is more insightful to look at the absolute bias in water table depth. It can be seen (Figure 6 of the manuscript) that these are low for water table depth up to 20 m below surface water level, particularly after the bias-correction.

We hope that the additional analyses and context provided have removed some of the reservations of Referee #1.

I looked at some previous publications related to this model, and similar concerns regarding accuracy have been raised (https://doi.org/10.5194/gmd-2022-226-RC1, https://doi.org/10.5194/egusphere-2024-1025-RC3). In those cases the authors argued that the performance cannot be expected to be high, given that the model was not calibrated. They argued that the model is still valuable because "The philosophy of these models is that they try to capture the right processes and do not rely on calibration to correct for errors in process representation, parameterization and/or meteorological forcing data." Now the

authors present a model that is calibrated, and wherein at least one of the inputs (groundwater recharge) in statistically downscaled to match observations.

We thank the reviewer for raising this point. We note that the reviewer comments cited from prior publications (RC1 on GMD-2022-226 and RC3 on EGUsphere-2024-1025) were ultimately addressed to the satisfaction of the handling editors, and those manuscripts were accepted for publication. We respectfully note that the 'uncalibrated model philosophy' cited from previous reviews was put forward as a justification for the inherent limitations of those specific uncalibrated model versions, not as a general argument that these process-based models should never be calibrated. The calibration applied here was limited to a coarse stratification based on lithology, using simple prefactors to hydraulic parameters over a limited number of permutations and bounded by ranges to maintain realistic aquifer responses. Calibration has improved the results, with respect to the original uncalibrated model of Verkaik et al. (2024), particularly the bias of shallow water tables that are of most interest to users. An extensive search method and a sophisticated regularization approach would have undoubtedly further improved the results (Doherty 2003), but the computational requirements are too extensive to do this. Hopefully in the near future, surrogate modelling could be used to make this possible (Triplett et al. 2025). We see this as future research and will point to these possibilities in the discussion of the revised manuscript.

The statistical downscaling of groundwater recharge addresses a known deficiency in the spatial resolution of the forcing data, not a deficiency in the model's process representation. Recharge downscaling was carried out independently of the groundwater head calibration, and the two correction steps target different sources of error. Using observationally constrained recharge as model forcing is analogous to bias-correcting precipitation inputs in climate-driven hydrological models, a standard practice that improves the quality of the forcing.

In addition, the authors use a machine learning model to adjust the predicted groundwater levels to match observations. So now the model is not conserving mass, and nevertheless the performance remains unsatisfying. I think it is time to either a) overhaul this model to try to achieve better global performance, b) focus only regions where the model matches historical observation, or c) return to a coarse-resolution model that might simulate regional dynamics without claiming 1-km resolution.

We wish to clarify that both the raw physically-based model output and the ML-corrected outputs will be provided to users. This was not made abundantly clear in the original manuscript and will be corrected in the revised version. In the original manuscript we validated both the raw and ML-corrected outputs with the intention that they are viewed as separate but complementary products. The ML correction is therefore not presented as a replacement for the physically-based model, but as an additional tool for users who require better predictions of water table depths for (e.g., ecological analyses). We have verified that the ML-corrected fields do not deviate substantially from the raw output in terms of spatial gradients and smoothness. Nevertheless, the use of the correction is optional, and should not be used when mass conservation is required. We will make the distinction between these two products, and the appropriate use cases for each, more explicit in the revised manuscript.

Regarding the suggestion to overhaul the model to try and achieve better global performance, in lines 66-89 of this rebuttal we demonstrate that the results here are comparable to other 1km global groundwater models. We therefore argue that the performance is sufficient. 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. Our arguments for maintaining 1 km resolution are provided on lines 31-38 and lines 391-407 of this response document.

The use of a threshold of -0.41 was proposed by Knoben et al. (2019) for the KGE, but it is not valid for the KGE-NP. The threshold should be 0. Suppose the GW head varies around 50 m with a standard deviation of 5 m. For the mean of observations, I get a KGE-NP of -0.0008. This occurs because the Alpha component goes to 1 if distribution is close to symmetrical and not close to 0. I encourage the authors to verify this with their observational data.

Thank you for this. This is a valid point which we will address in the revision.

The performance of the groundwater recharge downscaling (Section 2.1.2) is not reported. Was any cross-validation attempted? Moeck's (2020) data are not uniformly globally distributed. If you remove the Australian data from the training data, for example, can you predict those recharge rates with the regression model? In addition, limiting GWR_corrected to less than or equal to precipitation is not well-justified. In most cases this is probably a quite liberal constraint but in agricultural areas return flows can lead to recharge that exceeds precipitation. The choice of a multiplicative correction factor is also not justified.

We thank the reviewer for pointing this out. Estimating the performance of the recharge downscaling is a good suggestion. We will add this to the results in a revised version. Limiting the recharge estimates to the precipitation makes sense, since Moeck et. al. (2020) data are limited to non-irrigated sites. In hindsight, we should have refrained from correcting the groundwater recharge in irrigated areas, since PCR-GLOBWB does include additional recharge from return flows. This may possibly lead to overestimation of depletion in agricultural regions. We will report this omission in the discussion. A multiplicative correction makes sense to avoid negative corrected values.

The Machine Learning Bias Correction (2.4.3) should be also cross-validated regionally. The appendix indicates that the R-squared value is 0.6, which is already not very convincing as far as machine learning algorithms go. Can the model predict the bias for regions not included in training data? This is what

would be required to apply this bias correction globally. Also, the labels in Figure A1 (a) are not defined.

We thank the reviewer for this constructive comment. The ML bias correction was trained on 80% of the available observations and 20% was set aside to be used only for validation purposes. Observations were grouped into 75 spatial clusters via K-means clustering in geographic coordinates. The AutoML hyperparameter search used a group-based cross-validation, thus preventing overfitting to spatially autocorrelated data. We will elaborate on this procedure in the revised manuscript and make it clear that training involved cross validation.

While an r-squared of 0.6 may appear low, it is important to note that the residual bias is largely spatially unstructured and noisy in nature, meaning it does not reflect a systematic or spatially coherent error that could be meaningfully corrected by the model. Furthermore, an r-squared of 0.6 is consistent with recent comparable work; for example, Ma et al. (2026) report a similar model performance in their ML-based study over the US using ParFlow, suggesting that this level of explained variance is representative of the current state of the art for large-scale groundwater modelling with machine learning.

We have also updated the axis labels and legend definitions in Figure A1 as suggested, thank you for highlighting this.

My second major concern is the interpretation and discussion of results.

I would have expected to see more specific discussion of the future trends in different regions. What will happen to the major agricultural regions? Grouping by continents is not very informative. The fact that groundwater will rise in the Andes does not help northern Colombia and Central America, where groundwater levels could fall by ~50 m by the end of the century! Similarly for rising groundwater in Tibet and falling levels in the Ganges-Bramhmaputra. Also note that the majority of the rise seems to be happening in the Mountain regions where the authors say the model is less reliable. And what is causing the trend reversal in northwest North America*?*

This is a valid point, and we thank the reviewer for their comments on regional heterogeneity. We agree that continent-level aggregation masks local contrasts. However, a detailed regional comparison for agricultural zones and hydrogeological regions at the global scale entails a substantial additional analysis and discussion that falls outside of the scope of the current paper. Figure 10 is included to provide continental context for the global results, not to provide a comprehensive regional assessment. The sub-regional heterogeneity the reviewer identifies is the subject of future work currently in preparation.

We will update Figure 9 and Figure 9 and present the slopes aggregated by spatial units. The spatial units will be a combined water provinces (Straatsma et al. 2020) and WHYMAP aquifers (Chen et al. 2017) dataset. This will allow us to incorporate discussions on regional patterns in comparison with previous research. We believe that this is a more appropriate response than expanding the manuscript beyond its intended scope – the paper is also 30 pages long excluding references and 37 pages

including references. Below we have included a draft figure (Figure RC1 9) for the updated figure 9 and 8.

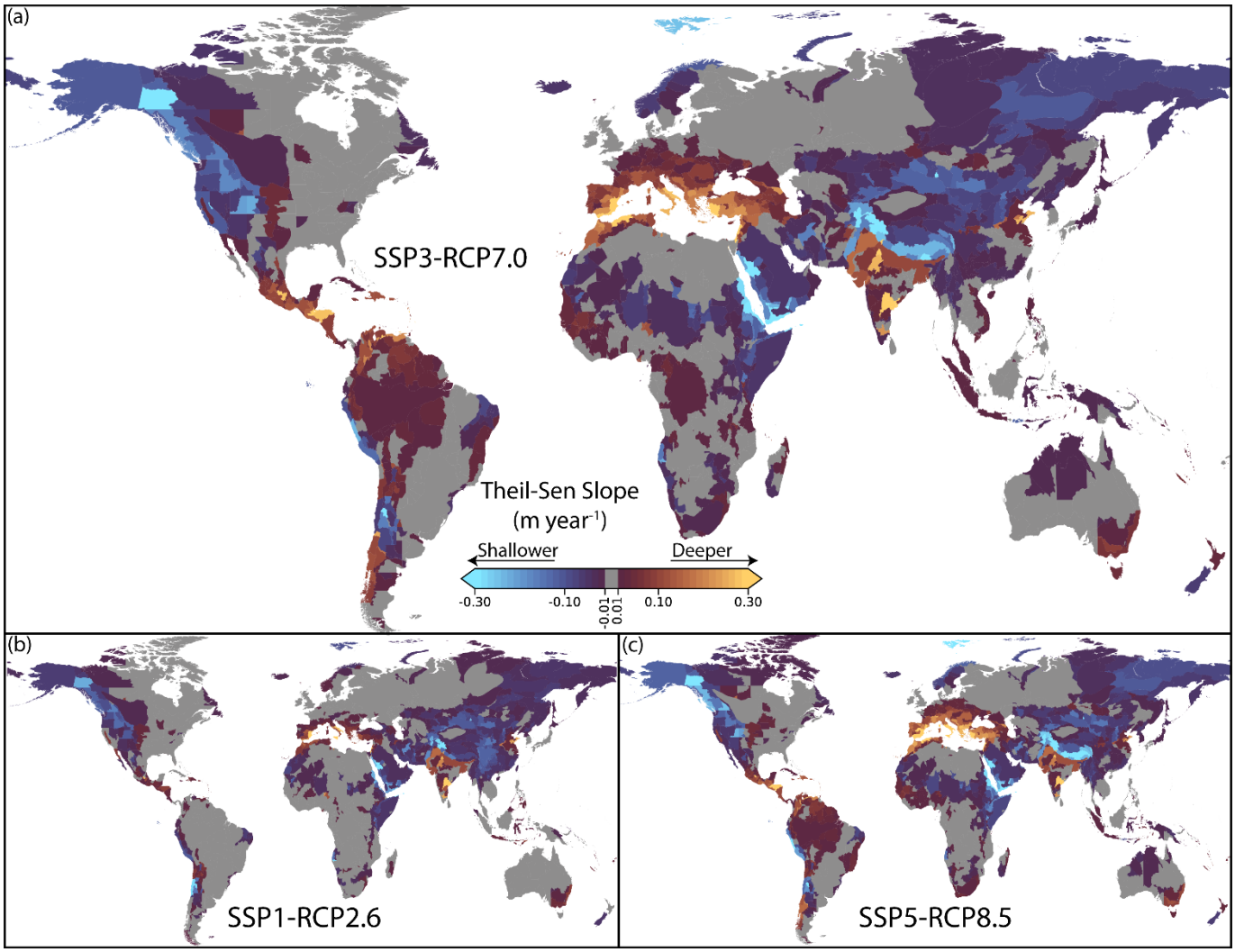

Figure RC1 9: Theil-Sen slope of annual water table depth (2014–2100) simulated by GLOBGM under three SSP-RCP scenario. Gray pixels indicate regions where slopes are either insignificant, less than 0.01m year−1 , or both. Note that this figure represents the lower model layer which is representative of unconfined and confined aquifers

A more robust discussion of uncertainty is needed, and the results should be compared to the literature. Given (a) the performance of the model in data-rich regions, (b) uncertain parameterization of the model in data-scarce regions, and (c) the uncertainty of the GCMs (you have included only five, and only one variant from each), how confident can we be in the projected trends? How do they compare to previous studies? For example, across the United States, Condon et al. (2020) predict deepening water tables across the US under warming. Meixner et al (2016) predicted decreased recharge over the southwestern US, little change of the northwest, and also that mountain recharge would decrease. Other regional assessments exist and should be compared as well.

We will extend the discussion to include the comparison to other studies and elaborate on the uncertainty in data poor areas. Regarding the choice of GCMs: as described in our Methods, we followed the ISIMIP 3b protocol. This entails using 5 bias-corrected GCMs that have been chosen from the complete CMIP6 ensemble (models and members) to encompass the variation in wet-warm, dry-warm, wet-cold, dry-cold models and members and a central one. This way, climate change uncertainty per scenario is taken into account with a small number of models making computation feasible.

I have some further minor comments and suggestions for the authors should they consider revising and resubmitting, in no particular order:

The authors state that forcing for the model are abstraction, recharge, and discharge (L227). Is this correct? If so, what is the model doing other than accounting for inputs and outputs? Why does the dynamic drainage elevation (L104) matter if groundwater discharge is already prescribed?

We are sorry for the misunderstanding. Discharge is river discharge, which is translated into surface water levels using Manning's equation. The monthly average surface water levels are then used in MODFLOW's river package as a boundary condition. We will improve the description of this process.

Equations 12-15: variables are not consistently labelled. What is alpha in equation 13? Is this different from the alpha in eq 10? Is WTDsim different from Wsim?

We agree that this is confusing. We will use another symbol for the alpha in equation 13 and WTDsim throughout.

There is no confining layer over Canada at all? Surely this could be improved.

We agree that this could be improved. We are working on a new global hydrogeological schematization. This is a significant task worthy of a separate paper and not ready for some time. In the meantime, we relied on the same schematization as developed by De Graaf et al. (2017).

Two of the GCMs (IPSL-CM6A-LR and UKESM1-0-LL) in your ensemble are 'hot' - they have and Equilibrium Climate Sensitivity (ECS) and Transient Climate Response (TCR) above the assessed 'very likely' ranges estimated by the IPCC. This means that their projected warming is probably too pessimistic for a given scenario. The recommendation is, if the warming trajectory is important (which I think it is here) to use only models that lie within the likely range (Hausfather, 2022). Consider reporting the ensemble mean just for the three models that do lie within the 'likely' range, and including the others in an appendix.

We understand the reasoning behind this, but we are very hesitant to start attaching likelihoods to individual GCMs and compute weighted ensemble means. We elect to stick to the ISIMIP3b protocol that represent the full envelope of the CMIP6 models. Also note that in the ISIMIP approach, the GCMs are bias-corrected for temperature, monthly precipitation totals and wet day frequency. This may mitigate some of the

deviations of the GCMs mentioned. We will however mention this issue in the Discussion and also state the possibility of selecting individual GCM runs which we have archived separately.

The model struggles to simulate well-based GW levels and trends, particularly in places where GW use is high. Maybe that is not surprising, given that your water use data (Lange et al, 2021) is originally at 0.5 degree resolution, and the recharge is also based on downscaling coarser-resolution data. Perhaps simulating well-based trends is too difficult a task at the global scale. Does the model accurately simulate regional trends? I suggest the authors perform a simple experiment: aggregate the well-based data at increasingly coarse resolutions (say, 1 km, 2 km, 10 km, 50 km, and 100 km) and do the same for the model data, and then calculate your performance metrics at each resolution. I would expect you'll find performance will improve with coarser resolutions. Then focus on reporting results at the finest resolution that provides an adequate match to observations.

We thank the reviewer for this suggestion. We acknowledge that the downscaling introduces additional uncertainty. However, we respectfully disagree with the above aggregation experiment for demonstrating model validity. Spatial averaging to coarser resolutions mathematically reduces variance in both observed and simulated datasets, which will inevitably improve KGE-NP performance. This is not because model physics has improved but because the validation target has been simplified. Reporting performance at the coarsest resolution that yields acceptable metrics would risk overstating model skill and hiding limitations. The objective of this study is explicitly to evaluate model performance at 1 km resolution which is the scale at which the model is applied. Aggregating to coarser resolutions to find a resolution where metrics become acceptable does not validate the model at its operational scale; it merely identifies the scale at which its errors become invisible. From the analyses shown above, we show that validation results are better than previous global model attempts at similar and coarser resolutions and that trends largely follow those that we observe from the piezometric data. Also, we have now done the upscaling exercise when comparing our trends with GRACE and representing regional trends in Figure 8 and 9.

Eq. 17: Is this missing a month index?

The month index is k which is within the summation, but not outside.

Figure 6a: The color scale for this figure should be the same as for the uncorrected WTD bias (Figure 2).

That is a good suggestion. We will change this.

Figure 7: It's unclear why these regions were chosen for insets. In any case they provide only about 2X magnification. I suggest removing the insets.

Thanks for the suggestion. We will change this.

Discussion: The discussion of modelling choices, calibration, and advances is somewhat incongruous with ESD. I would expect this in a journal like Geoscientific Model Development but I think for ESD the discussion should be more general, and focus more on the implications.

We will refocus some of the discussion to implications, and include the reference to other studies as asked for previously.

L424 - The beginning of this paragraph seems to be missing. "[Major rivers], such as..."?

Thanks for noticing. We will correct this.

L435: Are these rising water tables in the north robust? Do they match observations? The authors state that 'This could conceivably be due to climate change enhancing precipitation and groundwater recharge dynamics'. It seems to me there is no need to speculate here - are those two processes actually occurring in the model?

That is a good point. There are unfortunately hardly any observations in these regions. On the other hand, the rising water tables in these regions occur more clearly in the future scenarios. For these scenarios, we see increased precipitation and thus recharge in the North and there is no groundwater use, so we can be certain about the drivers. We will change the wording in the text.

Line 245: You could report the number of CPU-hours also. In addition, it would be responsible to report the CO2 footprint of these computations. This can be estimated as:

(node-hours) * (12 nodes/ # nodes at Snellius) * (power usage at Snellius) * (carbon intensity of Dutch grid)

Based on a quick search I get:

551 h * (12/1557) * (1200 kW) * 235 gCO2e/kWh = 1200 kg CO2e, about equal to a round trip flight ticket from Amsterdam to Beijing.

We will include the node hours in the text and estimate the energy use and associated CO^2 footprint where feasible.

3.2.2 - Are these values for the ML-corrected data or the raw model output?

These are the raw model outputs. We will make this more clear in the text.

If the model purports to include anthropogenic influences on the water table, why are regions with anthropogenic influence excluded from calibration?

We calibrated the steady state model for conductivity, anisotropy and riverbed conductance. The steady state model can be well informed by these resistances while steady state allows much more model evaluations that take significantly less

computation time than the transient runs. We did not further change the storage coefficients with transient calibration, as initial experiments showed limited sensitivity to changing these. Although groundwater abstraction can lead to a steady state if it is smaller than recharge, we chose to calibrate the groundwater model without groundwater abstractions using only long-term average recharge and surface water levels as forcing, to be sure that steady state can actually be reached everywhere. Consequently, during the calibration we excluded the observation wells in the areas with significant groundwater withdrawal where persistent trends can be expected. Thereafter, we performed the spinup and simulations with the transient model which was evaluated against the observation time series. We will check if this description can be improved in the paper.

L451: I would not call this 'disagreement' - rather, divergence in scenarios, or you could say the direction of change is scenario-dependent.

Agreed. We will update the text.

L288 'The performance of the model' - which model? GLOBGM or the XGBoost model?

Here we are referring to the XGBoost model, we will clarify the text.

Table 4: it would be more useful to report the absolute bias, so negative and positive biases do not cancel each other out. Did the calibration reduce the mean absolute bias?

The bias, weighted by depth class is mean absolute bias, this will be made abundantly clear in the updated manuscript.

Section 3.1.1. This section seems more like methods than results.

We will critically look at this section.

L390 and L473: The shallower depths show better correlations for monthly data but not for yearly. That suggests that at shallow depths you can better capture variation and change. Seasonal variation is probably only strong for shallow wells, so it makes sense to me that the difference in performance disappears in the yearly time series. I therefore disagree that "seasonal dynamics are more challenging to capture than inter-annual trends".

We agree with this point and will revise the text accordingly. As stated above, interannual variability is more difficult than seasonal dynamics.

Figure 10: is the y-axis in m? What does the uncertainty represent - standard deviation of the five GCMS? The GSWP3-W5E5 simulation should also be plotted on these graphs for comparison.

The y-axis is m and this will be added. The uncertainty bounds are indeed standard deviation of the 5 GCMs. We will add the GSWP3-W5E5 simulation in the revised manuscript.

References:

Bäthge, A., Vargas, C. R., Lischeid, G., et al.: A global-scale time series dataset for groundwater studies within the Earth system, Sci. Data, 13, 401, https://doi.org/10.1038/s41597-026-06966-1, 2026.

Chandanpurkar, H. A., et al.: Unprecedented continental drying, shrinking freshwater availability, and increasing land contributions to sea level rise, Sci. Adv., 11, eadx0298, https://doi.org/10.1126/sciadv.adx0298, 2025.

Chen, Z., Goldscheider, N., Auler, A., and Bakalowicz, M.: World Karst Aquifer Map (WHYMAP WOKAM), https://doi.org/10.25928/b2.21_sfkq-r406, 2017.

Condon, L. E., Atchley, A. L., and Maxwell, R. M.: Evapotranspiration depletes groundwater under warming over the contiguous United States, Nat. Commun., 11, 873, https://doi.org/10.1038/s41467-020-14688-0, 2020.

de Graaf, I. E. M., van Beek, R. L. P. H., Gleeson, T., Moosdorf, N., Schmitz, O., Sutanudjaja, E. H., and Bierkens, M. F. P.: A global-scale two-layer transient groundwater model: Development and application to groundwater depletion, Adv. Water Resour., 102, 53–67, https://doi.org/10.1016/j.advwatres.2017.01.011, 2017.

Doherty, J.: Ground water model calibration using pilot points and regularization, Groundwater, 41, 170–177, https://doi.org/10.1111/j.1745-6584.2003.tb02580.x, 2003.

Fan, Y., Miguez-Macho, G., Jobbágy, E. G., Jackson, R. B., and Otero-Casal, C.: Hydrologic regulation of plant rooting depth, Proc. Natl. Acad. Sci. U.S.A., 114, 10572–10577, https://doi.org/10.1073/pnas.1712381114, 2017.

Hausfather, Z., Marvel, K., Schmidt, G. A., Nielsen-Gammon, J. W., and Zelinka, M.: Climate simulations: recognize the 'hot model' problem, Nature, 605, 26–29, https://doi.org/10.1038/d41586-022-01192-2, 2022.

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

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-327, 2019.

Krysanova, V., Zaherpour, J., Didovets, I., et al.: How evaluation of global hydrological models can help to improve credibility of river discharge projections under climate change, Clim. Change, 163, 1353–1377, https://doi.org/10.1007/s10584-020-02840-0, 2020.

Liu, R., Zhang, X., Wang, W. et al.: Global-scale ERA5 product precipitation and temperature evaluation. Ecol. Indic., 166, 112481, https://doi.org/10.1016/j.ecolind.2024.112481, 2024.

Ma, Y., Condon, L. E., Koch, J., et al.: High resolution US water table depth estimates reveal quantity of accessible groundwater, Commun. Earth Environ., 7, 45, https://doi.org/10.1038/s43247-025-03094-3, 2026.

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, J. Hydrol., 534, 124–138, https://doi.org/10.1016/j.jhydrol.2015.12.027, 2016.

Moeck, C., Grech-Cumbo, N., Podgorski, J., Bretzler, A., Gurdak, J. J., Berg, M., and Schirmer, M.: A global-scale dataset of direct natural groundwater recharge rates: a review of variables, processes and relationships, Sci. Total Environ., 717, 137042, https://doi.org/10.1016/j.scitotenv.2020.137042, 2020.

Reinecke, R., Foglia, L., Mehl, S., Trautmann, T., Cáceres, D., and Döll, P.: Challenges in developing a global gradient-based groundwater model (G3M v1.0) for the integration into a global hydrological model, Geosci. Model Dev., 12, 2401–2418, https://doi.org/10.5194/gmd-12-2401-2019, 2019.

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., Dankers, R., Eisner, S., Fekete, B. M., Colón-González, F. J., Gosling, S. N., Kim, H., Liu, X., Masaki, Y., Portmann, F. T., Satoh, Y., Stacke, T., Tang, Q., Wada, Y., and Kabat, P.: Multimodel assessment of water scarcity under climate change, P. Natl. Acad. Sci. USA, 111, 3245–3250, https://doi.org/10.1073/pnas.1222460110, 2014.

Straatsma, M., Droogers, P., Hunink, J., Berendrecht, W., Buitink, J., Buytaert, W., Karssenberg, D., Schmitz, O., Sutanudjaja, E. H., van Beek, L. P. H., Vitolo, C., and Bierkens, M. F. P.: Global to regional scale evaluation of adaptation measures to reduce the future water gap, Environ. Model. Softw., 124, 104578, https://doi.org/10.1016/j.envsoft.2019.104578, 2020.

Triplett, A., Bennett, A., Condon, L. E., Melchior, P., and Maxwell, R. M.: A deep-learning based parameter inversion framework for large-scale groundwater models, Geophys. Res. Lett., 52, e2024GL114285, https://doi.org/10.1029/2024GL114285, 2025.

Verkaik, J., Sutanudjaja, E. H., Oude Essink, G. H. P., Lin, H. X., and Bierkens, M. F. P.: GLOBGM v1.0: a parallel implementation of a 30 arcsec PCR-GLOBWB-MODFLOW global-scale groundwater model, Geosci. Model Dev., 17, 275–300, https://doi.org/10.5194/gmd-17-275-2024, 2024.

Zamrsky, D., Ruzzante, S., Compare, K., Kretschmer, D., Zipper, S., Befus, K. M., Reinecke, R., Pasner, Y., Gleeson, T., Jordan, K., Cuthbert, M., Castronova, A. M., Wagener, T., and Bierkens, M. F. P.: Current trends and biases in groundwater modelling using the community-driven groundwater model portal (GroMoPo), Hydrogeol. J., 33, 355–366, https://doi.org/10.1007/s10040-025-02882-7, 2025.

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.