<u>Reviewer 1:</u>

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 and we maintain that our model is sufficiently accurate to assess regional trends in groundwater resources. In the following response and revised manuscript, 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 and often calibrated using local data on hydrogeology, groundwater levels and forcing. Global models are for global-scale assessments of past and future trends, which can be identified even if local 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 accuracy should lead to reducing the resolution to 5 arc-minutes 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. Moreover, the results shown here do a better job than transient model simulation with models at a 5 arc-minute resolution (see the comparison below). The main reason is that topography and surface-groundwater interactions are better resolved at higher resolutions. There is therefore no reason to coarsen the spatial resolution, as long as the accuracy of the results against observations are made clear.

That being said, we agree with Referee #1 that further improvements in the accuracy assessment and the discussion of the results are possible. We have addressed the specific concerns below but start by discussing in general the accuracy of the results in comparison to other global models. Please note that the line numbers here refer to the track-changes pdf.

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 in which, if multiple observation wells occur within a single grid cell, we choose the one with the longest temporal coverage. In the updated manuscript, we have updated validation procedure description in the methods (Section 2.4.2, lines 230- 231) and results (Section 3.2.2, lines 480-524).

Also, we compared the simulated trends with those of GRACE/GRACE-FO (Chandanpurkar et al. 2025). For fair comparison we applied a Gaussian spatial moving average with a standard deviation of 75 km, which approximately matches the GRACE/GRACE-FO footprint. Simulated TWS trends were computed from 5 arc-minute PCR-GLOBWB outputs, replacing the groundwater component with GLOBGM estimates. Groundwater storage trends from GLOBGM alone were also computed separately. The results show that similar trends can be observed, especially around the groundwater and abstraction hotspots (updated Fig. E1). The largest differences between GRACE/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). This has been added to the methods (lines 318-324), results (lines 517-524) and discussion (lines 679-692).

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

We have now included a section which compares the accuracy of GLOBGM v1.1 to previous global modelling efforts. We refer the reviewer to the revised manuscript for a description of the relevant methods (Section 2.5, lines 381-393), results (Section 3.3. lines 543-558) and discussion (lines 689-692). In the text that follows we summarise the main findings.

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). This shows that results are better than all of these models except Fan et al. (2017) when looking at steady-state estimates (updated Fig 7). We also compared Verkaik et al. (2024) and De Graaf et al. (2017), showing again that the results are better (updated Fig 7). Note that Fan et al. (2017) and Reinecke et al. (2019) provide steady-state results only. This leads us to conclude that our results are on par or better than previous modelling results, even though they are not perfect.

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

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 1).

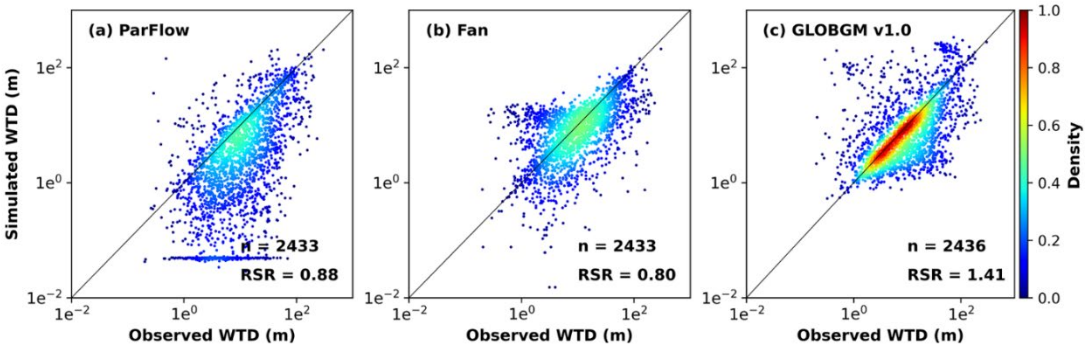

Figure RC1 1: Sourced from Yang et al. 2025: Scatterplots of simulations vs. observations of water table depth. The simulations are long-term average water table depths of 1981-2010 in the current work, of 2004-2014 in Fan's model, and of 1958–2015 in GLOBGM v1.0. The observations are the averages of 2018.

<u>The concerns regarding negative correlations:</u>

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 for 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 the original Figure 5 convey a pessimistic view of the model's 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 slight 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 simulated and observed Theil-Sen slopes fitted on groundwater heads.

We first show that for the locations with a negative spearman correlation and KGE-NP$_{skill}$ score < 0, 47-50 % of simulated and observed time series have negligible slopes (|slope| < 0.1 m year^{-1}; Jasechko et al. 2024; updated Fig D1a). In addition, when excluding these points from the overall distribution, we show that only 14% of annual and 7% of monthly locations show significant opposing direction correlation (updated Fig D1b).

To address this comment, we have included this additional context in the appendix, (appendix D), results (lines 497-508) and discussion (lines 682-704). We would also like to note that from the Figure 5 it is clear 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, at least partly, be attributed to the delays that occur in deeper unsaturated zones that are not modelled in our approach.

<u>Spatial overlap of abstraction and known limitations regions with performance:</u>

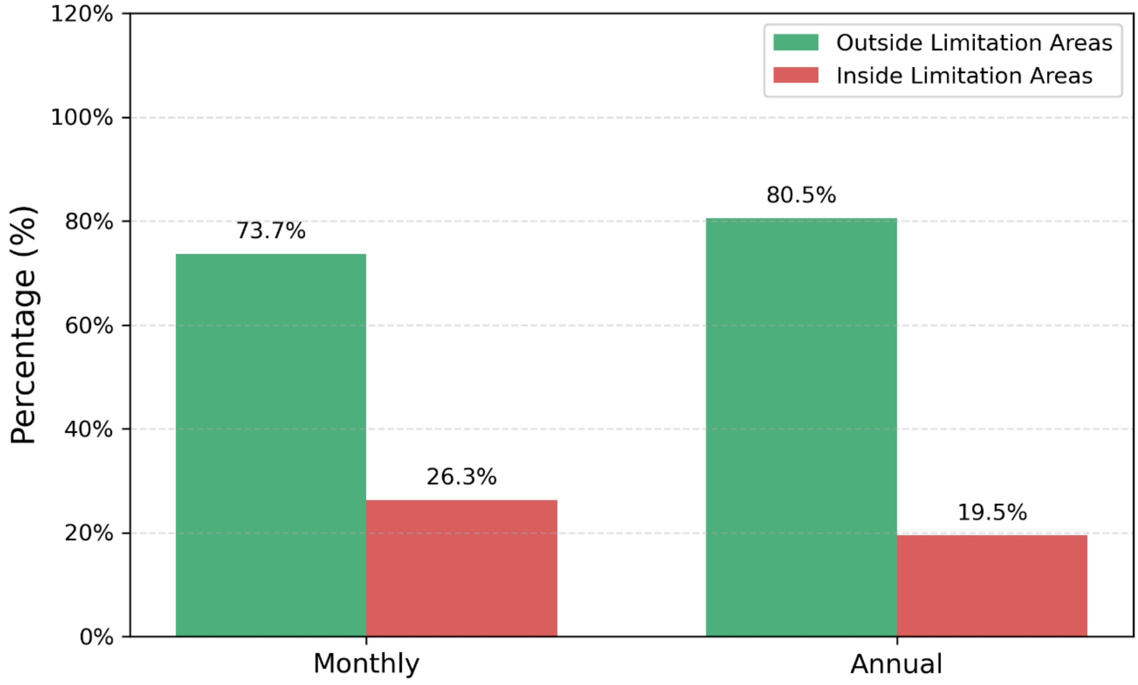

Figure RC1 2: Proportion of groundwater head observations located outside and inside the regions of known GLOBGM limitations (Figure 4), for both monthly and annual time series datasets.

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. The majority of observation wells occur 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 80% of all observations (both good and poor, Figure RC1 2) fall outside Figure 4 regions, then any spatial pattern of performance will predominantly appear outside these regions. The concentration of poor performing locations outside of Figure 4 is therefore a direct consequence of where the data exist. The spatial mismatch the reviewer identifies is not diagnostic of model behavior but of data locality.

Nonetheless, we have calculated box plots of KGE-NP_{skill}, r_spearman, alpha_fdc and beta separately for stations inside the regions marked in Figure 4, for the remaining land area, and for both combined (Figure RC1 3). This shows that the distributions of KGE-NP_{skill} 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 a artifact of where observations exist and not indicative of the model accuracy.

We also examined whether model performance differs systematically between high-abstraction and low-abstraction regions (Figure RC1 4). The distributions of KGE-NP_{skill} 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.

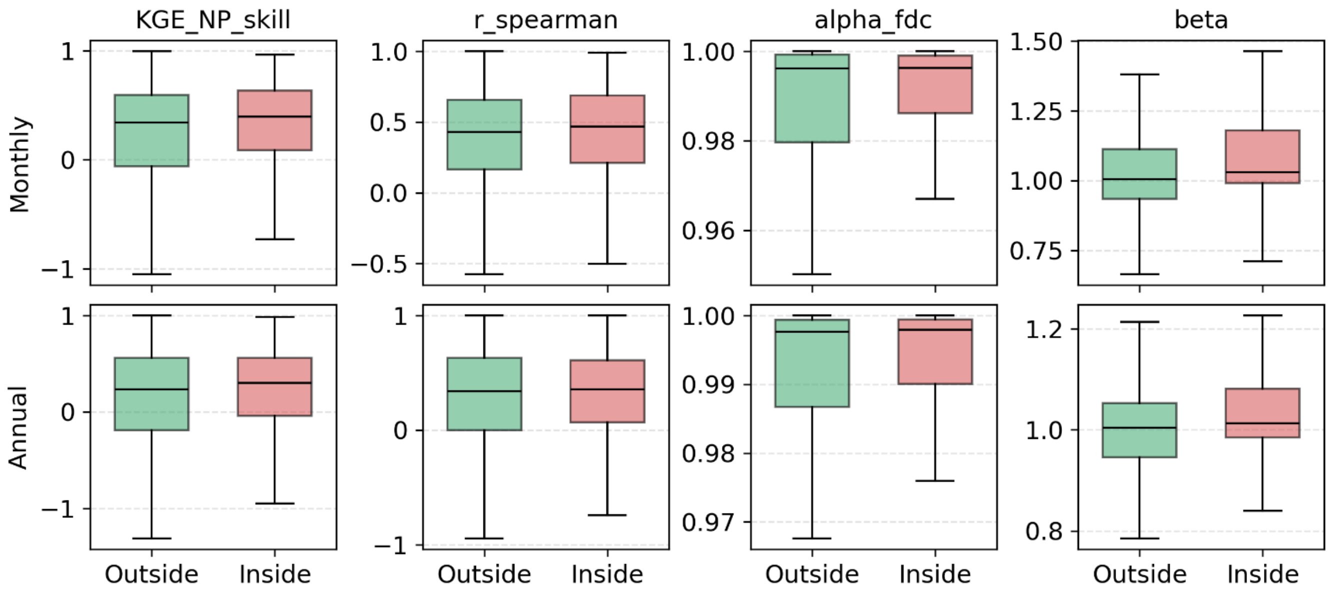

Figure RC1 3: Distribution of KGE-NP performance metric components, Spearman correlation (r_spearman), flow duration curve slope ratio (alpha_fdc), and bias ratio (beta), for observations located outside and inside the GLOBGM limitation regions (Figure 4), shown for monthly (top row) and annual (bottom row) datasets.

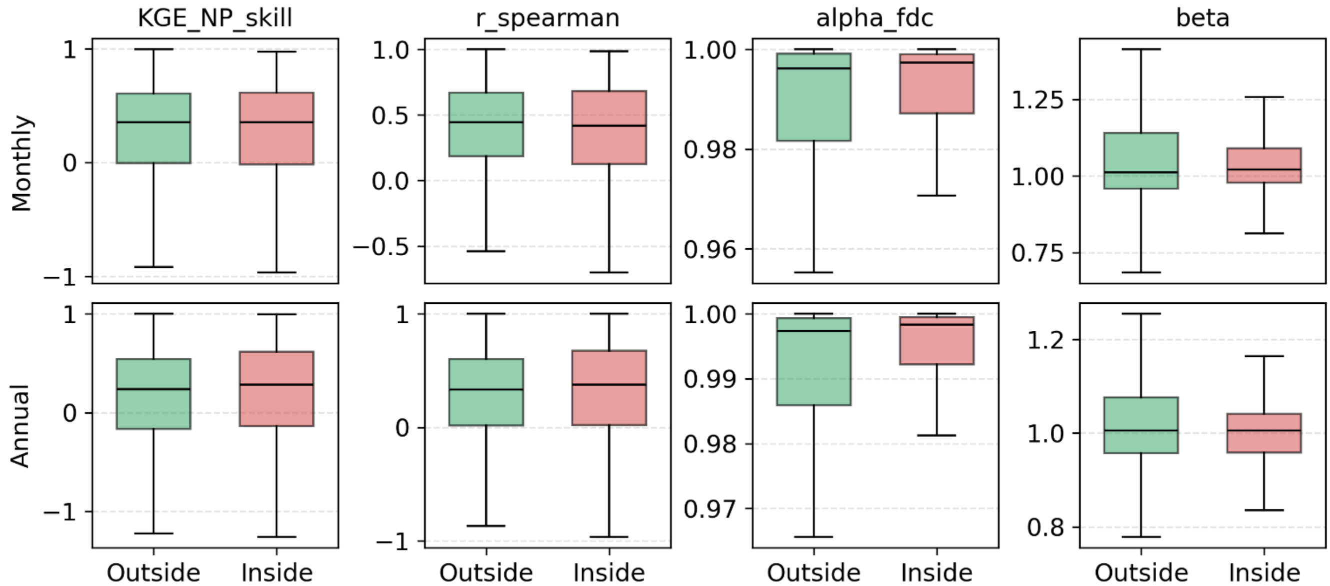

Figure RC1 4: Distribution of KGE-NP performance metric components, Spearman correlation (r_spearman), flow duration curve slope ratio (alpha_fdc), and bias ratio (beta), for observations located outside and inside abstraction hotspots (Figure 2), shown for monthly (top row) and annual (bottom row) datasets.

On the bias ratio:

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) 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 here and in the revised manuscript 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 high to make this feasible. 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 have pointed to these possibilities in the discussion of the revised manuscript (lines 622-664).

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 has been 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 have made the distinction between these two products, and the appropriate use cases for each, more explicit in the revised manuscript (lines 679-681).

Regarding the suggestion to overhaul the model to try and achieve better global performance, in lines 71-76 of this rebuttal we demonstrate that the results here are better than or at least comparable to other global groundwater models that have been used previously to assess groundwater level trends and projections of these trends under climate and socioeconomic change. 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 30-37 and 381-395 of this rebuttal.

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 have addressed in the revised manuscript. We aimed to replicate the rationale presented by Knoben et al. (2019) on the KGE-NP score. However, given that the KGE-NP of a mean benchmark simulation depends on the coefficient of mean absolute deviation of the observed record, a universal threshold cannot be defined. Instead of applying a global threshold, we rely on the KGE-NP_{skill} score to evaluate whether GLOBGM is better or worse than using a mean benchmark – this was calculated individually for each observation time series. Thereby we rely a on metric suitable for global-scale evaluation and multi-model comparison. We have detailed the derivation of the mean flow benchmark as it applies to KGE-NP in the Appendix B and lines 310 – 324.

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. In the revised manuscript, we have done a post-hoc analysis of the performance of GAM which estimates recharge. We employed a leave-one-out cross-validation approach, given the limited number of data points in Moeck (2020). Please see appendix section A which we point towards this section in the methods (lines 132-133). A multiplication factor was chosen to avoid the introduction of negative recharge values; we have included this in the revised manuscript (line 147).

We agree with the reviewer that the precipitation constraint may limit return flow contributions to recharge. For the steady-state simulation the influence of the constraint can be considered negligible since the simulation excluded anthropogenic influence and represents a pristine condition and there is no irrigation affects. For the transient simulation, the multiplicative correction factor, which was fitted on long-term mean estimates, produced unrealistically high recharge rates in wet months when applied at the monthly time step. Some form of bounding was necessary to prevent numerical instabilities and associated convergence problems. We chose monthly precipitation as the upper limit, which implicitly assumes that precipitation excess can transfer to recharge within a single month, which is a reasonable approximation in many regions, though admittedly conservative in arid environments during dry months. Alternatively, had we chosen annual precipitation as the constraint we would have avoided excessive suppression in dry months but at the cost of relocating recharge in a way that overestimates monthly fluxes. We have included this context in the updated manuscript and clearly stated the limitations (lines 147-157). In addition, the observation data are dominated by irrigated areas, and the resulting skill in comparison to GLOBGM v1.0 suggests that the net effect is positive.

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 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.

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.

In response to this comment, we have re-evaluated how we implement the ML bias approach. In the revised manuscript we have adopted a different approach. In short, we train a two-stage model using a LightGBM, where in the first stage the model predicts the sign of the bias and in the second stage the model predicts the magnitude. These two stages are combined to predict a final bias value. An advantage of this approach is that we have now explicitly incorporated information on spatial transferability of the ML model given that we train, evaluate and test spatially distinct datasets. We have elaborated on this approach in the methods (Section 2.4.3) and have made it more clear in the results that the performance is comparable between train, validation, and test sets - which suggest that the model is spatially transferrable, please see Appendix F and lines 527-530.

We have also updated the axis labels and legend definitions in Figure A1 as suggested, thank you for highlighting this mistake, please see Figure F1.

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 the scope of the current paper. Figure 10 (now Figure 11) was 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 have updated Figure 8 and 9 and present the slopes aggregated by spatial units (see updated Figure 9 and 10). The spatial units are a combined water provinces (Straatsma et al. 2020) and WHYMAP aquifers (Chen et al. 2017) dataset. This has allowed us to incorporate discussions on regional patterns in comparison with previous research (Section 4.3 & 4.4). In addition, the original figures have been moved to the appendix (Appendix G). We believe that this is a more appropriate response than expanding the manuscript beyond its intended scope – the manuscript was 30 pages long excluding references and 37 pages including references. With the updates, its is now 45 without references and 54 pages with references.

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 have extended 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 ISIMIP3b 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. We have now included additional context on line 263-265.

In the updated manuscript we have included a model comparison to other global models which we discuss in Section 4.2. In addition, we have included a descriptive comparison between the current results and previous studies for both the historical reference simulation (lines 706-730) and scenario-based simulations (lines 745-758).

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 apologise 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 have improved the description in the text, please see line 100 and line 105-108.

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 have replaced alpha with omega (see lines 374-379) and WTD_{sim} 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 attach 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 have however mentioned this issue in the revised manuscript, see lines 264-265.

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 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 (Appendix E) and are representing regional trends in Figure 9 and 10.

Eq. 17: Is this missing a month index?

We thank the reviewer for this comment. Upon reflection, our previous response was incorrect: k indexes years, not months, and the original equation did not account for monthly variability. We have updated Eq. 17 to include an explicit double summation over both months (m) and years (k), with the denominator adjusted to 12N. The variable definitions have been updated accordingly. Please see line 399.

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 have ensured that Figure 2 and Figure 6a have identical colour maps and hexbin properties in the revised manuscript.

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 have increased the magnification factor for the inset regions. We prefer to keep the insets as they are linked to specific results which will not be directly apparent from the global map. Please see the updated Figure 8.

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 have refocused the discussion to implications and included the reference to other studies as asked for previously, please see Section 4.

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

Thanks for noticing. This has been updated, please see line 572.

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 have changed the wording in the text, please see lines 716-714.

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 have included the CO_2 footprint as calculated by green-algorithms.org (Lannelongue et al 2021), which equates to 767.01 kg CO_2 (lines 276-279) We were unable to verify the equation presented by the reviewer or the power usage at Snellius. Please see The Green Algorithms framework estimates the marginal carbon footprint of a computation based on the actual resources used data-centre efficiency, and local grid carbon intensity. As opposed to contributing a fixed share of the entire supercomputer's power draw, especially given that the system would be powered and maintained regardless of whether our specific simulations were run.

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

These are the raw model outputs. We have made this more clear in the text, please see lines 484-485.

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 many 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. 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 all of the observation time series, including those in areas with anthropogenic influence. We have elaborated on the rationale in the revised manuscript, please see line 184-185.

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

Agreed. The text has been updated. Please see lines 502 & 592.

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

Here we were referring to the XGBoost model. This sentence has largely been replaced by new text, but we have ensured that it is clear that performance in this context refers to the two-stage machine learning model trained to predict bias. See lines 344-359.

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?

This is a valuable suggestion, The bias, weighted by depth class is mean absolute bias, this has been made abundantly clear in the updated manuscript. Please see the updated Table 4.

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

Agreed, we have moved some of the explanatory text into the methods (lines 200-205) and focus only on the results of the calibration in section 3.1.1 in the updated text.

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 have revised the text accordingly. As stated above, interannual variability is more difficult than seasonal dynamics. We have corrected this argument, please see lines 685-689.

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 has been added. The uncertainty bounds are indeed standard deviation of the 5 GCMs. In the revised manuscript we have changed it to min max spread to show the full spread. In addition, we have added the historical reference line and made the units more clear in the plot. Please see the revised Figure 11.

<u>Reviewer 2:</u>

The manuscript describes further development of a global groundwater model (GLOBGM v1.1). This updated model incorporates key advances on its predecessor (GLOBGM v1.0, Verhaik et al., 2024) that include model calibration and downscaling of groundwater recharge. The manuscript is well structured, written and presented.

Major comments:

As outlined by the authors, the updated model marks an advance over its predecessor. However, the more pertinent and central question, especially given the authors' arguments in the Introduction, Discussion and Conclusions, is whether this advance marks a progression such that the updated, calibrated model provides meaningful simulations of observed changes in groundwater storage over the historical period (1960–2014) and thus projected changes in groundwater storage under socio-economic scenarios (i.e. 2015–2100: SSP1-RCP2.6, SSP3-RCP7.0, SSP5-RCP8.5). What is meant by "meaningful" may well depend upon one's perspective but it would broadly be expected to demonstrate skill in representing historical observations. Objectively, the authors rely on the application of the non-parametric form of the Kling-Gupta Efficiency (KGE-NP). Beyond reference to the threshold value of -0.41 cited from Knoben et al. (2019), the authors need to make a stronger case as to why this threshold is appropriate, together with an evaluation of the implications of the uncertainty in defining this threshold. Indeed, greater explicit quantitative discussion is needed in section 4 of the consequences of the observed uncertainty in model simulations indicated by the CDFs and their (limited) spatial distribution in correlations in Figure 5.

We thank Referee #2 for the review and providing constructive comments and suggestions, which will help to improve the quality of the manuscript. We agree with the reviewer that "meaningful" performance is based on the reader's perspective. It should ideally depend on the intended use of the model, which for an Earth-system model as ours pertains to the ability to map patterns and space-time trends that are realistic and consistent with observations. As can be seen from the additional evaluation in response to Referee #2, this is mostly the case. Please note that the line numbers here refer to the track-changes pdf.

When it comes to judging whether a model development is deemed an advance compared to previous effort, we argue that the skill and accuracy of GLOBGM, or any model for that matter, should be judged relative to the current state-of-the-art. We acknowledge that this comparison was absent from the original manuscript. To address this directly we have added a model intercomparison to the revised manuscript please see the updated methods (Section 2.5) results (Section 3.3), and discussion (lines 682-704), in which we compare the performance of the present model against previous global groundwater models that have been successfully applied to predicting steady-state groundwater patterns and trends. This comparison demonstrates that the updated, calibrated model shows improved skill relative to the current state-of-the-art, providing a stronger basis for confidence in our simulations. We thank the reviewer for highlighting this gap.

The reviewer's concerns regarding the use of -0.41 threshold for KGE-NP is valid, since it was originally for the standard KGE. This is a valid point which we have addressed in the revised manuscript. We aimed to replicate the rationale presented by Knoben et al. (2019) for the KGE-

NP score. However, given that the KGE-NP of a mean benchmark simulation depends on the coefficient of mean absolute deviation of the observed record, a universal threshold cannot be defined. Instead of applying a global threshold, we rely on the KGE-NP_{skill} score to evaluate whether GLOBGM is better or worse than using a mean benchmark – this was calculated individually for each observation time series. Thereby we rely on a metric suitable for global-scale evaluation and multi-model comparison. Please see lines 310 – 315 for further context in addition we have detailed the derivation of the mean flow benchmark as it applies to KGE-NP in Appendix B.

We agree with the reviewer that the discussion requires more focus on the uncertainty in light of the limited spatial distribution of observed data. In the discussion we have dedicated a section for this (Section 4.2). In addition, we have included a comparison with GRACE/GRACE-FO which allows us to assess accuracy where point-based observations are lacking (lines 318-324, 517-524, Appendix E).

Given the hyper-resolution (~1 km) scale of the model, I can understand the authors' necessary focus on reconciling the GLOBGM v1.1 to measurements of groundwater head from piezometric observations. I am surprised, however, not to see an effort to reconcile the latest model to evidence of changes in terrestrial water storage, albeit at lower resolution (aggregating outcomes to larger scales), from satellite gravimetry (GRACE/GRACE-FO).

We agree with the reviewer that this is necessary and the suggestion to reconcile the model simulations with GRACE/GRACE-FO and have added this in the revised manuscript (lines 318-324, 517-524, Appendix E).

It is well accepted that what the authors seek to achieve (i.e. a global hyper-resolution groundwater model) is hugely challenging and of very considerable importance. It is also heartening to see groundwater specialists leading this effort. The question is whether GLOBGMv1.1 marks an important technical advance on previous global groundwater models or whether the model has now reached a stage where it can be applied to simulate meaningfully the impact of human interventions and climate change; the current text attests the latter but current outcomes suggest the former.

We thank the reviewer for the kind words and for drawing the distinction between technical advances that GLOBGM represents and the suitability of GLOBGM for assessing socioeconomic impacts. We agree that this is an important distinction to make.

We reiterate that an " important technical advance " must be assessed relative to the available state-of-the-art alternatives. The model comparison added in Section 2.5, Section 3.3 & Section 4.3 demonstrates that GLOBGMv1.1 performance is comparable to or better than existing global groundwater models, which are themselves widely applied to socioeconomic and climate impact assessment in the peer-reviewed literature.

Minor comments:

1. There appears to be imprecision in mapped features in Figures 1 and 4 that presumably derive from the databases used to generate these. In Figure 1 for example, aquifers throughout Bangladesh are depicted as confined whereas, in really, confining conditions are restricted to areas overlain by Pleistocene terrace deposits (Shamsudduha

et al., 2022). In the same figure in Canada, mapping of confined aquifers (overlain thick glacial tills, especially in the prairies) are absent. In Figure 4, permafrost regions – certainly for Canada - are of much greater extent than depicted.

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).

2. To better understand challenges to model performance, might the authors present (e.g. Supplementary Material) the employed spatio-temporal distribution in observations of groundwater head over from 1960 to 2019

This is a valuable suggestion, we have added this in Appendix C and point towards it on line 658.

3. In Figure 2b, it is curious to see the exclusion of the United Kingdom from the analysis of water table depth bias. As these data are publicly available, do these not feature in the dataset of 34 800 observations wells from IGRAC?

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 many 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. 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, including those from the UK (see Figure 5 and 6). We have elaborated on the rationale in the revised manuscript, please see line 184-185.

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.

Shamsudduha, M., Taylor, R.G., Haq, M.I., Nowreen, S., Zahid, A. and Ahmed, K.M., 2022. The Bengal Water Machine: quantification of freshwater capture in Bangladesh.

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.