Dear Dr. Cenlin He,

Thank you so much again for handling our manuscript. Below, we respond to the three reviewers' feedback. We have improved our manuscript in the several aspects below:

- We have revised the wording of the model's validation on WTD performance and discussions in the conclusion part;
- We have clarified several technical details of the model performance in terms of human gradients.

By incorporating these improvements, we hope that we have satisfactorily addressed all reviewer comments and that the revised manuscript will meet the standards for publication in GMD

Below, we provide our responses in blue text, while the reviewers' comments remain as black.

Sincerely regards,

Qing He

on behalf of all coauthors

Reviewer #1 (Dr. Robert Reinecke)

I applaud the authors for a comprehensive revision that addressed all my previous remarks. However, I also have a couple of new remarks that should be addressed before this can be accepted for publication.

Response:

Dear Dr. Reinecke, we sincerely thank you for the positive feedback, and we are grateful for the additional remarks provided in this round. We have carefully addressed each new comment point by point as below.

The abstract now states, "However, the model's WTD behaviour is reasonably well in densely populated and irrigated areas, demonstrating its validity for application relevant to human water use activities".

I would disagree with that statement. Which part of your analysis supports this conclusion? Also, what does reasonably well mean? Please provide a quantification.

Response:

We thank the reviewer for this valuable comment and agree that the previous conclusion was overstated. See our response to your next comment below.

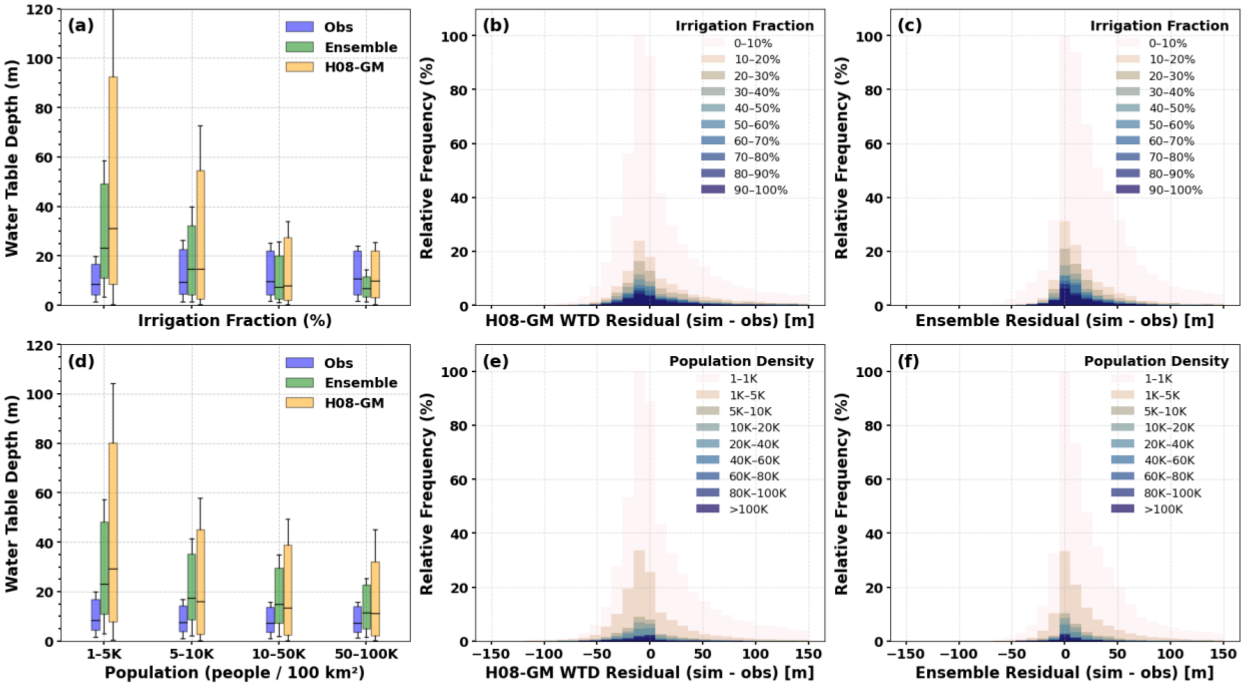
We have revised the abstract to remove any subjective assessment of model performance on WTD. The sentence now reads:

"Our analysis also reveals two complementary global relationships: one between groundwater depth and topographic slope, and another along gradients of human activity (irrigation and population), together demonstrating how natural and anthropogenic processes jointly control the spatial distribution of WTD." (Line 18-20)

I have the same concern about the description of Fig. 10. While it is interesting to see that the general spread of WTD is lower in highly irrigated and populated areas, this does not necessarily mean it is sufficient. Just because the mean is similarly close to a very uncertain ensemble is not enough. Also, even if the mean is close to the observed mean, it does not tell us anything about the model's validity in specific places or in general. I suggest removing this part of the analysis.

Response:

Thank you for this insightful comment and we fully agree that our earlier statement based solely on the comparison of mean values in the boxplots may have overstated the model performance. To better illustrate the information not captured by the boxplots, we have further added stratified histograms (below and in the revised Fig. 10). The results show that even if highly irrigated and populated regions exhibit a narrower spread of WTD biases, the number of samples with small residuals (low bias) also increases across both low and high human-influence groups, which confirms your concerns.



However, we prefer to retain this part of the analysis, as it does not aim to assess the model's validity at specific local scales, but rather reveals a systematic and interpretable relationship between human activities and the model's WTD biases at the global scale. The analysis is similar to the previously discussed WTD-slope relationship, but complement it with a focus on anthropogenic processes. We believe the analysis here is still valuable to provide insights into the model's performance under different climatic and human activity gradients, as Reviewer #3 already pointed out. A comprehensive site-level validation of model performance would indeed be valuable, and we intend to further explore this aspect in our future work.

Following your suggestion, we have revised the corresponding text in the manuscript to clearly state this intent and to avoid any implication that the model performance in these regions is "sufficient". (Line 494 - 500)

Line 21: magnitude of what?

Response: It is "the magnitude of the net groundwater lateral flow". Corrected now. (Line 22)

Fig. 6: All figures should contain a complete, standalone description that enables the reader to understand the figure.

Response: Thank you for this comment. We have added detailed captions in the revised manuscript (Line 417 - 420).

Fig. 11: Does this need to be in the main manuscript? I can't make out any differences on these small maps. I suggest moving this to the supplement.

Response: Thank you. We have moved it to the Appendix now.

620: de Graaf (2015) evaluated mainly the coefficient of variation of the model output as a sensitivity analysis. I disagree that this provided more insights than your OAT experiment. In Reinecke 2019 HESS, we provide a much more comprehensive analysis. I suggest citing both studies here.

Response: We sincerely thank the reviewer for this insightful comment and for recognizing the value of our sensitivity analysis. We have now cited both de Graaf et al. (2015) and Reinecke et al. (2019) and revised the text accordingly. The text now reads as:

"Compared with the earlier global sensitivity analysis by de Graaf et al. (2015), which mainly evaluated the coefficient of variation of model outputs, and the more comprehensive subsequent study by Reinecke et al. (2019b), which systematically quantified model sensitivity to both individual and combined parameter variations through an extensive set of 1,848 Monte Carlo experiments, our OAT sensitivity test provides a complementary but more limited perspective on parameter uncertainty." (Line 665 - 675)

631 and following: this needs to be rephrased. Also, avoid making a specific promise about a future paper in the conclusions. That should only be made if this is a multi-part paper from the beginning. I suggest a general remark on future research here.

Response: Thank you for your suggestion. We have revised the context accordingly. The text now reads as:

"In the next step, the temporal groundwater level variability and the human water withdrawal effect over the past 40 years should be investigated to help further advance our understanding of the important role of groundwater in supporting human water consumption, and the fundamental mechanisms behind the human-groundwater interactions.” (Line 685 - 687)

Reviewer #2:

Thank you for the revised version. The authors have addressed the previous comments thoroughly, and the new draft is well detailed and clearly presented.I have small comments, please clarify what “GHWMs” in line 60 refers to, and the citation in line 46has to be at the end of sentence.

Response: We sincerely thank the reviewer for the positive feedback and further comments. The term “GHWMs” in line 60 of the original manuscript was left from an earlier draft and has now been corrected to “GWMs” throughout the revised version for consistency. We have also moved the citation in Line 46 of the original manuscript to the end of the sentence in the revised manuscript (Line 49 and Line 64).

Reviewer #3:

I would like to thank the authors for their thorough and thoughtful revision. The manuscript has been substantially improved in clarity, structure, and scientific depth. Most of my previous comments have been carefully addressed, and I am overall satisfied with the revision. The inclusion of the additional validation in Section 3.2 is particularly valuable and provides new insights into the model’s performance under different climatic and human activity gradients.

Though I have two remaining points regarding the new Section 3.2.

Response: Dear Reviewer, thank you so much for your positive comments and we are glad to know our last round revision has met your expectation. Please see below for our point-to-point response to your new remarks.

Figure 10: The WTD boxplots for H08-GM show much wider ranges (25–75%) than those of the ensemble mean. Could the authors briefly explain why this spread is so large? For example, is it because the model captures more variability from the H08 forcing and lithology parameters, or because of larger uncertainty from the single-layer steady-state setup?

Response:

We thank the reviewer for this thoughtful question. The wider interquartile range (25–75%) of WTD in H08-GM compared to the ensemble mean can be partly explained by the averaging nature of the ensemble. Since the ensemble mean combines the outputs from four different global groundwater models, two of which (Fan et al., 2013; Reinecke et al., 2019) produce systematically shallower WTD (see Fig. 2 in Reinecke et al., 2024), the averaging process inherently smooths spatial variability and reduces the spread.

On the other hand, the H08-GM simulations retain more of the spatial heterogeneity arising from its specific forcing data, lithological properties, and parameterization, as the reviewer pointed out. Identifying which factors dominate this difference would require coordinated experiments under consistent simulation settings across all models, which we consider an important next step for future intercomparison work.

We have added the above discussion in the revised manuscript as well (Line 485 - 491.

Figure 9: The slope–WTD relationships look clear, but it would be great to add a simple number (like a correlation R or R^2) to quantify how well the model agrees with the observations and ensemble mean.

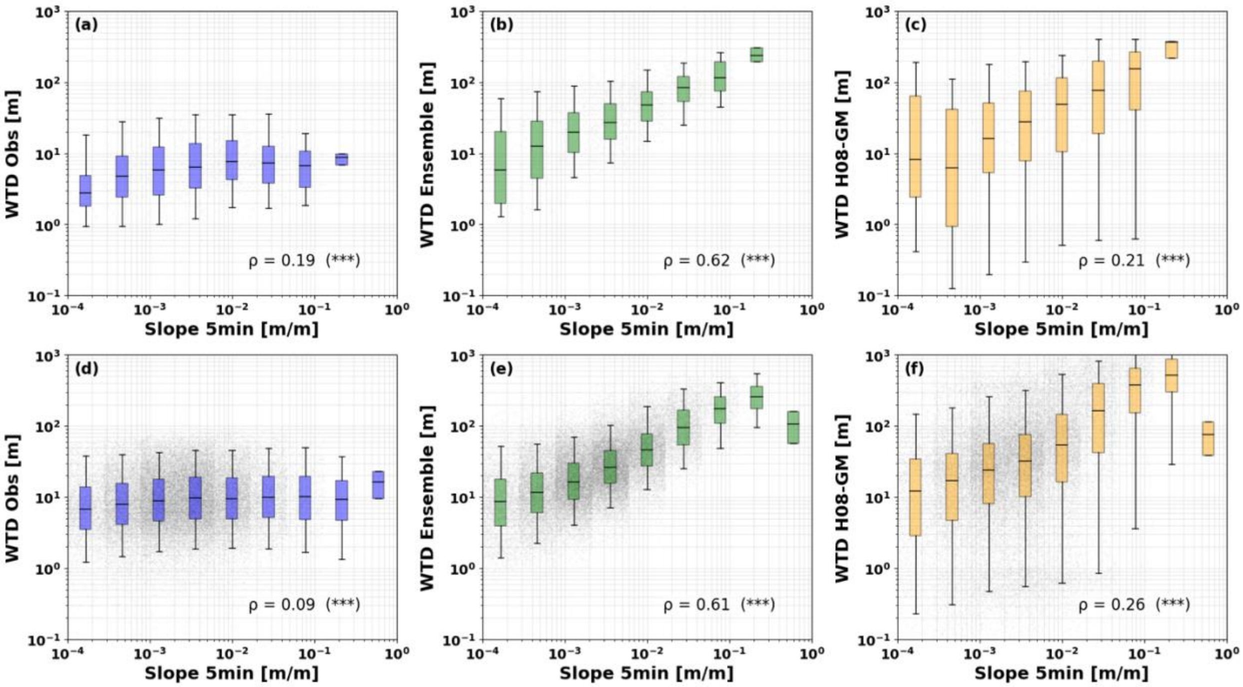
Response:

We thank the reviewer for this constructive suggestion. Following the advice, we have now added the Spearman correlation coefficients (ρ) and their significance levels (*** for $p < 0.001$) in each panel of the revised figure, as the relationships should be non-linear.

The results show that for observations (left column), the correlation between slope and WTD is generally weak ($\rho < 0.2$) under both energy-limited and water-limited conditions. In contrast, the ensemble mean exhibits much stronger correlations ($\rho \approx 0.6$), suggesting that the multi-model mean tends to emphasize a stronger slope–WTD dependence. The Spearman ρ of H08 are closer to those of the observations; however, it should be noted that the relatively low numerical correlations may partly result from the large variability of WTD within each slope bin. When looking at the medians, a stronger relationship with slope can still be observed.

This pattern therefore suggests that current global numerical groundwater models may all tend to overemphasize the “groundwater head follows topography” relationship; Or in other words, they potentially underrepresent the influence of other factors such as climate forcing and local aquifer properties.

We have added the relevant discussion in the revised manuscript (Line 464 - 472).



Overall, the revised manuscript is close to publication quality.

Response: Once again, we sincerely thank you for your valuable and constructive comments, which greatly help us improve the manuscript.