Reply on RC2

R2: The manuscript presents the most recent version of the WaterGAP model (v2.2e). Refinements, new algorithms, and updated data for calibration are presented. The manuscript is well-written, and the overall evaluation section is comprehensive and competently performed. My primary concern is the link between the extensive software modifications and data increment and the overarching scientific goal of assessing global and regional water resources. Even though the modifications, data additions, and model evaluation are comprehensively described, their scientific significance is not sufficiently justified in the context of water resources assessment. In this sense, I consider revisions (mostly related to rephrasing and further discussions) to be needed before publication. Please see below for my specific comments:

Answer: Thank you for your time to review the manuscript and for providing your comments and suggestions. We will reply to the referee comment, indicated by **R2** by our answer, indicated by **Answer**, and corresponding actions, indicated by **Action** and textual changes in *italic font*:

R2: Comment 1: The authors cite three papers to introduce the WaterGAP model. However, definitions necessary to understand the implemented modifications are missing. As a reader, I spent extra time reviewing the referenced papers to understand the main changes. Examples of these definitions include "naturalized mode" (L45), "standard runs" (L46-L47), and "local lakes algorithm" (L46), to name some just in section 2.1. Introducing the model's main features might be better so readers don't need to jump between papers (see related comment 2). As for GMD guidelines, for model description papers, it should be possible for independent scientists to build a model that, while not necessarily numerically identical, will produce equivalent results. In the current manuscript stage, the last is very hard to achieve.

Answer: Thank you. We agree that it is better readable when the introduction and the general concept is extended. So, we will extend the introduction while trying to find a good balance of providing the basis for understanding the model but avoiding repetition to the 2.2d description paper. The reader needs to read at least the comprehensive WaterGAP 2.2d description. GMD intends to publish further developments of models, so it can be expected that one paper does not describe the whole model, which is anyhow not possible in the case of complex models; and already in the 2.2d paper we had to limit the description to a certain extent. We believe that if we would repeat the lengthy description of the 2.2d paper we would not only run into duplication issues for many chapters but might receive referee comments that we should only highlight the differences of the specific model version. So, we intend to rely on the concept on describing mainly changes to the 2.2d paper.

With regard on your last point – we agree that the optimal goal of such model description papers should be to build a similar model based on these descriptions. Anyhow, this is hard to realize with such complex software that is in development since nearly three decades with it's thousands of lines of code. In a 3-year funded project, the model is currently being rewritten and re-implemented in Python (https://hydrologyfrankfurt.github.io/ReWaterGAP/) and this is already a tough task for us as WaterGAP developers due to the complexity; so it is probably unrealistic (and not practical) to aim for a reproduction of the software based on a description paper. Hence, the focus is more on describing the rationale and background of the model components.

Action: We further motivate the manuscript in the introduction and add further description; see also our reply to your Comment 2 below. To better allow the entry points for readers without pre-knowledge of the 2.2d paper, we have clarified the terms "naturalized" and "standard" by adding descriptions accordingly. More specific, we rename "standard" to "ant" (where appropriate) and "naturalized" to "nat" to be fully consistent to Müller Schmied et al, 2021. With regards to the local lakes algorithm as mentioned by the referee, we add the corresponding description section of the 2.2d paper to the end of Line 46.

In Line 18 we will add: "*Please note that this paper is not a thorough description of the whole WaterGAP model, it highlights only the modifications. For a comprehensive overview, the reader is referred to Müller Schmied et al. (2021).*

WaterGAP was developed to quantify global-scale water resources as well as water stress with focus on direct human impacts in terms of human water use and artificial reservoirs. The model framework (Fig. 1) consists of sectoral water use models that are linked in a submodel (GSWSUSE) to calculate potential net water abstractions from surface water as well from groundwater. This acts as an input for the WaterGAP Global Hydrology Model that calculates the water storages and fluxes as well as routes the streamflow to the basin outlet (Fig. 1). WaterGAP as described here operates with a spatial resolution of 0.5° x 0.5° and at daily time steps. The model can be run in a standard mode ("ant", including direct human impacts) and a variety of other variants in terms of human water use and reservoirs, e.g., a simulation of naturalized water flows and storages that would occur if there where neither human water use nor global artificial reservoirs/regulated lakes ("nat"). ".

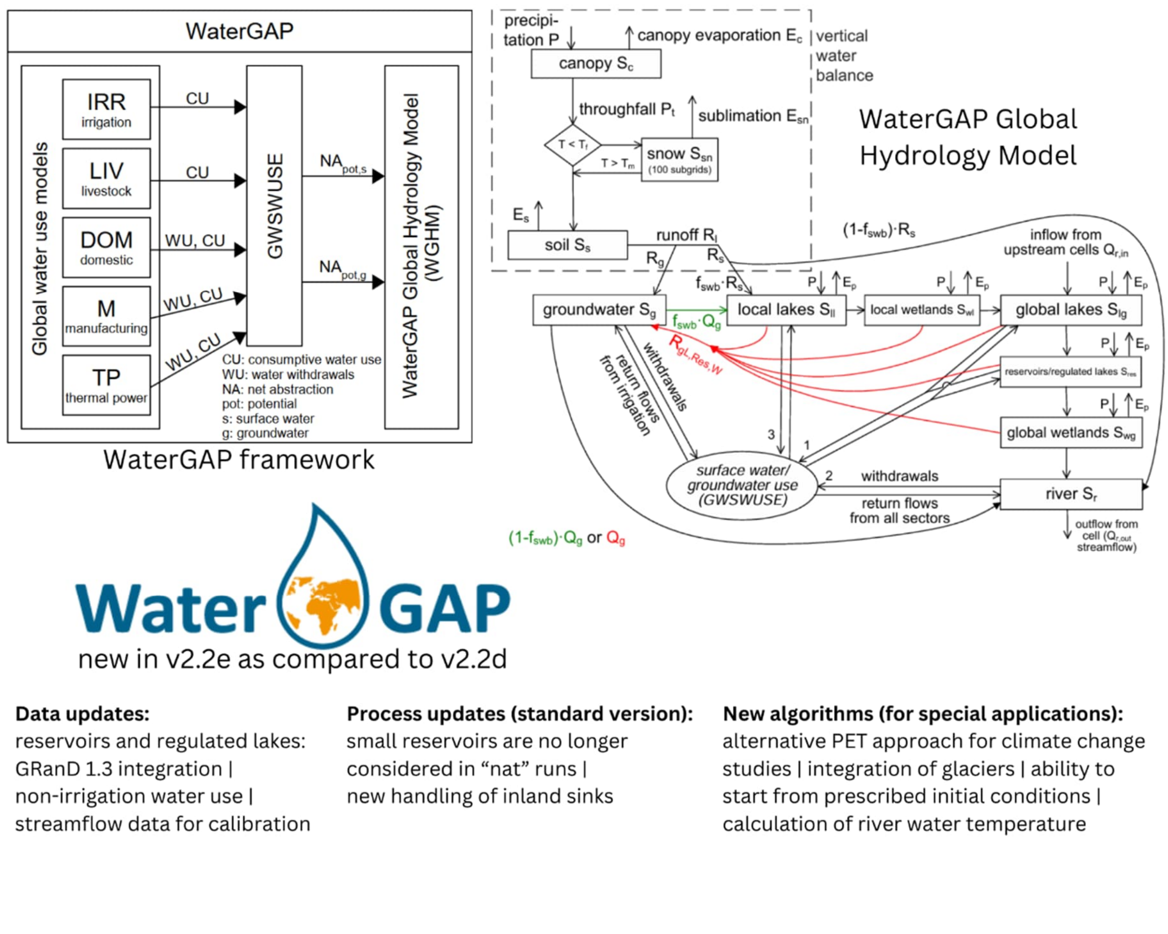

Fig. 1: Schematics of the WaterGAP framework and the WaterGAP Global Hydrology Model (both taken from Müller Schmied et al., 2021) and summary of data updates, process updates and new algorithms.

R2: Comment 2: The model modifications are widely discussed in Section 2. However, as a reader, it is hard to visualize the model as a whole and identify the modified components. A scheme presenting the model's overall structure with the modified stages highlighted might be helpful.

Answer: Thank you for this suggestion. We agree that showing the schemes is a good idea. However, the general structure has not changed as compared to the 2.2d paper, so we re-use these two figures (side-by-side) and highlight the new features in model version 2.2e in a summarized form inside this figure. We believe this also helps the reader for introduction (see our reply to your Comment 1).

Action: We add a new Fig. 1 to the introduction and describe the general concept (see above).

R2: Comment 3: Following an analysis of the simulated monthly time series of reservoir water storage to observations for 16 reservoirs in the United States, the authors decided to drop off the 85% maximum storage capacity assumption implemented in version 2.2d (Müller Schmied et al. 2021) (due to model underestimation in 11 of the analyzed cases). This implies that the decision to remove the original assumption is based on a local-based analysis where ~30% of the model results did not show underestimation issues. Thus, I am failing to see the reason for extrapolating these local results to the global implementation of the model. Furthermore, Figures S1 and S2 might indicate more pressing issues related to the model's accuracy (e.g., seasonality) that might not be improved by simply removing the assumption.

Answer: We have to apologize. The storage capacity threshold was already removed for the model version 2.2d (noted in Müller Schmied et al., 2021, Section A2.4 (last bullet point) and the inclusion in this manuscript was a result of an internal communication problem. We agree that the generic reservoir algorithm has several limitations and efforts are ongoing to improve the representation of reservoirs in large-scale hydrological models (Dong et al, 2023, Otta et al, 2023, Shrestha et al, 2024, Steyard & Condon, 2024).

Action: We have removed the update of the reservoir algorithm completely from this manuscript.

R2: Comment 4: Section 5 presents the effects of the model modifications on multiple areas and the impact of differential forcing. Overall, the effect of the implemented changes seems to be minor. However, I found it challenging to visualize the minor differences because most of the baseline results (from v2.2d) are presented in Supplementary Material. I would recommend summarizing the results as differences of v2.2e from the original implementation rather than presenting each version independently. Furthermore, this issue connects to my general comment about the lack of scientific justification for the model modifications. If model parametrizations and data changes lead to almost negligible changes, the scientific rationale for implementing them should be discussed further.

Answer: With regards to the general comment (scientific justification), please see our response to your Comment 1. Some standard model version updates were driven by ISIMIP3 requirements (e.g. inclusion of GRanD update and simulation of water temperature) and other updates serve to keep the model up-to-date (e.g. the spatio-temporal update of the calibration data basis). Overall it was rather expected that changes in model performance might not appear large if globally averaged but may be large for individual cells and basins. In particular for the standard version, not too many changes are expected. The more „severe“ modifications are not in the standard version. But we fully agree, it is hard to see differences especially from the global basin maps, and a direct comparison might help here.

Action: We created spatial maps that show differences of model performance on a basin level, for 2.2e vs. 2.2d. The motivation is to see, where the difference to the optimal value (in terms of model performance indicator) is reduced or increased between the two versions that are driven by the same climate forcing and calibration data. For all indicators we calculate this difference as the ratio of the absolute deviation from the optimal indicator value (here everywhere 1.0) as:

$$PR\_[IND] = \frac{|1.0 - IND_{2.2e}|}{|1.0 - IND_{2.2d}|}$$

where:

PR_[IND]: Performance ratio of the given indicator IND [-]

IND: indicator value (KGE and its components for streamflow; Amplitude ratio for TWSA and the ratio of Model divided by GRACE for TWSA trend) for the particular model version

The smaller the resulting PR_[IND] is, the better 2.2e is compared to 2.2d. For PR_[IND] values < 1.0, 2.2e performs better than 2.2d and vice versa. The closer PR is to 0, the better performs 2.2e against 2.2d.

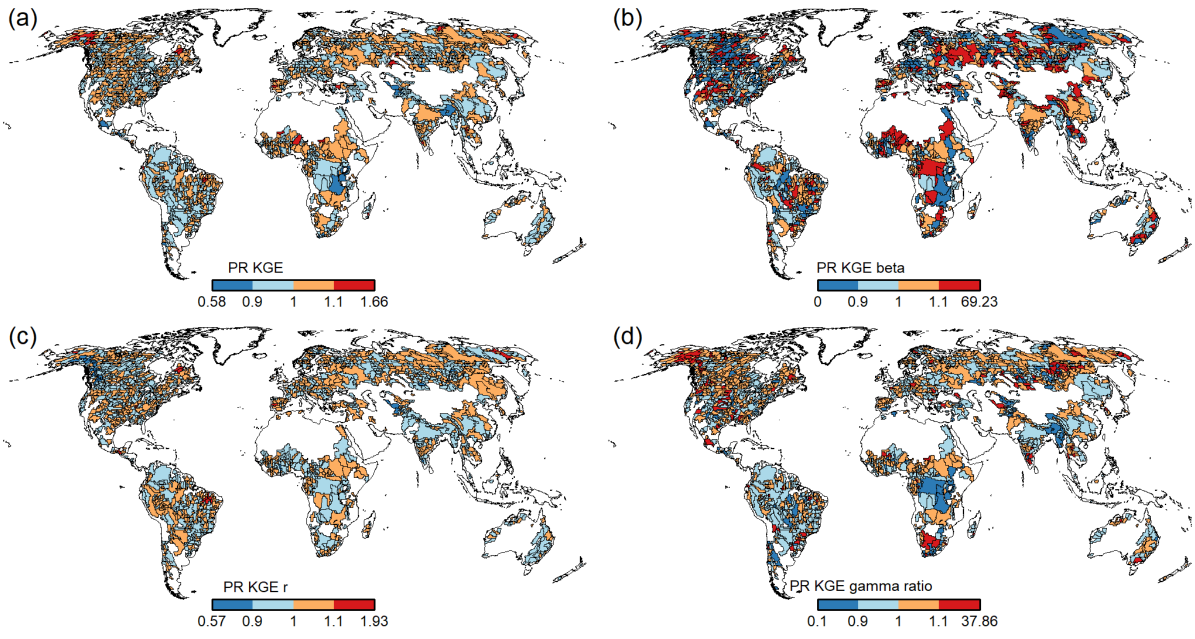

Fig. 2: Resulting PR of streamflow for the model version 2.2d and 2.2e as driven by gswp3-w5e5 for overall KGE (a), KGE beta (b), KGE correlation r (c) and KGE variability gamma (d). Bluish colours indicate that 2.2e is closer to the optimal parameter indicator value than 2.2d. Note that the calibration procedure forces KGE beta values to be close to the optimum value, hence the drastic colours are a result of only small differences to the optimum value.

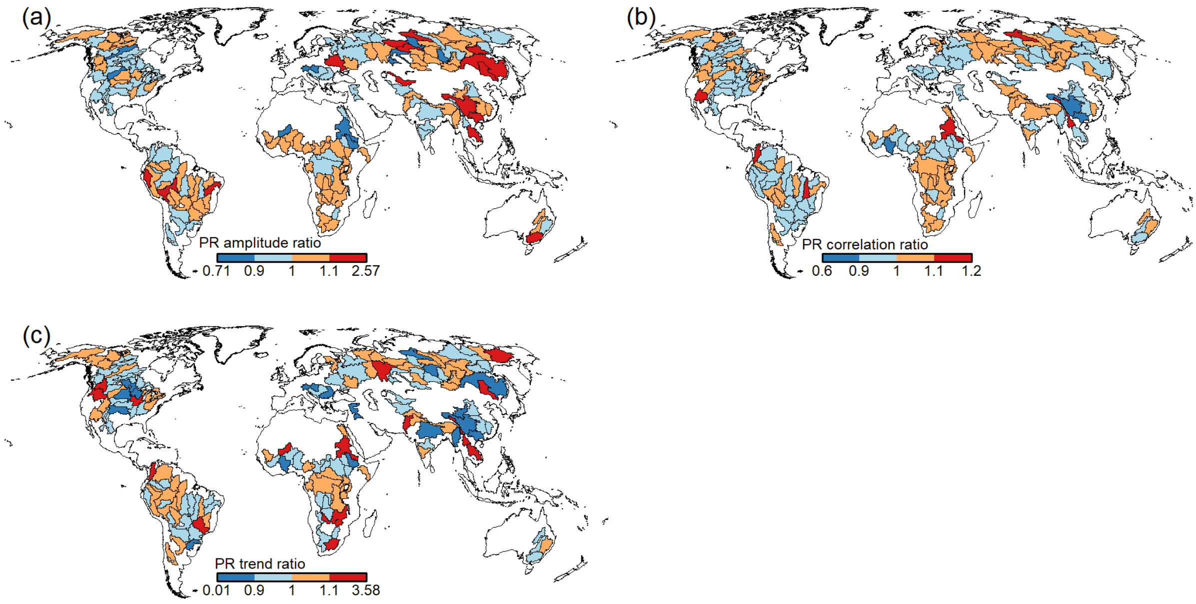

Fig. 3: Resulting PR of TWSA for the model version 2.2d and 2.2e as driven by gswp3-w5e5 for the amplitude ratio (a), correlation ratio (b) and trend ratio (c). Bluish colours indicate that 2.2e is closer to the optimal parameter indicator value than 2.2d.

Figs 2 and 3 will be included in Section 7.5.1 (WaterGAP 2.2e vs. WaterGAP 2.2d) with a brief description, the indicator description PR_[IND] added as Appendix C

References

Dong, N., Yang, M., Wei, J., Arnault, J., Laux, P., Xu, S., et al. (2023). Toward improved parameterizations of reservoir operation in ungauged basins: A synergistic framework coupling satellite remote sensing, hydrologic modeling, and conceptual operation schemes. Water Resources Research, 59, e2022WR033026. https://doi. org/10.1029/2022WR033026

Otta, K., Müller Schmied, H., Gosling, S. N., and Hanasaki, N.: Use of satellite remote sensing to validate reservoir operations in global hydrological models: a case study from the CONUS, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2023-215, in review, 2023.

Shrestha, P. K., Samaniego, L., Rakovec, O., Kumar, R., Mi, C., Rinke, K., & Thober, S. (2024). Toward improved simulations of disruptive reservoirs in global hydrological modeling. Water Resources Research, 60, e2023WR035433. https://doi.org/10.1029/ 2023WR035433

Steyaert, J. C. and Condon, L. E.: Synthesis of historical reservoir operations from 1980 to 2020 for the evaluation of reservoir representation in large-scale hydrologic models, Hydrol. Earth Syst. Sci., 28, 1071–1088, https://doi.org/10.5194/hess-28-1071-2024, 2024.