Hasan, H. M. M., Döll, P., Hosseini-Moghari, S.-M., Papa, F., and Güntner, A.: The benefits and trade-offs of multi-variable calibration of WGHM in the Ganges and Brahmaputra basins, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2023-2324, in review, 2023.

Response to Anonymous Referee #1

We thank you very much for your helpful comments and constructive suggestions for improving the manuscript. Below, each comment (in italics, indicated by "**RC**") is followed by our answer (normal font, indicated by "**AC**"). Proposed new text in the revised manuscript is written in bold.

RC: *I enjoyed reading the manuscript.*

My main concerns are; 1) the temporal only calibration of a distributed hydrologic model and 2) use of coarse meteo inputs while era5-land offers 0.1 inputs.

AC: 1) We agree that conducting a spatio-temporal calibration analysis would be the preferred approach from a conceptual point of view for a distributed model. However, such an approach would lead to a significantly expanded parameter space, which could render calibration impractical within a reasonable timeframe and acceptable accuracy range. Another major limitation is the insufficient capacity of observations to effectively constrain a large number of model parameters. As discussed in the introduction, most hydrological observations can effectively constrain only around 4 to 6 model parameters. While the hydrology modelling community is exploring methods to parameterize spatially distributed parameters for right reasons and we are also interested in conducting such a spatio-temporal calibration, our current study has its focus on basin-scale parameter calibration with more observables than usually performed.

We added a reference to the recently published paper Döll et al. (2024), a study where a multivariable calibration approach similar to the one in our study was applied for WaterGAP for both the whole Mississippi River Basin and five sub-basins. But different from our study, only streamflow and TWSA (but not evapotranspiration and surface water storage) were used as observational variables.

We inserted in line 152-159 of the introduction the following sentence:

Döll et al. (2024) also used observations of Q and TWSA to calibrate WGHM alternatively for determining pareto-optimal parameter sets for the Mississippi basin as a whole or individually for each of five sub-basins. The whole-basin approach improved the fit to sub-basin observations in all sub-basins as compared to the uncalibrated model (with the exception of one sub-basin for Q). It did not degrade the fit to TWSA for three sub-basins compared to the computationally more demanding sub-basin approach but this was only the case in one sub-basin regarding Q.

2) At present, the WaterGAP Global Hydrological Model (WGHM version 2.2e) can only operate at 0.5-degree resolution and thus with 0.5-degree meteorological inputs. We thus do

not expect significant added value of using higher-resolution forcing data that need to be aggregated to 0.5-degree resolution. However, if the model resolution will change in future, the methods employed in this study could be adapted and applied.

RC:

Other comments:

Section 3.3: More details on the SA should be provided. Morris is an elaborated SA method as compared to the one at a time local methods so that much more runs are required in Morris. How many runs were required for a 24 parameter model (Line 263).

AC: In the revised version, we provide the details of the Morris method in the supplementary materials (see section S2 of the supplementary materials). We reported that out of 24 model parameters, we excluded two parameters in the SA – EP-NM and P-PM (in Line 289-293). These two parameters directly modify model forcing, i.e., precipitation and net radiation, leading to very high changes in most target variables, which suppresses the relative influence of the other parameters. Thus, 22 parameters were considered in the sensitivity analysis (mentioned in Lines 432-433). For the 22 parameters, we needed to evaluate 23,000 samples for each basin. The number of model runs required in Morris's method is calculated as $r \times (m + 1)$, where m is the number of parameters and r is the number of elementary effects to be used. Additional details are provided in the method description in the supplementary materials. To specify the number of model runs in the sensitivity analysis, we added a statement after Lines 433-434. We also rephrased following statements according to this suggestion.

Instead of "For the sensitivity analysis, model simulations for the period 1990-2019 were used, with 1985-1989 taken as the model spin-up period and the first year of the spin-up was run 5 times to allow the water storages to fill up to an equilibrium state", we reformulated the statement as "**During the SA, a total of 23,000 samples were analysed for each of the river basins. Model simulations were conducted for the period 1990-2019, with the spin-up period from 1985 to 1989. The initial year of the spin-up was run five times to allow water storages to reach an equilibrium state**".

RC:

Can Morris identify effects of parameter interactions on the sensitivities like in Sobols' method? Why did you choose Morris instead of looking at Jacobian matrix in simple terms?

AC: The Morris method calculates the partial derivatives (i.e., elementary effects) at various points in the parameter space, similar to those in the Jacobian matrix. The sensitivity index is determined by averaging these partial derivatives. This approach provides a more accurate estimation of a parameter's effect compared to local methods like the Jacobian matrix.

Unlike the variance decomposition method of Sobol, Morris's method does not explicitly differentiate interaction terms. However, it does produce a variance term for the elementary effect that accounts for parameter interactions and the functional non-linearity of the model response. We utilized this variance term in the parameter selection process. We acknowledge the importance of providing comprehensive sensitivity analysis details and thus included them in the supplementary materials of our paper.

We added to the revised manuscript that the Morris SA is a global sensitivity analysis. According to this suggestion, we reformulated Lines 425-427 as follows: "**The sensitivity index of the EET method averages out the local influences by taking samples from many locations in the parameter space, making it a global sensitivity analysis method (Pianosi et al., 2016)**".

In addition, we added the following text to section 3.3 after Line 417:

"**While the Morris method does not explicitly show interaction terms, it produces a variance term for the elementary effect that accounts for parameter interactions and the functional non-linearity of the model response. We computed the standard error of the sensitivity index from this variance term and used it for parameter selection (Algorithm 4 in section S2 of the supplementary material)**."

RC:

L402: 5 times run of first year? I couldn't understand how? 1985-89 spin up run and one time should be enough to reach equilibrium, shouldn't be?

AC: The WaterGAP model offers two methods to achieve equilibrium in the state variables by spin-up runs: (i) repeating the simulation of the first year for multiple times, and (ii) starting the model from a sufficiently early point in time. Since there is no general guideline available for the WGHM model that specifies how many spin-up years are required to reach equilibrium states of the storage variables, we utilized both available options to reach equilibrium state.

RC:

Eq2: Why only NSE is used as performance metric? Why only temporal calibration is pursued for a distributed hydrologic model which can produce flux maps? How did you deal with unit differences from satellite AET (watt/m2) and model outputs at mm/day? The same may apply to Grace anomaly values and recharge output of the model.

AC:

We chose NSE as our performance metric because it is widely used in the field of hydrology, although there are significant concerns associated with its use. For example, NSE is sensitive to outliers, biases, and seasonality, and it uses the observed mean as the benchmark which may not be an adequate reference for most hydrologic variables (Schaefli and Gupta, 2007). Livneh and Lettenmaier (2012), however, noted that the NSE can be a useful indicator for inter-basin performance comparison since it normalizes the mean squared error (MSE) by the observed variance (σ_o^2) of each basin. While we acknowledge that we have not addressed all the limitations of NSE, we considered it sufficient for our study as our primary objective was to evaluate the benefits and trade-offs of multi-variable calibration. Nonetheless, our methodology allows for the use of alternative performance metrics. Please note that we used other commonly used indices such as RMSE and correlation for model validation (Table 9), and in the supplementary materials, we have provided the Kling-Gupta Efficiency (KGE) and its three components for the overall compromise solutions (Tables S10, S11, and S12).

The WaterGAP Global Hydrology Model (WGHM) indeed generates spatially distributed data of water fluxes. However, in our already complex study, we opted to utilize basin-scale

observations for all variables to obtain more accurate estimates of observational errors. Consequently, we were unable to leverage the flux maps and explore the potential use of spatial pattern-based metrics.

We used the LandFlux-EVAL multi-dataset synthesis ET product developed by Mueller et al. (2013), which reports ET values in units of mm/day. Similarly, for the GRACE anomaly, we incorporated basin-scale terrestrial water storage anomaly (TWSA) data processed in units of water height equivalent (mm). These GRACE TWSA data, including propagated errors, were prepared by the University of Bonn following the methodology outlined by Gerdener et al. (2020).

RC:

NSE is a bias sensitive metric and it might be necessary to use bias insensitive spatial pattern metrics in the calibration.

AC: As mentioned in the response to earlier comments, we did not use spatial pattern-based performance metrics because we employed basin-scale monthly average observations. As we also want to improve the simulated water balance for the target area at the basin scale, a bias-sensitive performance metric seems to be reasonable choice.

RC:

Introduction misses recent works on satellite based evaluation and calibration of the distributed hydrologic models using actual ET. Also, trade offs in multi objective Pareto calibration of hydrologic models have been studied in the literature. Please update your literature review with studies from 2018 to Jan 2024 from top journals (HESS and WRR). Compare your results with them in the discussions.

AC: In the revised manuscript, we updated the literature review in the introduction and included the following statements after Lines 162-164 of the current manuscript.

"Demirel et al. (2018) demonstrated successful enhancement of spatial pattern performance in a distributed hydrological model through multi-objective calibration using discharge and remote-sensing-based ET observations. Additionally, Demirel et al. (2024) provide a discussion on the trade-offs between temporal and spatial pattern calibration of the same distributed model using discharge and ET observations."

Additionally, we inserted the following text after Lines 179-182:

"Hulsman et al. (2021) utilized in-situ discharge, satellite-based evapotranspiration (ET), and GRACE Total Water Storage Anomaly (TWSA) data to calibrate a process-based distributed hydrological model in a large semi-arid basin in Africa, aiming to incrementally improve the process representation of the model."

RC:

Conclusion: Very different than conventional conclusion sections. Detailed results (numbers) should not be given here but just the conclusions drawn from the results should be provided in

bullets. It is lengthy and not easy to follow. Research Questions are repeated and probably not necessary.

The reader needs the main messages from the study and not the repetition of the results.

AC: We reformulated the entire conclusion chapter based on the suggestions in the revised manuscript, presenting the main findings clearly so that readers can quickly grasp the key messages. In the revised conclusion chapter, we avoided repeating the research questions.

References:

Campolongo, F., Saltelli, A., and Cariboni, J.: From screening to quantitative sensitivity analysis. A unified approach, Comput Phys Commun, 182, 978–988, https://doi.org/https://doi.org/10.1016/j.cpc.2010.12.039, 2011.

Demirel, M. C., Koch, J., Rakovec, O., Kumar, R., Mai, J., Müller, S., Thober, S., Samaniego, L., and Stisen, S.: Tradeoffs Between Temporal and Spatial Pattern Calibration and Their Impacts on Robustness and Transferability of Hydrologic Model Parameters to Ungauged Basins, Water Resources Research, 60, e2022WR034193, https://doi.org/https://doi.org/10.1029/2022WR034193, 2024.

Demirel, M. C., Mai, J., Mendiguren, G., Koch, J., Samaniego, L., and Stisen, S.: Combining satellite data and appropriate objective functions for improved spatial pattern performance of a distributed hydrologic model, Hydrol. Earth Syst. Sci., 22, 1299–1315, https://doi.org/10.5194/hess-22-1299-2018, 2018.

Döll, P., Hasan, H. M. M., Schulze, K., Gerdener, H., Börger, L., Shadkam, S., Ackermann, S., Hosseini-Moghari, S.-M., Müller Schmied, H., Güntner, A., Kusche, J. (2024): Leveraging multi-variable observations to reduce and quantify the output uncertainty of a global hydrological model: evaluation of three ensemble-based approaches for the Mississippi River basin. Hydrol. Earth Syst. Sci., 28, 2259–2295. https://doi.org/10.5194/hess-28-2259-2024

Gerdener, H., Engels, O., and Kusche, J.: A framework for deriving drought indicators from the Gravity Recovery and Climate Experiment (GRACE), Hydrol Earth Syst Sci, 24, 227–248, https://doi.org/10.5194/hess-24-227-2020, 2020.

Hulsman, P., Savenije, H. H. G., and Hrachowitz, M.: Learning from satellite observations: increased understanding of catchment processes through stepwise model improvement, Hydrol. Earth Syst. Sci., 25, 957–982, https://doi.org/10.5194/hess-25-957-2021, 2021.

Liu, D., Li, L., Rostami-Hodjegan, A., Bois, F. Y., and Jamei, M.: Considerations and Caveats when Applying Global Sensitivity Analysis Methods to Physiologically Based Pharmacokinetic Models, The AAPS Journal, 22, 93, https://doi.org/10.1208/s12248-020-00480-x, 2020.

Livneh, B. and Lettenmaier, D. P.: Multi-criteria parameter estimation for the Unified Land Model, Hydrol Earth Syst Sci, 16, 3029–3048, https://doi.org/10.5194/hess-16-3029-2012, 2012.

Morris, M. D.: Factorial Sampling Plans for Preliminary Computational Experiments, Technometrics, 33, 161–174, https://doi.org/10.1080/00401706.1991.10484804, 1991.

Mueller, B., Hirschi, M., Jimenez, C., Ciais, P., Dirmeyer, P. A., Dolman, A. J., Fisher, J. B., Jung, M., Ludwig, F., Maignan, F., Miralles, D. G., McCabe, M. F., Reichstein, M., Sheffield, J., Wang, K., Wood, E. F., Zhang, Y., and Seneviratne, S. I.: Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis, Hydrol. Earth Syst. Sci., 17, 3707–3720, https://doi.org/10.5194/hess-17-3707-2013, 2013.

Pianosi, F., Sarrazin, F., and Wagener, T.: A Matlab toolbox for Global Sensitivity Analysis, Environmental Modelling & Software, 70, 80–85, https://doi.org/https://doi.org/10.1016/j.envsoft.2015.04.009, 2015.

Schaefli, B. and Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21, 2075–2080, https://doi.org/10.1002/hyp.6825, 2007.

Hasan, H. M. M., Döll, P., Hosseini-Moghari, S.-M., Papa, F., and Güntner, A.: The benefits and trade-offs of multi-variable calibration of WGHM in the Ganges and Brahmaputra basins, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2023-2324, in review, 2023.

Response to Anonymous Referee #2

We thank you very much for your helpful comments and constructive suggestions for improving the manuscript. Below, each comment (in italics, indicated by "**RC**") is followed by our answer (normal font, indicated by "**AC**"). Proposed new text in the revised manuscript is written in bold.

RC: *This study presents a very thorough analysis of multi-variable calibrations considering different variables for a global hydrological model. The model was applied on two exemplary basins using in-situ and multi-satellite data. The authors did an excellent job in considering a variety of aspects that are important for modelling (e.g., required number of model runs, Pareto frontier, parameter sensitivity/importance). I also liked the selection of the authors of the data that was considered for the multi-variable calibration scenarios.*

AC: Thank you for your positive and encouraging feedback on our manuscript.

RC: *I am recommending minor revision because even though the modelling analyses seem thorough, the presentation for the reader could be improved. As a reader, it was rather difficult to extract the main aspects of this research as the text was not very concise. Additionally, a slightly unconventional structure regarding the results and discussion, and conclusion was used. I recommend shortening the manuscript or summarizing several points into one point to make it easier for future readers to follow it and to get the main points of the study. This is a very general recommendation, so I have picked out examples for you to explain what I mean.*

AC: To concisely report the main outcomes of our work, we reformulated the entire conclusion chapter. To shorten the manuscript, we removed three tables from the main text. Additionally, we made the following changes in the revised manuscript that contribute to a more concise writing and to shorten the manuscript.

We rewrote the following statements according to this suggestion:

Instead of "In total, 4.8 million samples were evaluated during the study which approximately consumed over 3.2 million CPU hours of execution time for the WGHM model to assess those samples", we wrote "**Overall, the study involved the evaluation of 4.8 million samples, requiring approximately 3.2 million CPU hours of model run time**".

We also reformulated following statements:

Instead of "We obtained a good number of non-dominated solutions, i.e. Pareto-optimal parameter sets, in most of the multi-objective calibrations (Table 5). The cardinality (number of solutions) of the non-dominated solution set of a multi-objective calibration depends mainly on the shape of the Pareto frontier (PF) and the crowding distance of the members. The crowding distance is controlled in the Borg algorithm by the epsilon parameters which was 0.005 for all objectives. The greater solution cardinality in the Ganges basin experiments, when compared to those in the Brahmaputra basin, already indicates heightened trade-offs among the

objectives, especially between NSE_Q and NSE_{SWSA}, as well as between NSE_{SWSA} and NSE_{TWSA}",

we wrote "**A high cardinality, i.e., a high number of solutions in the non-dominated Pareto solution set, was obtained in most multi-objective calibrations (Table 4). The cardinality depends on the shape of the Pareto frontier (PF) and the allowed crowding distance, which was constant (0.005) for all objectives in all experiments. A wider PF resulting in high cardinality reflects a high trade-off between the objectives. The high cardinality observed in the Ganges experiments indicates marked trade-offs among objectives, especially between NSE_Q and NSE_{SWSA}, as well as between NSE_{SWSA} and NSE_{TWSA}.**"

In the revised manuscript, we deleted the following statements: "This is a common occurrence in multi-objective optimization scenarios (Meyer Oliveira et al., 2021; Livneh and Lettenmaier, 2012). However, this comes at the cost of performance loss for the other variables that were not considered for calibration. The standard calibration of WGHM for mean annual streamflow (Müller Schmied et al., 2021) resulted in poorer results for all performance criteria in both the Ganges and Brahmaputra basins than the uncalibrated model for both basins. The mean NSE of all four objectives ($\mu_{NSE,ALL}$) was used as a simple indicator of the overall performance of an experiment."

We reformulated following statements:

Instead of "Different from the Brahmaputra, calibration against only Q in the Ganges basin (both the calibration method presented here and the standard WGHM method) resulted in worse fits to all three other variables as compared to the uncalibrated model version. Multi-variable calibration, however, works best if streamflow observations are included because the average fit to all observations is, in the case of both 2-objective and 3-objective calibration cases, highest if NSE_Q is one of the calibration targets (**Error! Reference source not found.** and **Error! Reference source not found.**)",

we wrote "**Different from the Brahmaputra, calibration against only Q in the Ganges basin resulted in worse fits to all three other variables as compared to the uncalibrated model version. Multi-variable calibration, however, works best if streamflow observations are included. Excluding NSE_Q as an objective in any calibration resulted in significantly poorer performance in streamflow simulation (Table 5 and Table 6)**".

We deleted following statement from the manuscript: "However, in the majority of the calibration cases, the performance in streamflow simulation was very poor when the model was not constrained by streamflow observations."

We also deleted following statements: "In contrast to the apparent trade-offs among objectives, there could be other non-traditional interactions among the objectives. For instance, in all replications of the calibration with only NSE_Q in the Ganges basin, we observed negative NSE_{TWSA} (not shown). But using only NSE_{TWSA} as the calibration objective, we consistently observed very high values in NSE_Q for all replications. Likewise, when NSE_{ET} is used as the only calibration objective, NSE_Q exhibited a significant decrease across replications in the two basins. However, when NSE_Q is employed as the only objective, reasonable performance in ET simulation is observed. Hence, the nature of the association between a pair of objectives, when attempting to describe the trade-offs, is neither unidirectional nor easily traceable through correlation analysis. Furthermore, there could be three-way trade-offs and so forth in a high-

dimensional objective space, making them challenging to detect. While the association and causality of such relationships are indeed intriguing, examining the nature of trade-offs among the objectives is beyond the scope of the current study."

We also deleted following statement: "This indicates the high reliability of our findings regarding the trade-offs among objectives discussed in the earlier paragraphs".

RC:

● *Make your sentences more concise: Lines 554-566 could for example be shortened into something like this: "Several parameters influence most or all response variables across various signatures. However, certain parameters affect only one or two signatures of the response variables. For instance, the Runoff Coefficient (SL-RC) significantly influences monthly means (MM) of ET in the Ganges basin and MTS of streamflow. Similarly, the snow melt temperature (SN-MT) is important for some cases in snow-dominated catchments in the Brahmaputra basin. These parameters may also affect other response variables and signatures to some extent but do not meet the defined threshold for calibration selection (Figure 4)."*

AC: According to the suggestion, we rephrased the above statement in the revised manuscript:

"Several parameters influence most or all response variables across various signatures. However, certain parameters affect only one or two signatures of the response variables. For instance, the Runoff Coefficient (SL-RC) – which is one of the parameters considered in the standard WGHM calibration – significantly influences monthly means (MM) of ET in the Ganges basin and MTS of streamflow. Similarly, the snow melt temperature (SN-MT) is important for some cases in snow-dominated catchments in the Brahmaputra basin. These parameters may also affect other response variables and signatures to some extent but do not meet the defined threshold for calibration selection (Figure 4)."

RC:

● *You created ten (lovely) figures and twelve tables. This is nice in the sense of replicability. However, in my opinion, this is too much to present in the main text. Please consider moving some of the tables that are not essential for the main outcomes of this study to the supplementary. Table 4 could be deleted entirely as it does not contain additional information to Figure 4.*

AC: In the revised manuscript, we moved those tables to the supplementary materials and we deleted Table 4 of the old manuscript.

RC:

● *Regarding the structure of the paper: For me the content of the conclusion chapter would be (the main) part of the discussion. Overall, the discussion part had become a bit short by being combined with the results. I recommend renaming the current conclusion chapter and writing a more common conclusion chapter. This will help the reader a lot to understand what you did. Also shorten the current conclusion chapter and do not present the results again.*

AC: To better communicate our main findings, we replaced the entire conclusion chapter in the revised manuscript.

RC: *Minor comments or examples:*

AC: We greatly value all the insightful comments. We were committed to incorporating all the suggestions which contributed to greatly improve the revised manuscript.

RC:

1 Introduction:

● *You switch between the terms multi-variable, multi-signature, and multi-objective throughout the manuscript. Please clearly define them in the introduction and use the terms consistently throughout the manuscript. E.g., line 595: "multi-objective" and in the title of the manuscript: "multi-variable". I assume the same is meant in both cases.*

AC: We used the terms 'multi-variable', 'multi-signature', and 'multi-objective' in their literal meanings in the manuscript. We utilized 'multi-signature' specifically in sensitivity analysis (SA), where the effects of parameters on multiple aspects of each variable were explored. The term 'multi-objective' was employed in the context of calibration exercises when more than one objective was utilized in the calibration process. Conversely, the term 'multi-variable' was applicable both in SA and calibration, as these analyses involved multiple variables. While 'multi-objective calibration' and 'multi-variable calibration' are not always synonymous—given that multiple objectives can be associated with a single variable, and multiple variables may contribute to a single composite objective—in our study, we used them interchangeably because each objective corresponds to a separate variable. To clarify the meanings of these terms, we included the following paragraph in the revised manuscript in Line 190-197.

"**The terms 'multi-objective' and 'multi-variable' are not always interchangeable, as multiple objectives can stem from the same variable and multiple variables can contribute to a single composite objective. We use these terms contextually based on their literal meanings. Our multi-objective calibration analyses involve multiple objectives and multiple variables, with one objective corresponding to each variable. A 'signature' of a data series consists of quantitative metrics or indices that describe its statistical or dynamic properties (McMillan, 2021). In this context, the term 'multi-signature' refers to a scenario where multiple quantitative properties of a data series are considered simultaneously**."

In addition, for clarity, we reformulated lines 173-179 of the old manuscript as follows:

"In this study, we present a comprehensive multi-objective calibration framework for estimating optimal basin-specific parameter values for a global hydrological model by taking into account observations of multiple model output variables. The framework consists of 1) an approach for selecting model parameters that is based on a global sensitivity analysis and considers multiple signatures of each variable and 2) a multi-objective parameter optimization that includes multiple variables. We apply the framework to WGHM and estimate, for the Ganges and the Brahmaputra basins of the Indian subcontinent, the most important model parameters using multi-variable multi-signature sensitivity analysis and multi-variable parameter optimization."

RC:

● *Line 34: "T" not explained anywhere.*

AC: The statement was corrected by replacing "T" with "TWSA" in the revised manuscript.

RC:

● *Lines 53-53: local or regional hydrological models*

AC: By the statement "Even more than local to regional hydrological models, GHMs suffer from high predictive uncertainties ...", we intend to compare all models that lie between local and regional scales to the global scale models. If, however, the expression is not clear in the statement, we may change it to "local or regional hydrological models".

RC:

● *Lines 62-64: Abbreviations are placed inconsistently. Maybe do: For example, the Water - Global Assessment and Prognosis (WaterGAP) Global Hyrdological Model (WGHM, Müller...)..*

AC:

To avoid inconsistencies in the abbreviation and full name of WGHM, we used "the WaterGAP Global Hydrological Model (WGHM)" as used in the reference model description paper for WGHM by Müller Schmied et al. (2021) and deleted the full form of WaterGAP according to this suggestion.

RC:

● *Lines 70-75: Add a reference*

AC: In the revised manuscript, to support those statements we referred to the study of Cheng et al. (2005) in which they mentioned that "...with more parameters, it takes longer time to accomplish the optimization procedure. This may result in premature termination of the optimization process which will adversely affect the quality of the results."

RC:

● *Lines 88-91: Sentence is a bit difficult to follow. Please rephrase.*

AC: The statement was reformulated in the revised manuscript as follows:

"The equifinality thesis proposed by Beven (1993) challenges the notion of a singular optimal model – whether in terms of structure, input, or parameters – particularly in the presence of multifaceted uncertainties. Instead, it suggests that there can be alternative models that exhibit comparable predictive capabilities while differing in their specific configurations."

RC:

● *Lines: 138-144: Maybe mention earlier!?*

AC: We moved the above mentioned section up and merge them with the paragraph that starts at Line 89. The paragraph reads as follows: "The equifinality thesis proposed by Beven (1993) challenges …… specific configurations. In the context of multi-objective calibration, Efstratiadis and Koutsoyiannis …."

RC:

● *Line 175: First time using the term "signature". Mention and explain it before.*

AC: In the revised manuscript, we added the definition of 'signature' in the introduction chapter.

RC:

● *Lines 191-193: Delete.*

AC: We deleted those lines in the revised manuscript.

RC:

2 Study area:

● *Table 1: Over which period are the means calculated (e.g., mean summer temperatures)?*

AC: We included a note in Table 1 indicating that the temperature means were estimated using data from 1969 to 2004.

RC:

● *How did you decide on the two basins? What are the differences between the two basins? What was the reasoning for not choosing two very different basins (regarding climate, geology, water abstractions etc.) to see the influence of these characteristics on the modelling scenarios?*

AC: The Ganges and Brahmaputra basins were selected for this study due to their significant geopolitical importance. Both basins are transboundary and characterized by very high population densities and substantial water demands. They are situated in a critical region where climate change poses a serious threat to water availability, with potentially severe impacts on human lives. Despite these challenges, these basins also exhibit numerous distinct features of scientific interest. The hydrological processes governing these basins differ substantially. The Brahmaputra is dominated by snowmelt, whereas the Ganges basin encompasses a wide range of climatic zones from arid to semi-arid to humid. Agricultural water use exerts the most significant influence on human-nature interactions in the basin. Although the impacts of various geomorphological and physiographic characteristics are intriguing, our study's limited scope prevents us from exploring these interactions further.

RC: *Highlight these differences or similarities between the basins also in the interpretation and comparison of the modeling results of the two basins. Why were different parameters*

selected between the different basins (Figure 4)? E.g., lines 693-696: Why do you think that is the case? Is there any explanation for that?

AC: We observed different sensitivities in the two study basins due to their distinct dominant hydrological processes. For instance, the Brahmaputra basin, being snow-dominated in its upstream parts, exhibits sensitivity of most response variables to snow parameters, with all four snow parameters were selected as important parameters for the basin. In contrast, snow parameters are not significant for the Ganges basin, as only a small fraction of the basin is affected by snow processes.

Regarding the differences in the Pareto fronts between the basins, we were unable to provide a definitive explanation. Apart from differences in dominant hydrological processes, variations in the error structure of observations could also contribute to the differences observed in the shape of the Pareto fronts. We believe that an in-depth investigation is necessary to elucidate the potential causes underlying these interactions.

RC:

3 Data and methods

● *Why not present available data in the chapter study area?*

AC: We intended to present the available data in a separate chapter due to the length of the text created by detailing the sources, processing, and the descriptions of the error information for each observation series. This approach enhanced the readability of both 'study area' and 'data and methods' chapters. Also, because of the importance of the different observables used for model calibration in this study, we though that a separate chapter on the data was justified. Moreover, as we partly use observables that were available at the global scale and not only at the basin scale, we had not considered subsuming the data description under the study site description.

RC:

● *Line 343: Title of chapter 3.2.5 Water balance closure is a bit confusing as it's a subchapter of 3.3 observations. Water balance closure is not an observation. Maybe call it storage change (which is also not exactly an observation, but sill might fit better)?*

AC:

The subchapter '3.2.5 Water balance closure' discusses an inherent imbalance in the water balance components of the observational data, pointing to lower quality observations and the presence of inconsistencies that could impact the calibration process. To enhance clarity, we renamed this subsection from '3.2.5 Water balance closure' to '**3.2.5 Water balance closure of observations**'. We need to stress that water balance closure as discussed in the section is not the storage change as derived from Precipitation-ET-runoff, but the imbalance of the components – Precipitation, ET, and storage change, all individually derived from observations.

RC:

● *Lines 436-448: Move them after line 465.*

AC: In the revised manuscript, we moved those lines to the suggested place.

RC:

● *Line 473: One comma too many*

AC: The comma was deleted in the revised manuscript.

RC:

4 Results and discussion

● *Lines 554-566: Explanation of parameters could also be in the method section when parameters are being presented or is this meant as a discussion?*

AC: The explanations of the parameters are given in Table-2 within the 'data and methods' chapter; however, the detail descriptions of these parameters and their physical (and/or hypothetical) meanings are not presented in the text. The readers are directed to Müller Schmied et al. (2021) for further explanation of those parameters.

With those statements of Lines 554-566 of the old manuscript, we intented to discuss some of the results of our sensitivity analysis.

RC:

● *Line 583: Maybe add a short sentence why P-PM was added later or refer to the method section (lines 263-268). Why is EP-NM not added?*

AC: We added the following statement in the revised manuscript to explain why P-PM was selected for calibration:

"**Nevertheless, P-PM was selected as an additional calibration parameter because precipitation forcing data, in contrast to radiation data, contain high uncertainties and biases which need to be corrected during model calibration, if possible. Recently, Goteti & Famiglietti (2024) pointed out the underestimation of precipitation in data sets of India that need to be corrected (here by P-PM) to avoid non-physical or process-based compensation by calibration of other parameters**."

RC:

● *Line 648: Livneh and Lettenmaier (2012)*

AC: We corrected the statement. The statement now reads "In their study, Livneh and Lettenmaier, (2012) concluded that …".

RC:

● *Table 6 and 7: Are those the NSE values of the calibration? I am not sure if I got that correctly, but due to data scarcity you could not calculate the NSE for all variables for the validation period (only for Q and TWSA). Is that correct?*

AC: The tables report the mean and standard deviation of the performance metric NSE of 8 compromise solutions of each calibration experiment. We reran the WGHM model updating the parameters of the compromise solutions and computed the NSEs of all target variables to obtain the performance of those solutions across all variables. The NSE values remained unchanged for the variables that had been used as objectives in the calibration experiments. For enhance clarity, we renamed the title of the table as follows:

"Table 4: Mean and standard deviation of model performance indicator NSE for the compromise solutions (N = 8) of the calibration experiments in the Ganges river basin **during calibration periods**. **The WGHM model was rerun using parameters from the compromise solutions to compute NSEs of all variables.** The $\mu_{NSE, ALL}$ represents the mean NSE across all objectives over all eight compromise solutions per experiment. The highest NSE for each objective is highlighted using bold face, also the highlighted mean across objectives ($\mu_{NSE, ALL}$) show the highest value in each group (2-objective, 3-objective, and 4-objective). The objective obtained in the standard calibration and in the uncalibrated model is also shown."

The title of Table 5 remains unchanged in the manuscript.

The NSE values in Tables 4 and 5 were calculated based on the monthly values during the calibration period. As you correctly pointed out, during validation, we were only able to compute the efficiency score for Q and TWSA, as no data were available for the other variables during the validation periods (2010-variable years).

RC:

● *Figure 6 and Figure 7 are a bit small.*

AC: We resized the components of those figures to make the figures more readable in the revised manuscript.

RC:

● *The authors chose to have a combined results and discussion chapter. Sometimes an explanation as to why the results turned out the way they did was missing.*

AC: We provided explanations of why some of the results turned out as they did in the revised manuscript in appropriate places.

RC:

● *Lines 864-865: From the following text of that paragraph, I still did not understand why the Ganges and Brahmaputra basins had different identifiability regarding their parameter comparison. Could you please explain that more clearly?*

AC: The parameter identifiability has a typical inverse association with the dimensionality of the search space. In the Ganges experiments, relatively higher parameter identifiability was observed due to the use of a smaller number of parameters in calibration. We investigated the relationship between parameter identifiability and the sensitivity of the response variables. Additionally, we examined evidence regarding the impact of multi-variable calibration on increasing parameter identifiability. To clarify our statement, we reformulated the statement as follows:

"Due to the fewer parameters involved in the Ganges calibration experiments, better parameter identifiability is observed within the basin compared to experiments in the Brahmaputra basin. We investigated how individual observations influence parameter identifiability during calibration and explored the impact of sensitivity on parameter identifiability."

Related to this issue, we also reformulated the following statements. Instead of "In the Brahmaputra basin, **the identifiability of parameters tends to be lower than in the Ganges basin**. Four parameters (P-PM, SN-MT, SN-TG, and SL-RC) are constrained well (i.e., they have low coverage of their a-priori range in the compromise solution sets) with the variable Q, two parameters (SN-MT and SL-MSM) by the ET variable, and two (SN-TG and SW-RRM) by the SWSA observations", we wrote "In the Brahmaputra basin, four parameters (P-PM, SN-MT, SN-TG, and SL-RC) are constrained well (i.e., they have low coverage of their a-priori range in the compromise solution sets) with the variable Q, two parameters (SN-MT and SL-MSM) by the ET variable, and two (SN-TG and SW-RRM) by the SWSA observations".

RC:

● *Could you add an outlook at the end of this section considering the following points: What do you expect for other basins? Could this method be applied to other basins? What would be the challenges?*

AC: We revised our conclusion chapter and reformulated the entire conclusion chapter. In the revised conclusion chapter, we included a few statements that emphasize the methodologies employed in this study can be effectively applied to other basins and other global hydrological models.

RC:

5 Conclusion

● *Lines 1031-1032: repetitive*

AC: In the revised conclusion chapter, we tried to omit all repetitive statements.

References:

Cheng, C. T., Wu, X. Y., & Chau, K. W. (2005). Multiple criteria rainfall–runoff model calibration using a parallel genetic algorithm in a cluster of computers / Calage multi-critères en modélisation pluie–débit par un algorithme génétique parallèle mis en œuvre par une grappe

d’ordinateurs. Hydrological Sciences Journal, 50(6), 1087. https://doi.org/10.1623/hysj.2005.50.6.1069

Goteti, G. and Famiglietti, J.: Extent of gross underestimation of precipitation in India, Hydrol. Earth Syst. Sci. Discuss. [preprint], https://doi.org/10.5194/hess-2024-18, in review, 2024.

McMillan, H. K.: A review of hydrologic signatures and their applications, WIREs Water, 8, e1499, https://doi.org/10.1002/wat2.1499, 2021.

Müller Schmied, H., Cáceres, D., Eisner, S., Flörke, M., Herbert, C., Niemann, C., Peiris, T. A., Popat, E., Portmann, F. T., Reinecke, R., Schumacher, M., Shadkam, S., Telteu, C.-E., Trautmann, T., and Döll, P.: The global water resources and use model WaterGAP v2.2d: model description and evaluation, Geosci Model Dev, 14, 1037–1079, https://doi.org/10.5194/gmd-14-1037-2021, 2021.