We thank both reviewers and the editor very much for their helpful comments and constructive suggestions for improving the manuscript. Below, each editor's and reviewer's comment is followed by our answer (indicated by "AC"). The new text in the revised manuscript is written in bold.

Editor Ryan Teuling

E: First of all, I would like to apologize for the time it has taken to receive sufficient reviewers for this submission, and to reach an initial decision. As you have seen, the two anonymous referees, whom I both consider experts in the field of regional-scale hydrological modeling and uncertainty analysis, both see your work as interesting and potentially suitable for HESS. However they also identified a considerable number of issues, several of which were shared between the reports. Given the nature of these issues, it seems that a considerable amount of rewriting, as well as some new analysis, might be needed, and this translates into major revisions. When preparing a revised version, you can generally follow the approaches outlined in your replies posted in the online discussion. I want to ask you to pay specific attention to the scientific contribution and wider relevance of your results. HESS generally does not publish case studies on a particular basin or using a particular model. Hence, your findings for the Mississippi using WaterGAP should be used to provide a more generic insight.

AC: The intention of the study is not to provide new knowledge about the Mississippi River basin or the WaterGAP model but to test the three ensemble-based calibration and uncertainty approaches with the aim to find out how they can be applied globally, also by other global hydrological models, for utilizing observations of multiple model output variables to reduce and quantify the model output uncertainty. We agree that in the original manuscript we did not formulate clearly the relevance of our study results for global hydrological modeling in general group. We are thankful to you and the reviewers for making us aware of this. To better show the scientific contribution and the wider relevance of our results for other (global) hydrological modelers, we completely rewrote the abstract, Section 5.1 and the conclusions (also including a figure that summarizes our recommendations but also caveats, see our answer to you third comment below. We also revised the introduction thoroughly. For example, we reformulated the objective of the study as follows:

The objective of this paper is to analyze how the uncertainty of the output of GHMs can be reduced and quantified by parameter estimation that utilizes observations of multiple output variables and their uncertainties. For the example of the Mississippi River basin (MRB), the paper shows how Q and TWSA observations can be utilized to obtain one optimal parameters set (the "compromise solution") as well as ensembles of Pareto-optimal and behavioral parameter sets for the GHM WaterGAP, by evaluating the applicability of the three multi-variable calibration approaches POC, GLUE and EnCDA. It presents a method for defining performance thresholds for behavioral parameter sets based on observations and their uncertainties as well as the initial GLUE ensemble. In each approach, model parameters of all grid cells within so-called calibration-data assimilation (CDA) units, either the whole MRB or five sub-basin CDAs, were uniformly adjusted. We derive conclusions for multi-variable parameter estimation and quantification of model output uncertainty in global-scale hydrological modeling, answering the following research questions:

E: Also, I do not agree with your reply on the issue of using NSE vs KGE. These are not independent metrics, and the mathematics behind the dependency are discussed extensively in

Gupta et al. (2009). Following your reasoning, I could see the value of optimizing against one of the components of KGE, and evaluating on the other(s), but given the nature of the components i do not expect expect them to contain much mutual information on hydrologic behavior.

AC: We changed the reply on the issue NSE vs KGE, which now reads:

"NSE is used for parameter estimation only in the case of POC and GLUE, but not in the case of EnCDA (where the objective function cannot be freely set, see Table A1, but is a weighted RMSE), and we evaluated the results using both NSE and the three components of KGE (not KGE itself). We think it is particularly informative to evaluate the result of a parameter set optimization using a certain objective function/performance metric by another performance metric, in particular the three KGE components as these are well interpretable (correlation: temporal shifts; bias: difference in means; variability ratio: difference in temporal variability). In the revised version, we have added a sentence to section 4.1 to indicate how the improvement in the correlation coefficient by the calibration can be interpreted: **Thus, calibration mainly leads to improved timing of monthly streamflow and TWSA.**"

E: In addition, I would like to ask you to consider optimizing the display items (figures and tables). There is a lot of information being presented, but key figures that condense some of the findings in clear concepts are missing. This makes the manuscript currently read perhaps more like a scientific report than a focused manuscript. I believe condensing some of the results, and using the space for 1-2 new figures summarizing the main results, could help in addressing several of the referees' comments.

AC: Table 1 was removed from the main text and the text in sections 1, 2, 4 and 5 was slightly condensed. We added a figure (Figure 7) to the new Section 6 Conclusions that summarizes the main results of the study.

E: I am looking forward to receiving a revised version of your work. Given the nature of the comments, this new version will be returned to the referees for consultation

Reviewer 1

RC: The study by Doell et al. compares three different strategies to reduce parameter uncertainty for the global hydrological model WaterGap. The methods used are BORG, GLUE, and an ensemble Kalman filter, which the authors apply in a pilot study to the Mississippi basin. How we best estimate global water models is an interesting and relevant question to which the authors contribute. I do like the study and what the authors do and show, but I have some critical comments regarding how the work is currently presented and discussed. I outline my main comments below.

AC: Thank you for the positive feedback.

RC: [0] The authors' use of sensitivity analysis is very nice and interesting, but the results are hardly discussed. I would have liked to see more detail on these results. For example, the precipitation multiplier is not slected as important. Interesting, given that this parameter is often very relevant. Is rthis due to the monthly time step? The authors study a huge domain. How did sensitivity to the parameters vary across this domain? A lot of insights to be gained from this analysis, but they are not discussed. I think this would be worth including rather than some other parts as suggested below.

AC: In the revised version, we added a paragraph to section 3.2.4 in which we discuss the new table below, which was added to the supplement as Table S2. In this and the following paragraph, we show in more detail which output variables are sensitive to which parameters and how this differs among the CDA units.

Table S2. The most influential parameters for streamflow, TWSA, snow cover and local lake storage, covering together at least 50% of the total effect.

CDA Unit	Streamflow	TWSA	Snow cover	Local lake storage
I Arkansas	**SL-RC, SL-MSM, EP-PTh, SL-MEP, GW-MM**	**SL-RC, SL-MSM, NA-GM**	**SN-MT**	**SW-LD, SW-DC**
II Missouri	**SL-RC, SL-MSM, EP-PTh, SN-MT, NA-SM**	**SL-RC, SL-MSM, SW-WD, EP-PTh, NA-GM**	**SN-MT**	**SW-LD, SW-DC, NA-SM**
III Upper MRB	**SL-RC, SL-MSM, EP-PTh, SN-MT, GW-MM**	**SL-RC, SL-MSM, SW-WD, SW-DC, EP-PTh**	**SN-MT**	**SW-LD, SW-DC**
IV Ohio	**SL-RC, SL-MSM, SW-RRM, EP-PTh, GW-MM**	**SL-RC, SL-MSM, EP-PTh, GW-DC**	**SN-MT**	**SW-LD, SW-DC**
V Lower MRB	**SL-RC, SL-MSM, SW-RRM, EP-PTh, SN-MT**	**SL-MSM, GW-RFM, NA-GM**	**SN-MT**	**SW-LD, SW-DC**
MRB	**SL-RC, SL-MSM, SW-RRM, EP-PTh**	**SL-RC, SL-MSM, EP-PTh, NA-GM**	**SN-MT**	**SW-LD, SW-DC**

Note that although SW-WD was not selected in unit I, IV, V, MRB, we decided to select the parameter for all units due to effect on groundwater recharge from surface water bodies

Regarding the precipitation multiplier P-PM, we write in line 583 “P-PM was excluded from calibration even though it ranked 1st in the sensitivity analyses in all six basins for almost all four test variables because the precipitation input is perturbed in EnCDA, and an additional multiplier would lead to a double-counting of precipitation uncertainty.” So one reason for not including P-PM in POC and GLUE was that we wanted to compare all three calibration methods. The other reason was that different from other basins such as the Amazon or the Ganges-Brahmaputra basins, precipitation in the Mississippi River Basin is expected to be rather well represented by the climate data used as input to WaterGAP. The mean annual precipitation in the CDA units that was used to drive WaterGAP does not differ much from the values derived from the high-resolution (4 km) PRISM dataset for the USA. In the revised version, we extended the explanation for why we did not use P-PM as calibration parameters and referred to the new Table S1 below that was added to the supplement. The new text in section 3.2.4 reads:

P-PM was excluded from calibration even though it ranked first in all six CDA units for almost all four test variables, for various reasons. First, the precipitation input is perturbed in EnCDA, and an additional multiplier would lead to a double-counting of precipitation uncertainty. Second, mean annual precipitation in the CDA units of WaterGAP climate forcing does not differ much from the values derived from the high-resolution (4 km) PRISM dataset for the USA (Table S1).

Table S1. Comparison of mean annual precipitation in the CDA units for the calibration period 2003-2012 between GPCC-WFDEI used to drive WaterGAP and the high-resolution (4 km) PRISM* dataset for the USA [mm/yr]

CDA unit	GPCC-WFDEI	PRISM	Ratio PRISM/GPCC-WFDEI (potential P-PM)
I Arkansas	705	667	0.95
II Missouri	595	622	1.04
III Upper MRB	951	878	0.92
IV Ohio	1313	1242	0.95
V Lower MRB	1286	1254	0.97
MRB	839	829	0.99

***https://climatedataguide.ucar.edu/climate-data/prism-high-resolution-spatial-climate-data-united-states-maxmin-temp-dewpoint**

RC: [1] This is a very long paper with a lot of details on the model and the data that, at least to me as a reader, seems excessive and not needed to understand the main story presented. It makes reading the paper a bit tedious because most readers will not run WaterGap and they might not even be interested in the extensive background information on the data (as part of the main story).

For example, lines 500-508 discuss problems with the GRACE data and how others have gone about reducing them. Is this really something I need to know to follow the story? I think text like this can go into the supplemental material without reducing the strength of the story told. On the contrary, it would make it better because I do not have to read through this background information unless I want to.

Lines 466-508 discuss details of the GRACE data and their uncertainties in (excessive) detail. At the same time, the authors spent one sentence on stating that two studies considered streamflow errors of about 10%, while the next sentence states that this is maybe a possible average but the variability is very large. The authors spend over 60 lines discussing GRACE and 6 (ok 7) lines to discuss the other variable they use. I do not understand why the authors do not present a more balanced discussion given that both variables suffer from significant and potentially complex uncertainties.

AC: Regarding the description of the WaterGAP model in section 3.1, we constrained the information to the information that is necessary to understand 1) the meaning and importance of parameters that are to be estimated by the multi-variable calibration and 2) the differences between the multi-variable calibration presented in the manuscript and the (very simple) standard calibration of WaterGAP. Thus, we think that it is not beneficial to shorten the model description or move it to the supplement.

Regarding the description of the GRACE TWSA data, we agree with the reviewer that there is excessive detail in the main text. We moved the text on leakage problems and other aspects related to the uncertainty of GRACE TWSA data (lines 483 to 508 in version 1 of the manuscript) to the supplement as Text S1. To increase the readability by decreasing the length of the main text, we also moved section 2.4 “Comparison of the three calibration approaches” (lines 276-377, including Table 1) to the Appendix.

RC: [2] Starting from the back, i.e. the Outlook section, I wonder what transferrable knowledge the authors contribute that is unrelated to using WaterGap (and potentially the traditional approach to calibrating WaterGap)?

My impression is that most of the conclusions are rather specific to the use of WaterGap. I do not think that this is a problem per se, but it would be good if the authors would be clearer about general outcomes and those specific to WaterGap. One problem in this context is that Discussion and Conclusions are jointly discussed and that this section is 7 pages long. I think these sections can be joined if this part of the paper is short, but here it is very long. A long discussion followed by a very short conclusions and outlook section would make it much easier for the reader. There the authors could also easily separate specific and general conclusions.

AC: We have followed your suggestion to split the “Discussion and Conclusions” section and organize the last part of the manuscript as follows:

5 Discussion (with sections 5.1 to 5.6)

6 Conclusions (which includes, in revised form, what was 5.7 Outlook).

We completely rewrote the conclusions (and partly the abstract), where we now focus on clearly saying what other (global) hydrological modelers can learn from our study, regarding the application of the three alternaive calibration approaches and caveats, and added a figure to the conclusions that concisely represents the proposed approach (**Figure 7: The proposed approach for reducing and quantifying model output uncertainty of GHM by multi-variable parameter estimation and main recommendations and caveats for applying the approach in global hydrological modeling.**)

RC: [3] The final recommendation to include uncertainty in climate change impact projections related to freshwater is good, but this is already widely done (see below). Can the authors be more specific regarding their recommendation? They could for example discuss this issue much more in the context of global models and the specific implications this has.

Just a few random examples from a quick online search:

https://www.nature.com/articles/s41598-019-41334-7

https://hess.copernicus.org/articles/21/4245/2017/

https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2011WR010602

AC: Thank you for making us rethink our recommendation on including parameter uncertainty in a multi-model ensemble of impact models to project climate change hazards. Focusing on global-scale climate change impact studies, we will replace the last paragraph of the main text

"We recommend including, in future freshwater-related climate change impact studies, a behavioral ensemble of parameter sets as determined by the GLUE approach even though this will require a significant computational effort. This should reduce the underestimation of modeling uncertainty by traditional multi-model studies. As shown in the multi-model/multi-parameter study for the Colorado River basin by Mendoza et al. (2016), parameter sets with a similar performance during the calibration period may provide very different projections of climate change hazards, and the impact of parameter uncertainty is similar to the impact of hydrological model selection. "

by the following:

Climate change impact studies for individual river basins have shown that parameter sets with a similar performance during the calibration period may provide very different projections of climate change hazards, and that the impact of parameter uncertainty can be similar to the impact of the selected climate or hydrological model selection (Mendoza et al., 2016; Her et al., 2019). Therefore, consideration of parameter uncertainty by running the hydrological model with a number of behavioral parameter sets helps to reduce the underestimation of the uncertainty of potential climate change impact. However, producing a global-scale ensemble of potential future changes in hydrological variables by combining not only multiple greenhouse gas emissions scenarios, global climate models and global hydrological models (as is currently done in ISIMIP) but also model-specific behavioral parameter sets is currently infeasible. The main reason is that behavioral (or even optimal) parameter sets have not yet been determined for any global hydrological model in a spatially explicit manner at the global scale. In addition, the computational effort for such a multi-model/multi-parameter ensemble is likely prohibitive.

RC: [4] The connection to existing literature is in places very extensive and in others very brief. All methods used here have been previously assessed widely. Maybe not exactly in this combination, but certainly individually or in combination with other methods. I would therefore have expected that the authors help the reader to start from a more informed level.

For example, the (poor) ability of GLUE to identify the best parameter set has been explored in the past (see link below) and thus is what should be expected. The issue now is rather what relevance this has for the study at hand.

https://backend.orbit.dtu.dk/ws/portalfiles/portal/9729153/MR2007_305.pdf

AC: With the help of both reviews, we noticed that we have not clearly described the role of GLUE (with a random ensemble of parameter sets) as compared to the role of POC (in which a search algorithm derives an ensemble of Pareto-optimal parameter sets). In the revised version of the manuscript, we revised the introduction and developed the storyline of the paper differently. Before formulating the study objective, we now state (citing Blasone et al. 2008) that optimal parameter sets are best identified using a search algorithm, as used in POC, while GLUE serves, in the face of equifinality, to identify behavioral parameter sets and thus to quantify the model output uncertainty. Connected to this, we changed the title of the manuscript from "Multi-variable parameter estimation for a global hydrological model: Comparison and evaluation of three ensemble-based calibration methods for the Mississippi River basin" to

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

We also revised the conclusions accordingly.

RC: [5] While I possibly sound rather critical, I think this is an interesting and relevant study. My comments are simply meant to help the authors communicate their work with the readers. Shortening the paper, being clearer about specific and general contributions, and a better connection with existing literature would make it much easier for readers to understand the study and its relevance

AC: We will direct our revision in this way.

Reviewer 2

RC: Döll et al. exploit three different approaches to identify parameters in a WaterGAP model of the Mississippi River Basin. This should provide insights for the calibration of global hydrological models. Although as a reviewer I aim to be constructive and to provide concrete recommendations, I have to admit that I found this difficult for the presented study. I hope I can make my points clear and that this provides enough guidance for the authors to search for other directions.

AC: Thank you for your critical feedback that will help us to better communicate the objectives, results and conclusions of our study.

RC: The long, quite unfocused, introduction seems to give the goal of this study (line 70), where the complex and long research question is already a preparation for the reader on what is coming. My summary of the goal of the study, if I understood correctly, would be to explore how global hydrological models can be calibrated in order to make better use of available observations.

AC: The manuscript does aim at showing how to make (better) use of available observations in global hydrological modeling beyond streamflow but it is not only about calibration in the sense of finding optimal parameter sets but also about estimation of model output uncertainty. In the revised version, we will therefore change the title of the manuscript from “Multi-variable parameter estimation for a global hydrological model: Comparison and evaluation of three ensemble-based calibration methods for the Mississippi River basin” to

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

We have condensed and restructured the introduction to better fit to the revised paper title and to clarify early on that GLUE is not applied to see whether is as efficient or better than POC (with an optimization algorithm) but to be able to determine behavioral parameter sets given the uncertainty of the observations and thus quantify model output uncertainty. Most importantly, we have changed the formulation of the research objective. We have replaced

“The objective of this paper is to assess the suitability of the three multi-variable calibration approaches POC, GLUE and EnCDA for identifying ensembles of optimal and behavioral parameter sets of the GHM WaterGAP by model calibration against observations of Q and TWSA, taking into account observation uncertainties. In addition, an approach for taking into account the observation errors for the definition of performance thresholds for behavioral parameter sets is presented. In each calibration approach, model parameters of all WaterGAP grid cells within so-called calibration-data assimilation (CDA) units were uniformly adjusted. Based on calibration exercises either for the whole Mississippi River basin (MRB) as one CDA unit or for its five sub-basins (four upstream basins and one downstream basin) as alternative CDA units, we will answer the following research questions:”

by

The objective of this paper is to analyze how the uncertainty of the output of GHMs can be reduced and quantified by parameter estimation that utilizes observations of multiple output variables and their uncertainties. For the example of the Mississippi River basin (MRB), the paper shows how Q and TWSA observations can be utilized to obtain one optimal parameters set (the “compromise solution”) as well as ensembles of Pareto-optimal and behavioral parameter sets for the GHM WaterGAP, by evaluating the applicability of the three multi-variable calibration approaches POC, GLUE and EnCDA. It presents a method for defining performance thresholds for behavioral parameter sets based on observations and their uncertainties as well as the initial GLUE ensemble. In each approach, model parameters of all grid cells within so-called calibration-data assimilation (CDA) units, either the whole MRB or five sub-basin CDAs, were uniformly adjusted. We derive conclusions for multi-variable parameter estimation and quantification of model output uncertainty in global-scale hydrological modeling, answering the following research questions:

RC: In my understanding and experience, one of the reasons why GHMs currently are not thoroughly or automatically calibrated is mainly because of computational demand, besides model complexity (leading to non-uniqueness). The argument of computational demand has, surprisingly, not been taken into account in any way in selecting calibration approaches for this study. Expensive algorithms and approaches, such as Borg-MOEA and EnKF are explored, already going towards computational limits for the basin explored here. How is this ever going to translate to a global application then?

AC: We wanted to explore whether it is possible to benefit from the advantages of EnKF (EnCDA), which - different from typical calibration of hydrological models - simultaneously adjusts system states (water storages) and model parameters. We hypothesize that this property can be of advantage in situations where a model has structural deficiencies that cannot be “absorbed” via parameter calibration. The goal of the study was to explore whether EnKF can be used to assimilate not only observations of total water storage anomalies, as has already been shown to be feasible and successful at the global scale (e.g., Gerdener et al. 2023) but also streamflow observations (which had not yet been demonstrated), while at the same time adjusting parameters (which also had not yet been demonstrated in this context). We hypothesized that taking into account, in EnKF, both the uncertainties of the climate input and 2) adapting storages, parameter estimates would stabilize towards the end of the calibration period; these parameters could then be used for periods without observation data. Our study has shown that in the current setting of this study, the EnKF approach was less successful in these aspects compared to POC and GLUE, and was thus not found as applicable for reducing and quantifying the uncertainty of output of WaterGAP. We would like to point out that compared to an uncalibrated run, in the present setting, EnCDA does show improvements. Increasing the numerical efficiency of the framework even with very large state vectors, as it could be the case for a global EnCDA, is still under development. After submitting this paper, the run time of the assimilation setup was already strongly improved by avoiding reading in and writing to the hard disc, reducing the run time of a global GRACE assimilation by 75%. It is well-known that EnKF performance relies very much on the proper representation of model state and, in this case, parameter correlations, and this in turn depends on ensemble size. Our EnKF may improve in the given setting for larger ensembles, but this is indeed computationally very demanding at the global scale. This has been expressed in section 5.7 (now section 6 Conclusions), where we write "Likely because of the very small ensemble size that was feasible in EnCDA due to its severe computational burden, parameter sets and model output uncertainties could not be estimated by EnCDA in this study, neither for the calibration period nor for projections for periods without observation data."

We are convinced that both GLUE and Borg-MOEA (for POC) are not too expensive to be applied in global hydrological modeling. With 20,000 ensemble members, the run times for six CDA units in our study were 72 hours and 53 hours for POC and GLUE, respectively, and, for 32 ensemble members in the case of EnCDA, 72 hours, in the parallel computing environments described in the manuscript (section 3.4). We are currently setting up a global POC for 712 calibration units (drainage basins) covering and based on runs with a small number of calibration units we estimate the total runtime in case of 20,000 ensemble members to be 15-20 days. In times of high-performance computing, computational demand for global-scale multi-variable parameter estimating is very high but not prohibitive. We have added the information about runtimes to the revised version of the manuscript, in sections 3.4.1, 3.4.2, and 3.4.3.

RC: But besides, there were other reasons to be surprised by the selected methods and approaches. The three methods seem to be presented as calibration strategies, but I would argue they are not. GLUE is presented as an optimization technique, while it is merely a way of evaluating a sample. Therefore, it should not come as a surprise that Borg-MOEA outperforms GLUE; the authors already write themselves that the search algorithm searches in the region of interest, while GLUE is just a sample across the whole parameter space. This conclusion, therefore, could have been drawn without doing all the computations. The same holds true for the EnCDA. An implementation of EnKF is used as a way of calibrating, but EnKF has never been developed to serve as a calibration algorithm. It is useful for real-time applications, it is useful to identify model structural errors, but it never claims a convergence towards an optimal parameter set. Therefore, no surprise that results drifted off in the validation period!

AC: We would argue that both Borg-MOEA and GLUE are calibration strategies; Borg-MOEA is a technique for identifying (Pareto-) optimal parameter sets, while GLUE is a technique for identifying behavioral parameter sets but can be also used to determine (in a sub-optimal way compared to POC) optimal (i.e., best-behaving) parameters sets. In this way, both are calibration techniques. GLUE approaches were called calibration, for example, in Marmy et al. (2016) and Wu and Jansson (2013).

With the help of both reviews, we noticed that we did not clearly formulate the role of GLUE (with a random ensemble of parameter sets) as compared to the role of POC (in which a search algorithm derives an ensemble of Pareto-optimal parameter sets). In the revised manuscript, we therefore modified the storyline and improved the insufficiently clear presentation of the roles of Borg-MOEA and GLUE. Before formulating the study objective, we now state (citing the additional reference Blasone et al. 2008) that optimal parameter sets are best identified using a search algorithm, as used in POC, while GLUE serves, in the face of equifinality, to identify behavioral parameter sets and thus to quantify the model output uncertainty. We also revised the conclusions accordingly.

Regarding EnKF, it is true that EnKF has never been developed as a calibration algorithm, and our research investigated whether EnKF can serve to estimate parameters of a global hydrological model using observations of streamflow and TWSA. EnKF has been demonstrated in various studies to improve the realism of global hydrological model simulations when compared to various observations, and this includes our own EnKF implementation at global scale with WaterGAP, which however, only takes into account TWSA observations (Gerdener et al., 2023). It is one of the standard techniques when multiple data sets, at different spatial scales and with possibly differing temporal or spatial coverage are to be combined, such as in meteorological or hydrological reanalyses. Various papers (e.g., Wanders et al., 2014, cited in the manuscript) have shown, typically in regional or local settings, that the EnKF variants are capable of estimating model parameters along with model states. Therefore, we believe it is perfectly reasonable to ask whether EnKF is able, at the same time, to estimate model parameters albeit maybe not as efficient as a dedicated calibration approach. Some of the reasons why EnKF has the potential for improved parameter estimation are provided in section 2 of the manuscript and our response to the previous comment.

We agree that from a parameter calibration perspective, deriving an optimal parameter set from EnKF seems complicated. POC and GLUE generate constant parameter sets. However, we hypothesize that the updates of the water storages could stabilize the parameters and compensate for model structure deficiencies and climate input uncertainties. EnKF generates a time series of estimates of parameter sets, which is often misunderstood as generating time-variable parameters. This time series which in the ideal cases converges may include typical seasonal signals, and such signals point towards model errors and are difficult to interpret. In this study, we decided to apply the parameter estimates of the last month of the calibration phase during the validation phase, to be able to compare the different ensemble-based approaches for reducing and quantifying uncertainty — which is the aim of this study. Future studies will investigate how seasonal signals in the parameter estimates can be used to (1) trace back model errors and (2) develop empirical error models, which can include parameterizations depending on the season.

RC: Besides these methodological issues, the study is hard to read and follow. Only at page 20 (!) I felt that I got a more concrete picture of what was done. And even then, it read a lot like a diary. For instance, first I was very very surprised at line 520 that also a multiplication factor for precipitation and net radiation were included as calibration parameters. Then I was not so surprised to find out that the multiplication factor for precipitation came out as most sensitive (l. 582), to then I was surprised again to learn that it was still left out of the calibration (l. 585). I know that there is an argument for documenting failures etc., but I don't think this is helpful at all at this level: just leave this kind of stuff out, don't bother the reader with it. Furthermore, there is some kind of strange mixed use of NSE and KGE. The NSE is optimized, but the KGE components are evaluated. Why not directly optimizing the KGE then? That would lead to different results compared to the NSE. Figure 3 shows NSE's if I read the axes, but the caption refers to some kind of KGE.

AC: To increase the readability, we decreased the length of the main text. We have moved section 2.4 “Comparison of the three calibration approaches” (lines 276-377, including Table 1) to the Appendix. Regarding the description of the GRACE TWSA data, we moved the text on leakage (lines 483 to 508) to the supplement.

Regarding the reviewer’s comment on the method descriptions reading as a diary, our goal was to make transparent to the reader the many decisions that need to be taken in parameter estimation. Regarding the process of deciding on whether to include the precipitation multiplier P-PM as a calibration parameter, it has nothing to do with documenting a failure but with explicating why it was excluded even though model results are sensitive. While we could remove this from the manuscript, reviewer 1 wanted to get a deeper discussion on the selection of calibration parameters and is interested in a more detailed explanation for the exclusion (R1 comment 0). We therefore revised the part on the precipitation multiplier as follows (section 3.2.4):

P-PM was excluded from calibration even though it ranked first in all six CDA units for almost all four test variables, for various reasons. First, the precipitation input is perturbed in EnCDA, and an additional multiplier would lead to a double-counting of precipitation uncertainty. Second, mean annual precipitation in the CDA units of WaterGAP climate forcing does not differ much from the values derived from the high-resolution (4 km) PRISM dataset for the USA (Table S1).

Table S1. Comparison of mean annual precipitation in the CDA units for the calibration period 2003-2012 between GPCC-WFDEI used to drive WaterGAP and the high-resolution (4 km) PRISM* dataset for the USA [mm/yr]

CDA unit	GPCC-WFDEI	PRISM	Ratio PRISM/GPCC-WFDEI (potential P-PM)
I Arkansas	705	667	0.95
II Missouri	595	622	1.04
III Upper MRB	951	878	0.92
IV Ohio	1313	1242	0.95
V Lower MRB	1286	1254	0.97
MRB	839	829	0.99

***https://climatedataguide.ucar.edu/climate-data/prism-high-resolution-spatial-climate-data-united-states-maxmin-temp-dewpoint**

Regarding the mixed use of NSE and KGE: NSE is used for parameter estimation only in the case of POC and GLUE, but not in the case of EnCDA (where the objective function cannot be freely set, see Table A1, but is a weighted RMSE), and we evaluated the results using both NSE and the three components of KGE (not KGE itself). We think it is particularly informative to evaluate the result of a parameter set optimization using a certain objective function/performance metric by another performance metric, in particular the three KGE components as these are well interpretable (correlation: temporal shifts; bias: difference in means; variability ratio: difference in temporal variability). In the revised version, we have added a sentence to section 4.1 to indicate how the improvement in the correlation coefficient by the calibration can be interpreted: **Thus, calibration mainly leads to improved timing of monthly streamflow and TWSA.**

We have corrected the typo (KGE) in the caption of Figure 3, it should read NSE.

RC: Finally, there is no conclusion-section, just a very extensive "Discussion and conclusion", which is already indicative that there are too many separate aspects that are aimed to be tackled in this study. This study aimed to serve the GHM community, but the kind of strategies and questions explored here have already been extensively addressed and investigated by regional scale models – with the same conclusions as this study. Now the challenge remains how to translate this to models applied to larger areal extends, and this study does not seem to contribute to that.

AC: We will follow your suggestion (and that of the other reviewer) to split the "Discussion and Conclusions" section and organize the last part of the manuscript as follows:

5 Discussion (with sections 5.1 to 5.6)

6 Conclusions (which includes, in revised form, what is now 5.7 Outlook).

To better show the scientific contribution and the wider relevance of our results for other (global) hydrological modelers, we have completely rewritten the conclusions, focusing on what was learned regarding methods for global-scale reduction and quantification of the output uncertainty of global hydrological models by the three approaches for multi-variable parameter estimation POC, GLUE and EnCDA. We also concluded that based on the experiences in our study, run times for a global-scale application for, e.g., 1000 basins are not prohibitive.

However, due to the high computational demand, those many behavioral parameter sets could not be used in climate change impact studies, neither if just WaterGAP were applied and certainly not in a multi-model ensemble with various global hydrological models. Regarding the consideration of parameter ensembles in global-scale climate change impact studies, we have therefore changed our conclusion and replaced the last paragraph of the main text

"We recommend including, in future freshwater-related climate change impact studies, a behavioral ensemble of parameter sets as determined by the GLUE approach even though this will require a significant computational effort. This should reduce the underestimation of modeling uncertainty by traditional multi-model studies. As shown in the multi-model/multi-parameter study for the Colorado River basin by Mendoza et al. (2016), parameter sets with a similar performance during the calibration period may provide very different projections of climate change hazards, and the impact of parameter uncertainty is similar to the impact of hydrological model selection. "

by the following:

Climate change impact studies for individual river basins have shown that parameter sets with a similar performance during the calibration period may provide very different projections of climate change hazards, and that the impact of parameter uncertainty can be similar to the impact of the selected climate or hydrological model selection (Mendoza et al., 2016; Her et al., 2019). Therefore, consideration of parameter uncertainty by running the hydrological model with a number of behavioral parameter sets helps to reduce the underestimation of the uncertainty of potential climate change impact. However, producing a global-scale ensemble of potential future changes in

hydrological variables by combining not only multiple greenhouse gas emissions scenarios, global climate models and global hydrological models (as is currently done in ISIMIP) but also model-specific behavioral parameter sets is currently infeasible. The main reason is that behavioral (or even optimal) parameter sets have not yet been determined for any global hydrological model in a spatially explicit manner at the global scale. In addition, the computational effort for such a multi-model/multi-parameter ensemble is likely prohibitive.

RC: Overall, the methods seem to be not in line with the goals that this study aims to achieve, and the written presentation requires substantial improvement.

AC: As described above, in the revised version we have changed the title and framed the three methods and study goals more clearly. And we have improved the presentation, mainly by reformulating, restructuring and shortening; we have fully rewritten the conclusions and strongly revised the abstract, the introduction, section 5.1 and the conclusions.

References

Gerdener, H., Kusche, J., Schulze, K., Döll, P., Klos, A. (2023): The Global Land Water Storage Data Set Release 2 (GLWS2.0) derived via assimilating GRACE and GRACE-FO data into a global hydrological model. J. Geodesy, 97, 73. https://doi.org/10.1007/s00190-023-01763-9.

Marmy et al. (2016): Semi-automated calibration method for modelling of mountain permafrost evolution in Switzerland. The Cryosphere, 10, 2693–2719. doi:10.5194/tc-10-2693-2016

Wu, S.H., Jansson, P.-E. (2013): Modelling soil temperature and moisture and corresponding seasonality of photosynthesis and transpiration in a boreal spruce ecosystem. Hydrol. Earth Syst. Sci., 17, 735–749. doi:10.5194/hess-17-735-2013