HESS-2023-18: 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

P. Döll et al.

We thank both referees 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

Two referees (of which one has also reviewed the initial submission) have now submitted their reports on the revision of your manuscript. As you will see, both assessments largely agree in that they recognize the improvements that you have made, but also point out a number of issues that need improvement. Specifically, both reviewers identify the need for a more thorough discussion, with a reflection on the literature mentioned in the Introduction. Anonymous referee #1 also lists a number of other issues that you will need to address. I think both reports will help you to further improve your manuscript. I am looking forward to receiving a revised version that addresses the issues and remaining concerns from the referees.

AC: We have further improved the manuscript by following the advice of the two referees. In addition to modifying the discussion, we revised the abstract, and to clarify the role of EnCDA (Referee 1), modifications were made in the introduction and methods section as well as in the conclusion. Below please find our responses to the comments of the referees.

Referee 1

RC: The authors significantly revised the manuscript. I like that some sections have been moved to the supplemental material or shortened. And I appreciate that some parts are better explained now. I still have some issues with the (new) discussion section and how the contributions are assessed. My main concern is that the discussion section makes hardly any effort to place the study's findings into the context of existing literature. Please see my detailed comments below.
AC: We are glad that with the previous resubmission we achieved an improvement in clarity and readability. We have now followed your advice and revised the discussion section. For details, please see our responses below.

(1) RC: You now state that you perturb the precip multiplier in the EnCDA runs, but not in the POC and GLUE runs. Is this a fair comparison then? It is surprising that EnCDA does not perform well if you perturb multiplier, parameters, and states, while you only calibrate the parameters with the other strategies. I am unclear why EnCDA worked well in previous studies where TWSA was assimilated, but not here. Would the required ensembles not be known from previous work by the authors?

AC: Thank you for this comment. The comparison is certainly not a fair one, but that would be extremely difficult to facilitate and it also may not reflect reality.
We conduct a specific model calibration exercise here, and we test two methods (POC and GLUE) that were specifically developed for model calibration, i.e. parameter estimation, together with a third one (EnKF) that was specifically developed for a different purpose, i.e. state estimation. The EnKF is slightly modified here to facilitate parameter estimation, by extending the state vector, which is new at least in this setting (global hydrology model calibration from GRACE and streamflow data), at least to our knowledge. We find here that the EnKF actually works quite well during the assimilation phase, and this is not at all in contrast with other studies and also not with our own work on assimilating just TWSA (Eicker et al., 2014; Schumacher et al., 2018; Gerdener et al., 2023. But it did not perform equally well in estimating model parameters with which we then run a free (uninformed by GRACE and streamflow) simulation in the validation period. We do not know of any comparable studies. EnKF is first of all developed to optimized state estimates (different from POC and GLUE) and here we try to see how far it can compete with dedicated calibration approaches, in a setting where POC and GLUE are expected to perform well. Put in other word, it would have been a big surprise if EnKF had been on par in the experiment performed here, even though we had hoped that due to e.g. balancing precipitation input uncertainties by state adjustment, parameter estimation might be improved. No study is known to us that had systematically compared model parameter estimation from TWSA data assimilation to other more dedicated calibration approaches, and thus certainly no study that would even compare multi-variable EnCDA to calibration approaches.

The other issue that we would like to recall is that EnCDA, as a variant of EnF, optimizes a weighted RMSE of observations, and this is in contrast to the POC and GLUE method, whereas "performance" in this paper is mainly understood in terms of NSE which is the metric optimized commonly in model calibration approaches (and in POC and GLUE here). That might be another reason why the EnKF does not compare so well.

To clarify the EnCDA approach, we revised the manuscript in numerous places, in the abstract, the introduction (e.g. lines 131-148 and extension of research question 1), Section 2.3 on the EnCDA method in general, Section 3.4.3 on the specific EnCDA implementation (adding one reference), the discussion sections 5.1 (adding two references) and 5.4 and the conclusions (Section 6, in particular lines 1287-1298).

(2) RC: You conclude (in the abstract) that POC is better than GLUE for finding optimal parameters, which has been shown many times before. More interesting is the issue that EnCDA did not work well. Why that was the case might warrant more space in the abstract than mentioning accepted knowledge.

AC: Please see our answer to RC (1). There may be multiple reasons here, and one of them is certainly the ensemble size. The ensemble size for Ensemble Kalman filters is generally much smaller as compared to those in model calibration approaches simply since one does not only perturb the parameters but the states and the forcing. Given limited compute time, we could increase the parameter ensemble in EnCDA greatly and likely obtain much better parameter estimates, if we would remove forcing perturbation and limit or even remove the state perturbation, but then we would simply not be running an EnKF anymore. The experiment might be considered more "fair" in the comparison with POC and GLUE, but that is simply not the real situation where we work with data assimilation. There are other reasons – the EnKF does not allow one to look at all at a Pareto front in hindsight, since one has to specify the weighting between GRACE and streamflow data right from the start. And the optimization functional of the EnKF is also different from many of the metrics considered

here. Then, we know that an Ensemble Kalman smoother (EnKS) would make better use of all data for providing the final state and parameter estimates, but the numerical expense would be twice as with the EnKF. Lastly, the EnKF and EnKS represent unbiased estimators only under limited conditions which are here clearly not met, in particular for streamflow simulation.

We have revised the section on the performance of the three approaches, which now reads:

The GLUE approach is almost as successful as POC in enhancing WaterGAP performance and also allows, with a comparable computational effort, the estimation of model output uncertainties that are due to the equifinality of parameter sets given the observation uncertainties. Our experiment reveals that the EnCDA approach performs similarly in most CDA units during the assimilation phase but is not yet competitive for calibrating global hydrological models; its potential advantages remain unrealized, likely due to its high computational burden, which severely limited the ensemble size, and the intrinsic nonlinearity in simulating Q.

(3) RC: The discussion of why EnCDA is so poor – given that it has more degrees of freedom than the other methods, is rather brief. (a) I would have assumed that the authors create a smaller testcase to see whether the problem is simply due to insufficient ensemble members. Does the problem go away if more samples are used? (b) Is the statement that Q simulations would have been better with transformed data or by ignoring zero flows correct. You would have obtained a better NSE value, because you reduce peak errors or include only a subset of the data. The simulations stay the same, you just have a different statistical metric that gives a higher value.

AC: Regarding (a): With increasing the ensemble gradually, e.g. to 64 or 100 members, results in EnCDA gradually improve but in this way the method would become unfeasible if one wants to compete with POC or GLUE in terms of ensemble size (several ten thousands). The problem does not go away. The smaller ensemble and the related loss of estimation power is the unavoidable price that we have to pay for using a recursive state estimator in contrast to the dedicated calibration method. The ensemble choice that we made here (n=32) corresponds to our earlier work and also to other studies (Zaitchik et al., 2018; Girotto et al., 2016; Kumar et al., 2016; Getirana et al, 2020a/b). With gradually increasing the size the estimators increase gradually in accuracy, but it will be never possible to use ensemble sizes as common in POC. We have added the above discussion in Section 5.1.

Regarding (b): The reviewers comment “The simulations stay the same, you just have a different statistical metric that gives a higher value” is not correct for the EnKF, at least not if we would have stayed with the usual least squares metric within the EnKF optimization. function. During the simulation the state and parameter updates would be different and therefore the simulation (after these updates) would be different. But we feel we cannot add this to the already very long manuscript – it has to wait until we publish an update.

(4) RC: The authors state that “However, the added value of any calibration is very low in the humid and hilly Ohio basin where the performance of the uncalibrated model is already good”. While this is interesting, it would be good to place this section into the context of existing literature. For example (just doing a quick scan across the literature), van Werkhoven et al. (2009, AWR, https://doi.org/10.1016/j.advwatres.2009.03.002) made similar conclusions regarding the high performance of a priori parameters for humid US regions when using the SAC-SMA model. Considerations of climate regimes for global water models calibration was studied in Yoshida et al. (2022, WRR, https://doi.org/10.1029/2021WR030660) and Kupzig et al. (2023, ERL, DOI 10.1088/1748-

9326/acdae8) . So, there seems to be a context of past studies in which this new result could be placed. The wide range of co-authors could easily add some relevant reference for discussion.

AC: Regarding the sentence cited by the referee, we did not think that it would add sufficient value to add a reference to back this rather logical and straightforward result that we do not consider to be very central for our study. Following the advice of the reviewer, we now added a reference to the study of Troy et al. (2008) as they found, in the calibration of the VIC model to streamflow only, the same behavior for the Ohio River (their Figure 3). The revised sentence in Section 5.2 now reads:
However, the added value of any calibration is very low in the humid and hilly Ohio basin where the performance of the uncalibrated model is already good; in their study on calibrating the VIC model for the USA using observed Q only, Troy et al. (2008) also found that modeled streamflow that fit well to observations before calibration, as was the case for the Ohio River, continued to do so.
We do not want to discuss this result in the context of climate (i.e. that hydrological models often work best in humid climates), because the point is just that rather obviously where an uncalibrated model is already doing a quite good job, calibration cannot help much. In fact, hydrological models also perform poorly in humid climates if the impact of surface water bodies on streamflow and TWSA is high. This is why we do not think that it would be useful to cite Yoshida et al. (2022). The papers of Van Werkhoven et al. and Kupzig et al. deal with sensitivity analyses to select the most relevant parameters from the whole model parameter set, while we chose not to discuss our selection of the most sensitive parameters (which was the same for all three compared approaches) to not further increase the length of the already very long manuscript. In addition, both papers use only streamflow observations and not TWSA for their sensitivity analysis, while Section 5.2 discusses the value of multi-variable calibration. In the revised version, the paper of Kupzig et al. (2023) is now cited in Section 5.6.

In Section 5.2, we also put our overestimation of summer low flows in the context of the Troy et al. study by adding:
In the study of Troy et al. (2008), the overestimation of summer low flows in the Arkansas basin, the basin that is affected most by this behavior in our study (Fig. 4), is reduced but not removed by the calibration.
In addition, we inserted text on the added value of multi-variable calibration at the end of Section 5.2:
The added value of multi-variable Pareto-optimal calibration of WaterGAP for 28 very large globally distributed basins using monthly time series of Q and TWSA was investigated by Werth and Güntner (2010). They found that improved simulations of TWSA and Q were achieved for most basins after calibration, but calibrated Q was still poor compared to the observed values in some basins; a better fit to GRACE TWSA did not necessarily lead to a better fit of simulated to observed Q. For the Mississippi basin, the relative RMSE was reduced by calibration by about 20% for both Q and TWSA. A multi-variable model calibration for the Lake Urumia basin (Iran) showed that satellite observations of time series of TWSA and irrigated area led to a good fit to observed TWSA and a reduction in the Q bias, but additional in-situ observations of Q were necessary to estimate parameter sets that lead to a good fit (Hosseini-Moghari et al., 2020). Both studies underline that model calibration should be based on both Q and TWSA observations.

(5) RC: The discussion of possible reasons for output uncertainty does also not contain any

meaningful link to existing literature. Only one paper on groundwater losses is mentioned, but nothing else. For example, other large-scale models such as VIC etc have been applied to the same domain (given runs for the US) (e.g. Troy et al. 2008 WRR, https://doi.org/10.1029/2007WR006513). Did VIC show comparable results (NSE values)? Given that this models also does (did?) not include transmission losses. I think the authors should make a bit more effort to provide context for their results. Zajac et al. (2017, JoH, https://doi.org/10.1016/j.jhydrol.2017.03.022) for example explored the influence of reservoirs/lakes on global streamflow simulations, incl. uncertainty – How does this link to the findings here in which the authors seem to explore/experience similar issues in their GLUE simulations?

AC: The study of Troy et al. does not provide comparable NSE values. However, we related our results qualitatively to their result in Section 5.2 (see our answer to RC comment 4). We do not see a link to the study of Zajac et al. (2017). This study just compared the difference in Q fit to observations by a model that does not take into account lakes and reservoirs at all and a model version in which major lakes and reservoirs are included. This is not comparable to our study with a model that, different to the model used in Zajac et al. (2017), its standard version already includes lakes and reservoirs (and wetlands) and for which parameters are to be optimized. To clarify the difficulty of simulating water flows and storages in the Prairie Pothole Region, we added a sentence (and reference) to the second but last paragraph in Section 5.3:
The Prairie Pothole Region contains between 5-60 wetlands per km^2, and their hydrological modeling relies on accurately characterized depth-volume relationships derived from detailed topographic surveys (Minke et al., 2010).

6) RC: The authors later discuss how their parameter equifinality with that of Harlin and Kung (1992). While I appreciate that a comparison is made, does the very different set-up (floods versus monthly flows, very different regions, …) make this a useful comparison? The huge number of existing GLUE might provide an example which is much more like the study performed here.

AC: While there is a large number of GLUE studies, most use only streamflow observations. Both Harlin and Kung (1992) and the study of Jost et al. (2012), which is now also referred to in Section 5.6, are examples of rare multivariable GLUE studies that also provide some information on the estimated parameter sets. In the revised version, we added the following sentence at the end of Section 5.6:
A multi-variable parameter estimation of a hydrological model for the upper Columbia River basin in Canada, which used observations of Q and glacier volume change, identified 23 rather different behavioral parameter sets that all led to very high NSE values for daily streamflow of at least 0.92 (Jost et al., 2012).

(7) RC: Because I do not want to discuss the same issue again and again, let me make a wider statement here. The discussion section significantly lacks references to other literature. You need to convince that the results and findings are not so specific to WaterGap that other users do not benefit. One way to do so is by showing how your results add knowledge to previous studies. The authors hardly do this, and several discussion sections contain no references at all. Let me stress that my point here is not for the authors to simply include the few references I mention, but to rather make some effort to scan the literature so that their results are placed in context. Also, there are some references mentioned in the introduction (e.g.

Scanlon et al. 2018/2019) but they do not come back in the discussion section, which I think should be the case.
AC: To our knowledge, this is the first time that a GHM has been calibrated using both streamflow and total water storage anomaly except for Werth and Güntner (2010) and Hosseini et al. (2020); and the focus of the discussion section is show/discuss what we have learned about to an appropriate methodology for doing multi-variable parameter and uncertainty estimation (using TWSA and Q) in global hydrological modeling, clarifying aspects that are of value for other large-scale to global hydrological modeling. This was the reason for not focusing our discussion on a comparison to studies presented in the literature that did not do multi-variable parameter estimation.

In the revised discussion section, we followed your advice and came back to some references from the introduction in the discussion and also tried to place our results more strongly in the context of (additional) publications. We added the following to the discussion Sections 5.1 to 5.6:

- 5.1 Advantages and disadvantages of the three ensemble-based multi-variable calibration approaches: We refer to two additional references regarding the ensemble size of EnKF data assimilation
- 5.2 Added value of multi-variable calibration as compared to the standard WaterGAP calibration for identifying one optimal parameter set: See our answer to (4) RC. And we included two sentences relating our study to Scanlon et al. (2018, 2019).
  Thus, where GHMs incorrectly simulate TWSA trends (Scanlon et al., 2018), multi-variable model calibration is likely to lead to more realistic simulated trends. However, at least for our CDA units, variability and probably also seasonality of simulated TWSA are not necessarily improved by such a calibration (Scanlon et al., 2019).
- 5.3 Estimation of output uncertainty: See our answer to (5) RC.
- 5.4 Trade-offs between optimal simulation of Q and TWSA: Nothing was added, as the study results had already been related to two other studies (Rakovec et al., 2016 and Schumacher et al., 2018).
- 5.5 Added value of sub-basin CDAs instead of one basin CDA: Nothing was added, as the study results had already been related to five other studies.
- 5.6 Characteristics of identified (Pareto-)optimal and behavioral parameter sets: See our answer to (6) RC. In addition, a sentence citing Kupzig et al. (2023) was added.
  SL-RC and SL-MSM, which affect the release of water from the soil and determine the maximum amount of water that can be stored in the soil, respectively, were found to be the most influential parameters for a number of Q metrics of the evaluated 347 global river basins (Kupzig et al. 2023).
-

References

Zaitchik, B. F., et al. (2008). Assimilation of GRACE terrestrial water storage data into a land surface model: Results for the Mississippi River basin. Journal of Hydrometeorology, 9(3), 535-548.
Girotto, M., et al. (2016). Assimilation of gridded terrestrial water storage observations from GRACE into a land surface model. Water Resources Research, 52(5), 4164-4183.
Kumar, S. V., et al. (2016). Assimilation of gridded GRACE terrestrial water storage estimates in the North American Land Data Assimilation System. Journal of Hydrometeorology, 17(7), 1951-1972.
Getirana, A., et al. (2020a). Satellite gravimetry improves seasonal streamflow forecast initialization in Africa. Water Resources Research, 56(2), e2019WR026259.

Getirana, A., et al. (2020b). GRACE improves seasonal groundwater forecast initialization over the United States. Journal of hydrometeorology, 21(1), 59-71.

Referee 3

RC: The authors have made a number of revisions based on the comments from the reviewers and the editor. While I agree most aspects pointed out have been properly addressed I think a few minor revisions are needed.
AC: Thank you very much for your careful reading and detailed suggestions.

Abstract:
RC: L23-24: "…ensemble-based multi-variable calibration approaches…" -> ensemble-based approaches (for brevity and consistency with the title)
AC: It is important to keep the "multi-variable" here, as it is the essential feature. In the title, this is covered by "multi-variable observations".

RC: L29, L353, L1157, and L1283: observation data -> observational data
AC: done

RC: L31-32: Delete ", which utilizes the Borg multi-objective evolutionary search algorithm to find Pareto-optimal parameter sets, "
AC: done

RC: L39-40: Delete ", in which both parameter sets and water storages are updated,"
AC: done

RC: L43 and the rest of the manuscript: validation -> evaluation.
AC: While we agree that the term validation is problematic and might be replaced by e.g. "testing" or "evaluation", we prefer the term "validation period" to the term "evaluation period", because it is widely in the hydrological literature to refer to the period outside the calibration period for which the model is run with the calibrated parameters. We also evaluate model performance during the calibration period, so the term evaluation period is also not perfect.

RC:L50: GHM abbreviation not defined in the abstract
AC: We changed the sentence and do not use the abbreviation anymore.

Introduction:
RC: L64: "… uncertainty due to uncertain …" -> uncertainty due to
AC: done

RC: L90-91: Delete "(some missing months before 2016 and a gap until the start of GRACE-Follow-on mission in May 2018)"
AC: done

RC: L144: "How and how well…" -> How and to which extent
AC: done

Section 2:
RC: L206-217: This last paragraph reads more as discussion than a description of the

approach, which is the intention as stated in L182. Thus, I suggest this paragraph to be removed from section 2.1.
AC: done

RC: L251-259: Same as above. I suggest to continue from L250 directly into "To achieve plausible and stable EnCDA results…"
AC: done

Section 3:
RC: L334: The sentence from L90-91 could come here to elaborate.
AC: not included for brevity.

RC: L451: "…, for various reasons." -> for two main reasons.
AC: done

RC: L522, L523, L527, L609: Is it Borg-MOEA or Borg MOEA? Pick one and use it consistently.
AC: Borg MOEA

RC: L549-550: in which environment did the GLUE runs take place? It is not clear/explicit from the text if the same environment described in section 3.4.1 was used or not.
AC: It is made explicit in the sentence before, which reads “All GLUE runs started in 1991 and were done on the same Linux cluster machine as the POC runs.”

RC: L565-566: plus/minus -> ±
AC: done

RC: L576: "… grids over …" -> grid cells over
AC: done

Section 4:
RC: Figure 4: I suggest increasing the legend font size. Also, in the caption, Table 36?
AC: Done, also for Figure S4

RC:
L741: Sect. -> Section
L770: correlates -> correlate
L791: Table 32?
L856: 5 m to 8 m -> 5 to 8 m
L893: "6 out of the 9 …" -> "Six out of the nine ..."
L965: quotation marks are opened but never closed. ("whole …)
AC: all done

Section 5
RC:
L1026: multi-variable calibration -> ensemble-based
L1061: RMS? Did you mean RMSE?
L1097: 9 -> nine
L1166: .. -> .
AC: done

Section 6

RC: L1280-1285: Is it possible to draw such conclusions if the EnCDA was limited to 32 (vs 20,000 for POC and GLUE) ensemble members?. As pointed out by the authors in L1051-1052, results might be an artifact of the small ensemble size.

AC: We had already written in the following sentence "Potential reasons are the severe computational burden of the EnCDA approach that only allowed setting up a very small ensemble and the intrinsic nonlinearity in simulating Q." For clarity, we revised the sentence to the following:

Possibly due to the severe computational burden of the EnCDA approach that only allowed setting up a very small ensemble, the multi-variable EnCDA approach that we followed in our pilot study is not suitable for application for GHM parameter estimation using Q and TWSA, as 1) its performance is lower during the calibration period than that of POC and GLUE, or for the large CDA unit MRB even lower than that of the uncalibrated WaterGAP and 2) its application during the validation period (without observational data) led to spurious results (Figs. 4 and S4). The intrinsic nonlinearity in simulating Q makes a multi-variable EnCDA that includes Q observations more difficult than an EnCDA that only includes TWSA or TWSA and other storage observations.