Reply to Review #2

We would like to start by thanking the anonymous reviewer for their constructive comments and insightful review. We believe strongly that the feedback will help strengthen the revised manuscript and thereby will be of value to the hydrological modelling community.

As mentioned in the reply to Review #1, we feel that the review may not fully capture the key message and central focus of our work, indicating that our framing requires significant improvement. In particular, we need to clarify the distinction between model developers and model users, a vital aspect for the proper framing of this study. As the authors of this publication, we belong to the model user category.

It is essential to clarify that our intention is not to overlook any structural issues in the inputs, models, or their applications. The central point of our study is: “How do users of models, whom are not themselves the developers of these models” have to interpret differences in model output in the light of discharge observation uncertainty. Our study provides these users of models with the workflow and statistical tools to make that decision. We make a clear distinction between model developers and model users, because in daily use of models this distinction is also present. We illustrate this workflow and its usefulness with the model runs done with the two models on the CAMELS-GB dataset. The models are employed in this study as intended by the developers to demonstrate inevitable short comings and strengths of these models. We recognize the need to address this framing issue throughout the publication after careful revision.

We are fully committed to addressing the specific concerns raised by the reviewer. In the following section, we will provide a point-by-point response to each concern and outline the steps we will take to address them adequately.

General comments:

“A significant overhaul is needed in my opinion. Mainly, the title says "observation uncertainty" but precipitation is also an observation and its uncertainty is left out. The paper should have said "observed discharge uncertainty" because that is the only thing it deals with.”

In light of your feedback, We recognize that the publication needs to distinguish between the observation uncertainty and observed discharge uncertainty. To accurately represent our work, we will modify the title to 'observed discharge uncertainty,' as this encompasses the specific scope of our study. We will make sure to address this distinction throughout the revised manuscript.

Specific comments:

Abstract:

What is temporal sampling and observation uncertainty?

Indeed, these terms are ambiguous for readers and need clarification. We will improve the abstract by clarifying the terms and by using “discharge observation uncertainty” instead of “observation uncertainty”. The term “temporal sampling uncertainty” will be clarified by stating that this refers to the uncertainty that stems from the sampling of the evaluation period for objective function calculation.

“L2-3: While mentioning that comparison studies are invalid, it should also be mentioned that all models (and the inputs used) are invalid to begin with. No model incorporates true nature. In my opinion, the point is more about finding out models that are useful for a given purpose.”

We agree with the statement that models are flawed by nature and are only fit for a given research question (purpose). As mentioned in the review, if a model's purpose is to accurately predict river discharge, as in our study, it is relevant to include discharge observation uncertainty in model comparisons. Adjustments will be made to include this in the revised manuscript.

L3-5: Regarding the problem of temporal sampling, same data is fed to all models. If some perform better than others then, isn't this what we are looking for?

The aim is indeed to identify the best performing model. In hydrology, model often determined by implementing a singular objective function that quantifies the agreement between simulations and observations over a prolonged time period. As discussed in Clark et al. (2021), we must not overlook the relevance of temporal sampling when comparing hydrological models. For instance, Fowler et al. (2018) pointed out that equally weighting each year in the calibration timeseries emerges as the most robust strategy. This approach ensures that a given wet year's influence, that may have a significant impact on the KGE-NP score (this study), is balanced with the potential value contained in dry years, particularly in the context of a changing climate. Furthermore, the work of Lamontagne et al. (2020) highlights an essential aspect of certain objective functions, which might exhibit low bias but still manifest considerable variability between samples of the streamflow time series. This variability can stem from the skewness and periodicity of the data. It is this variability that should be adequately captured by the hydrological models. Therefore, we highlight the problem of temporal sampling as evaluation procedures that only assess model performance on the whole temporal period can obscure objective judgement.

L10-13: Only two models are compared? I would have used may be 10 given how large the number of test catchments is. Gao et al. 2018 show many in their first table (both conceptual and physically-based). It would be interesting to see how the results change by taking more models.

There is a wealth of hydrological models available of which only a small selection is presented in Gao et al. 2018. In this study, we demonstrate the use of an easy to implement methodology that is agnostic towards the selected models. Therefore, we feel that the addition of more models will not further exemplify the proposed method. We decided on using distributed models given their relevance for other research. Minimizing redundant model runs and optimises time spent on novel research (Jain et al. 2022.) In addition, we utilized the eWaterCycle platform for model experimentation as it ensures reproducibility of this research. At the time of writing the number of available models on this platform is still limited, but actively being expanded upon. The analyses in this manuscript may be re-run once more models are available on the platform.

L11-13: For the inter-model case, please mention whether the models are calibrated or not.

Thank you for this comment, we will rectify this in the revised manuscript.

1. Introduction

L31: I am really really sorry for my nit picking but hydrologists were well aware of the challenging aspects of hydrological modeling long before 2018. It is a well-known problem. I think you can omit the citation.

We don't view this as nit picking, we will remove the reference in the revised manuscript.

L33-45: Very informative. Thanks.

Thank you for this comment.

2. Methodology

L71: Observation uncertainty is mentioned but temporal uncertainty isn't. Just add a few words for the sake of completeness.

We will add a few words on temporal uncertainty in the revised manuscript for the sake of completeness.

L94-99: Nicely summarized.

Thank you for this comment.

L103: Fine spatial resolution is used, but the problem of the daily temporal resolution is not treated. Small catchments (area < 1000 km2) have problems with time-of-concentrations. There, the peak precipitation and discharge take place at the same time step. Something that the model cannot solve and produces parameters that are unrealistic during calibration. The problem of a few values dominating the objective function is also the consequence of incorrect temporal resolution. At least in my experience. I have seen this problem for catchments of more than 4000 km2 size. And I have a sneaking suspicion that CAMELS-GB has smaller catchments inside it. A procedure has to be used that discards catchments where the precipitation peak and flow peak happen at the same time step, most of the times. Such a problem exists for larger catchment on daily time scales but to a much smaller degree. That is when a precipitation event happens near to the catchment mouth.

Thank you for this insightful comment. We will identify and report the basins with large temporal discretisation errors as well as their catchment size. An additional analysis will be conducted following the mentioned procedure that highlights catchments where the precipitation peak and streamflow peak coincide in the same time step. The results will be critically reflected upon in the discussion chapter.

Your feedback reinforces the importance of elucidating temporal sampling uncertainty. For instance, a hydrological simulations inability to capture peak flow might still yield a high objective function value when evaluated across an extended period. The potential penalization for missing peak flow can be mitigated by adequately capturing other components, such as baseflow. This underscores the intricate nature of the temporal sampling challenge, which we acknowledge in the study.

L105-106: It is not mentioned why only two models were chosen. I don't understand why legacy gave us two models only. In GB, Keith Beven has, for sure, used others.

This is an ill-used term that will be removed from the revised manuscript. Originally legacy, referred to the relevance for evaluating hydrological models in the context of "hyper resolution modelling" (e.g, Wood et al., 2011; Beven and Cloke, 2012; Bierkens et al., 2015).

L127: Yes, they will most probably lead to different conclusions. Hence, the recommendation of more than two models.

The term in question has been misapplied and will be removed from the revised manuscript. Initially, it was employed to denote its relevance in assessing hydrological models within the framework of "hyper resolution modeling".

L129-131: Please elaborate as to why additional calibration is not needed. I do not understand. Is it so that model parameters are somehow known already? Comparing uncalibrated models to calibrated ones is unfair in my opinion.

It's important to reframe the study to underscore that the authors are model users, not developers. We applied the models in accordance with the developers' intended use, meaning they were already calibrated and ready for direct application. Thus, evaluating the advantages of further calibration remains relevant. Although this comparison might appear biased, our curiosity lies in assessing

whether enhanced model performance remains meaningful when considering discharge observation uncertainty.

L145-151: Interesting. I am glad that this study relies so much on other's work and doesn't try to reinvent the wheel.

We agree with the notion that adopting or relying on other's qualitative work should be done when possible.

L161-174: Here, I have a major problem with this study. The problem being that input uncertainty is not taken into account. Normally, observation locations are not enough to capture the point of the maximum precipitation which consequently leads to underestimation (in some cases overestimation) of the precipitation volume. This problem was demonstrated recently in Bárdossy and Anwar 2023. From the methodology explained here, I do not see any mention/treatment of this major problem till now.

To address this significant concern, we will implement two key changes. Firstly, we will reframe the study to focus on discharge observation uncertainty, which is more accurate in capturing the essence of our investigation. Secondly, we acknowledge the necessity of addressing the aforementioned issue explicitly. We will introduce a discussion on this major problem in both the introduction and methodology sections. This will include a reflection on the challenges of performing demonstrations such as in Bárdossy and Anwar (2023) on a large-sample of catchments. Additionally, a dedicated section in the discussion will encompass a comprehensive reflection on all sources of input uncertainty. This will emphasize that our study constitutes just a fraction of a broader challenge, in which input uncertainty plays a substantial role.

Also, I saw in the CAMELS-GB paper that they give a mean value of precipitation over the catchment. This is problematic, if used for a distributed model.

Our study used distributed precipitation, temperature, and potential evapotranspiration fields as inputs. The CAMELS-GB paper used the same data to derive a single mean value per catchment.

However, dealing with input, model and discharge uncertainty is an ill-posed problem and doesn't seem to have any acceptable solution, as far as I know. A study that deals with observation uncertainty and leaves out precipitation will, in my humble opinion, lead to incorrect/invalid conclusions.

One could argue that all models are presented with the same input, and therefore it is not much of a problem. But then, why consider observation uncertainty as all models are evaluated based on the same discharge?

The point you raise about the uniformity of input might suggest that discharge observation uncertainty is less significant. However, it's important to note that while meteorological inputs might be similar, parameterization uncertainties and inherent model uncertainties vary between models. This leads to different propagation of uncertainties. The distinction in uncertainty propagation necessitates the consideration of discharge observation uncertainty, even when evaluations are based on the same observations.

Slightly off topic. I think the readers would benefit if the temporal uncertainty methods are summarized like other topics previously. Using words like bootstrapping and jackknifing are not so helpful. After all, it is what the study is about.

A short summary that does not rely on abstract terms would definitely benefit the reader. This will be adjusted in the revised manuscript.

L175-181: Don't all models struggle with the upper 5% of the distribution? I find it disconcerting that such an important detail is left out and is only mentioned now. These are the flows that cause actual problems; this is where major timing and volume problems exist and these are the time steps where the squared error dominates the objective function generally. I understand that not enough data was available but leaving the good stuff out is akin to ignoring the major problem at hand. Such details should be mentioned in the abstract as many are interested in the upper 5%. The low flows, I can forgive as they are contaminated by wastewater flowing in to the river which may or may not be originating, in terms of source, from the same catchment.

This is a very valid point, we will rectify this by extending the analysis to include the limited set of catchments simulations that contain estimates for the upper percentiles of the uncertainty distribution. We appreciate your perspective on the significance of this detail, especially given that these flows often pose real-world challenges. We agree that not addressing this aspect can be misleading and downplays a major issue.

Furthermore, we acknowledge the potential for this analysis to underscore both the importance of the temporal sampling issue and the relevance of considering the upper 5% in hydrological modelling assessments.

L183: What is model A and B in figure 2A? Are 1A and 2A showing the same event? There is no value of discharge on the y-axis. How is one supposed to tell, say, whether there was an actual peak when model B also shows a peak or just that model B spontaneously rose to a high value? And given that A is not as reactive as B, my guess is that something is very wrong with A.

To clarify, this figure was intentionally designed as an extreme example to illustrate our methodology, without cherry-picking specific events. We do recognize the need for better contextualization. In response to your feedback, we will enhance the figure's caption by explicitly stating that the illustration is an extreme case intended for methodological exemplification. This clarification will help readers better understand the purpose and context of the figure.

3. Results

L214-215: I find such a comparison to be meaningless. wflow_sbm default has some quasi-arbitrary parameters. These could have performed very good or very bad. In my opinion, if one has to take one single model, the calibrated model is the one because it the best we could do, assuming that the validation also shows improvement. Also, only one parameter was optimized. Which I find strange. If there is access to a supercomputer then, why not all (that may be optimized)? It would be interesting to see. I have no attachment with wflow_sbm or PCR-GLOBWB, but some sort of parameter optimization should also be done for this. At the end, it could be what the authors point out about how it routes flow. We would only know if optimization is carried out.

As mentioned earlier, this comes down to the distinction between model users and developers. We employed the models as intended by the model developers. The models are readily calibrated and already optimized on HPC infrastructure. The reason for optimizing only a single parameter for the wflow_sbm model is that this was identified as an effective parameter for calibration that increases baseflow and decrease peak flow (Imhoff et al., 2020; Aerts et al., 2022). We find it therefore of interest to see what this additional calibration step entails from a model user perspective.

L253-255: The relative low flow uncertainty could be higher due to wastewater being introduced into the streams as I mentioned earlier. And could also be due the presence of karst that is mentioned by the authors.

Thank you for this insight, we will search for wastewater estimates to confirm this notion. If not available, we will investigate the relationship between human influence and low flow uncertainty as

both are available in the CAMELS-GB dataset. The results of this additional analysis will be reported in the revised manuscript.

L258-267: Here, I would like to stress again that without considering input uncertainty, the conclusions are incorrect as the results are (heavily) influenced by input (precipitation). And the authors should acknowledge this.

We recognize the significant influence of input, particularly precipitation, on our results. To address this concern, we will include a dedicated section discussing input uncertainty and its broader implications. This discussion will underscore the interconnected nature of discharge observation uncertainty within a larger context.

4. Discussion:

L269-274: The authors finally mention the other sources of uncertainty this deep in the text. The reason why this problem is over-looked (by hydrologists that know about this problem) is that evaluating uncertainties is not trivial and requires much more data (which Coxon (2015) had the luxury of to some extent) computational power, and many assumptions (that likely remain unfulfilled or cannot be verified to hold). Working with uncertain data has been tried before but all end up at the same point i.e., if uncertainties have to be handled then the proper way is to take all types in to account simultaneously. This is a major problem. Uncertainty bounds of any variable are calculated, normally, using Gaussian-dependence. For precipitation, for example, advection and convection exists. Something that interpolation schemes cannot capture by considering only a subset of points in the catchment. Also, they are non-Gaussian fields. Radar shows some structure of the precipitation field but is also limited in its capabilities when it comes to precipitation volumes and is not always better than using gauge data.

Thank you for this view on working with uncertain data, especially in the context of precipitation. The suggested literature by Bárdossy and Anwar clearly demonstrates to us how this impacts rainfall-runoff models. While acknowledging the limitations of current uncertainty estimation methods, we recognize the value of data sources such as CAMELS-GB in exploring uncertainties. Despite relying on assumptions like Gaussian mixture models, these sources offer valuable insights into the broader uncertainty landscape. In our discussion chapter, we will provide a more comprehensive reflection on these limitations, as well as the importance of incorporating all relevant sources of input uncertainty.

References:

Aerts, J. P. M., Hut, R. W., van de Giesen, N. C., Drost, N., van Verseveld, W. J., Weerts, A. H., & Hazenberg, P. (2022). Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model. *Hydrology and Earth System Sciences*, *26*(16), 4407–4430. https://doi.org/10.5194/hess-26-4407-2022

Bárdossy, A., & Anwar, F. (2023). Why do our rainfall–runoff models keep underestimating the peak flows? *Hydrology and Earth System Sciences*, *27*(10), 1987–2000. https://doi.org/10.5194/hess-27-1987-2023

Beven, K. J., & Cloke, H. L. (2012). Comment on “Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth’s terrestrial water” by Eric F. Wood et al. *Water Resources Research*, *48*(1). https://doi.org/10.1029/2011WR010982

Bierkens, M. F. P., Bell, V. A., Burek, P., Chaney, N., Condon, L. E., David, C. H., de Roo, A., Döll, P., Drost, N., Famiglietti, J. S., Flörke, M., Gochis, D. J., Houser, P., Hut, R., Keune, J., Kollet, S., Maxwell, R. M., Reager, J. T., Samaniego, L., … Wood, E. F. (2015). Hyper-resolution global hydrological modelling: What is next? *Hydrological Processes*, *29*(2), 310–320. https://doi.org/10.1002/hyp.10391

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Knoben, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., & Papalexiou, S. M. (2021). The Abuse of Popular Performance Metrics in Hydrologic Modeling. *Water Resources Research*, *57*(9), e2020WR029001. https://doi.org/10.1029/2020WR029001

Imhoff, R. O., van Verseveld, W. J., van Osnabrugge, B., & Weerts, A. H. (2020). Scaling Point-Scale (Pedo)transfer Functions to Seamless Large-Domain Parameter Estimates for High-Resolution Distributed Hydrologic Modeling: An Example for the Rhine River. *Water Resources Research*, *56*(4), e2019WR026807. https://doi.org/10.1029/2019WR026807

Jain, S., Mindlin, J., Koren, G., Gulizia, C., Steadman, C., Langendijk, G. S., Osman, M., Abid, M. A., Rao, Y., & Rabanal, V. (2022). Are We at Risk of Losing the Current Generation of Climate Researchers to Data Science? *AGU Advances*, *3*(4), e2022AV000676. https://doi.org/10.1029/2022AV000676

Lamontagne, J. R., Barber, C. A., & Vogel, R. M. (2020). Improved Estimators of Model Performance Efficiency for Skewed Hydrologic Data. *Water Resources Research*, *56*(9), e2020WR027101. https://doi.org/10.1029/2020WR027101

Wood, E. F., Roundy, J. K., Troy, T. J., van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., de Roo, A., Döll, P., Ek, M., Famiglietti, J., Gochis, D., van de Giesen, N., Houser, P., Jaffé, P. R., Kollet, S., Lehner, B., Lettenmaier, D. P., Peters-Lidard, C., Sivapalan, M., … Whitehead, P. (2011). Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, *47*(5). https://doi.org/10.1029/2010WR010090