Point-by-point response Review #1

Dear Prof. Beven,

We wish to express our gratitude for your comprehensive and insightful review of our manuscript. Your feedback has been invaluable, serving not only as a guide for substantial revisions but also as a catalyst for enhancing the manuscript's overall clarity, depth, and scientific value. Below, we first detail the extensive revisions made followed by a point-by-point response to your comments.

Clarification of the study's aim, positioning, and structure:

- The title is revised to "On the importance of discharge observation uncertainty when interpreting hydrological model performance" and the publication only refers to "observation uncertainty" as discharge observation uncertainty. In addition, the words "outliers" and "return flow" are removed from the publication.
- The study's aim is clarified by clearly stating that this study is performed to provide model users with an assessment of the effect of omitting discharge observation uncertainty while interpreting model performance differences. *L1-9, L83-85*.
- To strengthen this point, 3 use cases are introduces that also improve the structure of the publication. *L103-115*
- The publication states and the results emphasize that discharge observation uncertainty and input uncertainties need not be included in model calibration and model evaluation efforts. *L20-23, L373-378, L418-420, L452-455*
- The various sources of uncertainty within hydrological modelling and the concept of equifinality are now extensively discussed in the introduction (*L26-59)* and discussion *(L373-379)*.
- The relevance of conducting a temporal sampling of the simulation and observation pairs for model performance interpretation is clarified. *L351-354, L399-413.*
- Input uncertainty is mentioned in the introduction and the importance of including input uncertainties is stressed in the discussion section.
- A section in the discussion reflects on the practical implications of the results for model users. L415-425.222
- A section is added in the discussion that reflects on the limitations of this study. *L426-435*

Methodological Revisions:

- Catchments are excluded from the analysis that indicate temporal discretization errors in small catchments due to peak precipitation and peak discharge occurring at the same time step. This is done by calculating the cross-correlation between observed discharge and precipitation for a range of lag times. Catchments that predominantly show less than 1 day of lag between observed discharge and precipitation are excluded.
- A new use case is introduced that compares the model structural uncertainty of six additional conceptual models. The model structural uncertainty is expressed in the maximum model discharge simulation difference and is subsequently compared to the discharge observation uncertainty estimates. *L273-296, L329-368*

General revisions:

- Improvements to overall written text.
- We extended the body of literature.
- We extended conclusions.

- Improvement of Figure 2 by selecting a more relatable example.
- Addition of model structure use case to Figures 3,4 and Table 2.
- Addition of Figure 6, comparing six individual conceptual hydrological models.
- Addition of Figure 7, indicating the effects of temporal sampling uncertainty for 6 conceptual hydrological models.

Point-by-point response:

- *"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."*
  "Input uncertainties should not be ignored, as has been done here, since they can be an important source of disinformation in calibration data sets."

- We mention the various uncertainty sources, there potential implications for hydrological modelling, and the equifinality concept at the beginning of the introduction and in the discussion section. The publication states and the results emphasize that discharge observation uncertainty and input uncertainties need not be included in model calibration and model evaluation efforts. *L20-23, L373-378, L418-420, L452-455*

- *"The authors refer to "outliers", but the authors do not differentiate between whether outliers might be the most important events in distinguishing models (e.g. Singh and Bardossy, AWR 2012), or whether they might in some cases introduce disinformation into the calibration process (e.g. the runoff coefficients greater than 1 of Beven and Smith, JHE ASCE 2015; Beven, PRSL 2019; Beven et al., HP 2022, 2023)."*

  We removed the term "outliers" as well as the term "return flow". In addition, we have included the suggested publications and mention the invalidation sites concept.

- *"Is the temporal sampling issue (in terms of different sampling periods) really relevant? Should we not give models the maximum chance to fail over extremes (after assessing for possible disinformation) by using as much data as possible (Shen et al., WRR 2022 also recently suggested this as the most robust calibration for use in prediction)."*

  We have improved the justification and clarification of the temporal sampling uncertainty analysis. *L351-354, L399-413.*

- *"There are, however, other temporal sampling issues in terms of the discretisation error of using daily time steps on some rather small UK basins. This really should be taken into account in that for some events it might be more significant than the rating curve error depending if an observed peak falls on one day or another relative to the model prediction."*

  Catchments are excluded from the analysis that indicate temporal discretization errors in small catchments due to peak precipitation and peak discharge occurring at the same time step. This is done by calculating the cross-correlation between observed discharge and precipitation for a range of lag times. Catchments that predominantly show less than 1 day of lag between observed discharge and precipitation are excluded. *L124-230*

- *"The authors recognise, as in other large sample modelling studies, that there is a significant percentage of catchments for which models perform badly. In terms of considering the potential impacts of observation uncertainties, why is this not taken more seriously (it has also been ignored in all the recent machine learning studies). Is it not important to learn why that is the case (and no it is not all down to chalk catchments – and if your perceptual model already informed you that your models would not perform well on chalk catchments why did you make the applications? Needs justification, and not just because everyone else has included as many catchments as possible)"*

  The results section is restructured and we now refrain from discussing individual model performance as this is outside of the scope of the study. Nonetheless, we mention this as a limitation in the added limitations section in the discussion.

- "L45. Correlation is not causality – it might actually be better to search for understanding at the local level."

  We agree and removed this statement.

- *"L130. Does not require additional calibration? Why not – surely it could benefit????? Should you not just say it was applied without additional calibration. And why is a 30 year steady state based on average daily values an appropriate initial condition for the start of 2008? Would not seem appropriate for either baseflow dominated (chalk) catchments or flashier catchments? OK, at least 2008 was discarded so not too important."*

  We reframed this statement making it clear that this is common practice by the model developers.

- *"L186/187. This is really unclear – averages of uncertainty bounds? Why do not these come simply from the rating curve uncertainties at each time step?"*
  "L189/190. T-test? But these are not independent values?"

  Both are clarified in the methodology and the absence of rating curve uncertainties is mentioned in the limitations.

- *L228. The relevance of which is debatable? Really??? The models are really poor for these sites – THAT is important - it is he relative values of just how bad are that is not so relevant."*

  Very ill-posed and removed from the publication.

The detailed revisions outlined above, undertaken in direct response to your feedback, have improved the manuscript's clarity, depth, and scientific value. We believe these revisions have comprehensively addressed your concerns, contributing to a manuscript that offers valuable insights and advancements to the hydrological modeling community.

On behave of the co-authors,

Jerom Aerts

Point-by-point response Review #2

Dear Reviewer,

We extend our sincerest thanks for your thorough review and the valuable feedback provided on our manuscript. Your detailed comments have prompted us to undertake a comprehensive revision process, significantly improving the manuscript's clarity and scientific value. In what follows, we detail the revisions made in response to each of your comments.

Clarification of the study's aim, positioning, and structure:

- The title is revised to "On the importance of discharge observation uncertainty when interpreting hydrological model performance" and the publication only refers to "observation uncertainty" as discharge observation uncertainty. In addition, the words "outliers" and "return flow" are removed from the publication.
- The study's aim is clarified by clearly stating that this study is performed to provide model users with an assessment of the effect of omitting discharge observation uncertainty while interpreting model performance differences. *L1-9, L83-85*.
- To strengthen this point, 3 use cases are introduces that also improve the structure of the publication. *L103-115*
- The publication states and the results emphasize that discharge observation uncertainty and input uncertainties need not be included in model calibration and model evaluation efforts. *L20-23, L373-378, L418-420, L452-455*
- The various sources of uncertainty within hydrological modelling and the concept of equifinality are now extensively discussed in the introduction (*L26-59)* and discussion *(L373-379)*.
- The relevance of conducting a temporal sampling of the simulation and observation pairs for model performance interpretation is clarified. *L351-354, L399-413.*
- Input uncertainty is mentioned in the introduction and the importance of including input uncertainties is stressed in the discussion section.
- A section in the discussion reflects on the practical implications of the results for model users. L415-425.222
- A section is added in the discussion that reflects on the limitations of this study. *L426-435*

Methodological Revisions:

- Catchments are excluded from the analysis that indicate temporal discretization errors in small catchments due to peak precipitation and peak discharge occurring at the same time step. This is done by calculating the cross-correlation between observed discharge and precipitation for a range of lag times. Catchments that predominantly show less than 1 day of lag between observed discharge and precipitation are excluded.
- A new use case is introduced that compares the model structural uncertainty of six additional conceptual models. The model structural uncertainty is expressed in the maximum model discharge simulation difference and is subsequently compared to the discharge observation uncertainty estimates. *L273-296, L329-368*

General revisions:

- Improvements to overall written text.
- We extended the body of literature.
- We extended conclusions.

- Improvement of Figure 2 by selecting a more relatable example.
- Addition of model structure use case to Figures 3,4 and Table 2.
- Addition of Figure 6, comparing six individual conceptual hydrological models.
- Addition of Figure 7, indicating the effects of temporal sampling uncertainty for 6 conceptual hydrological models.

Point-by-point response:

- *"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."*

  The title has been adjusted.

- *"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 have removed this statement from the publication as we agree with your comment.

- *"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?"*

  We have improved the justification and clarification of the temporal sampling uncertainty analysis. *L351-354, L399-413.*

- *"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."*

  We have added 6 conceptual hydrological models to the analysis.

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

  We have removed the use of intra- and inter-model cases and substituted these for use cases for clarification and better structure.

- *"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."

Catchments are excluded from the analysis that indicate temporal discretization errors in small catchments due to peak precipitation and peak discharge occurring at the same time step. This is done by calculating the cross-correlation between observed discharge and precipitation for a range of lag times. Catchments that predominantly show less than 1 day of lag between observed discharge and precipitation are excluded. *L124-230*

- *"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."*

  The description of calibration and justification is now better described in the methods section.

- *"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."*

  The various sources of uncertainty within hydrological modelling and the concept of equifinality are now extensively discussed in the introduction (*L26-59)* and discussion *(L373-379)*. In addition, input uncertainty is mentioned in the introduction and the importance of including input uncertainties is stressed in the discussion section.

- *"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."*

  We now clearly state throughout the publication that the lower and upper 5% of flow are not considered in the analysis. In addition, this is mentioned in the added limitations section in the discussion. *L426-435*

- *"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."*

We have selected a more relatable example hydrograph for this exemplary figure in the methodology.

The results section is restructured and we now refrain from discussing individual model performance as this is outside of the scope of the study.

The detailed revisions outlined above, undertaken in direct response to your feedback, have improved the manuscript's clarity, depth, and scientific value. We believe these revisions have comprehensively addressed your concerns, contributing to a manuscript that offers valuable insights and advancements to the hydrological modeling community.

On behave of the co-authors,

Jerom Aerts