Reply to Review #1

First and foremost, we would like to express our sincere gratitude to Prof. Beven for the valuable insights and constructive comments on our publication.

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.

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.

Point-by-point response:

"Observation uncertainties should not be separated from model calibration as has been done here – since any optimal parameter set depends on the particular sequence of errors in the observations (as well as whatever objective function is used). In fact it is pretty obvious that allowing for ANY source of observation error as part of the calibration process means that the concept of an optimal parameter set has no real meaning (as discussed at least since 1989)."

We agree with the importance of incorporating observation uncertainties in model calibration, as emphasized by Beven & Binley (1992), Beven & Freer (2001), Beven & Smith (2015), McMillan et al., 2018, Beven & Lane (2019), and Westerberg et al. (2022). It is crucial to note that the hydrological models employed in our study were initially calibrated by the model developers. As users, our primary focus is not solely on highlighting shortcomings in the models. Instead, our main emphasis lies in identifying and highlighting instances where differences between model outputs fall within the bounds of streamflow observation uncertainties.

We appreciate the reviewer's attention to this aspect and will ensure that the framing and messaging of our publication clearly reflect the objectives of our study. In addition, we will include the necessary literature to support this view in the discussion section of the publication.

"*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 appreciate the reviewer's acknowledgment of the importance of considering input uncertainties in the calibration process. The works of among others Kavetzki et al. (2006a, 2006b) and Westerberg et al. (2022) indeed highlight the significance of not disregarding input uncertainties. It is true that including input uncertainties in the calibration process is important when aiming to demonstrate the best possible hydrological model setup. Given the focus of this publication and the absence of input uncertainties (i.e. rainfall and PET) for large-sample hydrology studies, we employed the perspective

that using the same inputs for both models can help minimize the impact of these uncertainties. By forcing both models with the same set of uncertainties, we create a more direct comparison. Therefore, we promote the use the same forcing inputs for other hydrological modelling efforts that use the CAMELS-GB dataset as the case study.

We want to emphasize that the consideration of input uncertainties remains an important aspect that should not be overlooked. While using the same inputs for both models helps reduce the potential bias introduced by different input assumptions, it is still valuable to acknowledge and discuss the implications of these uncertainties in the context of model comparison and evaluation. In the revised version of our publication, we will make sure to explicitly address the role of input uncertainties in the discussion. In addition, we will use the term "observed discharge uncertainty" instead of " observation uncertainty" throughout the revised manuscript.

"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 acknowledge the concerns raised regarding the term 'outliers' and its potential negative connotation in the context of science and hydrology. Therefore, we will consider adopting an alternative term to refer to these data points. Denoting 'outliers' with 'heavy tails' instead, better describes the presence of heavy tails in the residuals of the squared error distribution of the observation and simulation pairs. In addition, we want to clarify that our intention is not to downplay the importance of these data points. We rather emphasize the need for careful inspection before drawing conclusions based solely on streamflow-based model performance. In our publication, these data points represent the heavy tails of the residuals between observation and simulation pairs. It is important to note that these points can have a disproportionally significant effect on the objective function, as discussed in Clark et al. (2021). As highlighted in the review, these data points can indeed often hold important information and serve as the most important events in distinguishing models and the introduction of disinformation.

To address this issue more comprehensively, we will adjust the terminology to 'heavy tails' in the introduction and discussion sections of our paper. By more critically reflecting on the nature and implications of these specific data points, we aim to present a clearer and more accurate representation of our findings. Additionally, we will further support our discussions with references to relevant literature such as Lamontagne et al. (2020) and Shabestanipour et al. (2023).

"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 agree that giving models the maximum chance to fail over extremes is fair. We achieve this by evaluating the models over a prolonged period. Nevertheless, we must not overlook the relevance of temporal sampling when comparing hydrological models. As pointed out by Fowler et al. (2018), equally weighting each year in the calibration data 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, 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 streamflow data.

Therefore, we argue that it is indeed important to demonstrate and account for the impact of temporal sampling when performing extensive hydrological model evaluations. By considering this aspect, we can obtain a more comprehensive and reliable understanding of model performance across diverse hydrological conditions.

"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."

Indeed this is clearly present in the hydrological model results. We will more critically reflect on this in the results and discussion of the publication. In our analysis, we encountered that the wflow_sbm model, originally developed for small-scale applications, is less affected by the temporal discretisation error in small UK basins compared to the larger scale PCR-GLOBWB model. We recognize the significance of this finding, and we will emphasize it in the results section.

Additionally, we will identify and report the basins with large temporal discretisation errors. As suggested by the second reviewer we will implement a procedure that highlights catchments where the precipitation peak and streamflow peak coincide in the same time step.

"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)"

In our publication, we do indeed highlight instances where the model performance is unsatisfactory, leading to the conclusion that the models are not fit for purpose in these specific catchments. We believe it is crucial to present not only the success stories but also the complete picture of model performance when evaluating using large-sample datasets. By doing so, we provide a more accurate representation of the models capabilities and limitations.

Regarding the application of models to catchments where the perceptual model suggests potential poor performance, we acknowledge the importance of justification. While it is true that the model developers perceptual model might have indicated potential limitations for certain catchments, we believe there is value in exploring these cases further. By giving models the opportunity to fail, we allow for more comprehensive model evaluation and gain valuable insights into the reasons behind their poor performance. As pointed out earlier, these cases of seemingly poor model performance can be the most distinguishing and informative results. Our study, therefore, identifies potential areas for model improvement and guides future research efforts. We will perform further analysis to better understand the reasons why the models fail in particular catchments using a wider range of catchment descriptors.

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

We agree with this statement and will adjust this in the revised manuscript by reflecting on the fact that correlation is not causation.

"L49. Some important papers missing here - Beven and Smith JHE ASCE 2015; Beven HSJ 2016, PRSL 2019, Beven and Lane HP 2022, Beven et al HP 2022. L52. There were earlier papers – e.g. Liu et al JH 2009; Blazkova and Beven, WRR, 2009"

We will extend the referenced literature based on these suggestions and will extend the literature further.

"L114. Aggregated how? In a way consistent with the hydrological process descriptions?"

The hydraulic parameters are upscaled using the method presented in Eilander et al. (2021). The parameter upscaling of the wflow_sbm model is based on the work by Imhoff et al. (2020) that used point-scale (pedo)transfer-function following the MPR technique by Samaniego et al. (2010). Parameters are aggregated from the original data resolution with upscaling operators determined by a constant mean and standard deviation across different scales. Fluxes and states are checked for consistency. See van Verseveld et al. (2022) for further information. We will better clarify the parameterization and aggregation of parameters in the revised manuscript.

"L118/119. Return flow has a specific meaning in hillslope hydrology that is different? And surely water use and water demand are not included in this data set so are not used here?"

We will refrain from using the term "return flow". Only a few large catchments in the Thames basin and parts of Scotland model water use and water demand by the PCR-GLOBWB model. This will be included in the Appendix of the revised manuscript.

"L124. Respectively is the wrong way round!"

Thank you, this will be adjusted in the revised manuscript.

"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 agree that (any) model can be benefit from additional calibration. We implemented the model as intended by the model developers. Admittedly, the 30-year simulation period might be considered relatively short for reaching steady state conditions in certain hydrological systems, if not all. As mentioned, to address this issue we remove the year 2008 from our analysis. By doing so, we aimed to minimize any potential transient effects.

In our revised publication, we will explicitly state that the model has been calibrated by the developers. Furthermore, we will acknowledge the potential benefits of conducting additional calibration, which could lead to further improvement in the model results. By providing this clarification, we aim to ensure transparency and emphasize the importance of considering calibration choices in our study.

"L140 why would you expect that lateral Ks should be much greater than vertical Ks? Are not most macropores in the near surface vertical. Is this an indication that the process representations are inadequate and sufficient to reject the model (e.g. subsurface celerities not being handled properly). And does a value of 100 already mean 100x or a factor of 1. Needs more discussion/clarification."

We do not expect lateral Ks to be much greater than vertical Ks. This is a severe limitation of the wflow_sbm model that highlights deficiencies in process representations. Therefore, we do not expect an amplification factor to be necessary to correct the model. One notable concern is the reliance on a topographic gradient drive procedure, which may not fully account for the complexities of pressure

driven flow in hydrological system. Additionally, the models lack of preferential flow representation can lead to unrealistic parameter values that lack physical meaning.

In our study, we take the role of model users and aim to distinguish ourselves from the model developers by addressing these limitations. We will more explicitly describe the model calibration approach in the methods section.

"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?"

This is due to the data limitations that are available in the CAMELS-GB dataset. We accept this limitation as we promote the use of existing dataset to ensure community participation into implementing the suggested evaluation procedure in other studies.

"L189/190. T-test? But these are not independent values?"

We applied a pair-wise T-test for dependent values. We will clarify this in the methods.

"Figure 3. Something seems wrong here. On both plots the calibrated model has worse values than the default model for the values > 0"

In Figure 3A, we can observe a slight overlap and lower performance among the wflow_sbm models. This can be attributed to two causes.

Firstly, the optimal parameter value obtained during calibration are often found to be similar to the default value for certain parameters. This suggests that the model calibration may not significantly improve the performance over the default configuration. The ever so slightly lower performance is due to the calibration process, where the model is calibrated using the average of individual water years over the entire calibration period. As a consequence, some parameters may end up with values that do not deviate significantly from their default settings.

Secondly, during the calibration process, we optimized the model for a single objective function (KGE-NP). While this approach allowed us to achieve a favourable performance with respect to the chosen objective, it may lead to variations in other objective functions. The results in Figure 3B demonstrate the duality of optimizing for a single objective function by how it affects the NSE objective function.

Given the importance of capturing multiple aspects of model performance, we found it necessary to include multiple objective functions in our analysis. By doing so, we could gain a more comprehensive understanding of the models capabilities and limitations, accounting for the trade-offs and interdependencies between different metrics.

"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."

We recognize that the term "relevance" might have caused confusion. Our intention was to convey that models failing to outperform a simple benchmark, such as taking the mean of the observed flow, should be excluded from the analysis. In other words, we aim to exclude models that do not demonstrate a meaningful improvement over the most basic representation of observed flow.

We will be more explicit in the revised manuscript.

"L230. See comments above about outliers and disinformation."

We will implement changes as mentioned in the comments above about outliers and disinformation.

"L254. Not clear here – if you have taken average percentage uncertainties by flow class and multiplied by the flow then how is there such variation?"

Here, we only refer to the uncertainty percentages without multiplication with flow. This will be clarified in the revised publication.

"L265/266 is that not the inverse of what you started at the start of this paragraph?"

Thank you, this will be corrected in the revised manuscript.

"Section 4.1. See comments above about temporal variability, outliers and disinformation"

We will implement similar changes as mentioned in the comments above about temporal variability, outliers and disinformation.

"L294/295. But equifinality has been discussed for more than 20 years now – but does not get an explicit mention in the text anywhere?"

We agree that this should be mentioned. We will discuss "equifinality" in the introduction and discussion supported by literature.

"L300/301 So only take your best cases??? is it not more important to understand what is happening at these sites and allow for that understanding in what you use to predict? Might these be model invalidation sites (see Beven and Lane, HP 2022)"

Indeed this is contradictory to what we mentioned earlier in this rebuttal. We will adjust this using the suggested concept and reference regarding model invalidation sites.

"Section 4.3. See Beven HP 2023 benchmarking paper for an alternative view."

We will incorporate the findings of the Beven (2023) commentary in the discussion section.

"L346 experiment"

Thank you for the technical corrections.

References:

Beven, K. (2023). Benchmarking hydrological models for an uncertain future. *Hydrological Processes*, *37*(5), e14882. https://doi.org/10.1002/hyp.14882

Beven, K., & Binley, A. (1992). The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes*, *6*(3), 279–298. https://doi.org/10.1002/hyp.3360060305

Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*, *249*(1), 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

Beven, K., & Lane, S. (2019). Invalidation of Models and Fitness-for-Purpose: A Rejectionist Approach. In C. Beisbart & N. J. Saam (Eds.), *Computer Simulation Validation: Fundamental Concepts, Methodological Frameworks, and Philosophical Perspectives* (pp. 145–171). Springer International Publishing. https://doi.org/10.1007/978-3-319-70766-2_6

Beven, K., & Smith, P. (2015). Concepts of Information Content and Likelihood in Parameter Calibration for Hydrological Simulation Models. *Journal of Hydrologic Engineering*, *20*(1), A4014010. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000991

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

Eilander, D., van Verseveld, W., Yamazaki, D., Weerts, A., Winsemius, H. C., & Ward, P. J. (2021). A hydrography upscaling method for scale-invariant parametrization of distributed hydrological models. *Hydrology and Earth System Sciences*, *25*(9), 5287–5313. https://doi.org/10.5194/hess-25-5287-2021

Fowler, K., Peel, M., Western, A., & Zhang, L. (2018). Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function. *Water Resources Research*, *54*(5), 3392–3408. https://doi.org/10.1029/2017WR022466

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

Kavetski, D., Kuczera, G., & Franks, S. W. (2006a). Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005WR004368

Kavetski, D., Kuczera, G., & Franks, S. W. (2006b). Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2005WR004376

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

McMillan, H. K., Westerberg, I. K., & Krueger, T. (2018). Hydrological data uncertainty and its implications. *WIREs Water*, *5*(6), e1319. https://doi.org/10.1002/wat2.1319

Samaniego, L., Kumar, R., & Attinger, S. (2010). Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, *46*(5). https://doi.org/10.1029/2008WR007327

Shabestanipour, G., Brodeur, Z., Farmer, W. H., Steinschneider, S., Vogel, R. M., & Lamontagne, J. R. (2023). Stochastic Watershed Model Ensembles for Long-Range Planning: Verification and Validation. *Water Resources Research*, *59*(2), e2022WR032201. https://doi.org/10.1029/2022WR032201

van Verseveld, W. J., Weerts, A. H., Visser, M., Buitink, J., Imhoff, R. O., Boisgontier, H., Bouaziz, L., Eilander, D., Hegnauer, M., ten Velden, C., & Russell, B. (2022). Wflow_sbm v0.6.1, a spatially distributed hydrologic model: From global data to local applications. *Geoscientific Model Development Discussions*, 1–52. https://doi.org/10.5194/gmd-2022-182

Westerberg, I. K., Sikorska-Senoner, A. E., Viviroli, D., Vis, M., & Seibert, J. (2022). Hydrological model calibration with uncertain discharge data. *Hydrological Sciences Journal*, *67*(16), 2441–2456. https://doi.org/10.1080/02626667.2020.1735638