Response to reviewers

Dear editor, dear reviewers,

Thank you for your work on handling and reviewing our paper. Please find our responses to the reviewers' comments below in blue.

On behalf of all co-authors,

Kind regards,

Wouter Knoben

Reviewer 1

This is my first review of the manuscript, "How many models do we need to simulate hydrologic processes across large geographical domains?" I appreciate the authors' work, which is both relevant and holds significant technical implications for the hydrological community. Sampling uncertainty is a critical issue, well-known to some but often overlooked by many. This study effectively highlights its importance within the context of large-sample hydrology, which, thanks to the widespread availability of the CAMELS dataset, is rapidly gaining traction.

Thank you for your comments. It has been particularly helpful to us that you outlined clearly which part of the methods did not work well on your first readthrough, and which elements were helpful to you to understand what we are doing in this paper. We have tried to clarify the methods to make things easier for future readers.

The introduction is technically strong, well-written, and cites relevant literature. However, I suggest that the authors define "model" within the paper as a specific model configuration rather than a "modeling framework." Clarifying this distinction early on would prevent confusion, ensuring readers understand that the paper focuses on specific model structures rather than broader frameworks.

We have added sentences to clearly identify the use of the word "model" in the paper. See below where we discuss your individual comments.

The methods section, however, falls below the expected standard for this type of paper. I found it challenging to follow the results due to the methods being insufficiently described. Some concepts became clear only later, in the limitations paragraphs of the discussion section. To enhance clarity, I recommend adding a schematic of the procedure and clearly defining key terms like "model equivalent" and "best model," which only become fully understood in the results section. The methods currently feel hastily written; a clearer presentation would significantly enhance the paper's accessibility and relevance (I have included specific comments in the annotated PDF). Additionally, the problem statement is vague in places (notably in the initial statement), and it would be beneficial to restate the specific answers to these questions in the conclusion, effectively closing the problem statement.

The suggestions in this comment can be summarized as follows:

- Improve the methods section, possibly by lifting elements from the first Discussion section into the methods, and/or adding a schematic of the procedure used. Also provide definitions of critical terms.
- Improve the problem statement.
- Return to the problem statement in the Conclusions.

Briefly, we:

- Substantially expanded to the Methods section, adding an explanation of the gumboot procedures, the model selection procedure, and an extra figure describing the most complex steps in the methodology. To the Introduction, we added definitions for the main terms used throughout the manuscript.
- We rewrote the problem statement (research questions) to use the new definitions. These, in turn, are provided immediately before the research questions are presented to the reader.
- We updated the text in the Conclusions to specifically return to the posed questions. Because this comment does not appear again below, we provide the changed text here. Changes in bold:

> Accounting for this sampling uncertainty reveals that it is often very difficult to distinguish among competing models. If we were to select models based on their validation KGE scores only, almost all of the investigated 36 models would be selected 360 for at least a handful of basins. When we account for the sampling uncertainty in the KGE scores**, we find that every model is close to the performance of the best model (i.e., performance-equivalent) in at least 50 and up to 350+ basin (Research Question 1; Fig. 4a). Conversely, we find that in almost all basins at least 2 and up to all 36 models have similar performance (Research Question 2; Fig. 4b). Finally, when we account for the sampling uncertainty in the KGE scores,** the number of models required to get performance-equivalent simulations on all basins drops drastically compared to selecting models based 365 on the best KGE in each basin. Only four models are needed to get simulations with acceptable accuracy in 95% of basins. 100% coverage is obtained with ten models, without accounting for further complicating factors such as data and parameter uncertainty **(Research Question 3; Fig. 5)**. These findings hold for a different objective function aimed at low flow simulation: the KGE of the reciprocal of flows. Compared to almost all models being selected based on validation KGE(1/Q) scores, only eight

> models are needed to cover 95% of basins and 19 models for 100% coverage **(Fig. 6).**

The results are sound, and the discussion is well-articulated and engaging.

Thank you, it's good to hear that our points come across well.

Based on these points, I recommend major revisions prior to publication.

Reviewer 1 – pdf comments

Line 23-27: Given the ability of many current models to work in multiparameterization mode (e.g., Noha-MP, NewAge GeoFrame, Summa) the authors shall mention probably the definition of model here, that I believe refers to a specific version of this large modelling framework.

We have added the requested clarification (changes in bold):

> The need for robust predictions of water availability and threats across large spatial scales (i.e., national, continental, global) requires models that work well across a variety of landscapes, discretizations, and purposes. There are two main streams of thought on how this can be achieved. The first is the idea that a single model **instantiation (i.e., a single set of equations)** will be able to give accurate predictions everywhere, and that the main challenges in large-domain modeling are related to our ability to parametrize, initialize, configure and run models at ever finer resolutions (e.g., Freeze and Harlan, 1969; Wood et al., 2011; Bierkens et al., 2015; Arheimer et al., 2020). The second is the idea that there are limits to our ability to measure and model the real world, suggesting that the main challenges in large-domain modeling are related to our ability to select and parametrize appropriate models for different places under varying data availability (e.g., Kirchner, 2006; Clark et al., 2011, 2016, 2017; Addor and Melsen, 2019; Horton et al., 2022). This is sometimes referred to as the "uniqueness of place" (Beven, 2000)**, and in modeling terms suggests that one will need different models (i.e., different sets of equations) in different places depending on each location's dominant hydrologic processes.**

Line 62-64: This is much less true today. With Earth Observation data chances are that this limitation is attenuated (but not disappeared of course).

We have clarified this statement. While we do have access to vast amounts of geospatial data, what we lack is a coherent way to derive a region's dominant hydrologic processes from this data, and understanding of the dominant processes is the key piece that's missing for model selection. Changes in bold:

> First, it is possible to use detailed understanding of a specific place to refine perceptual models of that location's hydro- logic behaviour (e.g., Mcglynn et al., 2002) and from such understanding derive models that strike an appropriate balance between realism and accuracy of the resulting simulations (e.g., Kirchner, 2006; Fenicia et al., 2008, 2016). However, despite encouraging progress on synthesis efforts of perceptual models used by hydrologists (McMillan et al., 2023), we currently lack a detailed understanding of **how** the spatial variability in the drivers of hydrologic behaviour (i.e., climate, topography, land cover and subsurface properties) **translates to the spatial variability of dominant hydrologic processes across large domains**. This prevents the use of these model development approaches for geographical domains much larger than individual research basins.

Line 113-114: Can you be more precise here. It is a bit difficult to grasp

We have tried to clarify this statement by adding a different explanation in addition to what was already there. The first explanation seems clear to us (but clearly was not to the reviewer), so it seems best to keep both. Changes in bold:

> For example, Clark et al. (2008) show that for a specific basin, only 10 out of approximately 4000 time steps contribute some 70% of the total model error **(or, in other words, that 70% of the total model error is concentrated in 0.25% of the time steps)**.

Line 132-133: a bit cryptic. I do not understand it. Can you improve it?

Based on this and your other comments, we have added three key definitions to the start of Section 1.3 Problem Statement, before the research questions are introduced. This has necessitated some forward-referencing to the methodology section where the calculation of the uncertainty intervals is described, but it seems more helpful to have these definitions here and leave the details till later. New text:

> In the remainder of this section and the paper, we rely on the following definitions:
>
> - Unless otherwise specified, *"best model"* for a given basin refers to the model with the highest performance score during model validation. Here, that performance score is the KGE.
> - *"Uncertainty bounds"*, *"(sampling) uncertainty interval"* and related terms refer to the 5th to 95th objective function sampling uncertainty interval calculated for the *best model* (details on how this is done can be found in Section 2).
> - *"Performance-equivalent"* and related phrases refer to any model with a validation KGE score that is within the *uncertainty bounds* of the *best model* in a given basin. In other words, *performance-equivalence* is meant to indicate that when objective function sampling uncertainty is considered, two models are

> effectively indistinguishable in terms of their performance scores because one score falls within the uncertainty interval of the other.

Given the newness of the work we are doing, we have also added alternative phrasing to the research questions, so that the reader is given two possible ways to understand the point of each question. Changes in bold:

> **These definitions of sampling uncertainty may not be intuitive. In an effort to enhance clarity, we therefore phrased each research question twice using different combinations of the definitions listed above:**
>
> - For a given model, in how many basins does that model's performance score fall within the uncertainty bounds of the best model in that basin? **Phrased differently, in how many basins is each model performance-equivalent with the best model in that basin?**
> - For a given basin, how many models show performance scores within the uncertainty bounds of the best model in that basin? **Phrased differently, how many models are performance-equivalent in each basin?**
> - What is the minimum number of models needed to obtain simulations with performance **that is within the sampling uncertainty interval of the best model in each basin? Phrased differently, what is the minimum number of models needed to obtain performance-equivalent simulations across the full domain?**

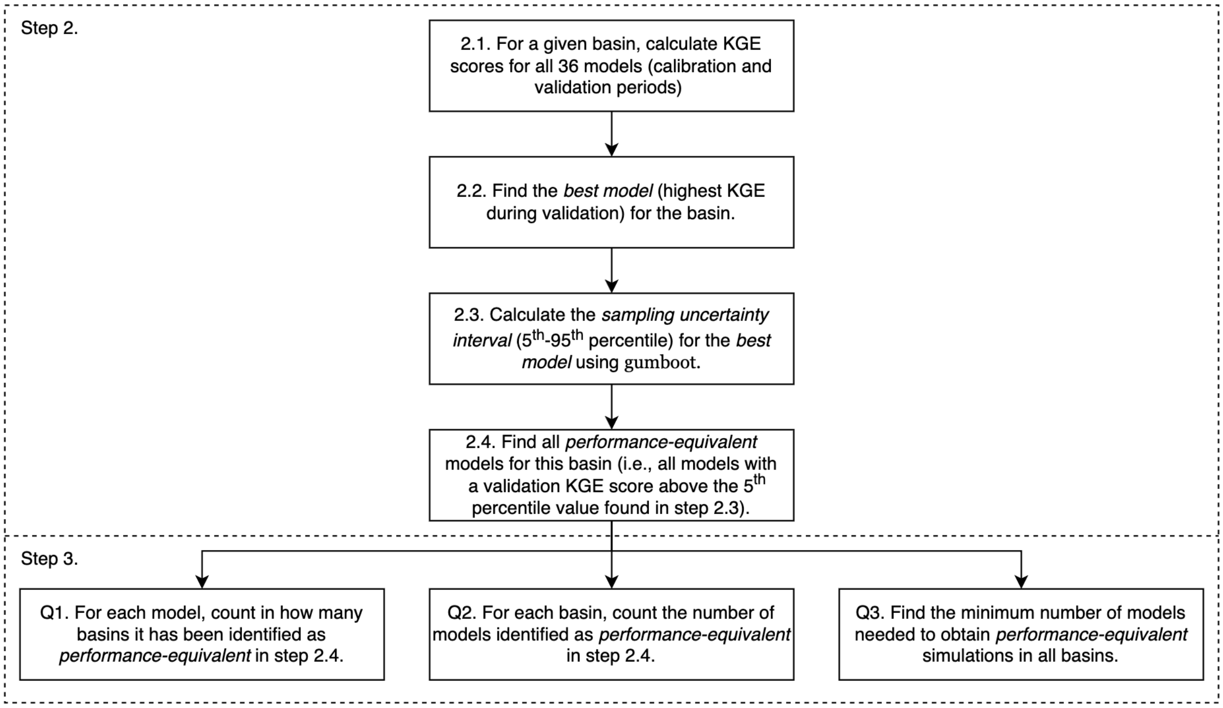
Line 151: provide some at least.

We added some. The interested reader can find the full list by looking at the references we list in the end of this paragraph, but it seems a bit much to expand the paper by the 50 or so citations needed to cover all models. Changes in bold:

> The 36 models mimic existing published models **such as IHACRES (Littlewood et al., 1997), TOPMODEL (Beven and Kirkby, 1979) and HBV-96 (Lindström et al., 1997),** and **thus** cover a wide range of configurations, varying from a simple 1 store (i.e., state variable), 1 parameter bucket model to relatively complex configurations with up to 6 stores and 15 parameters.

Line 157: It would be helpful to include a scheme for the procedure. This is not always straightforward.

We added a figure that provides more detail about steps 2 and 3 in the methodology. Steps 1 (convert streamflow into mm/day) and 4 (repeat steps 2-3 with a different objective function) don't seem as complex as steps 2 and 3, and were therefore not included in an effort to keep the paper uncluttered. New figure:


Step 2.
2.1. For a given basin, calculate KGE scores for all 36 models (calibration and validation periods)
2.2. Find the *best model* (highest KGE during validation) for the basin.
2.3. Calculate the *sampling uncertainty interval* (5th-95th percentile) for the *best model* using gumboot.
2.4. Find all *performance-equivalent* models for this basin (i.e., all models with a validation KGE score above the 5th percentile value found in step 2.3).
Step 3.
Q1. For each model, count in how many basins it has been identified as *performance-equivalent* in step 2.4.
Q2. For each basin, count the number of models identified as *performance-equivalent* in step 2.4.
Q3. Find the minimum number of models needed to obtain *performance-equivalent* simulations in all basins.

Line 164: So this is considered the best model right?

Yes. We changed the text a bit to more clearly use these definitions. Changes in bold:

> For simplicity, we only calculate the KGE uncertainty bounds for **the best model (i.e., the model with the highest validation KGE score)** in each basin, …

Line 168-170: Did you set a threshold on how many models to retain for a basin to be considered meaningful for the analysis. That is if you have only 5 models with score above the 5th of the best model in that basin, did you consider that basin for the analysis?

We did not set such a threshold because how many models are within the uncertainty bounds of the best model is part of the questions we wish to address in this paper. If only 1 (out of 36) model is within the uncertainty bounds that *is* a meaningful finding, because it suggests that in this basin there is a clear best model. We don't perform any statistical analysis that would rely on a certain minimal sample size that could be affected by the number of performance-equivalent models for a given basin. Hopefully the addition of definitions and clarifications to the research questions and methodology have been enough to clarify this.

Line 172-173: Can you be more precise here?

We substantially expanded on the initial description and added a reference to the GitHub repository that contains the exact code we used for this paper. Changes in bold:

> The sampling uncertainty obtained from gumboot provides enough information to answer the first two research questions. To answer the third research question, **we need to identify the minimum number of models needed to get performance-equivalent simulations in each basin. One way to find this minimum combination of models is to iteratively trial every possible combination of models, and identify the first combination of models for which we obtain performance- equivalent simulations in all basins. Such a brute-force approach is guaranteed to be accurate but slow, and proved infeasible for this work. A faster way is to rewrite the problem as a linear programming problem, where the goal is to find the minimum number of subsets needed to provide coverage for the full set. In our case, we have 36 subsets (one for each model) where each subset contains the basin identifiers where a given model is performance-equivalent with the best model. The full set contains the identifiers for all 559 basins, and the optimizer is tasked with finding the smallest number of subsets (i.e., models) needed to cover the full set (i.e., all basins). We refer the reader to our GitHub repository for further implementation details (Knoben, 2024).**

Line 173-175: I have some confusion here as you previously excluded models below 5th percentile but now you calculate the minimum number of models are within the best model uncertainty bounds so I would say that this step is linked with the previous one

Agreed. The changes listed in response to your previous comment, as well as the definitions we now provided earlier in the paper and the new methodology figure, should (hopefully) clarify this issue.

Line 194: Improve the legend. Grey lines are not in it.

The grey lines are merely guidelines that extend the 5th to 95th interval across the full width of the figure. We updated the caption the clarify this. Changes in bold:

> Figure 2. Example basins to illustrate the methodology. (a) Observations and simulations for the gauge with the lowest KGE sampling uncertainty within the 555 tested basins. (b) Model scores during validation, as well as sampling uncertainty ranges for the best model in basin 12035000. **Grey lines show how the uncertainty range compares to each individual model's KGE score.** (c) Observations and simulations for the gauge with the highest KGE sampling uncertainty within the 555 tested basins. (d) Model scores during validation, as well as sampling uncertainty ranges for the best model in basin 08082700. **Grey lines show how the uncertainty range compares to each individual model's KGE score.**

Line 204: Maybe you could unify panel a and c in a single figure representing the uncertainty as size of the marker. This would improve the visualization of the results.

We think the 2x2 layout works well and worry that merging two of the figures into one would lead to information overload for the reader. Because the exact mapping of maximum KGE and associated uncertainty into geographical space (i.e., on a map) is not the focus of this study we prefer to keep the figure as is. The more general pattern that there is a (rough) inverse relation between maximum model performance and associated uncertainty can already be seen in Figure 2d.

Line 205-206: It would have been better to define this in the method section as this concept returns many times below.

Agreed. See responses to your earlier comments. We changed the text here to remove the definition and ensure better flow. Changes in bold:

> Figure 4a shows the number of basins in which each model achieves a performance score that falls with the uncertainty bounds of the best model for each respective **basin or, in other words, the number of basins for which a model is performance-equivalent with the best model in a given basin.**

Line 209-213: Can the bootstrapping have an impact here when it samples from snowy and non-snowy periods?

No, because the bootstrapping samples blocks of full water years. The changes we made to the Methods section in response to reviewer 2's comments clarify this.

Line 218-222: You could exclude snowy dominated basin in a separated experiment

We expect this comment was mostly informed by the lack of clarity about how the bootstrapping works. If the bootstrapping were to sample blocks of less than a year, than it might be biased either towards snow or rain-dominated conditions and thus give a skewed impact of each model's relative suitability for a basin. However, the bootstrapping samples water years (see reply to reviewer 2) and there is thus no chance of a model without a snow module being judged acceptable in a snow-dominated basin: the bootstrap samples of a snow-dominated basin will always include snow accumulation and melt processes. Given this, we think separating basins into snow- and rain-dominated ones will likely make the paper more complicated (and add an arbitrary choice of where to draw the line between snow- and rain-dominated), without adding any clarity. The results wouldn't change. Hopefully our new explanation of the bootstrapping procedure (see reply to reviewer 2) sufficiently addresses this concern.

Line 291: Thanks to this explanation I have clearer now some steps of the paper.

We designed out methodology figure partly with the text in this paragraph in mind. Hopefully this increases clarity for future readers.

Line 309-312: As mentioned, you could do a separated exercise considering only rainy dominated basins.

See above.

Reviewer 2

This study is a synthesis of (1) observed and simulated data from a study using the CAMELS dataset and a subset of models from the Modular Rainfall Runoff Modelling Toolbox (MARRMoT) [1], and (2) the gumboot-methodology for postprocessing the residuals errors models using a mixture of Bootstrap and Jacknife methods [2] of the calibration and validation periods based on NSE and KGE performance metrics. The postprocessing reveals a high variability of the sampling uncertainty among the models. This can be used as an additional criterion to assess the model quality, and it supports the selection process when large domains are modeled with a lower spatial resolution. The results of this study are particularly significant, as the single use of integrated metrics such as NSE and KGE often leads to significant equifinality among potential models, which makes model selection difficult. The statistical method used to analyze differences in performance and sampling uncertainty may improve model selection and, thus, good modeling practice in the future. This study shows evidence of the applicability of the concept for large domains modeled with a lower spatial resolution.

The paper is within the scope and very interesting for the readers of HESS. The authors address a topic of high relevance, which significantly contributes to improving good modeling practice.

The authors have done a commendable job presenting the scientific results concisely and well-structured. I have only minor issues which should be addressed before publication:

Thank you for your comments. It is good to know you see merit in this work. Please see our responses to your individual comments below.

INTRODUCTION:
- I see “Bayesian model averaging and selection” as a paradigm of equal importance as the “single model approach” and the “multi-model mosaic approach”. The latter differs from the more rigorous “multi-model Bayesian paradigm” because it seems based more on professional expertise than statistics. So, the Bayesian paradigm should already be discussed in Section 1.1.

We see the difference between “one model for all places” and “different models for different places” as the main theme of Section 1.1. Bayesian model averaging and selection seems more related to the question of whether to use one or multiple models for a single place (basin). We therefore think the mention of Bayesian methods is better placed in Section 1.2 where we outline different approaches to model selection.

- The introduction mainly focuses on the challenges when only streamflow observations are considered output variables. This limitation should be highlighted here or in the LIMITATIONS-Section.

We added this to the final paragraph of the introduction to emphasize this for the reader. Changes in bold:

> Despite increasing attention for model structure uncertainty, and improved understanding of which models work well in different locations, selecting appropriate model structures across large geographical domains is an open challenge and actual implementations of a "multi-model mosaic" paradigm for hydrologic prediction are still rare. Our aim with this paper is to highlight a core challenge such implementations will need to overcome in order to fulfil their goal of providing locally relevant, realistic, and optimally performant simulations. **We focus our analysis on streamflow simulations only, but the concepts discussed in this work could be applied more broadly to hydrologic model evaluation.**

LINE 165:
- Reformat “gumboot”

Done.

LINE 161:
- Please give the full configuration of the application of the gumboot-methodology, such as time period, block size, number of blocks, number of samples ... Is the time period different from the one used in [2]?

We use the defaults as provided by [2], and have clarified this in the text. We also used this opportunity to clarify reviewer 1’s questions about the sampling strategy w.r.t. snow processes. Changes in bold:

> Using the observations and model simulations, we can calculate the KGE scores and quantify their associated uncertainty with the gumboot package (Clark and Shook, 2021; Clark et al., 2021). Briefly, gumboot returns various statistics, such as the 5th, 50th and 95th percentile estimates of the KGE score **through a “non-overlapping block” bootstrapping method that creates a sample of water years based on the data period provided. Each bootstrapped realization is based on random sampling with replacement of water years in the data period. Using water years as the non-overlapping blocks in the bootstrap ensures that each**

> **bootstrapped realization consists of hydrologically independent sub-periods. We use gumboot's default settings determined by Clark et al. (2021). This creates 1000 bootstrapped realizations, with October as the first month of the water year; a water year must contain at least 100 valid (larger than 0) flow values.** For simplicity, we only calculate the KGE uncertainty bounds for the best model (i.e., the model with the highest validation KGE score) in each basin, and use these bounds to inform our analysis. We can use gumboot in 555 out of 559 basins. The remaining four basins contain years with no observed flow, and this interferes with the computation of standard deviations and correlations during gumboot's bootstrapping procedure. These four basins are excluded from further analysis. For the remainder of basins, we exclude all models with efficiency scores below the 5th percentile estimate of the KGE score of the best model in a given basin from further analysis, and thus only keep those models with efficiency scores that fall within the uncertainty bounds of the best model in each basin.

LINE 171:
- Please give a formal definition of the linear program solved here.

We substantially expanded on the initial description and added a reference to the GitHub repository that contains the exact code we used for this paper. Changes in bold:

> The sampling uncertainty obtained from gumboot provides enough information to answer the first two research questions. To answer the third research question, **we need to identify the minimum number of models needed to get performance-equivalent simulations in each basin. One way to find this minimum combination of models is to iteratively trial every possible combination of models, and identify the first combination of models for which we obtain performance- equivalent simulations in all basins. Such a brute-force approach is guaranteed to be accurate, but slow. A faster way is to rewrite the problem as a linear programming exercise, where the goal is to find the minimum number of subsets needed to provide coverage for the full set. In our case, we have 36 subsets (one for each model) where each subset contains the basin identifiers where a given model is performance-equivalent with the best model. The full set contains the identifiers for all 559 basins, and the optimizer is tasked with finding the smallest number of subsets (i.e., models) needed to cover the full set (i.e., all basins). We refer the reader to our GitHub repository for further implementation details (Knoben, 2024).**

LINE 181:
- I suggest moving the following lines to RESULTS-Section.

This part of the text and the accompanying figure is meant to help the reader understand the methods, and not a main result in its own right. We have updated the text to make this clear. Changes in bold:

> **To aid in understanding the methods used here,** Figure 2 illustrates the sampling uncertainty and model filtering described in step 2 **for** the basin with the lowest (Fig. 2a,b) and highest (Fig. 2c,d) sampling uncertainty respectively. Figure 2a shows the observations and simulations in a basin with a strongly seasonal and relatively regular flow regime.

FIGURE 2d:
- Could you highlight the best “performance-equivalent” models in red?

The models in this figure are only the best model in each basin. There are only 559 dots in total (one for each basin), not 36*559. We updated the caption to clarify this, using the definitions we added in response to reviewer 1’s comments. Changes in bold:

> Figure 3. Model results for 555 CAMELS basins. (a) maximum KGE score obtained per basin for the evaluation period. (b) Model that obtains the maximum evaluation KGE score shown in (a). (c) Sampling uncertainty in KGE scores obtained from gumboot, with the color axis capped at either end for clarity. (d) Scatter plot showing the relation between the **KGE of the best model in each basin** and its associated uncertainty **interval**. Borders here, and in later Figures, from Commission for Environmental Cooperation (CEC) (2022).

I suggest that the authors consider the above points before final publication. This will ultimately benefit the manuscript and the overall study.

[1] KNOBEN, W. J. M.; FREER, J. E.; PEEL, M. C.; FOWLER, K. J. A. & WOODS, R. A.: A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments. In: Water Resources Research 56 (2020), Nr. 9

[2] 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.: The Abuse of Popular Performance Metrics in Hydrologic Modeling. In: Water Resources Research 57 (2021), Nr. 9