Dear reviewers,

Here we provide a point-by-point response of the changes we made based on the reviews. Thank you for the valuable feedback. We believe that this greatly benefitted the quality of the publication.

Best Regards,

Jerom Aerts

Point-by-Point response Review #1

Major Comments:

The streamflow performance of the different model instances is evaluated through the KGE score. Although the authors state in L150-L152 that they assessed the KGE score for both a calibration and an evaluation period, it seems that the results are mainly focused on the evaluation period: the CDFs of Figure 7 correspond to the evaluation period, and at least the map in Figure 8d also corresponds to the evaluation period according to the figure caption. It is not clear if Figures 8a, b and c also correspond to the evaluation period.

We clarified the distinction between calibration period and evaluation period results by adding a first sentence to all relevant figure captions. In addition, we added the headers “3.1 calibration period results” and “3.2 evaluation period results” to add structure to the results section.

The calibration results briefly appear in Figure 3 for an example basin, but I consider this insufficient. Therefore, my recommendation is to include the CDFs for the calibration period in Figure 7 (see also next two comments), and clearly distinguish between calibration and evaluation scores in the figure captions.

We added Figure 4 that shows the KGE CDF of the calibration period results and the 3 KGE components. These results are described in lines 245 – 255 and included in the discussion.

Similarly to the NSE score, KGE can be decomposed into three parts: the coefficient of correlation, the ratio of the mean values and the ratio of the standard deviations (Gupta et al., 2007; Knoben et al., 2019). All these CDFs should be present in the manuscript, as they will help understand why the KGE values are as they are. Apart from the CDFs for KGE, Figure 7 should collect the CDFs for these three components (not necessarily for the MARRMoT ensemble, although it would be more than welcome). These new results should be discussed as well.

We have added the CDFs of the 3 KGE components to the newly created calibration period results Figure 4 (described in lines 245-255) and to the existing evaluation period results Figure 8 (described in lines 289-307). The new results are discussed in lines (395 – 401) of the discussion section.

The two-fold statistical benchmark (one for the mean and one for the median) produces a poor performance (Figure 6d) that wflow_sbm can easily beat for most of the basins (Figure 6b). Although this is not a problem, I feel curious about why the KGE values are so low for the statistical benchmark. Then, the decomposition of the KGE score mentioned above should also be done for the statistical benchmark and should be incorporated into Figure 7 (a multi-panel figure where the plotted lines can be differentiated from each other may be the best way to show all this). This will help understand why the "mean statistical benchmark" outperforms the "median statistical benchmark" (Figure 6c). In particular, the ratio of the mean values will provide an interesting insight: is the ratio of the mean values closer to one for the "mean statistical benchmark"?

We added the statistical benchmark CDFs of the KGE score and its individual components in Figure 6. The results are described in lines 267-276. Indeed this provided an interesting insight regarding the bias (mean values) component. These are closer to 1 for the mean statistical benchmark. This is discussed in lines 364-369.

The Discussion section is not structured and is written as a single block. It can be clearly divided into two parts: one part discussing the benchmark selection and one part discussing the spatial scaling effect. For sure, the new CDFs will strengthen the results and will enrich the discussion.

We have added much needed structure to the discussion section. The new structure is now as follows: 4.1 Benchmarks, 4.2 Streamflow estimates and uncertainty, 4.3 Relative model instance differences (spatial scaling), 4.4 Computational cost.

I also miss in the discussion some recent and important references for the CONUS domain: for example, Mizukami et al. (2017) (already cited in the Introduction) and Rakovec et al. (2019) also carried out a large-domain calibration exercise and followed a benchmark approach to evaluate their results for the CONUS basins. Are the results of this study similar to their results?

We have added the references for the CONUS domain in the discussion lines (377 – 379). Due to the many differences between studies we find it not possible to compare results. This is one of the reasons why we advocate for clear guidelines for modelling studies to facilitate future comparative work.

Minor Comments:

Title

The title is extremely long and sounds like a sentence extracted from the abstract or the conclusions. I would suggest a more concise title, something like "Large-sample assessment of spatial scaling effects on the streamflow estimations of a distributed hydrological model". The reader will find that "finer spatial resolution does not necessarily lead to better streamflow estimates" in the abstract. In any case, I will leave this open to the authors.

We changed the title to: "Large-sample assessment of varying spatial resolution on the streamflow estimates of the wflow_sbm hydrological model". Note that we removed the term “spatial scaling” as reviewer #2 pointed out that a spatial scaling assessment goes beyond varying spatial resolution.

Section 2.1.1 The CAMELS data set

The authors point out three reasons behind failed runs: errors during parameter derivation, errors during run time and missing streamflow observations. While the last one is clear, the other two are not properly described. What do the authors mean by “errors during parameter derivation”? Is this related to the parameter estimations from external sources prior to calibration? Or is it related to the calibration procedure? On the other hand, what do the authors mean by “errors during run time”? I suggest a more detailed description.

We clarified the reasons behind failed runs (lines 100 – 103). Besides missing streamflow observations errors occurred due to parameter estimation errors that in some cases became clear during runtime. Therefore we now only state parameter estimation error as a reason. These occur during drainage network delineation, either when the basin outlet consisted of a single grid cell that results in a model coding error or when inconsistencies occurred in the local drainage direction layer.

Section 2.2.3 Model Runs & Calibration

The parameter KsatHorFrac is the only parameter subject to calibration, and the rest of the parameters are derived from external sources. Firstly, the parameter range for KsatHorFrac should be indicated here and not in L198 when the results are presented. Secondly, it is not clear if the selection of this parameter is based on prior studies, on calibration recommendations for wflow_sbm, or on a sensitivity analysis carried out by the authors. Some information is provided in L60-L62, but I find it confusing to read this in the introduction. I suggest mentioning this information in section 2.2.3 as I feel it belongs here.

We have added a clearer description as to why we calibrate the parameter KsatHorFrac in lines 164 – 171 in section 2.2.2 Model Runs & Calibration.

How is the model calibrated? Do the authors use a calibration algorithm? Is it based on a Montecarlo experiment? No details are given on the calibration procedure, only L153-L154 state that “the calibration procedure finds an optimal parameter value based on the KGE objective function of streamflow estimates at the basin outlet”. The calibration procedure should be properly described.

The calibration routine is now more extensively described in lines 164 - 178. We manually calibrated the parameter based on an interval ranging between 1 – 1000 KsatHorFrac values. The best performing model run based on streamflow estimates at the outlet is evaluated using the modified KGE score and subsequently used for the evaluation period.

The following minor comments will be adjusted in the publication:

- *The last sentence in L187-L188 seems incomplete, or at least has no cohesion with the previous sentence*

- *Instances of “Figure 7” throughout the paragraph seem to refer to Figure 6.*

- *“Figure 5” in L249 seems to refer to Figure 7.*

- *The colorbar in Figure 8c should indicate “KGE difference” or “âˆ † KGE”. “KGE value” is not correct.- Should “their” in L303 be “there”?*

These minor comments have been resolved in the revised manuscript.

Point-by-Point response Review #2

The review covered lots of ground and not always specific parts of the publication therefore we provide a more generic point-by-point response.

Comments:

References

We agreed that the reflection on the use of references should be improved and that the references need to be extended to include the broader land-surface community past and present. Therefore we extended the cited literature in the introduction by including competing modelling philosophies, bodies of work from the land surface community, the parameter identifiability and transferability problems, and the representative elementary watershed concept. As mentioned in the review we adjusted the reference to the MPR methodology. This resulted in lines 15-36 being added to the start of the introduction.

Lakes and reservoirs

We reported the presence of lakes and reservoirs (lines 136-138). In 25 of the 567 basins, lakes and or reservoirs were included in the model parameters given a threshold area of 1 km2 and 10 km2 respectively. Due to the small amount this did not alter the conclusions of the study.

Basin selection

We clarified the reasons behind failed runs (lines 100 – 103). Besides missing streamflow observations errors occurred due to parameter estimation errors that in some cases became clear during runtime. Therefore we now only state parameter estimation error as the reason. These occur during drainage network delineation, either when the basin outlet consisted of a single grid cell that results in a model coding error or when inconsistencies occurred in the local drainage direction layer.

Calibration methodology

The calibration routine is now more extensively described in lines 164 - 178**.** We manually calibrated the parameter based on an interval ranging between 1 – 1000 KsatHorFrac values. The best performing model run based on streamflow estimates at the outlet is evaluated using the modified KGE score. This model instance is subsequently used for the evaluation period model run.

We have added the description as to why we calibrate the parameter KsatHorFrac in lines **164 – 171** in section 2.2.2 Model Runs & Calibration.

Objective function relevance (KGE 0.22)

We have added the sampling uncertainty method of Clark et al. (2021). This is included in the methods (lines 201-207), the results (lines 314-320), and the discussion (lines 419-416). The results altered the main conclusion based on the testing of the statistical difference in objective function distributions. The difference in results is now considered to be too small in the context of sampling uncertainty. We believe that this great suggestion from the reviewer helped with the discussion of the results based on objective functions.

Forcing and model resolutions

The effect of native forcing resolution is discussed in more depth in (lines 447 – 457) and reflects on the findings in literature. We included the need for testing this effect further in future research (outlook section).

Scaling in hydrology

We agree with the reviewer that spatial scaling encompasses more than only varying spatial resolution in hydrological models. Therefore we adjusted the use of the term scaling in the publication and now use varying spatial resolution instead. This is reflected in the title of the title of the publication and further discussed in the discussion and outlook sections.

Model selection

The reason for selecting the wflow_sbm model is now more clearly defined in the introduction. In addition, we added the benefits of using alternative delineation methods (vector-based) to the discussion section.

Overview of Changes:

1. Title:

    - Better describes content of the study and uses the term varying spatial resolution instead of scaling.

2. Abstract:

    - The abstract is updated based on the extra analyses we performed.

3. Introduction:

    - Restructured.
    - Includes competing basin discretization approaches.

- Includes a larger body of cited literature that refers to the land surface community, parameter identifiability problem, parameter transferability, and the representative elementary watershed.
- Clearly stated that we are not using the MPR method.
- We now use “the effects of varying spatial resolution” instead of scaling as this is only a part of scaling.

4. Methodology:

- Better description as to why basins are excluded from the analyses.
- Better description of the calibration methodology.
- Description of the sampling uncertainty of the KGE score method of Clark et al. 2021.

5. Results:

- Restructured.
- Included calibration period CDF of the KGE score and the decomposed components.
- Added the evaluation period KGE score components to the CDF figure.
- Added objective function uncertainty section. Includes the analyses similar to Clark et al. 2021. Added a table summarizing these results based on the evaluation CDF.

6. Discussion:

- Restructured. Added headers for readability.
- Refer to other large domain studies.
- Discuss the sampling uncertainty results and what this means for the statistical KS-test results.
- Discuss the effect of using coarse meteorological forcing products.
- Discuss how vector-based discretization has major benefits when it comes to computational cost and topographic discretization.

7. Outlook:

- Expended on what needs to be added to this study in order to do a complete scaling assessment as opposed to only a spatial scaling assessment.

8. Conclusions:

- State that the sampling uncertainty is large and therefore the conclusion based on the KS-test are inconclusive.

- Added “at the outlet” to the conclusion regarding finer spatial resolution does not always leads to better streamflow estimates.