Response to Review #1

Dear Referee #1,

We would like to thank you for reviewing our publication and for the constructive comments. As we believe that the comments will greatly increase the quality of the manuscript, we agree with all the suggestions and will adjust the publication accordingly. This will include the following points:

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. 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 agree that adding information on the KGE values of the calibration and comparing it to the validation adds insight for the readers. The publication will include CDFs of the calibration period similarly to those of the evaluation period (Figure 7). These results will be placed in section 3.1 and include the three parts that compose the KGE score. In addition, a separate section in the discussion chapter discusses the effect of calibration given the insights gained by including the three components.

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.

Yes, we will adjust the figures to include the three components of the KGE score. These results will be included in the discussion. This will include the MARRMoT ensemble results.

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"?

This is a very interesting point and will help explain the value of the statistical benchmark and its performance. We greatly appreciate the suggestion concerning the visualization of these results. This will be implemented accordingly for Figures 6,7, and the new calibration CDF figure. All figures will include the decomposed parts of the KGE score.

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 agree that the discussion section needs more structure and therefore we will subdivide the section. The subsection will follow the order of the results section and will have clear distinctions between results. This will help with the argumentation and the readability of the publication.

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?

At an earlier research stage we have looked at these results and found "similar" behaviour between studies. We will include this in the discussion. This includes a comment on the differences between both studies.

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.

Initially we decided on a "Nature Journal" style of title that describes the main conclusion of the study. Given this remark and that of Referee #2 we decided on restructuring such that the title is more in line with the HESS journal.

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.

This part needs more elaboration in the final publication. The errors stem from parameter estimation from external sources. These may occur during parameter estimation or come to the surface during runtime. Therefore they should be grouped together. This will be clearly described in the publication.

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 will clarify the calibration routine by firstly stating the parameter range clearly. Secondly, by referencing previous sensitivity analyses (e.g., Imhoff et al. 2020). This information will be inserted in section 2.2.3.

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 model is calibrated "manually" by predefining a parameter range. The parameter corresponding to the highest KGE score at the basin outlet is then selected. The parameter range is selected based on the sensitivity of the parameter to the KGE score. We decided on manually calibrating the hydrological model as this greatly reduces the amount of compute time while still finding a close to optimal parameter value. This information will be included in the publication.

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"?*

Best Regards,

Jerom Aerts