Discussion comment 2

General comment The manuscript is to some extent a sequel to the paper Beck et al. (2016) published in Water Resources Research; in both papers, the performance of global hydrological models (those that are included in a research project) is evaluated against time series of streamflow in small basins with areas of less than 5,000 km2, e.g. less than two (out of globally 67000) 0.5◦ grid cells. In the submitted manuscript, the number of performance indicators has been increased and the impact of forcing data on results has additionally been investigated. While there is some added value to this, the results obtained in Beck et al. (2016) have, in my opinion, not been sufficiently used in designing the study and writing the new manuscript, and conclusions are not well founded. My major concern is that given the overall poor capability global-scale models for estimating runoff in small basins (given e.g. the uncertainty in climate data), which however is not clearly shown in the manuscript but in Beck et al. (2016), the quality of the models even with respect to runoff generation cannot be compared well with the selected evaluation approach (streamflow in small upstream basins) (See major point 3). At least this problem has to be clearly shown and discussed. In addition, there are various points that need clarification.

We sincerely thank the reviewer for his comments and are glad that he/she believe the study provides added value.

We do not fully agree that the “poor” (a very subjective term) capability of the models is “not clearly shown“. Quite the contrary, in fact: we tried to be as transparent as possible in presenting the performance scores by explicitly showing the results for all five signatures and all six correlation coefficients for all models and all catchments (see Tables 4 and 5 and Figures 1 and 2 in the m/s, and Figures 1–14 in the Supplementary material). We prefer to leave it to the reader to determine whether this constitutes poor performance.

In describing the models’ performance as ‘poor’, the reviewer may be drawing a comparison between global models and locally-calibrated hydrological models in such small catchments. While potentially correct on average (in individual cases it would surely depend on the quality of model, forcing and calibration), we feel that this is not the main insight to be gained from our analysis since (i) these local models are only available in certain regions and thus not a viable choice for researchers interested in large-scale hydrological simulations in predominantly ungauged regions, and (ii) this is already widely documented (we provided references in the text, see page 17 lines 7–8).

Specific comments

1) One major conclusion of the manuscript is that “more effort should be devoted on calibrating and regionalizing the parameters of macro-scale models” which the authors base on the fact that the four models that are calibrated (in very different ways) show better values for the selected performance indicators than the six non-calibrated models. However, the model comparison in

Beck et al. (2016), which included not only the calibrated/regionalized version of the model HBV-SIMREG but also a version where all 14 model parameters were globally uniform (not even considering independent land cover and soil information, e.g. rooting depth), showed that this model version has a better performance than all or most (depending on metric) of the other more complex calibrated or non-calibrated models (comp. Tables 7 and 8 of Beck et al. 2016). I therefore suggest to include the HBV-SIMREG version with spatially uniform parameters into the analysis. HBV-SIMREG runoff that is computed as an ensemble mean of 10 model runs with different parameters sets. Then, conclusions regarding the benefits of calibration/regionalization of should be formulated more carefully.

While we can see the value of such an analysis, we could not include HBV with spatially-uniform parameters in the current m/s since it is not part of the eartH2Observe collection of models (noting that the objective of the m/s is to evaluate the eartH2Observe collection of models).

Table 7 of Beck et al. (2016) shows that HBV with spatially-uniform parameters performs overall worse than two models but better than seven models. Thus, while HBV with spatially-uniform parameters performs indeed quite well among the models, it certainly did not perform beyond the range of the other models. Accordingly, the fact that HBV-SIMREG outperforms the other models is really mainly attributable to the calibration and regionalization, and our conclusion would not change if we were to include HBV with spatially-uniform parameters in the current analysis.

2) The study of Beck et al. (2016) also indicates that performance of the HBV-SIMREG model results that are not derived as the ensemble mean of 10 runs with 10 different parameter sets but just 1 (derived from the most similar donor catchment) perform only slightly worse than the ensemble mean and better than the other models (Tables 6 and 7 of Beck et al.). Therefore, the conclusion that the fact that HBV-SIMREG with 10 runs performs better than the ensemble mean of all models tentatively suggests that a multi-parameterization ensemble for a single, sufficiently flexible model could replace multi-model ensemble studies (p. 20), is not backed by the analysis in the manuscript. I suggest including the HBV-SIMREG variant with 1 run/parameter set only, and consider the result when formulating such a conclusion.

The reviewer suggests that our *speculation* that multi-parameterization models may substitute multi-model ensembles is not supported by the results. We certainly agree that the results here do not unequivocally support our speculation and have therefore used cautious terms like “may”, “tentatively”, “if this is confirmed”, and “speculate”. However, our conclusion that HBV-SIMREG with 10 runs performs better than the ensemble mean is certainly not unfounded. The good performance of HBV-SIMREG is certainly the combined effect of the calibration/regionalization and the use of ensembles, as Beck et al. (2016) demonstrate unambiguously. We prefer to retain the statement that multi-parameterization models *may* substitute multi-model ensembles, as we hope to encourage research and development in this very promising direction.

In addition, it should be taken into account (and explained very clearly in the manuscript) that HBV-SIMREG only computes runoff in 0.5◦ grid cells and not river discharge, as grid to grid lateral

routing including the impact of lakes and wetlands as well as water abstraction are not simulated by this model. I suggest adding a table in which the scope of the different models as well as well as the specific calibration/regionalization approach (including number of adjusted parameters) are listed (in section 2).

We thank the reviewer for the suggestion. In the m/s we state that “three of the models did not not simulate river routing (JULES, SWBM, and HBV-SIMREG)” (page 5 line 5). HBV-SIMREG is thus not the only model without a routing routine. However, whether the models have a routing routine is less relevant given that we analyze daily non-routed specific runoff (mentioned on page 4 lines 14–15). Besides HBV-SIMREG, four other models also do not account for lakes, while five other models also do not account for water use. However, this is also less relevant, as streamflow observations for catchments with intensive irrigation and/or large reservoirs were excluded (page 5 lines 9–12).

A comprehensive table with model specifics can be found in Dutra et al. (2015). Since the current m/s is already quite long we prefer not to repeat all this information, but refer to Dutra et al. (2015) instead.

And to clearly state that global hydrological models currently cannot and do not aim at representing reality at scales below 5000 km2.

While we cannot speak for all model developers, some of the models have definitely been designed to represent reality at scales <5000 km^2 (e.g., LISFLOOD, WaterGAP3, and W3RA, which were run at 0.1°, 0.08°, and 0.05° spatial resolutions, respectively). Furthermore, the results demonstrate that some of the models indeed do represent reality to a certain degree at these scales. LISFLOOD and HBV-SIMREG, for example, exhibit high monthly correlations (>0.8) for the large majority of the catchments (Figure 2), indicating that they simulate monthly flows with satisfactory skill at scales <5000 km^2.

Given the overbearing importance of precipitation in predicting runoff, we would expect the ability to represent reality at smaller scales depends more on the precipitation forcing than on the models per se, and indeed there is considerable published evidence for this.

3) The third information of Beck et al. (2016) that has not been made good use of for at least framing the extensive performance comparison done in the submitted manuscript is the information provided on Nash-Sutcliffe efficiency NSE for the 10 models (Table 8 in Beck et al. 2016). Different from the aggregate objective function AOF, the wellknown NSE allows the reader to understand the overall very poor performance of all global models at the scale of small basins. For example, daily NSE of all 10 models (for 1113 catchments) was negative for all models and the ensemble mean (so mean discharge would have been a better estimator than the models, while monthly NSE varied between -1.16 and 0.17! Therefore, information on the NSE should be added. However, it appears to me that one cannot really say that a model achieving a NSE of e.g. 0.17 is better than a model achieving a NSE of -0.17, they are both just very poor predictors. So

what can we really learn from the comparison? This has to be deduced more carefully. I would also suggest to add to the supplement the hydrographs of e.g. 4 selected calibration basins (observed and 10 model results) and a table with the pertaining performance indicators (like Table 5) so that the reader can see the meaning of these performance indicators.

As requested, NSE scores for the models have been added to the Supplementary material. In addition, we have added hydrographs including statistics for four catchments. With respect to NSE scores, we note that this measure does not provide an adequate summary of overall model performance. The NSE has been criticized in several previous studies, as has been explicitly mentioned in the current m/s (page lines 25–26) as well as in Beck et al. (2016), for being overly sensitive to peak flows. For this reason Beck et al. (2016) did not use the NSE for the main analysis, but rather an aggregate performance metric very similar to the one used here.

If the reviewer argues that a NSE of 0.17 is not better than an NSE of –0.17 than we take this to mean that the reviewer also does not see NSE as a good measure of model performance.

4) The conclusions regarding underestimation of snow precipitation need to be better supported. You should take into account that WFDEI precipitation includes an undercatch correction. However, for the USA, WFDEI mostly overestimates PRISM precipitation ("with sophisticated corrections for undercatch) and mean runoff. Maybe the WFDEI undercatch correction is overestimated in the USA and underestimated elsewhere? Maybe you could analyse the uncorrected precipitation data that went into WFDEI and see if they are already higher than the PRISM values. You should also discuss why two models do not show the early bias. One of the adjusted HBV-SIMREG parameters is snow undercatch, which may take care of this problem, but does it? And what may be the reason in case of ORCHIDEE?

Thank you for the suggestions. We have added the following text to Section 4.3 in an effort to better highlight the sources of uncertainty in the WFDEI *P* data: "It is conceivable that biases are present in the WFDEI *P* data, because: (i) the CRU dataset, which has been used to correct the WFDEI dataset, is based on only a subset of the available gauges and does not explicitly account for orographic effects; (ii) in sparsely gauged regions the correction using CRU is more likely to deteriorate rather than improve the *P* estimates; and (iii) the Adam and Lettenmaier (2003) gauge undercatch correction factors are based on interpolation of a very sparse sample of gauges and thus subject to considerable uncertainty." We now also mention that PRISM is based on considerably more stations than CRU.

The snowfall correction parameter of HBV-SIMREG could indeed be responsible for the better runoff timing, we appreciate the suggestion. We have added the following text to Section 4.7: "The fact that HBV-SIMREG performs well in this regard is probably attributable to the snowfall gauge undercatch correction factor of the model." For ORCHIDEE we are unsure about the cause for the good runoff timing.

5) Try to explain more the behavior of the different models (to the extent this is possible)

With the current results we cannot draw further robust conclusions regarding each model's behavior. A detailed evaluation of each model independently would allow such a discussion, but this is beyond the scope of the current m/s.

Other/technical comments

P1L1: Replace “runoff” by “streamflow”

Done. Thanks for the suggestion.

P6L14: I find the mean (over 966 or 641 basins) difference between simulated and observed runoff signature D not very informative and suggest adding the standard deviation of this differences in Tables 4 and 5. Maybe do this also to the temporal correlations.

The mean *D* provides information about the average deviation between simulated and observed signature values. We feel this information is very useful to identify, for example, if a particular model tends to produce insufficient or excessive baseflow (using the BFI signature), or if it consistently under- or overestimates peak flows (using the Q1 signature).

We prefer not to include standard deviations. First, they are difficult to interpret, being influenced by the simulated as well the observed signature distributions. Second, they require the distribution of *D* values to be (more or less) normally shaped, which they are not. Finally, they would clutter the tables and deteriorate the readability.

P7L10: I would not say that the Spearman rank correlation coefficients evaluate the ability to simulate “the spatial variability” but just “the variability among the observation basins”.

Thanks for the suggestion, we have made the change.

P12L25: AET in WaterGAP can exceed PET due to calibrating against mean annual discharge; while this may be unphysical, it may correct for a wrong PET estimates. So it is not the evapotranspiration routines that need to be re-evaluated but the PET (or P) estimates.

Thank you for pointing this out. We agree and have changed the text accordingly.

P17L34: How many basins coincide?

This information is unfortunately not available.

Fig. 5: Use color to indicate snow-dominated catchments and/or to color by latitude.

Thank you for the comment, we have added a color scale to the figure.