Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-124-RC2, 2016

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Global evaluation of runoff from ten state-of-the-art hydrological models" *by* H. Beck et al.

Anonymous Referee #2

Received and published: 8 July 2016

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.

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.

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

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 well-known 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.

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 precip-

itation (“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?

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

Other/technical comments

P1L1: Replace “runoff” by “streamflow” 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. 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”. 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. P17L34: How many basins coincide? Fig. 5: Use color to indicate snow-dominated catchments and/or to color by latitude.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-124, 2016.