Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-176-AC4, 2020

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Behind the scenes of streamflow model performance" *by* Laurène J. E. Bouaziz et al.

Laurène J. E. Bouaziz et al.

laurene.bouaziz@deltares.nl

Received and published: 17 July 2020

We thank the anonymous referee 3 for his/her comments and provide an answer to each point below. However, we are surprised and puzzled by the review as we think that most of the points raised by the referee are covered in the manuscript. The overall assessment and relatively minor comments do not seem to correspond with the associated evaluation report of the referee.

Comment 1:
This manuscript proposes a multi-objective model evaluation to compare a number of different hydrological catchment models. While this is certainly a valuable task, I

honestly have very split feelings about this study. The general idea of multi-objective testing is not new, but very important, and the comparison of several models is an interesting novel aspect. However, I have a number of fundamental concerns, which would require new data and computations to be addressed.

The study is based on only three catchments. Several studies have shown how variable results between catchments can be, and these days with more and more data sets being available, the use of just three catchments seems a bit surprising for this type of study.

Reply 1:
The overall objective and novel aspect of this study is to analyze and quantify the differences in the magnitudes and dynamics of multiple internal state and flux variables of multiple models that provide similar performance characteristics when exclusively evaluating them against observed streamflow. We will further emphasize this in the introduction. More specifically, in addition to streamflow, we quantify the differences of five model internal state and flux variables for twelve models in three catchments. The primary aim of our study is to demonstrate and underline that models that are calibrated to streamflow can generate similar, high performance levels in reproducing streamflow, but that they use different "pathways" to do so, i.e. all representing the system in a different way. A secondary objective is to benchmark the internal state and flux variables against remote sensing data. We will clarify this in the introduction of the revised manuscript.

As in previous comparison studies, there needs to be a trade-off in what can be done in one single experiment. This is not only a question of computational capacity and time restrictions but also a matter of how results can be analyzed and communicated in a feasible and meaningful way. Already now, with five internal variables from twelve

models and three catchments, the sheer amount of results produced makes it difficult to identify and communicate the most relevant points. Extending such a study to say 10 or 50 catchments will add an additional layer of results, which needs to be interpreted and discussed in addition. This will lead to a very unfocussed paper, in which the reader will struggle to find in-depth results. Such trade-offs in the analyzed factors are common in comparison studies and we are in fact not aware of any study that combines an analysis of many models with many variables and many catchments. For example, in their comparison study, Holländer et al. (2009) used ten models in one artificial catchment and assessed evaporation and discharge. The distributed model intercomparison project (DMIP; Smith et al. 2012) worked with 16 models, 17 catchments but mainly assessed streamflow and soil moisture. Noh et al. (2015) and Koch et al. (2016) compared three models with respect to seasonal variability of soil moisture. Orth et al. (2015) used three models to assess streamflow and soil moisture in eight catchments. Le Moine et al. (2007) used two models in 1040 catchments to focus on intercatchment groundwater flows. Rakovec et al. (2016) studied three internal state and flux variables in 400 catchments using a single model. Very recently, Knoben et al. (2020) investigated differences in performance of 36 models in 559 catchments with respect to streamflow as single variable. Their analyses are based on general performance metrics of daily streamflow. The conclusions remain general due to the considerable volume of data produced, allowing for less detail on process-relevant insights. Each of these studies has a specific focus and this is similar for our study. To our knowledge, a study with strong focus on internal model dynamics for multiple models in more than one catchment has not been done in this way before. We deliberately chose to balance depth with breadth and perform a thorough analysis of the set of **twelve models** and **five variables** in the **three catchments** in this study. We will stress this motivation in the revised version of the manuscript.

Comment 2:
The study addresses different storages, including snow storage. However, the impor-

tance of snow in the test catchments is minor. I could not find any information on the relative importance of snow (the info of about one month of snow cover is incomplete as this does not say anything about the amount of water stored as snow). Still, my general understanding is that snow does not play any major role in these catchments. This is probably also the reason why the authors can get away with not using any elevation zones for modelling snow processes.

Reply 2:
We agree that snow is not a major component of the streamflow regime within these catchments (as briefly mentioned in Section 2). Most of the precipitation falls as rain and the models have high streamflow model performance even without including a snow module. The amount of water stored as snow is shown in Figure 5b and 5c of the manuscript with maximum annual amounts of less than 20 mm. Despite these relatively small amounts, snow can be very important for specific events. In 2011, rain on snow caused widespread flooding in these catchments. The elevation range of the Ourthe Orientale upstream of Mabompré, where we are evaluating snow processes, is approximately 370 m (from 294 m to 662 m). We believe this can be treated as a single elevation zone, in particular as 65% of the catchment falls into a narrow 100 m elevation band (see Fig. 1 of this reply). In any case, given the absence of detailed observed temperature lapse rates, the assumption of a stable environmental lapse rate of e.g. 0.6 degree C/100m, required for an elevation stratification remains very speculative and thus not really warranted by the available data. We will clarify this in the revised version of the manuscript.

Comment 3:
Each of the storage estimation used for model testing is associated with significant observation uncertainties. There is also a scale-mismatch which results in additional uncertainties. These issues have to be considered!

Reply 3:
We completely agree that there is considerable and effectively mostly unknown uncertainty in the used remote sensing data. This was the underlying reason why we did not use the remote sensing products for model calibration nor for any type of quantitative model evaluation. We rather only treated these data as additional information against which to indicatively compare the modeled internal state and flux variables. We cannot and do not consider the remote sensing data to be a reliable representation of real-world quantities. However, they are useful to detect potential outliers. The uncertainty associated to remote sensing data should not restrain us from using them at all. However, we will clarify the uncertainties associated to the use of each remote sensing product in the revised version of the manuscript. In the next version of the manuscript, we will define a set of soft criteria to evaluate not only how consistent the model internal dynamics are amongst each other, but also if they provide a consistent behavior with what we expect from expert knowledge and remote sensing data.

Comment 4:
Another point that seems to be missing is that each of the models of course also is affected by parameter uncertainties (which will influence the simulated storages). Perhaps I am missing something, but as I understand, single parameter sets are used for each model. This is not sufficient; we know that the same model can result in very different internal simulations because of parameter uncertainty. This leaves me wondering how much of the differences presented here are due to parameter uncertainty rather than due to model differences.

Reply 4:
We absolutely agree that parameter uncertainty can cause differences in model

internal behaviour. Therefore, we of course use an ensemble of feasible parameter sets to account for parameter uncertainty. This is briefly mentioned in Section 4.1 and we will clarify this further in the revised version of the manuscript. The error bars and/or boxplots and/or ensemble of lines in Figure 3, 4, 5, 6, 7, 8 and 9 represent the ensemble of feasible parameter sets, as also mentioned in each caption. The narrower uncertainty ranges of some models are related to the use of different search strategies of the parameter space (see caption of Figure 8).

References
Holländer, H. M., Blume, T., Bormann, H., Buytaert, W., Chirico, G. B., Exbrayat, J. F., et al. (2009). Comparative predictions of discharge from an artificial catchment (Chicken Creek) using sparse data. *Hydrology and Earth System Sciences*, 13(11), 2069–2094. https://doi.org/10.5194/hess-13-2069-2009

Knoben, W. J. M., Freer, J. E., Peel, M. C., Fowler, K. J. A., Woods, R. A. (2020). A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. *Water Resources Research*. https://doi.org/10.1029/2019WR025975

Koch, J., Cornelissen, T., Fang, Z., Bogena, H., Diekkrüger, B., Kollet, S., Stisen, S. (2016). Inter-comparison of three distributed hydrological models with respect to seasonal variability of soil moisture patterns at a small forested catchment. *Journal of Hydrology*, 533, 234–249. https://doi.org/10.1016/j.jhydrol.2015.12.002

Le Moine, N., Andréassian, V., Perrin, C., Michel, C. (2007). How can rainfall-runoff models handle intercatchment groundwater flows? Theoretical study based on 1040 French catchments. *Water Resources Research*, 43(6), 1–11. https://doi.org/10.1029/2006WR005608

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., Zappa, M. (2015). Does model performance improve with complexityâĂŕ? A case study with three hydrological models. *Journal of Hydrology*, 523, 147–159. https://doi.org/10.1016/j.jhydrol.2015.01.044

Noh, S. J., An, H., Kim, S., Kim, H. (2015). Simulation of soil moisture on a hillslope using multiple hydrologic models in comparison to field measurements. *Journal of Hydrology*, 523, 342–355. https://doi.org/10.1016/j.jhydrol.2015.01.047

Rakovec, O., Kumar, R., Mai, J., Cuntz, M., Thober, S., Zink, M., et al. (2016). Multiscale and multivariate evaluation of water fluxes and states over european river Basins. *Journal of Hydrometeorology*, 17(1), 287–307. https://doi.org/10.1175/JHM-D-15-0054.1

Smith, M. B., Koren, V., Zhang, Z., Zhang, Y., Reed, S. M., Cui, Z., et al. (2012). Results of the DMIP 2 Oklahoma experiments. *Journal of Hydrology*, 418–419, 17–48. https://doi.org/10.1016/j.jhydrol.2011.08.056

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-176, 2020.

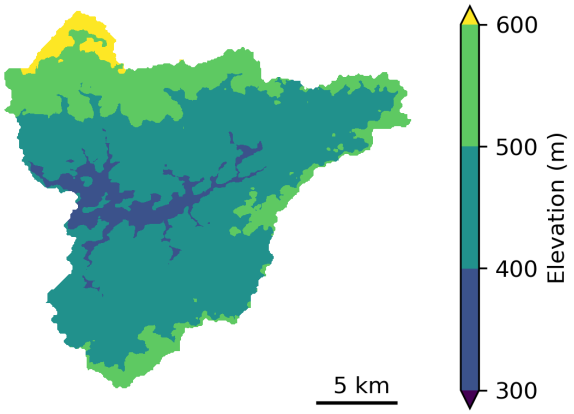

Fig. 1. Elevation contour lines of the Ourthe Orientale upstream of Mabompré