Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-176-AC2, 2020

Hydrology and Earth System Sciences
Open Access
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: 16 June 2020

We thank the anonymous referee 2 for carefully reading our manuscript and providing interesting suggestions. We provide an answer to each comment below.

Comment:
Bouaziz et al. (2020) evaluate 12 hydrologic models for three medium-sized Belgian catchments, all established and calibrated by eight research groups. Although the spatially aggregated streamflow performance differences among models are negligible, the internal model states and processes (can) differ significantly. This paper is an interesting diagnostic study, with nice figures. I have some minor comments which the

authors should consider to address.

First of all, it is nice to see the huge collaborative efforts across many institutes behind this model inter-comparison study. This study with many details shows large differences among 12 hydrologic models and even larger differences against different remotely sensed products. Something what a reader would expect. I encourage the authors to stress more clearly, what is the main "take-home" message of the main paper.

Reply:
Thank you, this is a good suggestion. In the revised version, we will more clearly stress the main take-home message, which is to underline and demonstrate 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. In the next version, we will also emphasize on the use of remotely-sensed products in combination with expert knowledge to evaluate if models can be considered behavioral.

Comment:
Because the authors did not use an ensemble of model structures from a modular framework (e.g., FLEX, FUSE), which could properly address those differences or individual model deficiencies step by step (by identifying individual hypothesis), in their study they cannot clearly separate and identify, which hidden hydrological processes can help improve model functioning against those reanalysis products. Could you please comment on this?

Reply:

We agree that the set of models does not easily allow for a step-by-step identification of differences in individual hypothesis, as they are mostly full-grown models. Perhaps only the subset of FLEX models M2 to M5 allows us to identify stepwise differences in internal model representations. However, we grouped models with similar parametrizations in Tables 2 and 3 and focused our analyses on model components that were present in most models (evaporation, snow, root-zone soil moisture, total storage). One important systematic difference that we found amongst the models is the significant drying-out each summer for some models. In the next version of the manuscript, we will hypothesize on the model parametrization that leads to this behavior. We believe that these specific findings can help to identify some model functioning aspects that can be improved by adapting model parametrization. In the revised version, we will also include more detailed findings on the plausibility of model behavior for a selection of criteria related to the remote sensing data and expert knowledge (as suggested by referee Prof. Keith Beven).

Comment:
Line 80,140+: evaporation => "evapotranspiration"? Please don't forget about the plants! Hargreaves-Samani formula is for evapotranspiration, not for evaporation only.

Reply:
We will clarify that we have used the term "evaporation" to describe the sum of all evaporation components (including transpiration, soil evaporation, interception, sublimation and open water evaporation when applicable). It is perhaps a matter of taste, but following Savenije (2004) and Miralles et al. (2020), we are using the term evaporation instead of evapotranspiration for all evaporative fluxes. We will make sure to clearly state this in in the text and in Table 1 to avoid confusion.

Comment:

Line 85: streamflow => "runoff", because of the unit

Reply:
We strongly prefer using "streamflow" to describe the flow of water in the river and have consistently applied this terminology throughout the manuscript, irrespectively of the unit. In our view, runoff is more generic and can refer to (sub)surface flow, which is not yet in the river.

Comment:
Line 86: You should start this sentence that this is a headwater basin of ID1

Reply:
Agree, we will add this in the revised version.

Comment:
Line 101: I guess the authors could have used a bit more advanced method for interpolation rain gauge observation instead of the Thiessen polygons, to better account for input error uncertainty, e.g. kriging or its variants. The uncertainty in the meteorological inputs is not mentioned in the manuscript.

Reply:
We agree that there is always uncertainty in meteorological input data and will mention this in the revised version. Another method to interpolate precipitation could also have been feasible. However, the number of available precipitation stations would likely not be enough to perform a meaningful Kriging interpolation. The advantage of Thiessen polygons is that extremes are not averaged out, which would occur in any other type of interpolation. Many threshold processes are controlled by these extremes. Besides,

our primary aim was to make sure that the same forcing data was used by all research groups.

Comment:

Line 114: PET method is based on Priestley Taylor, which is different from section 3.1. How is it compatible with section 3.1 and overall results?

Reply:

We believe that the different methods to estimate potential evaporation should not impede us from testing the consistency between the resulting total modeled actual evaporation E_A and estimated E_A from GLEAM. This is also supported by the findings of Oudin et al. (2004), who reported similar model performance irrespective of the method applied to estimate potential evaporation. Additionally, we do not consider E_A from GLEAM to be representative of the truth, but the comparison can enable us to detect outliers (either one/several models or the remote sensing data).

Comment:

Line 143: how did you spatially average soil moisture?

Reply:

We calculated the mean soil moisture over all SCATSAR-SWI1km pixels within the Ourthe catchment. We will clarify this in the revised version.

Comment:

Line 153: I guess, your entire study domain is just a single GRACE pixel. I am quite skeptical for using it at all, as it's beyond the limits of its usability. The original raw GRACE signal is based on a much larger region (3degrees). You may better wait for

the GRACE-FO, which has much finer native resolution...

Reply:

Yes, the catchments indeed fit in single GRACE pixels. At this small scale, we agree that we must be careful with possible 'signal leakage' from surrounding areas, which increase the uncertainty. We believe that the GRACE signal is still informative and the best currently available, despite the larger errors and uncertainties at this small scale compared to large scale spatial averages. It should also be noted that we are not using GRACE for model calibration. Instead we are testing if the modeled regional seasonal water storage anomalies are consistent with GRACE estimates. Additionally, GRACE signals in small scale catchments were shown to provide valuable information for hydrological modeling (Rakovec et al. 2016, Nijzink et al., 2018). In the next version of the manuscript, we will use total storage anomalies provided by the three processing centers instead of taking the mean of the three to better account for the uncertainty. In the future, it would surely be interesting to work with GRACE-FO, but this is unfortunately not available for our study period.

Comment:

Section 4.1 I guess all models were applied in spatially lumped manner, i.e. no spatially distributed mode, isn't it? Please write down explicitly in this section.

Reply:

We will clarify if models are lumped or (semi-)distributed in Table 2. The wflow-hbv model is the only fully distributed model, but parameter values are mostly uniform over the catchment area. The FLEX-Topo model is a semi-distributed based on hydrological response unit within each Thiessen polygon. All other models are lumped.

Comment:
Line: 179-181 This analysis was done here, or in previous study? Not clear, please specify, and provide link to the transferability results. Curious to see them.

Reply:
We will clarify this part in the next version of the manuscript. The analysis was done in the previous study (de Boer-Euser et al. 2017), in which the models were calibrated for the Ourthe at Tabreux and parameter values were transferred to two neighboring and two nested catchments (including the Semois at Membre-Pont and Ourthe Orientale at Mabompré). The previous study covered the study period 2001-2010. In the current study, we use the previous calibration of the Ourthe at Tabreux for the three catchments and ran the set of models for an additional period from 2011 to 2017.

Comment:
How many behavioral parameter sets per model were used? Is the number same per hydrological model? Here referring to error bars in Fig. 3, Fig. 4 and elsewhere.

Reply:
For each model, we retained 20 feasible parameter sets. However, the width of the error bands varies due to the different calibration strategies applied by the modelers.

Comment:
Figure 9: Is it possible to rank the models according to their performance? Which one seems to be most relevant and how it compares to e.g. an operational model, if that's available? Please think about putting some implications to the paper.

Reply:

Thank you for this interesting suggestion. In Figure 9, the models are currently ranked from the highest to the lowest performance according to the Euclidean distance of Nash-Sutcliffe efficiencies of the streamflow and the logarithm of the streamflow (see Figure 3 and Table 2). In the next version of the manuscript, we will adapt Figure 9 to rank the models for each criterion.

In the revised version, we will mention that GR types of models (as GR4H) are used for operational purposes in France and that a lumped version of the HBV model is currently used by the Dutch operational system. In fact, each of these models could potentially be used operationally.

In the next version of the manuscript, we will introduce some soft criteria to rank and evaluate models in terms of how plausible it is to consider them behavioral based on the remotely-sensed data and expert knowledge (following the suggestions of referee Prof. Keith Beven).

References

de Boer-Euser, T., Bouaziz, L., de Niel, J., Brauer, C., Dewals, B., Drogue, G., et al. (2017). Looking beyond general metrics for model comparison ndash; Lessons from an international model intercomparison study. *Hydrology and Earth System Sciences*, 21(1), 423–440. https://doi.org/10.5194/hess-21-423-2017

Miralles, D. G., Brutsaert, W., Dolman, A. J., Gash, J. H. (2020). On the use of the term "Evapotranspiration." *Earth and Space Science Open Archive*, 8. https://doi.org/10.1002/essoar.10503229.1

Nijzink, R. C., Almeida, S., Pechlivanidis, I. G., Capell, R., Gustafssons, D., Arheimer, B., et al. (2018). Constraining Conceptual Hydrological Models With Multiple Information Sources. *Water Resources Research*, 54(10), 8332–8362. https://doi.org/10.1029/2017WR021895

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., Loumagne, C. (2005). Which potential evapotranspiration input for a lumped rainfall-runoff model? Part 2 - Towards a simple and efficient potential evapotranspiration model for rainfall-runoff modelling. *Journal of Hydrology*, 303(1–4), 290–306. https://doi.org/10.1016/j.jhydrol.2004.08.026

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

Savenije, H. H. G. (2004). The importance of interception and why we should delete the term evapotranspiration from our vocabulary. *Hydrological Processes*, 18(8), 1507–1511. https://doi.org/10.1002/hyp.5563

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