Review of Manuscript

'Behind the scenes of streamflow model performance' (hess-2020-176)

by L. J. E. Bouaziz et al.

Dear Editor, dear Authors,

I have reviewed the aforementioned work. My conclusions and comments are as follows:

1. Scope

The article is within the scope of HESS.

2. Summary

The authors present an evaluation of a set of twelve conceptual hydrological models - all set up in the same Belgian watersheds. The key aspect of the study is that the models were calibrated against streamflow only, and are now evaluated in terms of several criteria (water balance, streamflow characteristics, runoff coefficient, evapotranspiration, snow storage, root-zone storage, total storage) against remote-sensing data taken as reference 'truth' - except discharge, which is based on local water level observations. The main hypothesis is that the models, showing comparably good performance in terms of streamflow, should do so based on similar internal representations of internal states and fluxes. This hypothesis is tested by determining the (dis-)agreement of model outputs and the corresponding remote sensing and gauge data, and comparing these (dis-)agreements among models. The main findings are that substantial disagreements exist for most models and several criteria, and that these disagreements do not agree among models. Also, no single-best model could be identified, and, taking into account the considerable uncertainties of the ground truthing data, no model could be rejected. With respect to the main hypothesis, the authors therefore conclude that good model performance with respect to discharge at the catchment outlet does not guarantee realistic and unambiguous internal model workings.

For future studies, the authors recommend multi-data calibration and validation (to reduce the equifinality associated with calibration against streamflow only), and that despite their uncertainties, remote-sensing data can play an important role in this, especially the dynamical patterns they provide. They also advocate multi-model, multi-parameter approaches to reveal uncertainties related to model predictions, and taking in-situ measurements to inform process studies.

3. Evaluation

Altogether, the study was conducted and presented in a very thorough and accessible manner, and it is a good example of a collaborative effort. I particularly welcome how the authors made uncertainty (of the forcing and ground-truthing data, and of parameter identification during calibration) a central part of their study, and how they do a multi-criteria evaluation – including soft criteria - of their models to gain an as-complete-as-possible picture of model performance.

That said, the substantial weakness of the paper is that there is nothing really new to learn from it. The main hypothesis the authors address is in fact no hypothesis, as the answer to it has already been given in the literature many times. In fact, the authors provide in the introduction a very good literature overview on the inevitable equifinality of model parameter estimation when using discharge as the single, aggregate evaluation criterion (and hence the impossibility to correctly address the 'individual hypotheses', see p. 3 l 57), on approaches to tackle this by multi-criteria calibration, or by using multiple models.

Furthermore, all models used by the authors are conceptual hydrological 'bucket' models (see Fig. 2), there is neither a real bottom-up physics-based model involved, nor a mainly data-based (such as LSTMs or other). For this relatively narrow selection of models, the authors rightly state in the introduction (p 2 l 20-23) that the representations of particular hydrological processes are quite similar. So given this narrow range of models, what we can learn from comparing them also only has a narrow range of applicability.

Furthermore, given the simplicity of the models used, most of the deficits of particular models with respect to particular processes (e.g. FLEX-Topo root-zone storage falling completely dry, see p 16 l 498) can be directly inferred from their structural/functional setup.

Taken together, the conclusion with respect to the main hypothesis is not wrong, but it comes as no surprise, and likewise the recommendation for future studies have also been made elsewhere (in fact, the authors give a good account of the related literature in the discussion section, e.g. p. 18 l 556-557, p. 18 l 560-563, p. 19 l 585-588).

My points of concern relate to the very core of the study, so I do not think they can be solved by a major revision. So it is either 'reject' or 'publish-as-is'. My recommendation - despite the substantial deficits of the study – is that the study deserves publication as a thoroughly done example of state-of-practice hydrology.

One last comment: Given the nice set of data the authors have put together, it would be interesting to reverse the approach: If the models are calibrated on the 'individual hypotheses' only, using the available remote sensing data (and not discharge), how good will they perform with respect to the 'aggregated hypothesis' (discharge)? This could be useful for cases where remote-sensing data are available, but not discharge data (PUB).

Yours sincerely,

Uwe Ehret