Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-176-AC1, 2020

Hydrology and Earth System Sciences
Discussions
Open Access

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: 2 June 2020

We thank Prof. Keith Beven for his detailed and constructive review. In the following, we express our view on the raised points, and illustrate our plan on how to improve the paper.

Comment:
This paper takes a diverse collection of hydrological models, previously calibrated to the Oerthe basin, and subjects them to comparison with estimates of evapotranspiration, soil moisture, snow cover and GRACE total estimates. The models all produce "reasonable" streamflow calibrations (I assume, since it seems that none of them have

been rejected in the first calibration part of the study). The conclusion is that they do so in different ways, and still none of them are rejected. Now I understand why it is diplomatic when working within an international project to be kind to all the groups who are participating, but doing so does not produce an outcome that is in any way useful. The models are just shown to be different. Why are these models not being tested as hypotheses about how the catchment system is working? Indeed, we could rather say on the basis of the evidence presented that none of them are really fit for purpose when the additional variables are taken into account.

Reply:
Indeed, in the current version of the manuscript, our primary objective is to quantify the internal differences between the set of models with similar streamflow output and only a secondary objective is to benchmark the internal state and flux variables against remotely-sensed estimates. However, we agree that deriving qualitative or quantitative measures to evaluate, rank and potentially reject models using the remotely-sensed data in combination with expert knowledge could add value to the study. In response to this criticism, we will introduce quantitative measures to rank and evaluate the models in terms of how plausible it is to consider them behavioral, both in view of the independent remotely-sensed data, and based on expert judgment (e.g. whether model storages are always filling or running empty).

Besides data uncertainty (see comment and reply below), the main reasons we are reluctant to formally and explicitly reject models are the following:

All interpretations and conclusions here (and in any modelling study, really) are also conditional on the individual parameter selection strategies chosen by the individual contributing institutions. The use of different and/or more calibration objectives and/or criteria may have resulted in considerably different model results and associated conclusions. The same is true for the use of different search strategies of the parameter space. In addition, and quite obviously, a rejection of a model (i.e. the combination of model structure, numerical implementation, parameter selection strategy, etc.) would only be valid for the study catchment.

In our opinion, an explicit rejection of one or more models may give the *impression of their general unsuitability*.

Thus, a rejection is therefore not (only) a diplomatic question but, when perceived as *general*, may be unjustified, as these rejected models may be the most suitable models elsewhere.

We admit that this is of course a communication issue. But in the moment a model is formally and explicitly rejected in a paper, it will be perceived as generally useless by many even if it is emphasized that a rejection can essentially always only be conditional on the above points (and many others, such as, needless to say, data uncertainty).

In other words, the label “rejected” will stick, and not the reasons and circumstances. We would really like to avoid that because we think it is neither fair nor justified.

Comment:
Except that it is not quite that simple, because ALL of the additional variables used in this comparison are subject to significant uncertainty and commensurability issues. And without taking some account of those uncertainties no real testing is possible (it is also worth noting that no account is taken of uncertainty in the original calibration exercise either – why not in 2020? It has been recognized as an issue in model

calibration for more than 30 years!).

Reply:
We agree that assessing the uncertainty of such remote sensing products is important, but not easy. In principle, uncertainty estimates should be provided by the remote sensing models, of which we are final users, but these estimates are usually not available. Even if they were, using them in a meaningful way would uncover many questions, which would go beyond the scope of this work. For this reason, we are highly reluctant to use these data to determine hard rejection thresholds. Rather, we will use them to provide a “soft” assessment of the relative merits of the various models in form of an overall ranking guided by criteria formulated based on “soft”, “expert judgement” of trustworthiness of the individual types of remotely-sensed data.

Comment:
The section on knowledge gaps at the end should be moved to before the model comparison is presented, and should explicitly consider the uncertainty and commensurability issues. Nowhere is there any mention of the uncertainties arising in verification studies of these additional variables, but that is surely significant.

Reply:
In the next version of the manuscript, we will make the uncertainties and commensurability issues of the evaluation variables clear before using them to evaluate the models.

Comment:
To give a particular example: models with and without interception storage. This is an example of why more thought is required about what is actually being compared here. One of the reasons why models choose NOT to have an interception store is

to reduce the number of parameters required to be calibrated or estimated a priori. But how this works will also depend on how potential evapotranspiration is estimated. Does it include the effects of a wet canopy – especially over rough canopies. This can be really important (and subject to significant uncertainties in effective roughness and humidity deficits because of sensitivities under such conditions). Here the Hargreaves PE formula does not explicitly consider wet canopy conditions, but GLEAMS, with which model outputs are being compared, does). So in what sense (or degree of uncertainty) are these comparable?

Reply:

This is a very interesting point. We believe that it should not matter too much how potential evaporation is estimated, as this goes as input to hydrological models, and we are comparing the resulting total actual evaporation (E_{A}) between models and testing if E_{A} is consistent with E_{A} of GLEAM. In our models, E_{P} is used as forcing and models will either calculate E_{A} by explicitly accounting for interception or not. Hence, the E_{A} from the hydrological models combines, implicitly or explicitly (e.g. when an interception reservoir is included), all forms of evaporation. In this sense, we believe it is comparable with E_{A} from GLEAM. In the revised version, we will clarify that we do not consider E_{A} GLEAM to be representative of the truth. However, we believe that the comparison is still valuable to detect outliers, which can be either one/several of the models or the remote-sensing product itself and understand their behavior. GLEAM interception is calculated using precipitation and vegetation characteristics (Miralles et al. 2010). For models with a separate interception module, we also test the consistency of modeled interception E_{I} with GLEAM E_{I}. In the next version of the manuscript, we will use validation studies of GLEAM (Miralles et al. 2011, Miralles et al. 2016) to describe the uncertainties associated with GLEAM estimates.

Comment:

Similarly for the soil moisture comparison. The satellite derived estimates really only deal with near-surface moisture (with a depth that varies with wetness) and that in itself is associated with uncertainty, especially near saturation. There is some discussion here about the issue of comparing relative moisture content in the root zone when the different models parameterize that in different ways and a rather odd correlation analysis with the T parameter – can you not think about how (and if) that data can be used as a hypothesis test. There are clearly similar issues with GRACE and snow cover data (e.g. is fact that some models do not predict snow storage on a day important if snow covers are small)

Reply:

We agree that remotely-sensed soil moisture products provide estimates of near-surface soil moisture, while the represented variable in our models is root-zone soil moisture. The Soil Water Index provides estimates of root-zone soil moisture but requires an estimate of the characteristic time length T. In our study, we show that the models have a different representation of root-zone soil moisture dynamics as shown by the variability in optimal T-values. As also mentioned in the manuscript, the absolute ranges of remotely-sensed estimates of root-zone soil moisture are hardly comparable with relative soil moisture simulated by our models and many studies apply data matching techniques to rescale the product range towards the model. This implies that only the similarity in dynamic patterns can be used to evaluate the models. One specific aspect of interest for hypothesis testing in this study, is to use the remotely-sensed Soil Water Index to identify periods of drying out, where root-zone soil moisture remains constant at its lowest values. In the next version of the manuscript, we will also define a criterion based on GRACE estimates of total storage anomalies to test the behavior of models in terms of drying out of the catchment.

Even if mean annual snow storage is relatively small, snow can be important for specific events. A highly relevant example in the study region in the 2011 rain- on-snow event that caused widespread flooding in the Ardennes. Using criteria for the recall and precision (Figure 5d,e), we will identify behavioral models to derive plausible snow characteristics of the catchment, which can be confronted with expert knowledge. We will account for the fact that a frequent error of the MODIS product is the snow/cloud discrimination (Hall Riggs 2007), which could lead to an overestimation of MODIS snow days and therefore a relatively high ratio of miss over actual positives (1-recall).

Comment:
So rather than have a "so what?" outcome to this paper, I would suggest instead that it should be reformulated into a hypothesis testing framework (EAWAG might be able to make suggestions about how this should be done). This is a real opportunity to frame the issue in this way. Not that because of the uncertainties and commensurability issues that does not imply that any or all of the models will be rejected. That will partly depend on what assumptions and expert knowledge are made in the analysis (– see L450, except that no expert knowledge has really been used in the study as presented). Effectively what you have here are some indices of dynamic behavior with which to evaluate the models – the expert knowledge needs to come in as to how (or IF) those indices (with all the issues with them) can be used to test the models in any way rigorously. This would require very major revisions to the analysis but would make the whole project so much more worthwhile in advancing the modelling process.

Reply:
Yes, we gladly take up this advice and we agree that the study can benefit from your suggestions to go one step further to answer the "so what?" question. In the revised version, we will define a set of (soft) criteria, in the spirit of behavioral modeling, to evaluate not only how consistent the model-internal dynamics are amongst each other but also if the models provide a consistent behavior with what we expect from expert

knowledge in combination with remotely-sensed data. For each retained parameter set, we will evaluate if models can be considered behavioral. This will provide us with an indication of the plausibility of each model to describe several retained aspects of catchment functioning. These results will be presented in a way to allow us to identify potential trade-offs in model capabilities and understand if certain aspects of the parametrization cause a specific model behavior.

Comment:
L35 There are other variables that have been used (and much earlier than the studies cited) – eg. saturated contributing areas (Beven and Kirkby, HSB 1979; Güntner et al., HP 1999; Blazkova et al., HP 2002) and patterns of water tables in many piezometers (Seibert et al., HP 1997; Lamb et al. AWR 1998; Blazkova et al. WRR 2002).

Reply:
Yes agreed, we will add these early studies.

Comment:
L228 drop infinitely – this is misleading. Theoretically yes, but it is directly related to baseflow outflow by water balance and you do not expect baseflow to go to zero in droughts for these catchments. It can also have the advantage of reducing number of parameters required.

Reply:
Yes, it was indeed a theoretical 'infinitely'. We will drop it to avoid confusion.

Comment:
L288 How can GLEAM potential ET be less than the annual estimate cited in L271

(and how undertain are those estimates)

Reply:
The reason why GLEAM potential evaporation (E_{P}) is less than the annual actual evaporation estimate is because GLEAM total actual evaporation ($E_{\mathrm{A,GLEAM}}$) is calculated as $E_{\mathrm{A,GLEAM}} = E_{\mathrm{I}} + S * E_{\mathrm{P}}$, with E_{I} is interception and S is a stress factor depending on the root-zone available water and dynamic vegetation information (Miralles et al. 2011). GLEAM interception is calculated separately and only depends on precipitation and vegetation characteristics. We will clarify this point in the revised version of the paper.

GLEAM does not provide uncertainty ranges, but there are studies that compare GLEAM evaporation estimates with other evaporation products and FLUXNET stations (Miralles et al. 2011, Miralles et al. 2016). These studies will be used to describe the uncertainty of the remotely-sensed evaporation estimates.

Comment:
L397 ecosystems have adapted – but these ecosystems have not surely. In this area they have been affected by forestry and agricultural practices for thousands of years.

Reply:
Yes, the area is affected by commercial forestry and agricultural practices since many years. However, approximately half of the catchment is covered by forests and it seems very unlikely that transpiration is reduced to almost zero for several days in a row each year over the entire catchment. This is also not supported by the satellite-based evaporation and soil moisture products. The MODIS Normalized Difference Vegetation Index (NDVI) data may provide some additional useful information. We will reformulate

this point in the next version.

References
Hall, D. K., Riggs, G. A. (2007). Accuracy assessment of the MODIS snow products. *Hydrological Processes*, 21(12), 1534–1547. https://doi.org/10.1002/hyp.6715

Miralles, D. G., Gash, J. H., Holmes, T. R. H., De Jeu, R. A. M., Dolman, A. J. (2010). Global canopy interception from satellite observations. *Journal of Geophysical Research Atmospheres*, 115(16), 1–8. https://doi.org/10.1029/2009JD013530

Miralles, D. G., Holmes, T. R. H., De Jeu, R. A. M., Gash, J. H., Meesters, A. G. C. A., Dolman, A. J. (2011). Global land-surface evaporation estimated from satellite-based observations. *Hydrology and Earth System Sciences*, 15(2), 453–469. https://doi.org/10.5194/hess-15-453-2011

Miralles, D. G., Jiménez, C., Jung, M., Michel, D., Ershadi, A., Mccabe, M. F., et al. (2016). The WACMOS-ET project - Part 2: Evaluation of global terrestrial evaporation data sets. *Hydrology and Earth System Sciences*, 20(2), 823–842. https://doi.org/10.5194/hess-20-823-2016

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