Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-176-RC1, 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.

Keith Beven (Referee)

k.beven@lancaster.ac.uk

Received and published: 30 April 2020

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

Printer-friendly version

Discussion paper

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.

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!). 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.

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?

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)

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.

Some particular points

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).

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.

L288 How can GLEAM potential ET be less than the annual estimate cited in L271 (and how undertain are those estimates)

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.

Keith Beven

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