Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-176-AC3, 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: 30 June 2020

We thank Shervan Gharari for the interesting discussion on our manuscript and we reflect on each comment below.

Comment 1:
Following the comments by Prof. Beven, I would like to ask the authors a more direct question: "is there any practical use in exploiting the remote sense data in constraining the hydrological models at the scale of interest?" In some applications, bucket-style models are constrained based on evaporation products. I understand that the evaporation products can be used for practical purposes and possibly as

a preliminary benchmark, however, my concerns are: (1) the reduced uncertainty coming from confronting the model with another set of products might result in an "illusion of certainty" in simulation and patterns. As an example, refer to Wang et al., 2015a to see the possible uncertainty in the transpiration/evaporation products from a model. (2) the whole modeling purpose is to predict unknown and come up with the temporal and spatial prediction of some states and fluxes. We then set up a model, say it does or doesn't get the spatial pattern, train it with the result of another model, and then it gets the spatial pattern probably right. What is the end goal of this practice? We can probably join efforts with the developers of the already existing products to improve their products rather than just being a user. Or use a machine-learning algorithm to capture what the patterns in the products are. My question in short; do we learn? Or do we produce a similar product (hopefully a better one)?

Reply 1:
Thank you for raising this interesting discussion. However, it should be clear that we are not using the remote sensing data to calibrate the models. The primary aim of our study is to quantify the differences in internal process representation for a set of models with similar streamflow performance. A secondary objective is to benchmark the internal state and flux variables against remote sensing data. We do not consider the remote sensing data to be representative of the truth, but we use the data to detect potential outliers (being the data itself or one/several models).

It should also not be forgotten that streamflow data also relies on a model with associated uncertainty, namely the rating curve. Of course, there is also (and probably more) uncertainty in evaporation data based on remote sensing, but these products are often also validated against in-situ FLUXNET stations. The use of remote sensing data is valuable as additional independent source of information. Therefore, hydrological applications could benefit from the use of remote sensing data for calibration,

depending on the purpose of the application (e.g. flow predictions in a poorly-gauged catchment).

Comment 2:
GRACE is rather coarse for the basins of interest. It is suggested that the GRACE data should be used for catchment above 150,000 square kilometers (Rodell et al., 2011). This might be counterintuitive; visualization of GRACE over a large area will show that the data is more diffused that its actual resolution. Also, GRACE is very uncertain in itself, using a mean value of its three or more variations may result in deliberate killing of uncertainty (Scanlon et al., 2018).

Reply 2:
We agree that we should be careful with using the GRACE signal over a single pixel, as possible signal leakage from surrounding areas can increase the uncertainty. Even if the catchment area fits within one GRACE pixel, we hypothesize that the signal is of interest as benchmark against which to evaluate the modeled regional seasonal water storage anomalies. In the next version of the manuscript, we will mention these issues in the Data section instead of the Discussion. We will also show the uncertainty of total storage anomalies provided by the three processing centers.

Comment 3:
Checking the consistency of input data is essential before starting the modeling phase. The knowledge gap Section, in my point of view, can be moved earlier in the manuscript and can be populated by quantified evaluation of the available data sets for forcing/calibrating the models. Basically, from the data sets, you have all the components that you need to close the water balance.

$e = sum(P - E - Q) \triangle t - \triangle S$

Can you get e close to zero over a month or a year? (similar experiment to Wang et al., 2015b) Do you have a sense of uncertainty or disagreement between the precipitation and rain gauges?

As low-hanging fruit, it is possible to have an understanding of approximate interception/transpiration for this region. Any flux towers? Study sites from Luxembourg might be helpful? It seems the product you used in this study for evaporation underestimates interception significantly. From their website it seems interception is set to 10% globally (if I interpreted it correctly). Do you perhaps know this ratio for the region of study from this data (model outcome)? This seems to contradict some earlier paper by co-authors on the global uncertainty of the interception/transpiration. Soil evaporation from the used product may include many assumption or simplification similar to land models (Bohn and Vivoni, 2016). Another low-hanging fruit! Can we perhaps estimate the recession coefficient from the hydrograph and compare it with calibrated values in the models to see actually which model structure allows for a more accurate estimation of recession coefficient when calibrated and why [as doctor-father always says]. For example, land models are not suited for this recession inference (Gharari et al., 2019).

Reply 3:
We agree that checking the consistency of the input data is essential and a standard procedure in any modeling practice. We will mention the uncertainty in the evaluation variables in the Data section rather than in the Discussion. We limited the scope of our study to the evaluation of internal variables against remote sensing data. Comparing catchment scale averages with point measurements of evaporation also faces substantial commensurability issues. However, it is good to know that GLEAM evaporation was evaluated using FLUXNET data (including the station of Lonzee

which is close to the study area). For this station, a correlation coefficient between GLEAM and FLUXNET of 0.91 for the daily time step and similar annual rates are reported in Miralles et al. (2011). In Miralles et al. (2016), the likely underestimation of interception evaporation and the likely overestimation of total actual evaporation compared to other products is discussed. We will make use of these evaluation studies to describe the uncertainty of the evaluation variables and derive soft criteria to evaluate the plausibility of the set of models.

Comment 4:
The area is mostly agriculture, is there any regulation on the stream than may affect your inference. Referring to section 5.3, the area is mostly agriculture, to my understanding, the Sumax/root zone storage co-evolution is hypothesized for forests (that has a life of more than a year). Agricultural lands do not follow any of that logic, it does what farmers do (there might be some correlation). Maybe you can argue around the rain-fed nature of agriculture in this region but still, crops have a lifespan of a season. Land models can see the variation of leaf area index (LAI) and with some modification even variation of root zones over period of time. That can be a better testbed for exploring root zoon hypothesis than bucket-style models.

Reply 4:
We agree with your suggestion of looking at vegetation indices for additional information on this matter. Given that almost half of the catchment of the Ourthe at Tabreux is covered by forests (46%), we do not expect transpiration rates to drop to approximately zero for several days in a row each year over the entire catchment.

Comment 5:
"The T-value has previously been positively correlated with root-zone storage capacity, assuming a high temporal variability of root-zone soil moisture and therefore a low

T-value for small root-zone storage capacities $S_{\mathrm{R,max}}$ (Bouaziz et al., 2020)". Possible that I totally get it wrong, but if I understand correctly, Bouaziz et al., 2020 used a hydrological model in combination with satellite observation. Is this a model result that is used for intercomparing rather than the satellite observation itself? Maybe separate the data (products) into groups of "directly observed" and "inferred based on a model".

Reply 5:
The modeled root-zone soil moisture of each model is evaluated against satellite observations of the Soil Water Index. The Soil Water Index is provided by the Copernicus Global Land Service for several T-values. The Soil Water Index is derived from near-surface soil moisture to represent root-zone soil moisture but requires an estimate of the T-value. In Bouaziz et al. (2020), we found a link between the optimal T-value and the root-zone storage capacity. So, to answer the question, yes, we are using the by Copernicus provided satellite data for the comparison, and yes the data is inferred based on an algorithm. However, all the evaluation variables we are using are relying on algorithms, even streamflow is not directly measured.

Comment 6:
Upscaling of snow cover to basin level is a tricky business. Snow storage, snowpack extent may not be uniform over an area (Cherkauer et al., 2003). Also, the snowpack can persist with temperature much higher than the phase-change temperature identified in the model. Snowpack may also stays longer under canopy. The phase-change temperature can have a range, for example, for the VIC model this is a transitional span of temperature (for example from -2C to 2C) that can be tuned for the same reason (snow precipitation). I would suggest checking the snow extent versus the temperature first. This might give insight into whether or not any model can simulate the observed snow extent given the temperature. Also, snow under canopy may stay longer, does the product you use capture that?

Reply 6:
Thank you, it is indeed a good suggestion to compare snow extent and temperature. As the snow cover relies on MODIS imagery, it is unlikely that snow under the canopy is captured by the product.

Comment 7:
As a modeler that might be interested to model the basins of interest, what is the take-home message for me. I assume one of the aims of an inter-comparison project is the knowledge mobilization of already known facts about basin(s) to the wider community. This can be done better in this manuscript I would say. Perhaps, identifying the target audience. Is the manuscript targeted for catchment hydrology? Or Large-scale hydrology? The current manuscript does not server any.

I would say, as coordination of the large team takes a lot of efforts and work, maybe give a new dimension to your paper by elaborating the organizational efforts put into this study (why did you initiate this inter-comparison, why the current list of models and authors, what made you to choose them? what effect it might have on real-world application, etc).

Reply 7:
We will clarify that the take-home message of the manuscript 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 our opinion, this is relevant for catchment hydrology and large-scale hydrology as both rely on model selection. In this context, the selection of the Ourthe catchment is more a case study

to demonstrate the different internal process representation.

In the next version of the manuscript, we will also use expert knowledge in combination with the remote sensing data to evaluate the plausibility of model behavior for a selection of criteria.

The study is a joint research effort of institutes and universities gathering each year at Liège Université for the symposium on hydrological modelling of the Meuse basin to exchange knowledge and work together on the Meuse basin, as also mentioned in the discussion. In the revised version, we will discuss the challenges inherent to such a comparison study.

Comment 8:
I would suggest the authors clarify their research equations in the beginning and come back to the research questions in the conclusions. In the current version, there are no tangible research questions. For example, "Haddeland et al. (2011) and Schewe et al. (2014) compared global hydrological models and found that differences between models are a major source of uncertainty." I think this is what you can reflect/elaborate on in your conclusions (hopefully quantitively)?

One collusion from this study can be for example, "a two-bucket model with snow component is sufficient enough to get the dynamic of the data we selected". Can this be one of your conclusions?

Some studies from the land modeling community can be helpful in this regard. For example, Bets et al., 2015 provided a structure for the comparison (including evaluation, comparison, benchmarking, fit for purpose, utilizing the available data,

etc). Following this structure or similar structures can hopefully clarify the manuscript more. One benchmarking strategy could be ensemble simulation of all models within their prespecified parameter ranges. This can be the basis for comparison when the model is calibrated on the streamflow and subsequently on other data sets such as evaporation in a stepwise fashion. This seems to be not a lot of work as the models are already set up.

Moreover, the land model studies can provide more insight into large scale modeling and their related issues. For example, Crow et al. (2003) is a classic example. In my point of view far ahead of its time and not very well received [the same work nowadays would probably have 20 authors with the same citation level in a single year and will be magically highlighted!]. Another great example to show uncertainty in large scale models in reproducing mean and variability (Koster, R.D. and P. Mahanama 2012) with a very simple model. Another example is Hurkmans et al. 2008. These studies and similar works may provide a better understanding of the exploitation of additional data in large scale modeling and associated uncertainties.

Reply 8:
We will sharpen the research question and conclusion. In this study, we test the hypothesis that models with similar streamflow performance have similar internal process representation, as stated in the last paragraph of the introduction. We conclude by saying that models have different internal representations of water storage and release. This suggests that all models can't simultaneously be different and close to reality.

In the next version of the manuscript, we will go one step further by evaluating the plausibility of the models and identify behavioral models in view of the remote sensing data and expert knowledge.

Although an interesting suggestion for follow-up studies to constrain the models using the remote sensing data, this is outside the scope of the current study.

Comment 9:
Concerning FLEX-Topo. It seems to be the only semi-distributed model among all the other models. Have you properly constrained the component of this model (or do you have enough expert knowledge to do so)? It would be good to highlight the advantage of the semi- distributed model here if any. The control over the different components of FLEX-Topo becomes increasingly hard if the code is written separately for each landscape (different structures). I tried to have a similar code for each landscape and recreate the desired structure just by adjusting the parameters. That provides better control over the performance of each landscape. For example, did you check the transpiration of each landscape? Sometimes it is the case that soil moisture from one landscape is empty and the other landscapes are evaporating at the maximum rate.

Reply 9:
The calibration of the models was done in the previous study (de Boer-Euser et al., 2017), in which we found that models had similar streamflow performances based on commonly used metrics. Here, we are consistently using the previous calibration to assess internal model representation. Of course, model deficiencies appearing in this study are helpful to improve the parametrization of the model but going into such details for each model is outside the scope of the current study.

Comment 10:
I didn't know that FLEX-Topo got a sublimation component. How that is implemented? Is sublimation a major process in the region of study? I would not say so. Sublimation

is also a tricky process; a magical one! it can account for uncertainty in snowpack similar to the transpiration for soil moisture. There is also a refreezing formulation for one of the models. Interested to know how that happens in a model that may not close the energy balance. It would be good to include all the model formulation in the Appendix if not too much work.

Reply 10:
The raised discussion is interesting. The sublimation component implemented in the FLEX-Topo model is described in de Boer-Euser (2017). It is a simple representation that allows evaporation from the snowpack at potential rate, provided there is enough snow storage. As precipitation mostly falls as rain and not as snow, we agree that it is not a major process for the study region. Additionally, the magnitude of sublimation is likely to be rather low due to the limited direct radiation in winter in Belgium. We are providing the most relevant equations for our study in Table 2 and 3 and we refer to the references on the model descriptions for more details.

Comment 11:
The figures presenting the results are very hard to follow. I am not sure if I understand most of them. I would suggest simplifying them.

Reply 11:
Unfortunately, given the general character of this comment, it is unclear which aspects of the figures are very hard to follow. Following the comments of the referees, we intend to adapt Figure 9 in the next version.

Comment 12:
A question from Prof. Beven and maybe the authors; is that possible to even reject

a model in large scale modeling? From my experience and due to the issue of scale (and observation at that scale), most of the models can be accepted. For example, in a recent modeling effort that we have done (Gharari et al., 2020), the VIC model with the only micropore and with only macropore water movement yields the same result when calibrated (exploring the inclusion of macropore water movement in land models; aligned with Beven and Germann 1982, to Beven 2018). How should I justify macropore versus micropore at that scale for a colleague whose entire career is focused on how to properly/mathematically represent micropore water movement? What is the path forward? I appreciate your thoughts on that.

Reply 12:
Thank you for raising this interesting comment. In our study, we are reluctant to reject models because any rejection is conditional on many factors, including the individual parameter selection strategies chosen by the individual contributing institutions. And model rejection would only be valid for the study catchment. And these specific circumstances could be forgotten if taken out of context and the label “rejected” would remain, even if this is not justified. Additionally, there is a large uncertainty in the used evaluation variables. The uncertainty estimates of remote sensing data are usually not available and using them in a meaningful way would uncover many questions, which is outside the scope of the current study.

Instead, we propose to evaluate and rank the models and define soft criteria to test the plausibility of model behavior in terms of the remote sensing data and expert knowledge. We will include this in the next version of the manuscript.

Regarding the specific case you are mentioning on macropore versus micropore flow, some answers might be provided in Zehe et al. (2014). We believe this also depends on the study area and on the availability of evaluation variables and their uncertainty

estimates.

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

de Boer-Euser, T. (2017). Added value of distribution in rainfall-runoff models for the Meuse basin. Doctorale thesis TU Delft Repository. https://doi.org/10.4233/uuid:89a78ae9-7ffb-4260-b25d-698854210fa8

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

Zehe, E., Ehret, U., Pfister, L., Blume, T., Schröder, B., Westhoff, M., et al. (2014). HESS Opinions: From response units to functional units: A thermodynamic reinterpretation of the HRU concept to link spatial organization and functioning of intermediate scale catchments. *Hydrology and Earth System Sciences*, 18(11), 4635–4655. https://doi.org/10.5194/hess-18-4635-2014

Beven, K. and Germann, P., 1982. Macropores and water flow in soils. Water resources research, 18(5), pp.1311-1325.

Beven, K., 2018. A Century of Denial: Preferential and Nonequilibrium Water Flow in Soils, 1864-1984. Vadose Zone Journal, 17(1).

Bouaziz, L. J., Steele-Dunne, S. C., Schellekens, J., Weerts, A. H., Stam, J., Sprokkereef, E., Winsemius, H. H., Savenije, H. H., and Hrachowitz, M.: Improved understanding of the link between catchment-scale vegetation accessible storage and satellite-derived Soil Water Index, Water Resources Research, https://doi.org/10.1029/2019WR026365, 2020.

Best, M.J., Abramowitz, G., Johnson, H.R., Pitman, A.J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P.A., Dong, J. and Ek, M., 2015. The plumbing of land surface models: benchmarking model performance. Journal of Hydrometeorology, 16(3), pp.1425-1442.

Bohn, T. J., and E. R. Vivoni, 2016: Process-based characterization of evapotranspiration sources in the North American monsoon region, Water Resour. Res., 52(1), 358-384, doi:10.1002/2015WR017934.

Wang, S., Huang, J., Yang, D., Pavlic, G., and Li, J., 2015. LongâĂŘterm water budget imbalances and error sources for cold region drainage basins. Hydrological processes, 29(9), pp.2125-2136.

Wang, S., Pan, M., Mu, Q., Shi, X., Mao, J., Brümmer, C., Jassal, R.S., Krishnan, P., Li, J. and Black, T.A., 2015. Comparing evapotranspiration from eddy covariance measurements, water budgets, remote sensing, and land surface models over Canada. Journal of Hydrometeorology, 16(4), pp.1540-1560.

Crow, W.T., Wood, E.F., and Pan, M., 2003. Multiobjective calibration of land surface model evapotranspiration predictions using streamflow observations and spaceborne surface radiometric temperature retrievals. Journal of Geophysical Research: Atmospheres, 108(D23).

Scanlon, B.R., Zhang, Z., Save, H., Sun, A.Y., Schmied, H.M., Van Beek, L.P., Wiese, D.N., Wada, Y., Long, D., Reedy, R.C. and Longuevergne, L., 2018. Global models underestimate large decadal declining and rising water storage trends relative to GRACE satellite data. Proceedings of the National Academy of Sciences, 115(6), pp.E1080-E1089.

Rodell, M., McWilliams, E.B., Famiglietti, J.S., Beaudoing, H.K., and Nigro, J., 2011. Estimating evapotranspiration using an observation based terrestrial water budget. Hydrological Processes, 25(26), pp.4082-4092.

Haddeland, I., Clark, D. B., Franssen, W., Ludwig, F., Voß, F., Arnell, N. W., Bertrand, N., Best, M., Folwell, S., Gerten, D., Gomes, S., Gosling, S. N., Hagemann, S., Hanasaki, N., Harding, R., Heinke, J., Kabat, P., Koirala, S., Oki, T., Polcher, J., Stacke, T., Viterbo, P., Wee- don, G. P., and Yeh, P.: Multimodel estimate of the global terrestrial water balance: Setup and first results, Journal of Hydrometeorology, 12, 869–884, https://doi.org/10.1175/2011JHM1324.1, 2011.

Schewe, J., Heinke, J., Gerten, D., Haddeland, I., Arnell, N. W., Clark, D. B., Dankers, R., Eisner, S., Fekete, B. M., ColoÌĄn-GonzaÌĄlez, F. J., Gosling, S. N., Kim, H., Liu, X., Masaki, Y., Portmann, F. T., Satoh, Y., Stacke, T., Tang, Q., Wada, Y., Wisser, D., Albrecht, T., Frieler, K., Piontek, F., Warszawski, L., and Kabat, P.: Multimodel assessment of water scarcity under climate change, Proceedings of the National Academy of Sciences of the United States of America, 111, 3245–3250, https://doi.org/10.1073/pnas.1222460110, 2014.

Koster, R.D., and P. Mahanama, S.P., 2012. Land surface controls on hydroclimatic means and variability. Journal of Hydrometeorology , 13(5), pp.1604-1620.

Hurkmans, R.T.W.L., De Moel, H., Aerts, J.C.J.H. and Troch, P.A., 2008. Water balance versus land surface model in the simulation of Rhine river discharges. Water resources research, 44(1).

Gharari, S., Clark, M., Mizukami, N., Wong, J.S., Pietroniro, A., and Wheater, H., 2019. Improving the representation of subsurface water movement in land models. Journal of Hydrometeorology, (2019).

Gharari, S., Clark, M. P., Mizukami, N., Knoben, W. J. M., Wong, J. S., and Pietroniro, A.: Flexible vector-based spatial configurations in land models, Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-111, in review, 2020.

Cherkauer, K.A., Bowling, L.C. and Lettenmaier, D.P., 2003. Variable infiltration capacity cold land process model updates. Global and Planetary Change, 38(1-2), pp.151-159.

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