A short comment on Bouaziz et al., 2020 and a question from Prof. Beven:

I think the work has a lot of potentials and there is room for improvement which motivated me to write this short 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)?
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).
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.
   $e = \text{sum}(P\text{-}E\text{-}Q)\Delta t\text{-}\Delta S$
   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).

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.

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 SR,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".

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 -2°C to 2°C) 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?

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).

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.

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.

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.

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.

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.

I am confident this manuscript will be an interesting one. Hope that my comments are helpful.

Finally, please note that any reference to my work was just for explanation, discussion and clarification. I don't quest for citations.

With kind regards,
Shervan Gharari

Reference:

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. Long-term 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., Colón-González, 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. J*ournal 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.