

Interactive comment on “Adaptive clustering: A method to analyze dynamical similarity and to reduce redundancies in distributed (hydrological) modeling” by Uwe Ehret et al.

Shervan Gharari (Referee)

shervangharari@yahoo.com

Received and published: 19 March 2020

The manuscript introduces an adaptive clustering for grouping the hydrological elements of a distributed model. I enjoyed reading the manuscript and I think the work is interesting and eventually can be published. However, I should raise few overall points mostly regarding the presentation and the aim of the study.

1- The use of English language is far from being perfect. It took me more than usual to read through the manuscript due to insufficient use of English language. The sentences are very long, the wording are sometimes very awkward. For example, the first sentence of the abstract is very hard to follow. It is amalgamation of information which,

C1

at the end, does not say much about the intention of this study (the intention being representative model with lower computational demand). Here I give examples from the text. Page 4, line 4, “offers full code control”; full code control is very subjective. Page 4, line 21, “. . .straightforward non-iterative forward-in-time numerical scheme”. Page 7, line 19 “majority vote”. Page 11, line 6 “for adaptive clustering to make sense”. Page 13, line7 “gold standard” instead of synthetic. Page 14, line 3, “keep things simple”. I would strongly recommend the authors to have this manuscript proofread by a native speaker.

2- The Introduction seems to be superficial. I would say the paper is about representation of the system in a model vs computational time/resource. In land [surface] modeling community there is significant body of literature devoted to the effect of grid size (computational burden vs spatial representation) and example of them can be Melsen et al., 2016 (and many more). This is the case in hydrological rainfall/runoff models as well (Liu et al., 2016 and many more). In its current form the Introduction starts with general reflection on sophisticated processed-based models; then moves to concept of co-evolution (which is not directly relevant to the message this study wants to convey) and then comes down to clustering. I would suggest to re-organize the Introduction to reflect on pervious works on computational burden vs spatial representation, clustering and its application in hydrological similarities and finally make it clear what the reader should expect from this paper.

3- Section 2 can be better organized. There is still mention of the CATFLOW in this section. I would suggest the authors to shorten the text in section 2.1 and directly explain the model. I suggest using transpiration coefficient instead of crop coefficient in Table-1 as the region is partly covered with forest. It seems to me that the routing is only a linear reservoir (equation-11). Can this representation simulation lag function or unit hydrograph which is often used in the models? I would say no, therefore the highest streamflow peaks are the same as precipitation peaks. The are k_r and k_b identifiable/related? it seems to a redundancy in the two processes/parameters. It is

C2

unclear how the computational units are set up, based on superposition of all the geo-spatial data or at the sub-basin level (later in Table-7 it becomes more apparent that the setup is at sub-basin level). Which data is directly used in setting up the model? I would suggest providing a table or bullet points for that; now it is very scattered around. I suggest removing equation 12, to a separate Section with entropy as measure of performance and uncertainty. Why there is NS values report in this Section? is the satellite based evaporation a result of more sophisticated model (such as a land model)? Section 2.2 is the heart of this manuscript. I would say it should be presented separately in a Section before the model is presented as this approach is model independent. Don't give sub-section "main steps" title (2.2.1). why section 2.3.1 is called "hydrological system analysis"; it is about entropy as a measure of uncertainty (as NS is a measure of performance). Section 2.3.2 is again called adaptive clustering, similar to section 2.2. and again, in Section 2.3.2 the authors are referring to CATFLOW and MIKE SHE, etc.

4- Section 3 is also rather hard to follow as well. The primary message of the manuscript is about tradeoff between spatial representation and computational resources (time). It is best to start from Figure-6 and 7 and then move to Figure-5 for example. Page 20 line 9, why it is "striking" that the entropies are lower than the uniform? I would always expect so. It is also expected that the entropy is lower for the recession and higher for rising discharge. This is kind of similar to the heteroscedasticity assumption on the error as well (more diffused with higher discharges). If only observation is used with varying error assumption, higher streamflow will have higher entropy and lower streamflow will have lower entropy. I don't really see the information the first and second paragraph of Section 3.1 provide. The third paragraphs, starting with "In Fig.6", is also kind of obvious. Are these what we have trained the model to do? The result is like a self-fulfilling prophecy (I will elaborate on this later).

5- The final Section is lacking proper discussion on what we have read in this manuscript. What can be the take home message for a modeler who want to model this basin in the future. A bullet point conclusion of this study is also appreciated.

C3

6- For me personally, moving from the world of conceptual models to land models, I would like to question the motivation of this study. Although saving time is valuable but having method that needs model re-run or updating for more complex model is terribly cumbersome. This is the reason why the authors have chosen to use SHM rather than CATFLOW for example. I also didn't really understand the model set up, are we looking at comparison of models with clustering with a synthetic case? If that is the case, then the comparison of what is the best representation in a specific time of the year is dependent on the most elaborative set up and its simulations (basically the most computationally expensive model should be set up and simulated). Also, one might say there are easier time stepping approaches for lower computational costs. For example, for the high flow the model temporal resolution can be maximum (1 day) and during recessions it can be up to couple of days. These type of approach to reduce computational time is much easier to implement. Moreover, if computational time is considered, I would say Matlab is not the best choice. Land models implemented in C or Fortran can handle hourly simulation (with much more variables, IO traffic) much faster. BTW, I also missed the modeling time stepping; please clarify. I would suggest the authors to clarify their methodology of the clustering. Please make more time and maybe better visualization to convey the message here. As I said earlier, I would suggest the author to allocate a full Section for the clustering method they have proposed.

Overall, the manuscript is interesting but I think much more work is needed to meet the publication standard. I will suggest Major revision for this manuscript. I hope my comments can help the authors better present their good work.

With kind regards,

Shervan Gharari

Liu, H., Tolson, B.A., Craig, J.R. and Shafii, M., 2016. A priori discretization error metrics for distributed hydrologic modeling applications. *Journal of hydrology*, 543,

C4

pp.873-891.

Melsen, L., Teuling, A., Torfs, P., Zappa, M., Mizukami, N., Clark, M. and Uijlenhoet, R., 2016. Representation of spatial and temporal variability in large-domain hydrological models: case study for a mesoscale pre-Alpine basin. *Hydrology and Earth System Sciences*, 20(6), p.2207.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., <https://doi.org/10.5194/hess-2020-65>, 2020.