This is a review of “Simultaneous calibration of hydrological models in geographical space” by A. Bárdossy, Y. Huang, and T. Wagener.

The general idea of this paper is to separate the model parameters into two groups: One reflecting water balance and the other the runoff dynamics. The water balance parameters are transformed into a new variable “n”. The “n” parameter is determined on ungauged basins (offline) and the dynamics parameters are determined by regional calibration. The authors show that the dynamical model parameters are good even when transferred to very different catchments.

While I like the innovative idea of separating the parameters into 2 groups and attempting to identify from where the model skill originates, I have noted many problems with the methodology, discussion and setting of the work performed. My recommendation is “Major revision”.

There are two main issues with the paper:

1- The authors do not do a sufficient job putting their work in context. The literature review is outdated and not very useful in setting the paper in the current context.
2- The discussion lacks in depth. Results should be compared to other studies and discussed. As o now, the discussion is mainly a recap of the results.

Also, in many places, English proofreading should be performed as some sentences are difficult to understand and interpret.

Specific questions / issues:

How does the loss in performance compare to other regionalization methods? Is the robustness gained worth it if many catchments offer suboptimal performance compared to a multi-donor regionalization approach?

How does catchment similarity impact performance in calibration/validation? The paper states that the climate data dominates over catchment characteristics, but can the authors quantify the correlation or relationship to catchment descriptors?

Table 1: I do not feel that relative humidity is an acceptable physical catchment descriptor. Perhaps change to “physioclimatic” or something of the sorts to indicate that there is also climate data taken into account. Also, using base flow index as a descriptor while working with ungauged basins seems like it is cheating. Perhaps clearly indicate that catchment descriptors are not used for the parameter transfer. In this manner there will be no conflict.

Introduction:

References are dating, lots of research has been done in the past few years regarding this subject.

It would be nice to see a range (histogram perhaps) of the 10000 calibrated parameter sets. For example, in figures 5 and 6, the large spread of values would lead to believe that the NSE values are very heterogeneous. In figure 12, we see that NS skill ranges from 0.2 to 0.8. What would the difference be if the best (0.8 NS) parameter set was selected?

11226 Lines 22-23 : Missing “is”

11227 Line 9 : Make a single sentence out of the two.

Landscapes are formed during long time through climate, and are thus in a kind of quasi equilibrium.

How about 2 very different catchments? Do you expect water dynamics to be similar for a steep catchment vs a flat catchment? Must there not be a pre-processing of similarity index for the catchments? While at it, why not go one step further and do physical similarity regionalization?

11229

I do not understand this part of the sentence (totally > 1010-year discharge calculations)

11233

Line 14: “...This is necessary as it is thought to establish correct water balances”. But what do you make of equifinality? Surely this equation will produce different n values depending on the calibration parameter set.

11235

How do you compute the long-term water balance if the catchment is ungauged? The way I see it, there are two options. Either the n parameter is adjusted based on the actual gauged data (biasing the results since the parameter set received will need to conform to the n parameter) or there is another way to estimate the value of n at an ungauged site, namely using other regionalization techniques. It is imperative that this be discussed beforehand.

11236

Can you explain the differences observed? What happens when the “good basin” parameters are transferred to the “bad basin” and that the modelling fails? What do you observe in the hydrograph? Why is this not seen in the reverse order?

11241

I do not understand the sentence: “This is as expected that there is less common behavior of a large set of catchments as for a few”

11242

“But for the Rottweil catchment, model performance is worse than for the Fils catchment. It indicates that there is some skill in the transferred parameters, but the differences are substantial. Figures 15 and

16 show part of the observed and the modeled hydrographs using the NS performance measure. We can see the transfer is reasonable and the dynamics of the discharge are similar to the US case. This experiment demonstrated that even very distant and different catchments may behave similarly.

// Not sure that this is what is implied from the text. The second sentence says that the differences are substantial, whereas the last sentence says that the catchments may behave similarly. Also note the strong underestimation of peak flows.

11246 - Conclusion

Lines 7-11 : I do not agree with this assessment. Are the authors implying that very different catchments (mountainous vs flat, forest vs grasslands, difference in lithography and geology, etc.) react the same to similar rainfall? Could it simply be that by selecting the lowest common “acceptable” parameter set, the method neglects key differences, thus skewing the results towards this conclusion? More details are needed to justify this point.

Discussion: The discussion must be improved significantly and expanded:

9.1 -> Does this “deepest parameter set” have stronger ties to physical catchment descriptors than other parameter sets?

9.2 -> It is critical that the authors discuss the estimation of n at ungauged sites. How is the parameter estimated if there is no streamflow? Does it use observed streamflow to estimate properly and then only the dynamics parameters are fitted? If so, how does conditioning the dynamic parameters to the n parameter impact the result? What if we use a “bad” n?

9.3 -> Ok, place 9.3 before 9.2 or talk about this point much earlier. It is absolutely critical for understanding the paper. Also, if the n parameter is easily regionalized through space based on proximity, why not use the spatial proximity regionalization method for the other parameters? One can also combine spatial proximity and physical similarity with multiple donors to improve performance, such as described in Oudin et al. 2008 and applied in Zhang and Chiew 2009; Arsenault and Brissette 2014; Zelelew and Alfredsen 2014, etc.

Speaking of which, the authors should point out explicitly how regional calibration instead of direct regionalization, based on past results (Parajka et al. 2007; Ricard et al 2013, Gaborit et al. 2015) which discuss regional calibration and its strengths/weaknesses.

How would traditional regionalization methods fare if allowed the advantage of forcing the “n” parameter as in this case?

In my opinion, the discussion needs to be improved substantially and should have references to the current state-of-the-art to better relate the results in this paper to the literature.

References:

Gaborit, E., Ricard, S., Lachance-Cloutier, S., Anctil, F. And Turcotte, R. (2015). Comparing global and local calibration schemes from a differential split-sample test perspective. Canadian Journal of Earth Sciences, 52(11): 990-999, 10.1139/cjes-2015-0015

Ricard S, Bourdillon R, Roussel D, Turcotte R. 2013. Global calibration of distributed hydrological models for large-scale applications. Journal of Hydrologic Engineering 18: 719-721

Parajka, J., Bloschl, G., and Merz, R., 2007. Regional calibration of catchment models: Potential for ungauged catchments, Water Resour. Res., 43, 1–16, doi:10.1029/2006WR005271.

Arsenault, R. and Brissette, F. (2014) Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. Water Resour. Res. 50 (7) , 6135-6153.

Zelelew, M.B., and Alfredsen, K., 2014. Transferability of hydrological model parameter spaces in the estimation of runoff in ungauged catchments. Hydrological Sciences Journal, 59 (8), 1470–1490. http://dx.doi.org/10.1080/02626667.2013.838003

Zhang, Y. and Chiew, F.H.S., 2009. Relative merits of different methods for runoff predictions in ungauged catchments. Water Resources Research, 45, W07412.