Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC2
https://doi.org/10.5194/hess-2021-271-RC2, 2021

Comment on hess-2021-271

Anonymous Referee #2

Referee comment on "Quantifying the Regional Water Balance of the Ethiopian Rift Valley Lake Basin Using an Uncertainty Estimation Framework" by Tesfalem Abraham et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2021-271-RC2, 2021

This manuscript was challenging to assess. The transferability of model parameters calibrated at gauged locations to ungauged locations using a regionalization approach where parameters are estimated using catchment properties able to estimated at the ungauged catchment is, in many ways, well-worn territory, as the authors also note (L45-58). Much of the discussion in Section 5.2 also points to results that are consistent with previous studies. In my opinion, there has been inconsistent success demonstrated in previous studies as to the utility of this approach and the results presented here are no different than previous studies have found.

The question then is both whether the approach presented here represents such a difference from past studies as to be a substantial departure from past practices that it would be of value to report the results and that the study area and catchments are sufficient to make broader conclusions about this potential new approach.

From what I am able to understand about the approach and the catchments, neither of these meet the criteria so as to make a substantial and broader contribution to our understanding of why or how we might improve on regionalization approaches for parameter estimation at ungauged locations.

My recommendation is based on a number of what I see as serious methodological and evaluation questions as well as a highly complimentary presentation of a limited application of the approach to only a small number of catchments. I describe these issues in more detail below. If the manuscript does receive a recommendation other than Reject, I also offer additional minor and editorial comments that the authors need to consider in their revision.

(1) Broader contribution of the work

(1a) The use of weighted least squares (L200), although not necessarily a substantial advance, is what I believe to be the novel aspect of the study. Perhaps if this were emphasized more in the introduction and concentrated in more detail with the existing studies, it might become more clear that this is a more substantial contribution that the impression I was left with. Otherwise, this being mentioned is more detail so later in the discussion paper (in the methods) contributes to this point being lost. I would also be more explicit as to how this work differs from Wagener and Wheater (2006) and the follow-on studies that have cited that paper.

(1b) I do not agree that this work is novel because these approaches have only been applied in data-rich regions (L56-58). In my opinion, the reason these methods have been applied in data-rich areas is to test the limits of these approaches. Even then, the results have certainly been mixed. Certainly, you could have chosen a more data-rich area to test this approach and then removed streamgauges to understand the effects of gauges on the performance of the method.

(1c) Linked to Comment 1b, it is difficult to make broader conclusive statements about the utility of this approach when only 16 (or 14) catchments are being used. Either way, for a regionalization study, 14-16 gauges is a very limited number. I realize that 2 catchments were removed because they were poor performing, which reduced the number of catchments to 14. I am not sure if removing these 2 catchments was the correct thing to do here; are they poorly performing because the underlying model is not a good representation? Were these locations removed just to improve your own study results? It seemed as though there was not a solid technical reason to remove these gauges from the study.

(2) Methodological and evaluation concerns

(2a) I missed where non-linear regressions are being used in conjunction with weighted (linear) least squares (L200)? I see later in L218 that the non-linear regression is discussed but with not much justification or explanation as to why this is the case.

The form of Equation 10 looks like the form of a regression equation when the regression was performed in log space and then transformed back to normal space. In other words, the logs of the response and predictor variables were taken to linearize the relation between them (to better ensure the assumption of a linear relation for the regression) and then the regression was performed on the log-transformed variables.

Of course, an additive model in log space is a multiplicative model in normal space. So to get the values back to normal space, Equation 10 is what the regression equation looks like when the additive linear model is re-transformed back to normal space.

Seeing that you do not mention anywhere that you performed the regression on the logs of the response and predictor variables, I am not understanding why you would apply the non-linear equation shown in Equation 10 for this reason. More justification is then needed for the application of Equation 10 to the data.

(2b) Keep in mind NSE values less than 0 have the interpretation that the mean of the data is a better model than the model being proposed (in this case, the regionalization model is worse than simply using the mean of the data as the model). NSE values less than 0.5 are likely poor fits and those less than 0.25 are approaching the case where would have been better off using the mean of the observed data instead of the regionalization approach. You make the statement on L383 that "79% of the catchments had a NSE > 0"; however, I do not believe this is a statement that puts the method in a positive light. Surely you could find a simpler model (even the drainage area ratio, perhaps) that would achieve the same success as having 80% of the model results better than using the mean of the data. The reverse of the statement on L383 means that 3 catchments (20%) of the 14 catchments did have an NSE < 0 using this regionalization method. How would one in practice guarantee that they were applying the regionalization to an ungauged location where the method would not provide a worse estimate than the mean of the data?

(2c) In calculating the NSE based on the actual values of flow, what were the range of flow values? If no attempt to balance the weight of the high and low flows in the NSE calculation, the NSE itself would be most affected by the fit of the model at the highest flows, and thus the NSE may only be a reflection of how well the parameters are estimating flows for the largest flows. For example, a difference of 0.1 cms and 5 cms would be a poor fit but if your high flow values are large (on the order of 100s or 1000s of cms) a difference of 4.9 cms would register as an excellent fit for NSE and this fit - simply by the numerical calculation of the NSE - would swamp any of the fits at the low flows since the differences squared would be so much less. Would it not be better to compute the NSE on the logs of the streamflows? Or at least split the flows into high, low, and mid flows so that these issues of scale are not affecting the interpretation of fit?

(2d) There are no regression equations provided or regression diagnostics for the equations so that one could assess whether these are valid regression equations with statistically significant explanatory variables. To use these regression equations in prediction mode and calculate uncertainty and prediction intervals (which is done in Section 4.2), the behavior of the regression equations must adhere to the properties of a linear regression (statistically significant explanatory variables, homoscedastic residuals, uncorrelated and normally-distributed residuals, and uncorrelated explanatory variables).

(2e) In Equation 8, the weights are described as 1/CV (the reciprocal of the CV; L213). I was having difficulty understanding this. The CV = standard deviation / mean; the reciprocal is then mean / standard deviation. The weights in a weighted least squares regression are, ideally, 1 / variance. How then were you able to achieve a weight equal to 1 / variance by using the inverse of the CV? This needs to be clarified in more detail so the reader can follow along.

(2f) For insensitive parameters (Figure 4), such as Mmaxeas, it seems it would be advantageous to incorporate this knowledge somehow into your regionalization scheme, although it would be unclear how this would hold up for ungauged locations. On L432-433, the statement is made "Our study shows the insensitivity of model parameters to be related to catchment properties." I am not sure how that can be. If a parameter is insensitive to model calibration, then it would have no preference for the value; therefore, why would one expect this parameter to be estimable or predictable? Would it not be better to just simply randomly generate a value for this parameter from a uniform distribution of values given the parameter range in Table 3?

Then in an ungauged location, how would one be able to predict whether this was a catchment that was insensitive to the parameter Mmaxeas or if it was one of the 3 catchments (figure 4) that was highly sensitive to this parameter?

Could you simplify your regionalization by only regionalizing sensitive parameters and then assigning a random, uniformly distributed value to the insensitive parameters?

(2g) The use of the word "stable" parameter set is not very clear. The definition of the "stable" parameter set is the set of parameters that "shows the smallest difference between the calibration and validation NSE". But this does not consider also picking the parameter set with the highest NSE as well. It also does not explain how the validation period has a higher NSE than the calibration period for some catchments. Lastly, how does this criteria help in determining the best parameter set for regionalization? What is the benefit of transferability when you have a "stable" set of parameters at one location? In other words, what would be the guarantee that a parameter set will work well at another location just because it is "stable"?

(2h) Section 3.5: I am not understanding the validation and selection of the parameter sets (L229-233). From what I could understand, the parameter sets are tested on the validation phase and in leave-one-out mode. What is the leave-one-out method not sufficient itself to assess the performance of the method? Also, it would seem a longer period of calibration (one that includes both of what you term the calibration and validation periods) provide better parameter estimates? I am not understanding why the leave-one-out approach to measure uncertainty is not enough to evaluate the approach?

It is also unclear in the methods when calibration and validation are used. You could use Figure 2 to clarify this. From my reading, in Figure 2, you could modify the box "regional regression" to read "regional regression using calibration parameters" and then "evaluation of the regression procedure using validation and leave one out". Although, as I note, I do understand why the validation and leave-one-out are both used.

(3) For these methodological reasons given in (2), there are a number of questions related to the results and interpretations:

(3a) Figure 3 shows that the model performs better in the validation phase for some catchments, which is quite puzzling. Why would parameters perform better under validation rather than calibration for some catchments? I believe this needs to be explained thoroughly, unless I am not understanding the methods, in which case, this needs to be better explained in the methods.

(3b) The sections on elasticity and uncertainties would need to be evaluated after the comments in (2) are addressed, as I am not sure the methods themselves were applied in a manner consistent with the assumptions of linear regression nor am I certain the non-linear regression was needed because a log-transform of the data did not appear to be used.

(3c) Figure 5a: Please add a 1-to-1 line on the figure so that the reader can determine for themselves how much worse the regionalization method performs. By presenting the x- and y-axes at different starting locations, it gives the impression that the methods are somewhat similar, unless the reader looks carefully at the axes values.

(3d) The conclusions discuss how identifiable parameters are able to be reasonably well reproduced but one cannot know a priori which parameters are identifiable at an ungauged location. How would one be able to apply this conclusion in practice then, when a leave-one-out approach is not possible? How would one know which parameters are sensitive and insensitive and at which catchments are there exceptions? Otherwise, this proposed method does not seem to very useful in practice.

(4) The data statement is inadequate. Please note the EGU data policy: https://www.hydrology-and-earth-system-sciences.net/policies/data_policy.html. Having the streamflow data "available upon request" is not consistent with the EGU data policy. If the data are not publicly accessible, a detailed statement as to why this is the case must be stated. Otherwise, the data needs to be placed in a public repository and cited.

Minor Comments:

L152: There should be a clear statement here that these 3 parameters are also calibrated, much like it is stated in line 169. Consider modifying it to read "Three calibrated parameters…"

Table 2: The headings are not formatted for easy readability and cut off mid-word.

Figure 3 - add the abbreviations Cal, Val, and Stb to the caption.

Line 361: Decrease of 0.4% in what?

Line 367: What model was 3 regressions? Were there not 6 parameters to estimate via regression? Or do you mean there were 3 explanatory variables in each regression model? If the equations were shown, this would help clarify the number of regressions.

L368: What is the "optimal regional model"? I have not seen this term defined anywhere else in the text?

L369-371: How is the "spatial cross-validation" different from the "Leave-One-Out method"? Only the leave-one-out method was described earlier as a validation method. In L377-378, how could a robust spatial cross-validation be completed with only 14 (or 16) catchments?

L373: The text states that "this method is more stable and more resilient to errors…" but an explanation would be needed here, as I am not convinced this is the case.

L379: Change to read: "A scatter plot of monthly NSE values between parameters estimated from the model calibration and parameters regionalized from the regression equations show…"

Lines 446-452: No evidence is offered to support these points.