Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-123-AC2, 2016

Hydrology and Earth System Sciences Discussions
Open Access
EGU

Interactive comment on "Delineation of homogenous regions using hydrological variables predicted by projection pursuit regression" *by* M. Durocher et al.

M. Durocher et al.

martin.durocher@uqtr.ca

Received and published: 1 June 2016

** see attached pdf**

Major

(1) The literature is not complete and does not state what other researchers have done in order to improve the flood estimation at ungauged sites. So, the authors should improve the description of the existing literature on the topic investigated. In particular, the manuscript should elaborate a little bit better on the evolution of the ROI method as the study focuses on the neighborhood approach for homogenous region delineation. The authors agree with the reviewer and a revised version of the manuscript will in-

Printer-friendly version

Discussion paper

clude an improved description of the existing literature on ROI. Notice, however, that much effort in the recent literature of ROI method concern the problem of estimating the model by generalized least squares, which is not the problem addressed in the present study. See for instance Reis, D., Stedinger, J., Martins, E., 2005. Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation. Water Resources Research 41. Griffis, V., Stedinger, J., 2007. The use of GLS regression in regional hydrologic analyses. Journal of Hydrology 344, 82–95. Kjeldsen, T.R., Jones, D.A., 2009. An exploratory analysis of error components in hydrological regression modeling. Water Resources Research 45, n/a–n/a. doi:10.1029/2007WR006283 Moreover, the main focus is more the improvement of the ideas behind the CCA method, for which the recent developments are included in the introduction.

(2.1) The methodology is blurred, difficult to follow, and contains some odd judgements. For instance, LSK was maintained because it is associated to better predictive performance, however, it is poorly predicted by the site characteristics (P11 L5 – 12). The authors believe that the proposed methodology has relatively simple steps. 1. Select the desired reference variables (RV) 2. If necessary predict the RV that are not available at the target site 3. Calculate the distance between the RV 4. Form the neighborhood based on the previous distance 5. Fit a regional model 6. Predict the target site As mentioned the points 3-6 are common procedure in RFA. The authors agree that the 3 main steps in page 6 could be decomposed in more direct steps as above. In a revised version of the manuscript the results section will provide clearer indication of which step the discussion takes place. The authors did not intend to impose a specific procedure for choosing the RV in its methodology. In the result section, the choice is made to adopt a backward stepwise selection procedure. This procedure is commonly used in regression modeling to select the explanatory variables. The effect of a RV on the final prediction is not straightforward and depends of the “interaction” with the other variables. For instance, two RV can be very well predicted, but if both contained the same information, one will be rejected by the procedure. In the present situation, the

LSK is not well predicted, but it still appears to bring few information that is not contained in the other RV, which improves the final prediction. Therefore, LSK is included. In a revised version of the manuscript the stepwise procedure will be introduced directly in the methodology section. (2.2) Also, the authors did not show some details such as the additional translation which necessary to avoid numerical difficulties of LSK and LKT due to negative values (P10 L19). LSK and LKT are skewed but not positive distribution. In the revised version of the manuscript the following sentences will be modified to provide a mathematical formulation of the actual transformation (P10 L19). "These reference variables are transformed to obtain near symmetric distribution and standardized to obtain zero mean and unit variance. More precisely the transformation for L1 and LCV is the logarithm and for LSK and LKT, the transformation is , where is the minimum of the reference variables. The translation is necessary as LSK and LKT are not positive distribution." (3.1) Although the authors introduced a complicated methodology, they did not make enough efforts to clarify the description of the results; such as confusing explanation of Fig. 4 (P12 L8 – 17) (e.g., why 80 sites? in P12 L12), and unclear Fig. 5 and its explanation (P12 L26 – 33). The Figure 4 presents 4 cross-validation criteria in respect of the different possible neighborhood sizes for calibration. The main point of this approach is to show that selecting a size implies a trade-off between the different criteria. Additionally to the Figures 4 and 6, the author will include in a revised version of the manuscript the following summary table, which may be an easier and more conventional way of presenting the same results. Model RRMSE NHS AHM ARE Index-Flood ROI 46.2 86.5 72.8 57.3 CCA 45.4 86.2 41.7 42.9 RVN-LM 45.0 87.1 14.5 36.9 RVN-HYB 40.1 86.2 16.5 43.1 Regression-based ROI 44.9 86.9 72.8 64.7 CCA 45.2 86.4 40.7 26.7 RVN-LM 42.6 87.3 14.2 34.3 RVN-HYB 42.3 86.2 21.2 51.1 Best criteria in bold The author agree that best NHS for RV-HYB at 80 sites is surprising. However, Figure 4a and 4b mostly shows that regionalization is not very useful in terms of NHS, but it is important in terms of RRMSE, which is the reason why the authors use RRMSE as a calibration criterion. In Figure 5, small changes in average can be difficult to assess by the naked eye. The red lines are smooth curves fitted between the residuals of the two models and are there to indicate which models in average have locally lower residuals. It can be seen that for the residuals above approximately 0.2 (i.e. 20% of the observed values) the red line is distinctively under the bisector (upper right corner) and this is true for several points. This means the sites that are overestimated are in average less overestimated by the RVN-HYB method. Actually, 8 out of 9 residuals located at the most right are better predicted by the RVN-HYB model. In a revised version of the manuscript the meaning of Figure 5 and the red line will be better explained and the arguments above will be added. I invite to see also the answer to the comments of the another reviewer where some aspects of the figure 5 are discussed in more details. (3.2) Furthermore, the presentation of the results of the regression-based model needs improvements to be clearer (P12 L35 – P13 L10). I recommend using the simple Q-Q plot to assess the compared methods regarding the estimation of regional flood quantile. The authors want to highlight the fact that the description of the steps of the regression-based model in the result section is voluntarily short because they are the same as the index-flood model. Based on the new 6 steps proposed in major comments (2.1), the only modified steps is the steps 5-6, which consist simply to fit a common linear model on the at-site quantile. A sentence will be added in the revised version of the manuscript to mention that more directly. The authors agree with the reviewer and will include a QQ plots in a revised version of the manuscript to improve the analysis of the regression-based model. (3.3) Also, the results should contain numerical tables to quantitatively clarify the differences between the considered methods. The authors can find a close example for the presentation of such results in the reference Gado and Nguyen (2016). Finally, comparing the results of the index flood and the regression methods would be valuable here. The summary table proposed by the reviewer is presented in major comment (3.1). This table will be used in the revised version of the manuscript to briefly compare the index-flood and the regression-based model. However notice that this comparison will be coherent with what is already known in the literature, i.e. that both models performed similarly, except that the index-flood have more coherent quantiles regarding the different return periods. See for instance: GREHYS, 1996. Presentation and review of some methods for regional flood frequency analysis. Journal of Hydrology 186, 63–84. GREHYS, 1996. Inter-comparison of regional flood frequency procedures for canadian rivers. Journal of hydrology(Amsterdam) 186, 85–103. Haddad, K., Rahman, A., 2012. Regional flood frequency analysis in eastern Australia: Bayesian GLS regression-based methods within fixed region and ROI framework – Quantile Regression vs. Parameter Regression Technique. Journal of Hydrology 430–431, 142 – 161. doi:10.1016/j.jhydrol.2012.02.012

Minor (1) P1 L16 – 17. Which properties does the hydrological information in Regional Frequency Analysis enforce for a group of gauged stations? I suggest to add "desired properties".

The modification will be done in a revised version of the manuscript.

(2) P1 L18. Ungauged sites can be defined by site characteristics in the neighborhood delineation methods (e.g., ROI). Therefore, there is no a challenge for using neighborhoods in RFA regarding the unavailable hydrological information at ungauged sites. The study of Oudin et al. (2010) shows that pooling sites together based on the similarity between the physiographical variables does not necessarily lead to the same group of sites as if hydrological similarity was considered. Their study reports a case of 60% overlapping sites. In other words, it means that if you would like to pool together sites based on hydrological similarity to extrapolate the behavior of another site, but that you substitute it by physiographical information, then 40% the identified sites would not be the ones that will have been chosen if the hydrological information was available. Therefore by "challenge" the authors mean that physiographical cannot replace the missing hydrological information completely. In the revised version of the manuscript, the following sentence will be modified: "A challenge for using neighborhoods in RFFA is that hydrological information is not available at target locations and it cannot be completely replaced by the available physiographical information."

Oudin, L., Kay, A., Andréassian, V., Perrin, C., 2010. Are seemingly physically similar catchments truly hydrologically similar? Water Resources Research 46, n/a–n/a. doi:10.1029/2009WR008887

(3) P1 L23. The regional frequency analysis can be applied for flood or extreme rainfall or any other extreme events. Hence, it should be stated that the case study is for regional flood estimation. The authors agree to specify the scope of the study. In the revised version of the paper, the Acronym RFA for "Regional Frequency analysis" will be changed for RFFA for "Regional Flood Frequency Analysis".

(4) P2 L21. "the distance between hydrological variables". The distance is between locations not variables. The authors disagree with the reviewer. Different notions of distances can be constructed by considering different spaces. The distance between locations is the geographical distance and represent a measure of the separation of two points in the geographical space. Similarly, the hydrological distance is defined as the separation between two sites in the space of the hydrological variables. The sentence below will be modified accordingly in a revised version of the manuscript (P2-L20): "To identify the most similar gauged sites in terms of hydrological properties, a notion of distance is needed to evaluate the proximity, or relevance, of each gauged site to the target location and identify the most hydrologically similar gauged sites."

(5) P3 L4. "as an estimation model" The modification will be done in revised version of the manuscript.

(6) P9 L4. Please, define NH in equation 14. In Eq. (14), correspond to the product of the variable N and H, which are defined. corresponds to the number of gauged sites in the neighborhoods and is the heterogeneity measure in Eq. (13) calculated on all the available gauged sites. In a revised version of the manuscript a product symbol will be added to Eq. (14) to clarify that it is the product of two separate variables.

(7) P9 L14 – 15. How can a regression model fitted on two different neighborhoods, for the same target location, obtain identical values? A simple illustration would be to

consider sites with value: {1,2,3,4,5}. Two delineation methods lead to the group {1,3,5} and {2,3,4}. Here, both groups have predicted value 3 as their mean, but the first group has a variance of 4 and the second has a variance of 1. The following sentence will be modified in the revised version of the manuscript to clarify that "identical values" stands for "similar predicted values". "Notice that a regression model fitted on two different neighborhoods (for the same target location) can lead to very similar predictions, but with different levels of uncertainty."

(8) P10 L2. I don't believe that 15 years of data are enough to get statistically reliable results, why did authors choose 15 years as the minimum time series used in the study. This choice has not been made in the present study. As indicated in P10-L3: "Only a brief description of the data and the at-site frequency analysis is provided since the elements were presented in detail in previous studies (e.g., Chokmani and Ouarda, 2004)." Notice also that 15 years is a minimum. In general the time series are longer and the average is actually 31 years. The average length information will be added to a revised version of the manuscript.

(9) P10 L3. I think you should have at least a map showing the locations of the selected stations in the case study (Quebec). The authors agree with the reviewer and a map will be added to the revised version of the manuscript.

(10) P10 L7. Using the maximum likelihood for parameter estimation with small time series (e.g., 15 years) may cause convergence problems, I would recommend using L-moments instead. As mentioned in the minor comment (8) the at-site frequency analysis was performed and validated in previous studies, where the necessary precaution were taken to assure reliable estimates. Moreover, notice the use of the terms "including" and "In general" in the following sentences of the present manuscript to underline that only a brief description of the full methodology is provided: "The at-site distributions are selected among several families including: generalized extreme values (GEV), Pearson type III (P3), generalized logistic (GLO) and log-normal with 3 parameters (LN3). In general, the estimation of the at-site distribution was achieved

by maximum likelihood and the final choices of distributions are based on the Akaike information criterion."

(11) P11 L16. What does HYB denote for in "RVN-HYB"? The acronym "HYB" stand for hybrid, because hydrological and physiographical variables are used as reference variables. The following sentences will be added to the revised version of the manuscript to clarify this point (P10-L15). "The first group is based on L-moments only and the second is based on the combination of L-moments and site-characteristics. The acronym LM for L-moment and HYB for hybrid are used to differentiate them."

(12) P11 L19. "One of the objectives of RFA is to identify a proper family of distributions from regional information" This is not an objective of the RFA. I suggest to write one of the main steps. In a revised version of the manuscript the sentence will be modified as follows: "One of the main steps in RFFA is to identify a proper family of distributions from regional information, which is achieved here by analyzing the distribution of the gauged sites inside a neighborhood."

(13) P11 L23. Please, define here the Q(r) as the regional quantile. In the revised version of the manuscript the sentence will be changed as follows: "In this model, the quantile corresponding to a return period at a target location , where is the index-flood."

Please also note the supplement to this comment:
http://www.hydrol-earth-syst-sci-discuss.net/hess-2016-123/hess-2016-123-AC2-supplement.pdf

---

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-123, 2016.

Printer-friendly version

Discussion paper