Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-639-SC1, 2017

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Regional regression models of percentile flows for the contiguous US: Expert versus data-driven independent variable selection" *by* Geoffrey Fouad et al.

S. Mylevaganam

sivarajah@abzwater.com

Received and published: 13 January 2017

Title: Regional regression models of percentile flows for the contiguous US: Expert versus data- driven independent variable selection

Authors: Geoffrey Fouad, André Skupin, Christina L. Tague Journal: Hydrology and Earth System Sciences URI:http://research.abzwater.com/review/ABZR2.pdf

Review:

Flow duration curves, which describe the flow equaled or exceeded for a given percent of time, are used to make decisions for streamflow applications, such as hydropower, wastewater dilution, and water abstractions. Unfortunately, these applications are often

conducted without observed flow duration curves as most basins are ungauged. In this case, as per the authors, regionalization procedures are typically adopted to predict flow duration curves based on information from gauged basins.

As underscored in the literature, regional regression modeling has been used to predict flow duration curves in the US. However, as per the authors, the current literature is based on particular geographic regions of the US, such as southern New England, southern and central California, and the mid-Atlantic. Moreover, the regional regression equations published by the US Geological Survey are at the state level. Therefore, in this paper, with the availability of physical and climatic data for the contiguous US, the authors develop few regional regression models of flow duration curves for the contiguous US.

As per the current version of the paper, initially, 918 near-natural basins are clustered using a multi-variate clustering algorithm, namely, K-Means clustering algorithm. The clustering is based on the variables (e.g., mean annual precipitation, potential evapotranspiration, baseflow index) that explain the variation of flow duration curves. The number of clusters are determined based on the number of calibration basins per cluster available to develop subsequent regression models, and validity of the clusters in terms of their compactness and separation.

The approach used to select the variables may influence the performance of the regression equations. Therefore, in this paper, in fitting the regression equations, the authors consider two methods/approaches namely, expert assessment and data-driven, to select the initial set of variables that explain the variation in flow duration curves.

As per the authors, the expert assessment of the flow duration curves selects a small number of variables according to the physical understanding of the curves. Therefore, the mean annual precipitation (MAP), potential evapotranspiration (PET), and baseflow index (BFI) are considered to be the controlling factors in fitting the regression equations using the expert assessment method. On the other hand, the data-driven

approach is adopted to account for many possible relations to the flow duration curves using a large number of variables. Under the data-driven approach, two sets of variables are considered. A set of lumped variables was used to describe the averages of data for each basin, while distributed variables described both the average and distribution of the basin data in space and time.

Both of these approaches (i.e., expert and data-driven) are applied to create the regional regression models. The difference in performance is then evaluated to answer the following research question: How does the performance of regional regression models for predicting percentile flows differ when using an expert assessment to select a small number of variables versus a data-driven approach involving a large number of variables?

Based on this study, the authors conclude that the small set of variables selected through expert assessment produced similar, if not better, performance than the two larger sets(i.e., lumped and distributed) of variables. A parsimonious set of variables only consisted of MAP, PET, and BFI. Additional variables in the two larger sets of variables added little to no predictive information.

Based on this review, the following comments are made:

1) The current version of the paper does not show the statistical measures (e.g., F-stat) of the fitted regression equations.

2) The authors do not provide a substantial evidence (e.g., reference) to convince the adopted equation shown on page number 9. The authors should mathematically prove. If not, few graphs are required to show the trends.

3) In this paper, the authors develop regional regression models of flow duration curves for the contiguous US. Having said this, as per the authors, the current literature is based on particular geographic regions of the US, such as southern New England, southern and central California, and the mid-Atlantic. Therefore, the authors can verify

their results for those geographic regions that have already been researched.

4) The authors should provide some statistical measures (e.g., minimum/maximum/average geographical area) of the basins that have been analyzed in this paper.

5) This research (i.e., results and the discussion) relies on the tool developed by the authors. Therefore, the results may not be of useful unless the tool is developed accurately. Having said this, the readers may not be conversant with the programming language(s) to go through the supplementary material provided by the authors. Therefore, it may not be feasible for a reader to authenticate the results without going through the source code. Thus, a section to outline the development of the tool is required.

6) The methodology adopted to determine the optimum number of clusters is not crystal clear. With the current methodology, the basins that fall within a particular group may not be the same in the chosen methods (i.e., expert, lump, and distributed).This is also visible by observing the figures 3(a), 3(b), and 3(c). Therefore, with the methodology adopted to determine the optimum number of clusters, it is meaningless to evaluate the performance difference between the chosen methods.

7) The statistical measures on performance evaluation (e.g., coefficient of determination Nash and Sutcliffe efficiency) presented in this paper are for validation basins. The paper does not present the statistical measures on the fitted equations.

http://research.abzwater.com/review/ABZR2.pdf

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-639, 2016.

Title: Regional regression models of percentile flows for the contiguous US: Expert versus data-driven independent variable selection

Authors: Geoffrey Fouad, André Skupin, Christina L. Tague
Journal: Hydrology and Earth System Sciences

Review:

Flow duration curves, which describe the flow equaled or exceeded for a given percent of time, are used to make decisions for streamflow applications, such as hydropower, wastewater dilution, and water abstractions. Unfortunately, these applications are often conducted without observed flow duration curves as most basins are ungauged. In this case, as per the authors, regionalization procedures are typically adopted to predict flow duration curves based on information from gauged basins.

As underscored in the literature, regional regression modeling has been used to predict flow duration curves in the US. However, as per the authors, the current literature is based on particular geographic regions of the US, such as southern New England, southern and central California, and the mid-Atlantic. Moreover, the regional regression equations published by the US Geological Survey are at the state level. Therefore, in this paper, with the availability of physical and climatic data for the contiguous US, the authors develop few regional regression models of flow duration curves for the contiguous US.

As per the current version of the paper, initially, 918 near-natural basins are clustered using a multi-variate clustering algorithm, namely, K-Means clustering algorithm. The clustering is based on the variables (e.g., mean annual precipitation, potential evapotranspiration, baseflow index) that explain the variation of flow duration curves. The number of clusters are determined based on the number of calibration basins per cluster available to develop subsequent regression models, and validity of the clusters in terms of their compactness and separation.

The approach used to select the variables may influence the performance of the regression equations. Therefore, in this paper, in fitting the regression equations, the authors consider two methods/approaches namely, expert assessment and data-driven, to select the initial set of variables that explain the variation in flow duration curves.

As per the authors, the expert assessment of the flow duration curves selects a small number of variables according to the physical understanding of the curves. Therefore, the mean annual precipitation (MAP), potential evapotranspiration (PET), and baseflow index (BFI) are considered to be the controlling factors in fitting the regression equations using the expert assessment method. On the other hand, the data-driven approach is adopted to account for many possible relations to the flow duration curves using a large number of variables. Under the data-driven approach, two sets of variables are considered. A set of lumped variables was used to describe the averages of data for each basin, while distributed variables described both the average and distribution of the basin data in space and time.

Fig. 1.