A paradigm-shift from regional to global flood-frequency analysis in large-sample hydrology for prediction in ungauged basins

Francesco Dell'Aira, Ph.D. Student, University of Memphis, Memphis, TN, fdllaira@memphis.edu

Nischal Kafle, Ph.D. Student, University of Memphis, Memphis, TN, nkafle@memphis.edu

Antonio Cancelliere, Full Professor, University of Catania, Catania, Italy, antonino.cancelliere@unict.it

Claudio I. Meier, Associate Professor, University Memphis, Memphis, TN, cimeier@memphis.edu

Extended Abstract

Regional flood frequency analysis (R-FFA) for flood prediction has been often used as a tool to extend information from gaged basins to ungaged ones with similar characteristics. The original motivation for R-FFA approaches was to compensate for short flow records at a single gage station and limited data for basin characterization, at a time when only rudimentary regression models were available to study the relationship between basin characteristics and the magnitude of floods with different return periods. Given these limitations, regionalization offered some appealing advantages, including (i) pooling data from multiple homogeneous watersheds to improve the fit of statistical models, and (ii) breaking down the full extent of spatial variability as it affects the probability distribution of floods. A single, regional model could satisfactorily describe the hydrological behavior of extreme events for a subset of watersheds, which were considered homogeneous based on the limited available information about basin descriptors. In contrast, the definition of separate homogeneous regions represented an indirect way to deal with the variability in the probability distribution of floods at larger scales, whose interpretation would have otherwise needed a deeper basin characterization. In this framework, spatial proximity was often used to define the homogeneous regions, clustering together watersheds that may share similar climatic and geomorphic characteristics, when quantitative information on these characteristics was not as readily available as it is today. However, all these advantages of regionalization come at a cost, i.e., there is an underlying, limiting assumption that floods from homogenous basins must follow the same normalized probability distribution (index-flood method; Dalrymple 1960); additionally, subjectivity is introduced when defining and identifying homogenous regions (Hosking & Wallis 1997).

Despite its intrinsic limitations, regionalization remains one of the most used techniques in hydrology since it first appeared in the technical literature 70 years ago (Dalrymple 1960), still enjoying great popularity today. However, data availability for R-FFA has dramatically increased, both in terms of longer flow records at an increasing number of instrumented watersheds, as well as availability of information for basin characterization, ranging from climate to land use to geomorphic descriptors. In addition, advancements in machine learning

(ML) combined with exponentially greater computational power suggest alternatives to the traditional regression models for studying complex, non-linear hydrological relationships between basin and flood characteristics. This led us to wonder whether regionalization is still a prerequisite for hydrological applications aimed at transferring knowledge from gaged to ungaged watersheds, as many of the limitations that it originally addressed have been overcome. Furthermore, switching from a regional to a global approach (i.e., where a single, global model is developed instead of multiple regional ones) for FFA may represent an opportunity to achieve greater generalization of ML models by training them on a wider variety of dynamics between watershed characteristics and their hydrologic responses, as reflected in the probability distribution of peak flow events.

We investigated training artificial neural network (ANN) models over the CAMELS dataset (Addor et al. 2017), consisting of hundreds of minimally disturbed watersheds in the U.S., each characterized through a wide spectrum of information, including climatic, relief, and geomorphic descriptors, as well as hydrologic signatures. To empirically test the hypothesis that the spatial variability in the shape of flood distributions across regions is affected by climatic characteristics and basin properties, we excluded information on the geographical location of watersheds when training ANN models. This is a crucial difference with traditional regionalization, because the model is forced to learn a general mechanism to associate basin properties with the parameters of flood distributions, despite the high degree of heterogeneity across the case-study basins at the large spatial scale considered. We compared the global model obtained training the ANN on the full dataset against the regional models obtained by the well-known index-flood method. The performance assessment was based on the mean absolute error and the width of the interquartile range of errors in flood prediction on a series of test basins, obtained by k-fold cross validation. Our results show that:

1) ANN models can be used to perform spatially continuous predictions of the parameters of flood distributions, without the need to identify homogeneous regions first, and without knowing where each basin is located; however, our current, preliminary ANN models achieve a slightly lower accuracy compared to traditional regionalization in the prediction of flood quantiles, based on the index-flood method.
2) Overall, R-FFA tends to overestimate flood quantiles typically considered in hydrologic applications, such as the 50-year event, while ANN-based FFA tends to underestimate them.
3) ANN models can handle and benefit from a deeper watershed characterization, still without needing any information on the geographical location of the basin. In comparison, traditional regionalization needs information on the basin location and can handle only a limited number of basin characteristics for clustering watersheds into homogenous regions. If too many variables are used, the resulting clustering may be meaningless, unless variables are preliminary weighted to assign different levels of importance (Hosking and Wallis 1997), a procedure that introduces further subjectivity.
4) Using evolutionary optimization techniques coupled with ANN modeling to identify basin characteristics with the highest predictive power, we found that a larger number of basin descriptors, as compared to the features used for regionalization, is required to better frame the inter-regional variability in the shape of flood distributions, as defined by their parameters. More in detail, we find that the variability in the shape parameter is mainly governed by climatic characteristics, which agrees with previous studies. On the other hand, the variability in location and scale parameters is a more complex

phenomenon; both are primarily affected by the size of the watershed and river, but they also display some dependency on geomorphic and climatic properties.

5) Traditional regionalization excluded ~20% of the watersheds in the dataset because they could not be fit in any of the regions that were defined following state-of-the-art regionalization techniques (Hosking and Wallis 1997). In contrast, ANN can consider all the watersheds, with a small impact on the model performance, because of the spatially continuous mapping of flood distributions as function of the basin characteristics.

All these results suggest that a shift from traditional R-FFA to a more global approach for FFA (G-FFA), which would allow for spatially continuous predictions across widely different sites, is possible and worthwhile investigating. While past applications of ML techniques for RFFA can be found in the literature (e.g., Srinivas et al. 2008; Durocher et al. 2015; Ghaderi et al. 2019; Allahbakhshian-Farsani et al. 2020), these are all still based on a preliminary identification of homogeneous regions. However, an alternative, global approach like the one proposed in this work offers several advantages, such as: (i) eliminating subjectivity in the definition of the homogenous regions; (ii) relaxing the assumption required by the index flood method (Hosking & Wallis 1997) that similar basins have exactly the same normalized distribution of flood probability; (iii) a more straightforward and objective methodology for making predictions in ungaged basins, and (iv) eliminating the chances of dealing with "ambiguous" ungaged basins (i.e., hard to unequivocally assign to one region), or the fact that some basins do not fit in any homogenous region, with the consequent limitations in the applicability of the methodology to such "unusual" basins. In the regionalization framework, ambiguous ungaged basins would require the adoption of methods that consider the probability of region membership (Cowpertwait 2011).

The success of a global approach to associate basin and climatic characteristics with flood distribution parameters, without regional separation or information on basin location, also has some deeper implications, beyond the practical advantages listed above. To date, hydrologists have regarded distinct behaviors in flood frequency as regional, localized phenomena, specific to basins with similar characteristics. On the other hand, if a single model can successfully produce continuous predictions of the shape of flood distributions, then perhaps each regional distribution, determined for basins with similar characteristics, may be regarded as a particular instance of a "universal" mechanism that relates the occurrence and frequency of floods to the geographical characteristics of catchments. In other words, it seems that our ANN model has been able to approximate a general, nonlinear relationship between watershed properties and flood occurrence. Improvements in both basin characterization and deep learning modeling should pave the road towards increasingly better performances of the G-FFA approach, which would help achieve a deeper understanding of the hydrologic mechanisms behind the variability in flood frequency and magnitude across catchments in different geographical locations.

References

Allahbakhshian-Farsani, P., Vafakhah, M., Khosravi-Farsani, H., & Hertig, E. (2020). Regional flood frequency analysis through some machine learning models in semi-arid regions. Water Resources Management, 34, 2887-2909.

Cowpertwait, P. S. (2011). A regionalization method based on a cluster probability model. Water Resources Research, 47(11).

Dalrymple, T. 1960: Flood frequency analyses, USGS, Water Supply Paper 1543A, 11-51

Durocher, M., Chebana, F., & Ouarda, T. B. (2015). A nonlinear approach to regional flood frequency analysis using projection pursuit regression. Journal of Hydrometeorology, 16(4), 1561-1574.

Ghaderi, K., Motamedvaziri, B., Vafakhah, M., & Dehghani, A. A. (2019). Regional flood frequency modeling: a comparative study among several data-driven models. Arabian Journal of Geosciences, 12, 1-9.

Hosking, J. R. M. & Wallis, J.R., 1997. *Regional Frequency Analysis: An approach based on L-moments*. Cambridge University Press.

Srinivas, V. V., Tripathi, S., Rao, A. R., & Govindaraju, R. S. (2008). Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. Journal of Hydrology, 348(1-2), 148-166.