Designing flood forecasting systems using machine learning, feature importance measures and synthetic scenarios

Francesco Cappelli[1], Flavia Tauro[2], Ciro Apollonio[3], Andrea Petroselli[4], Emanuele Borgonovo[1], Elena Volpi[5] and Salvatore Grimaldi[2]

(1) Department of Decision Sciences, Università Bocconi, Milan, Italy
(2) DIBAF Department, University of Tuscia, Tuscia, Italy
(3) DAFNE Department, University of Tuscia, Tuscia, Italy
(4) DEIM Department, University of Tuscia, Tuscia, Italy
(5) Dipartimento di Ingegneria, Università Roma Tre, Roma, Italia

Designing early warning systems is a complex procedure that includes several crucial options related to the forecast tool, its calibration, and the monitoring network. In the present contribution we propose a framework based on hydrological-hydraulic synthetic scenarios for: (1) selecting the optimal machine learning tools for forecasting discharge values (2) calibrating it, and (3) applying feature importance measures for identifying the most influential sub-basins in which to install the discharge measurement instrumentations. The Tiber River watershed is selected as a case study for which is available a massive synthetic flood hydrograph database composed of about 20'000 simulated annual maximum hydrographs in 39 sub-basins and at the outlet. In the case study a proof of concept of the proposed framework is described. We compare four machine learning methods (*Linear Model, Gradient Boosting, Random Forest, Extreme Gradient Boosting*) and investigate on six feature importance measures (*Permute-and-Relearn importance, Shapley feature importance, ALE-based feature importance, First-order sensitivity measure, Density-based sensitivity measure, Cumulative distribution-based sensitivity measure*). The results are promising. Indeed, they suggest that focusing on only eight sub-basins a high predictive performance is achieved, thus providing useful feedback for designing flood forecasting systems.