Geophysical Research Abstracts
Vol. 20, EGU2018-4911, 2018
EGU General Assembly 2018

Estimating large amounts of missing precipitation data

Hector Aguilera, Carolina Guardiola-Albert, Carmen Serrano-Hidalgo, and Nuria Naranjo-Fernández
Spanish Geological Survey, C/ Ríos Rosas, 23, 28003, Madrid, Spain (h.aguilera@igme.es)

Accurate estimation of missing daily precipitation data remains a difficult task particularly for large watersheds with sparse rain gauge network and large amounts of missing records. Reliable and representative precipitation time series are essential for any hydrological or hydrogeological model. This study considers three techniques for filling in missing precipitation data: Spatio-Temporal Kriging (STK), Predictive Mean Matching (PMM) and Random Forest (RF). STK is an interpolation method that fills missing data taking into account the spatiotemporal correlations and minimizing the mean squared prediction errors. PMM is a semi-parametric method where imputed values are sampled only from the observed values of the respective variable by matching predicted values as closely as possible. RF is a non-parametric method based on the popular machine learning algorithm.

A dataset including 112 weather stations in the period October 1975 - May 2017 (15219 x 122 matrix) with and overall 63% of missingness was used. The stations are located in the 2640 km2 area covered by the Almonte-Marismas aquifer in SW Spain, connected to the Doñana National Park wetland system. Ten stations were selected to compare performance of the three methods, 5 with the highest amounts of missing values (90% to 98%) and 5 with the lowest degree of missing information (6% to 25%). Three different train/test splits were carried out in the series of available data of these stations for validation: first 50% for training and the last 50% for testing; last 50% for training and first 50% for testing; 50-50 random partition. Then each imputation method was applied to the large matrix using the training sets for the 10 stations and all available data for the remaining 102 stations.

In general RF perform best in terms of error, but with the disadvantage of not being able to detect all non-rainy days and adding some constant values in them. SKT simulates the distribution of precipitation more accurately at the expense of very high computing times compared to the other two methods. PMM outperformed RF and SKT in some cases, particularly in all three simulations for one station. Overall, encouraging results were obtained through the application of these techniques given the large degree of missingness. Further improvements in the case of RF are required to cope with cero rain.