Geophysical Research Abstracts
Vol. 18, EGU2016-16456, 2016
EGU General Assembly 2016

Spatially-based quality control for daily precipitation datasets

Roberto Serrano-Notivoli (1), Martín de Luis (1), Santiago Beguería (2), and Miguel Ángel Saz (1)

(1) Department of Geography and Regional Planning, University of Zaragoza,Zaragoza, Spain (rs@unizar.es), (2) Estación Experimental de Aula Dei, CSIC, Zaragoza, Spain

There are many reasons why wrong data can appear in original precipitation datasets but their common characteristic is that all of them do not correspond to the natural variability of the climate variable. For this reason, is necessary a comprehensive analysis of the data of each station in each day, to be certain that the final dataset will be consistent and reliable.
Most of quality control techniques applied over daily precipitation are based on the comparison of each observed value with the rest of values in same series or in reference series built from its nearest stations. These methods are inherited from monthly precipitation studies, but in daily scale the variability is bigger and the methods have to be different. A common character shared by all of these approaches is that they made reconstructions based on the best-correlated reference series, which could be a biased decision because, for example, a extreme precipitation occurred in one day in more than one station could be flagged as erroneous.
We propose a method based on the specific conditions of the day and location to determine the reliability of each observation. This method keeps the local variance of the variable and the time-structure independence. To do that, individually for each daily value, we first compute the probability of precipitation occurrence through a multivariate logistic regression using the 10 nearest observations in a binomial mode (0=dry; 1=wet), this produces a binomial prediction (PB) between 0 and 1. Then, we compute a prediction of precipitation magnitude (PM) with the raw data of the same 10 nearest observations. Through these predictions we explore the original data in each day and location by five criteria: 1) Suspect data; 2) Suspect zero; 3) Suspect outlier; 4) Suspect wet and 5) Suspect dry. Tests over different datasets addressed that flagged data depend mainly on the number of available data and the homogeneous distribution of them.