Geophysical Research Abstracts
Vol. 12, EGU2010-13357, 2010
EGU General Assembly 2010
© Author(s) 2010

First results from the COST-HOME monthly benchmark dataset with temperature and precipitation data for testing homogenisation algorithms

Victor Venema (1), Olivier Mestre (2), and the COST-HOME Team

(1) University of Bonn, Meteorological institute, Bonn, Germany (victor.venema@uni-bonn.de), (2) Meteo France, Ecole Nationale de la Meteorologie, Toulouse, France (olivier.mestre@meteo.fr)

As part of the COST Action HOME (Advances in homogenisation methods of climate series: an integrated approach) a dataset was generated that serves as a benchmark for homogenisation algorithms. Members of the Action and third parties have been invited to homogenise this dataset. The results of this exercise are analysed by the HOME Working Groups (WG) on detection (WG2) and correction (WG3) algorithms to obtain recommendations for a standard homogenisation procedure for climate data. This talk will shortly describe this benchmark dataset and present first results comparing the quality of the about 25 contributions.

Based upon a survey among homogenisation experts we chose to work with monthly values for temperature and precipitation. Temperature and precipitation were selected because most participants consider these elements the most relevant for their studies. Furthermore, they represent two important types of statistics (additive and multiplicative).

The benchmark has three different types of datasets: real data, surrogate data and synthetic data. The real datasets allow comparing the different homogenisation methods with the most realistic type of data and inhomogeneities. Thus this part of the benchmark is important for a faithful comparison of algorithms with each other. However, as in this case the truth is not known, it is not possible to quantify the improvements due to homogenisation. Therefore, the benchmark also has two datasets with artificial data to which we inserted known inhomogeneities: surrogate and synthetic data.

The aim of surrogate data is to reproduce the structure of measured data accurately enough that it can be used as substitute for measurements. The surrogate climate networks have the spatial and temporal auto- and cross-correlation functions of real homogenised networks as well as the exact (non-Gaussian) distribution for each station.

The idealised synthetic data is based on the surrogate networks. The change is that the difference between the stations has been modelled as uncorrelated Gaussian white noise. The idealised dataset is valuable because its statistical characteristics are assumed in most homogenisation algorithms and Gaussian white noise is the signal most used for testing the algorithms.

The surrogate and synthetic data represent homogeneous climate data. To this data known inhomogeneities are added: outliers, as well as break inhomogeneities and local trends. Furthermore, missing data is simulated and a global trend is added.

The participants have returned around 25 contributions. Some fully automatic algorithms were applied, but most homogenisation methods need human input. For well-known algorithms, MASH, PRODIGE, SNHT, multiple contributions were returned. This allowed us to study the importance of the implementation and the operator for homogenisation, which was found to be an important factor.

For more information on the COST Action on homogenisation see:
http://www.homogenisation.org/

For more information on - and for downloading of - the benchmark dataset and the returned data see:
http://www.meteo.uni-bonn.de/venema/themes/homogenisation/