Geophysical Research Abstracts,
Vol. 11, EGU2009-8860, 2009
EGU General Assembly 2009
© Author(s) 2009

The COST-HOME monthly benchmark dataset with temperature and precipitation data for testing homogenisation algorithms

V.K.C. Venema (1) and O. Mestre (2)

(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 is generated that will serve as a benchmark for homogenisation algorithms. Members of the Action and third parties are invited to homogenise this dataset. The results of this exercise will be 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 introduce this benchmark dataset.

Based upon a survey among homogenisation experts we chose to start our work with monthly values for temperature and precipitation. Temperature and precipitation are 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 will have three difference types of datasets: real data, surrogate data and synthetic data. Real datasets will 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 (non-Gaussian) exact distribution of 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.

Every scientist working on homogenisation is invited to join this intercomparison. For more information on the COST Action on homogenisation see:
http://www.homogenisation.org/
For more information on - and for downloading - the benchmark dataset see:
http://www.meteo.uni-bonn.de/venema/themes/homogenisation/