Geophysical Research Abstracts
Vol. 14, EGU2012-9310, 2012
EGU General Assembly 2012

Estimation, Classification and Analysis of Errors in Historical Upper-Air Observations for Use as a Data Validation Tool: A Concept

R. Wartenburger, A. Stickler, and S. Brönnimann
University of Bern, Institute of Geography, Bern, Switzerland (richard.wartenburger@giub.unibe.ch)

High quality 3-dimensional observational data sets are the cornerstone for a wide spectrum of geoscientific applications. In particular, high quality observations are a fundamental prerequisite for the generation of accurate reanalysis datasets, because any uncorrected error in the observations propagates into the derived data products. Following the minimum requirement for many research questions, modern reanalysis datasets now tend to include a numerical representation of uncertainty for each climate variable in 4-dimensional space, commonly expressed by an ensemble spread. But even if those estimates exist, they are generally not validated against independent datasets. Looking back in time to the data sparse mid-twentieth century, this lack of reliable error estimates becomes even more eminent. As large volumes of unassimilated upper air data sets from the very same period are now available to us and new data sets are still waiting to be digitized, it is time to develop a strategy of how to process those datasets in order to use them as a validation tool for present and possibly future reanalyses and other datasets.

This study demonstrates a universal concept of how to gain error estimates from upper-air datasets without presuming the availability of a comprehensive set of metadata. As a case study, we use a subset of the Comprehensive Historical Upper-Air Network (CHUAN, www.historicalupperair.org), spanning the years from 1948 to 1966. CHUAN is a global dataset of corrected upper air observations from various platforms containing the essential climate variables temperature, geopotential height, and wind on a synoptic scale. For each station with a sufficient number of neighboring stations, the observation and representivity errors are determined using a two-step least squares regression approach that is applied on each variable and altitude level. The root mean square error (RMSE) is determined for each linear model of synchronous observations from neighboring station pairs over a limited number of time steps. The determined errors are adjusted for climatological differences among the station pair. For each iteratively selected reference station, the adjusted RMSEs of all neighbor stations are regressed onto the distances to the reference station, assuming a linear relationship between station distance and magnitude of the errors. The slope and the intercept of the resulting linear fit can then be interpreted as a measure of representivity and observation error of the reference station for the selected time span, respectively.

The errors are then analysed with respect to the station network, the type of observation systems involved, the geographic location of the corresponding stations, the influence of the dominating synoptic situation, the seasonality, and long-term trends and step changes in the error time series, among others. An error covariance matrix is calculated for the entire data set in order to yield information on spatiotemporal interdependencies of the errors. As a next step, the resulting product will be applied to validate gridded upper-air datasets such as atmospheric reanalyses or statistical reconstructions.