Geophysical Research Abstracts
Vol. 17, EGU2015-6841, 2015
EGU General Assembly 2015

Big climate data analysis

Manfred Mudelsee
Climate Risk Analysis, Heckenbeck,Bad Gandersheim, Germany (mudelsee@climate-risk-analysis.com)

The Big Data era has begun also in the climate sciences, not only in economics or molecular biology. We measure climate at increasing spatial resolution by means of satellites and look farther back in time at increasing temporal resolution by means of natural archives and proxy data. We use powerful supercomputers to run climate models. The model output of the calculations made for the IPCC's Fifth Assessment Report amounts to $\sim$650 TB. The 'scientific evolution' of grid computing has started, and the 'scientific revolution' of quantum computing is being prepared. This will increase computing power, and data amount, by several orders of magnitude in the future. However, more data does not automatically mean more knowledge. We need statisticians, who are at the core of transforming data into knowledge. Statisticians notably also explore the limits of our knowledge (uncertainties, that is, confidence intervals and P-values).

Mudelsee (2014 Climate Time Series Analysis: Classical Statistical and Bootstrap Methods. Second edition. Springer, Cham, xxxii + 454 pp.) coined the term 'optimal estimation'. Consider the hyperspace of climate estimation. It has many, but not infinite, dimensions. It consists of the three subspaces Monte Carlo design, method and measure. The Monte Carlo design describes the data generating process. The method subspace describes the estimation and confidence interval construction. The measure subspace describes how to detect the optimal estimation method for the Monte Carlo experiment. The envisaged large increase in computing power may bring the following idea of optimal climate estimation into existence. Given a data sample, some prior information (e.g. measurement standard errors) and a set of questions (parameters to be estimated), the first task is simple: perform an initial estimation on basis of existing knowledge and experience with such types of estimation problems. The second task requires the computing power: explore the hyperspace to find the suitable method, that is, the mode of estimation and uncertainty-measure determination that optimizes a selected measure for prescribed values close to the initial estimates. Also here, intelligent exploration methods (gradient, Brent, etc.) are useful. The third task is to apply the optimal estimation method to the climate dataset.

This conference paper illustrates by means of three examples that optimal estimation has the potential to shape future big climate data analysis. First, we consider various hypothesis tests to study whether climate extremes are increasing in their occurrence. Second, we compare Pearson's and Spearman's correlation measures. Third, we introduce a novel estimator of the tail index, which helps to better quantify climate-change related risks.