Geophysical Research Abstracts
Vol. 16, EGU2014-3277, 2014
EGU General Assembly 2014
© Author(s) 2014. CC Attribution 3.0 License.

Random projections in reducing the dimensionality of climate simulation data

Teija Seitola (1), Visa Mikkola (2), Johan Silen (3), and Heikki Järvinen (4)

(1) Finnish Meteorological Institute, Helsinki, Finland (teija.seitola@fmi.fi), (2) Department of Mathematics and Statistics, University of Helsinki, Finland (visa.mikkola@helsinki.fi), (3) Finnish Meteorological Institute, Helsinki, Finland (johan.silen@fmi.fi), (4) Department of Physics, University of Helsinki, Finland (heikki.j.jarvinen@helsinki.fi)

Climate simulation data is often high dimensional with several variables and thousands of time steps and grid points. High dimensionality presents a problem by making I/O and post-processing expensive and time consuming. It also excludes use of some analysis methods.

Random projection (RP) is a dimensionality reduction method that has been earlier applied to high dimensional data sets, for instance, in image processing.

Here we introduce random projection as a dimensionality reduction method applied on simulated global surface temperature data (so called Millennium simulation data of MPI-M) and show how the projected data preserves the essential structure of the original data. We apply Principal component analysis (PCA) on original and randomly projected lower dimensional data to analyze how RP preserves structures when original data is compressed down to 10% or 1% of the original volume.

We also demonstrate the application of the RP method on very high dimensional data of the atmospheric temperature in three-dimensions.

Our experiments show that information is naturally lost in RP but the main spatial patterns (the principal components) and temporal signatures (spectra of time-dependent coefficients) can still be recovered from the randomly projected low-dimensional subspaces.

Our results imply that RP could be used as a pre-processing step before analyzing the structure of large data sets. This might allow investigating the dynamics of truly high dimensional climate data sets of several state variables, time steps and spatial locations.