Karlsruhe Institute of Technology

Steinbuch Centre for Computing (SCC)

Institute of Meteorology and Climate Research - Atmospheric Trace Gases and Remote Sensing (IMK-ASF)

Towards an optimised environmental data compression method for structured model output

Analysing environmental datasets regarding information space and density for designing a compression method

Ugur Cayoglu*, Peter Braesicke, Tobias Kerzenmacher, Jörg Meyer and Achim Streit

QR code for pdf version of this poster

Motivation

Climate models and powerful high-perfomance computers make it possible to generate even higher resolution output in the environmental sciences. A dedicated compression procedure for climate data can help to reduce the amount of space required for data storage, transfer and processing. In order to develop such a procedure, it is necessary to analyse which **temporal and spatial information** can help determine redundant data. For this, we looked at methods from different (related) fields of science and applied them to climate data.

Goal

The goal of a compression procedure is to **recognize redundant data** in a data set and either define these as irrelevant and save to disk (lossy compression) or encode the information in a more dense alphabet (lossless compression). These analyses will help develop a prediction-based compression method for structured climate data. **Prediction-based compression** methods try to generate a high compression rate by **using adjacent data to predict a date**. The difference with the actual value is then written to the disk. The better the prediction, the smaller the residual and the amount of bits to be stored.

Conclusion

Our results show that the optimal information space for the identification of redundant data depends on the type of the analyzed variable, the location on the earth (latitude, longitude and altitude) for which the information space is to be identified and the current season with the respective temporal resolution of the data set.

*Ugur.Cayoglu@kit.edu

Recurrence Analysis

Recurrence analysis used in neurosciences and can help identify temporal dependency in the data. The goal of this analysis is to find out how likely it is that a previous value will reccur in the future.

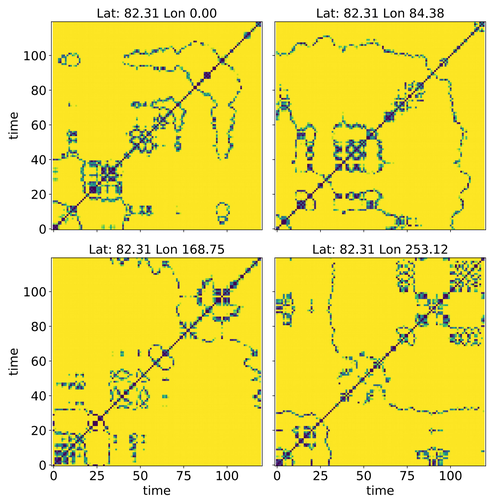

Figure 1: Recurrence analysis gives pretty good impression about the development of the variables over time (in this case zonal wind). It describes The y- and x-axis describe time. It shows, for a given moment in time, the times at which the value revisits roughly (within standard deviation) the phase space.

Video-Link of a recurrence plot for wind along longitude

Video-Link of a recurrence plot for wind along latitude

Entropy Analysis

The Shannon Entropy (Information theory) represents the average amount of information represented by a date of the dataset. The Sample Entropy (Neuroscience) provides information about regularity and predictability of fluctuations. The Shannon Entropy can be interpreted as the lower limit for lossless compression.

$$\mathrm{SaEn} = -log\left(\frac{d[D_i^{i+m+1}, D_j^{j+m+1}] < t}{d[D_i^{i+m}, D_j^{j+m}] < t}\right) \quad \mathrm{ShEn} = -\sum_i p_i log(p_i)$$

Figure 2: The Sample Entropy (left) calculates the distance `d` between two simultaneous data `D` which are less than a certain threshold `t` for an interval of length `m`. The Shannon Entropy (right) is defined as the negative logarithm of the probability mass function `p` for the value.

```
#Given
s1: [10,20,10,20,10,20,10,20,10,20,10,20]
s2: [10,10,20,10,10,20,20,20,10,10,20,20]

#Results
Shannon(s1) = 1.000     Sample(s1)  = 0.223
Shannon(s2) = 1.000     Sample(s2)  = 0.693
```

Figure 3: The Sample Entropy interprets the time series differently, since the fluctuation in `s1` is rather regular than pure noise. Lower values represent lower information density and therefore possibly higher compression rates.

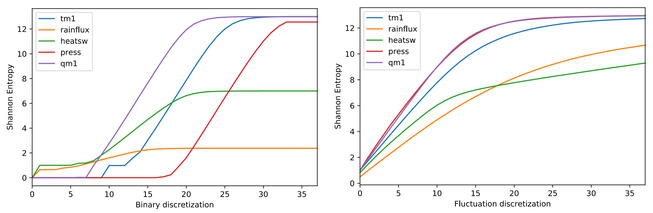

Figure 4: Since having two identical values in a floating-point dataset is unlikely the dataset to be analysed needs to be discretised. Here are two methods for discretisation and their respective Shannon Entropy shown: binary discretization (left) and fluctuation discretization (right).

```
rainflux  : 5.434E-03 (large scale rain precipitation flux)
gworo_dv  : 1.983E-02 (v tendency due to ORO GW DRAG)
heatsw    : 5.761E-02 (total shortwave heating)
geopoti   : 1.724E-01 (interface geopotential)
tm1       : 1.896E-01 (dry air temperature (tm1))
qm1       : 3.349E-01 (specific humidity (qm1))
dtdt_lw   : 3.855E-01 (longwave heating rate)
um1       : 4.320E-01 (u-wind)
vm1       : 5.680E-01 (v-wind)
vom1      : 6.921E-01 (vorticity)
press     : 1.852E+00 (pressure)
```

Figure 5: The Sample Entropy for different variables in the test data set. Lower values signal the possibilty to reach good compression rates. Variables plotted in Fig. 4 are emphasised.

Variance Analysis

The variance analysis provides information about the *temporal and spatial* dependency in the data set.

- Temporal variance **changes strongly across altitudes**. These differences are **mostly continous** and behaviour of one level can be predicted from those below or above.
- Temporal variance on horizontal level is consistent around the equator, but varies strongly at the poles.
- Altitude variance can be assumed to be **constant over time**. This might be a candidate for compression using class representation.
- Variance over latitude is continuous over levels. Sometimes it is pointwise sometimes spatially constraint. Grouping of several layers might be appropriate. Also a memory function could be used to improve update cycles of differences.

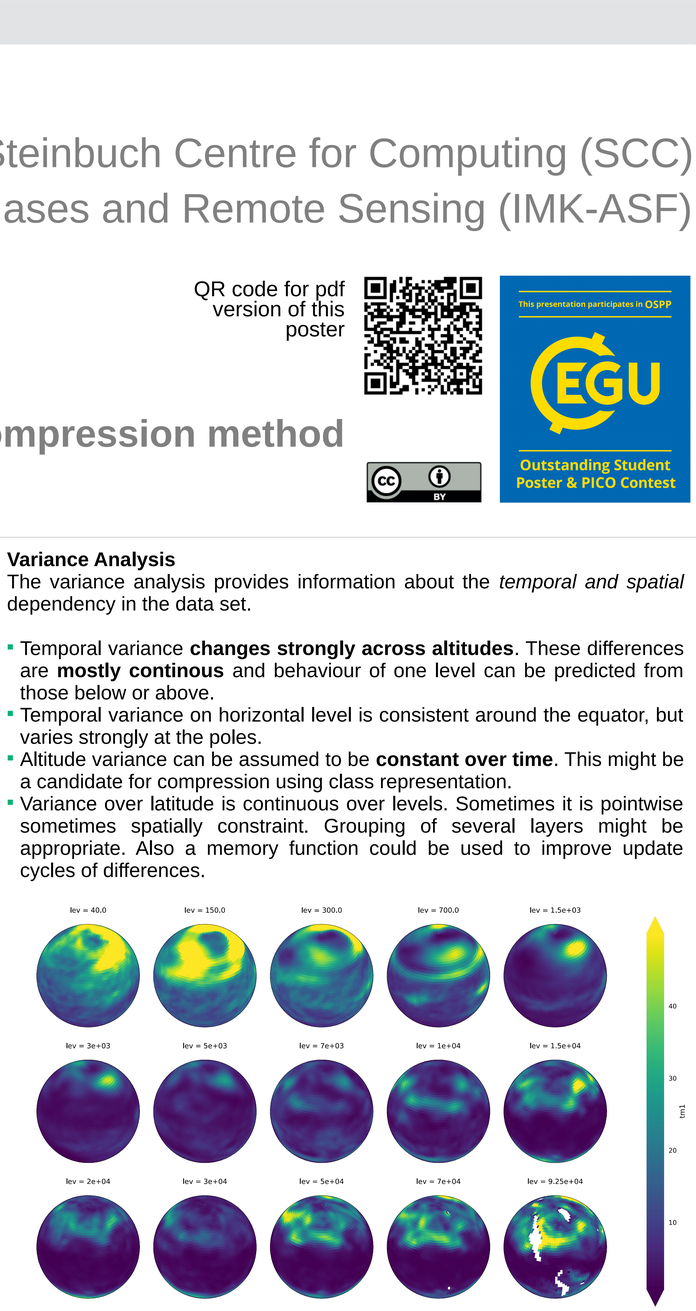

Figure 6: The variance of temperature over time at different altitude levels (Pa). It is visible that around the Equator the variance is lower than at the northern pole during November 2014.

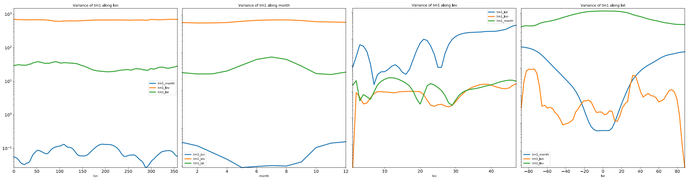

Figure 7: The variance of monthly temperature values in a test data set from 2000 to 2013. The variances (from left to right) are given for longitude, time, altitude and latitude along every other dimension.