Geophysical Research Abstracts
Vol. 13, EGU2011-8105-1, 2011
EGU General Assembly 2011

"Zipping" hydrological timeseries: An information-theoretical view on data compression as philosophy of science

Steven Weijs and Nick van de Giesen
Delft University of Technology, Water Resources, Delft, Netherlands (s.v.weijs@tudelft.nl)

Science and data compression have the same objective: by discovering patterns in (observed) data, they can describe them in a compact form. In the case of science, we call this process of compression "explaining" and the compact form a "theory" or physical "law". The similarity of these objectives leads to strong parallels between philosophy of science and the theory of data compression.

A formal description of these ideas was put forward in "A formal theory of inductive inference" by Solomonoff in 1964. Together with similar ideas, independently developed by Kolmogorov (1968) and Chaitin (1966), Solomonoff's theory was the start of the field of Algorithmic information theory (AIT). Using the length of programs that describe the data on a universal Turing machine as measures, the theory offers formal definitions of complexity and algorithmic probability and gives an explication of Occam's razor, i.e. the principle of parsimony.

Until present, these theories do not seem to have percolated into the hydrological community. This despite the fact that AIT offers a good overarching perspective on model complexity, probabilistic forcasting, Bayesian (multi-model) inference and uncertainty analysis. Because Solomonoff induction is a limiting idealized case, it shows the fundamental limits to model inference, but is also impossible to apply in practice due to incomputability. However, this does not preclude the possibility of using the basic ideas from this theory or approximations to it in practice. Notably, the connection between probability and description length has potential for application in hydrological modeling.

In this research, we applied several well-known general purpose data compression algorithms on hydrological time series, interpreting the resulting file sizes as indications of their information content. We compared compression of hydrological data with that of several artificially generated time series with specific statistical properties and relate compression ratio the information entropy of the data distributions. We also note the connection with our previous work, where we advocated the use of relative entropy or Kullback-Leibler divergence as a performance measure for forecasts. The logarithms of probability that appear in this information-theoretical measure can now be interpreted as the connection between probability and description length: where description lengths add, probabilities multiply.

Results from our practical study confirm that hydrological data often has a high temporal dependence, thereby lowering the total information content of a series of subsequent data points. This may be important in application of model complexity control methods, which aim at balancing model complexity with information content of the data. Although the conclusions from these preliminary results are not surprising, the method offers substantial potential as a tool to estimate information content in the context of model complexity control.

Concluding, the analogy between the quality of models and data compression is worth further exploring. The benefits can be in both directions. Firstly, compression algorithms may offer insight in information content of data and model performance, which is useful in model selection. Secondly, good hydrological models can be used to compress bulky hydrological data efficiently.