White paper for NSF workshop
Intelligent Systems for the Geosciences and EarthCube (IS-GEO)
March 26-27, 2015
by Imme Ebert-Uphoff (iebert@engr.colostate.edu)

A) BIG IDEAS IN LAST FEW YEARS WITH IMPACT ON GEOSCIENCES RESEARCH

1) Probabilistic models with fixed structure, e.g. Hidden Markov models, Bayesian Hierarchical Models. Structure is fixed, parameters are learned from data.
Applications: paleo-reconstruction, downscaling of daily rainfall, tracking climate models (choosing continuously between different climate models to achieve best results), simulating climate variability.

2) Complex network theory applied to climate science
Applications: climate networks, finding tele-connections, identifying and analyzing patterns in global climate, studying impact of El Nino, many other applications.

3) Providing access to big data - framework to allow local queries on server
Example: Gallileo, http://galileo.cs.colostate.edu/- also at CSU, but *not* my work.
Applications: Select and transfer *selected* data, using local data queries, thus cutting out all the unwanted data *before* the transfer.

4) Probabilistic models with learned structure (just emerging)
Applications: Causal discovery, e.g. generate hypotheses of cause-effect relationships from observed data; tracking interactions in the atmosphere, see Item B1 below.

B) ANTICIPATED FUTURE TRENDS WITH APPLICATIONS TO GEOSCIENCES

1) Tracking information flow, aka spatio-temporal structure learning
One can track *information flow*, and thus *interactions* in the earth' atmosphere, using causal discovery and related methods. Domain experts can study these interactions to yield new insights into dynamic processes of the atmosphere. I believe this area will expand greatly in coming years, using a much wider variety of methods (probabilistic graphical models, Gaussian models, Granger graphical models) and be applied to a wide variety of geoscience applications.

2) Knowledge representation and other frameworks applied to climate data
In Item B1 above one trains specific models, such as probabilistic graphical models, using observed climate data, not to use them for prediction or similar tasks, but to study their *structure*. Generalizing this idea, it is time to explore *other* types of models, e.g. other network types, statistical models, and models developed in the area of knowledge representation, for the same purpose. Namely, we learn the model, then *study its properties to learn about the properties of the underlying dynamic processes*. Given that each type of model represents the observed data in a different way we expect to gain from each model *new viewpoints and new insights into the underlying structure* of the climate system.

3) Development of new machine learning algorithms to meet geoscience needs
While many climate data sets are large in size, they are typically of high dimensionality, so actual sample size is usually very *small*. The high dimensionality comes from the fact that one sample often contains data for a large number of different locations, e.g. locations around the globe, and for different physical variables. Thus classic ML algorithms often do not work and there is a need to modify existing algorithms and/or to *develop new algorithms that can deal with a large number of variables and small sample size*.