Statistical Science: Contributions to the Administration's Research Priority on Climate Change

April 2014

A White Paper of the American Statistical Association's Advisory Committee for Climate Change Policy[1]

EXECUTIVE SUMMARY

Data are fundamental to all of science. Data enhance scientific theories and their statistical analysis suggests new avenues of research and data collection. Climate science is no exception. Earth's climate system is complex, involving the interaction of many different kinds of physical processes and many different time scales. Thus this area of science has a critical dependence on the examination of all relevant data and the application of statistics for its interpretation. Climate datasets are increasing in number, size, and complexity and challenge traditional methods of data analysis. Satellite remote sensing campaigns, automated weather monitoring networks, and climate-model experiments have contributed to a data explosion that provides a wealth of new information but can overwhelm standard approaches. Developing new statistical approaches is an essential part of understanding climate and its impact on society in the presence of uncertainty. Experience has shown that rapid progress can be made when "big data" is used with statistics to derive new technologies. Crucial to this success are new statistical methods that recognize uncertainties in the measurements and the scientific processes but are also tailored to the unique scientific questions being studied.

This white paper makes the case for the National Science Foundation (NSF) to establish an interdisciplinary research program around climate, where statisticians have the opportunity to collaborate with researchers from other disciplines to advance the understanding of the climate system (e.g., quantification of uncertainties, the development of powerful tests of scientific hypotheses). Although NSF supports basic and applied statistical research, these efforts often do not involve scientists and statisticians in partnerships or in teams to address problems in climate science. This program would also address the critical need for training a new generation of interdisciplinary researchers who can tackle challenging scientific problems that require complex data analysis by developing and using the necessary sophisticated statistical methods.

---

[1] Authors: Bruno Sanso, University of California, Santa Cruz (Chair); L. Mark Berliner, Ohio State University; Daniel S. Cooley, Colorado State University; Peter Craigmile, Ohio State University; Noel A. Cressie, University of Wollongong; Murali Haran, Pennsylvania State University; Robert B. Lund, Clemson University; Douglas W. Nychka, National Center for Atmospheric Research; Chris Paciorek, University of California, Berkeley; Stephan R. Sain, National Center for Atmospheric Research; Richard L Smith, Statistical and Applied Mathematical Sciences Institute; Michael L. Stein, University of Chicago. Affiliations are for identification purposes only and do not imply an institution's endorsement of this document.

BACKGROUND

Climate is the distribution of temperature, rainfall, air pressure, and other meteorological quantities over long time scales (usually 30 years or more). Averages over 30 years, say, of these variables are often used to describe climate, but the standard deviations and extreme percentiles (climate records) of the distributions can be just as important to track (e.g., for flood mitigation). Relationships between variables such as temperature and precipitation, or for a single variable at multiple times and/or locations, are a critical component of climate for which statistical expertise is particularly needed. For example, a drought at the level of a county can be a disaster for the farmers that live there, but it can be addressed through appropriate agricultural policies, whereas a drought at a continental scale could create national or international food shortages. Another statistical challenge is that our climate is presently in a transient state due to constantly changing greenhouse gas forcings, bringing into question the whole notion of climate as a characterization of meteorology over long time scales. Sophisticated statistical methods motivated by underlying physical models are needed to handle such non-stationary and highly multivariate spatio-temporal processes.

Both modern instruments and climate models can produce enormous amounts of information. Climate models usually project climate to the end of the century. Their output can be massive, possibly containing many realizations of weather, but varies by the climate model used. Electronic instrumentation of meteorological stations throughout the world is an important example of automated monitoring that yields increasingly large, multivariate, complex datasets. Remote sensing of Earth from above and, to a lesser extent, the oceans from below has led to a rich data resource that helps determine the drivers of climate change. Most standard software is overwhelmed by the complexity of the data and of the physical processes that drive climate models, leaving climate scientists without the benefit of powerful statistical inferences. Further, identifying the signal and filling in gaps in these noisy, massive, yet incomplete datasets is a problem that should be tackled by both climate scientists and statistical scientists in close collaboration.

Data products such as meteorological re-analyses result from a combination of observational data and a large-scale numerical model, and thus they require teams of researchers with a broad range of statistical, scientific, and computational expertise to provide the best estimates of the state of the environment and the uncertainties associated with those estimates. A new generation of data products that draw information from heterogeneous sources of information and that account for the uncertainties introduced by data scarcity and instrument errors are needed to better understand the evolution of Earth's climate. Highly efficient statistical methodologies can illuminate trends present in the system earlier and with greater probability than standard statistical approaches.

STATISTICS AND CLIMATE

Some compelling areas needing powerful innovative statistical methods are given below.

- **Spatio-temporal models.** Modern statistical techniques for spatio-temporal data go well beyond the old idea of "kriging" as a simple linear interpolating/smoothing technique, allowing the use of nonlinear dynamical models to describe the temporal evolution of the

system. They also incorporate such features as (temporal) changepoint and (spatial) boundary detection to accommodate different regimes, such as might correspond to bifurcations or tipping points in the underlying climate dynamics. Modern statistical computation techniques allow such models to be fitted to large datasets of the kind that arise both from observations and from climate models.

- **Climate extremes**: Many of the most urgent problems for climate change involve extremes, such as determining the extent to which extreme temperature or precipitation events will become more prevalent as a result of climate change, or whether hurricanes and tornadoes will become more frequent or of greater intensity in the future. An additional question is the extent to which these changes (if they exist) can be attributed to human-induced climate change. As a result of previous collaborative efforts between statisticians and climate scientists, many climate scientists are familiar with the specialized probability distributions used to characterize extremes, such as the generalized extreme value distribution. However, there are also questions of dependencies among extreme events, for example whether the Russian heat wave and the Pakistan floods of 2010 (which occurred at the same time) were in fact linked physically. Climate model projections raise the possibility of changes not only in the individual distributions of climate variables, but also of their dependence structure. Hence, new statistical methodology is needed for multivariate, spatial, and temporal extremes in order to answer the critical questions associated with climate change.

- **Data collection and homogenization:** The scientific community relies heavily on data products that contain a wide variety of observational records that are collected by a plethora of instruments. Some of the instruments/products fail to correct for important biases or use naive methods that ignore relevant correlations when infilling gaps. Statisticians can determine where data should be optimally recorded, how best to infill missing data, and in an era of shrinking budgets what data locations could be discontinued with the least loss of information. Records of temperature, precipitation, and other climate variables must be homogenized before trends are assessed. Specifically, any time a gauge is changed or a recording station is physically moved, the recorded values can shift (called a changepoint). US temperature stations average six changepoints per century, and the problem is worse in many other countries. Modern statistical methods allow one to blend information from various sources, consider changepoint effects and uncertainties, incorporate different types of error, include process dynamics, and provide ensembles that reflect uncertainties. These must be considered when developing new generations of data products and accurately assessing trends from instrumental records. A new approach to this problem is the use of data ensembles (multiple versions of a data product that account for these joint uncertainties).

- **Design and analysis of computer climate experiments**: Climate models are subject to uncertainty due to initial conditions, model structure, uncertain parameterizations and uncertain forcings. These factors should be considered in a systematic fashion when performing climate simulations, which is the basis of an area known as computer experiments. When using climate models for projection of future climate, a prime interest is

in how well the model captures changes in climate, including changes not just in means but also in variability and multivariate relationships. Thus, it is critical to focus on differences between models with regard to these changes, design computer experiments that provide the best possible information about such differences, and to do it with limited computational resources. Since the high-dimensionality of the output can obscure the signals, statistical methods are also critical in assessing the results of such experiments.

- **Decision making under uncertainty**: Modern statistical methods have the ability to propagate coherently the uncertainty generated by different sources of information. Dynamical uncertainty (sensitivity to initial conditions, a well-known property of nonlinear dynamical systems) is only one of numerous sources of uncertainty, others including model error, mis-specified model parameters, and measurement error when observational data are assimilated. When combining information from multiple sources, such as in integrated assessment modeling, the challenges from the statistical, scientific, and policy perspectives are immense and require effective collaborations between statisticians and, among others, climatologists, economists, and political scientists. There is also a need to further develop methods to visualize uncertainty and to communicate risk in climate systems. For example, rather than just observing averages of a climate variable over time or on a map, there is a need to explore distributional characteristics (e.g., standard deviations and extreme percentiles) of that variable spatially and temporally.

- **Hierarchical statistical models**: Hierarchical statistical models allow for the inclusion of multiple layers of information in a stochastic setting. Complex statistical models of climate processes can be generated using a series of building blocks: These are levels of different conditional probability models that link one process to another. Hierarchical models can synthesize multiple data sources, link data to models, link global to regional to local climate, and link weather to impacts. As an example, hierarchical models have been used to provide a different approach to multi-model ensembles, where different climate models employing possibly different physical assumptions are linked with each other and with observational data to provide probabilistic projections for future events. Physical relationships can be naturally incorporated into hierarchical statistical models, such as from atmospheric and ocean science, but also from ecology, biology, public health, and from social and behavioral science when understanding the impacts of climate change.

- **Data reconstructions, paleoclimate**: One way to understand climate processes is to collect and analyze data that provide insight into the past behavior of the climate system. Since most direct measurements of climate variables only extend back to the mid-19th century, there is considerable interest in using proxy data (e.g., from ice cores), along with reanalysis and direct measurements for paleoclimate reconstructions. These typically employ statistical models that produce estimates of past climate fields, with associated uncertainties. Modern statistical approaches, such as the fitting of hierarchical models, offer great promise for linking paleoclimate analyses to responses of nature in the past. Critical applications include understanding of climate-biotic interactions, predicting the future of species, including

humans and the diseases to which they are susceptible, and assessing impacts on national and global food security.

- **Statistical methods and computing for massive datasets**: Climate datasets, both observational and model output, are massive. Compounding this, statistical models that account for the relevant spatial and temporal dependences in the data generally require intensive computation even for datasets of modest size. In recent years, statisticians have made progress in fitting statistical models to large-scale climate datasets, but the methodology generally accounts for only some of the many important relationships in the datasets. Methods for characterizing complex patterns and features in climate data, such as empirical orthogonal functions/principal components analysis, self-organizing maps, and deep belief networks, provide a rich body of techniques upon which to build. Applying these methods on "big data" for critical climate applications involves exploiting modern high-performance computing and developing and improving computationally-efficient statistical algorithms.

CONCLUSIONS

Through the development of novel statistical methods to leverage understanding about the uncertainties inherent in complex climate processes, statistical scientists can greatly advance the administration's climate-research priority area. Indeed, interest in creating powerful statistical methods to provide insight into large complex datasets is booming. To best take advantage of the interest and the opportunity, we believe NSF should establish an interdisciplinary program to encourage and facilitate critical statistical research on climate, including but not restricted to climate change and its impacts. Besides the need for more meaningful interdisciplinary collaborations, an NSF program to promote such collaborations should also address the need for talented students interested in advancing statistical methods applied to large complex scientific problems.