Challenges in understanding the atmosphere

Doug Nychka *

Abstract

Atmospheric science provides a rich area for statistics due to the range of temporal and spatial scales of geophysical processes. This overview gives examples of three different areas where statistics is important: developing parameterizations and analyzing output of general circulation models, combining data and model output for weather forecasts and using model output to predict clear air turbulence.

Understanding climate change, forecasting the weather for next week or predicting turbulence for an aircraft flight. Each of these activities combines a knowledge of the atmosphere with data analysis and statistical science. The clear benefits of these activities to our society and economy is balanced by the fascinating interplay between the models that describe geophysical processes and need for statistics to quantify uncertainty and to assimilate observational data.

As an overview of statistics in the atmospheric sciences, this vignette has organized some examples around different scales, from large scale problems connected with our climate to the localized phenomenon of clear air turbulence. This scope will showcase the role of statistics, progressing from a support role in the development and assessment of numerical climate models, through a partnership in forecasting weather, to a central role in predicting small scale phenomenon such as clear air turbulence. As a companion to this short article, the reader is referred to [4] for a larger set of examples with more detail and an extensive bibliography. Before giving these examples, however, it is useful to outline some of the physical principles that underlie atmospheric science.

1 Physical equations for the atmosphere

The motion and evolution of the atmosphere can be determined from a system of several nonlinear partial differential equations. These equations, derived from classical results in thermodynamics and fluid mechanics on a rotating sphere, are well accepted as the fundamental description of atmospheric flow (see [7] for an introduction) and will be referred to as the primitive equations. So if a complete physical description is known why is statistics needed? The short answer is that these equations are simply too complicated to solve in

*Doug Nychka is Section Head for the Geophysical Statistics Project at the National Center for Atmospheric Research, PO Box 3000, Boulder, CO 80307-3000

an exact manner for any practical modeling of the atmosphere. A more deliberate answer is related to scales. The primitive equations afford links between large and small scale motions. Even if one wanted to focus on large scale motions, say in a climate study, it would be necessary to solve the equations at a much finer scale. This is because energy from smaller scales feeds back into the larger scale motion.

Another property of the primitive equations is a sensitive dependence on initial conditions. This feature comes under the rubric of chaos (see [2]) and was noted by E. Lorenz in the 1960's. Suppose one attempted to solve primitive equations using two slightly different initial conditions. The nonlinear components in the primitive equations would tend to amplify small differences at $t = 0$ into significant departures as t increased. Eventually the two solutions would appear to be independent of one another. Of course this has tremendous practical significance for forecasting weather. One never knows the state of the atmosphere perfectly and while this discrepancy may be small, differences between the forecast and the true state will diverge exponentially as a function of the forecast time. This property also has implications for variability among climate experiments because they also depend on initial conditions of the atmosphere and ocean.

2 Climate Change

2.1 General Circulation Models

A focus for climate studies is the use of a general circulation model (GCM). The development and evaluation of these models provides a rich area for statistical research. A GCM is an adaptation of the primitive equations to a spatial grid and discrete steps in time. The initials GCM are sometimes mistaken for global climate model but it is also an apt description. Long term changes in the earths atmosphere require modeling the large scale, global motions. It is important for the statistical community to realize that such models represent the confluence of many individual scientists efforts, running to many millions of lines of code and requiring high performance computing to run at useful levels of resolution. One example is the Community Climate Model [5] and is typically run on spatial grid of approximately 8000 points at 18 vertical levels with a time step of 20 minutes. A typical climate experiment might involve stepping the model every 20 minutes for hundreds of years, producing large and, depending on your definition, even massive data sets. Because of the necessity of dealing with model output, atmospheric scientists are very data savvy and statisticians can benefit from their experience manipulating large data sets. The statistics used to summarize model output are often rudimentary, however, and one challenge for statistical research is to provide more a comprehensive analysis of GCM output beyond means and variances. For example, the societal impacts of climate change might be tied to the occurrence of extreme events, not shifts in mean level. At a more fundamental level a GCM is a dynamical system. Delineating the attractor, finding regimes where the system is nearly linear or reducing the dimension of the state vector are important theoretical questions but are made difficult by the sheer size of the model.

2.2 Sub grid scale processes and parameterizations

Despite its large state vector a typical GCM, such as the Community Climate Model has grid boxes that are $280km \times 280km$ ($2.8^\circ \times 2.8^\circ$) at the equator. It is important to realize that features smaller than these horizontal dimensions can not be resolved by the model and are referred to as subgrid scale processes. Although some features of the atmosphere are of large scale there are crucial components that are smaller than this grid size. One example are the strong convection events associated with thunder storms. In the tropics these storms generate strong updrafts and are a mechanism for vertical transport in the atmosphere. Convection also produces clouds that in turn influence the amount of radiation reaching the surface and the amount that is trapped by the atmosphere.

Clearly any model of the atmosphere's general circulation must account for localized, strong convection and any model for radiation must include the influence of clouds. But both are subgrid scale processes not directly represented by the model state vector. These processes are accounted for through a technique known as *parameterization.* (This is different from a statistician's use of this term.) A parameterization is a function that expresses a spatial average (or higher moments) of the subgrid scale process in terms of variables that are part of the model's state vector. The parameterization is often grounded on physical arguments but can often involve simplifying approximations and empirically derived relationships. In many cases GCMs are sensitive to how subgrid scales processes are parameterized and in particular climate results hinge on how clouds reflect and absorb radiation.

Statistics has a clear role to play in improving parameterizations. Building on conventional approaches one can use observational data to estimate an empirical, regression relationship between subgrid scale processes and the grid values. One successful example of these ideas is using neural network models to parameterize the fraction of clouds in a grid box [1].

The fact that a statistical approach, and most other schemes, do not give perfect predictions of subgrid scale features raises a deeper conceptual issue concerning geophysical parameterizations. The sub grid scale processes may be more realistically modeled with the inclusion of a random component. At some level the subgrid scale effects and their resulting feedback to larger scales are not predictable and adding a random component is an efficient way to model this uncertainty. GCMs are deterministic and so adding a stochastic component would demand a different perspective. One potential benefit is that the increased variability in the model due to stochastic terms may provide better simulations for tail behavior and extremes of climate variables.

2.3 Assessing climate change

The ingredients for studying climate change involve both observational data and climate model output. (Current climate models combine a GCM with an ocean circulation model, a land surface model and also a model for sea ice.) A summary of an experiment using the NCAR Climate System Model [6] is given in Figure 1. A "fingerprint" of climate change, e.g. a spatial pattern of warming temperatures, is estimated from different runs of the climate model and the observed data is then regressed on one or more fingerprints. The strength of the regression relationship is used to draw inferences for climate change. There are many methodological issues in the carrying out a standard fingerprint analysis and an important recent extension is to consider the analysis within a Bayesian framework ([3]). The observational record for temperatures is a rich area for spatial statistics and statistical modeling. Just as the GCMs have model error, the observational record especially in 19^{th} and early 20^{th} centuries is both irregular and may include systematic biases. The compilation of monthly or annual means on a regular spatial grid with reliable measures of uncertainty is a difficult problem but it is an important one for any attempt to compare the climate of this century with climate model output.

3 Weather forecasting

The need for accurate forecasts of the weather is undisputed. In this short overview I would like to focus on the statistical foundations of weather forecasts rather than operational aspects of implementation or evaluation. Let $\boldsymbol{x}_t$ denote the state of the atmosphere at time t. A caricature of a forecast system is a complicated, nonlinear, function that evolves the current state forward one time step. $\boldsymbol{x}_{t+1} = g(\boldsymbol{x}_t)$. Without any other information one would use this relation to make a forecast at time $t+1$, plugging in our best estimate of the current state. Now suppose that atmospheric data, $\boldsymbol{y}_t$, is available at time t and follows the model $\boldsymbol{y}_t = h(\boldsymbol{x}_t) + \boldsymbol{e}_t$. Here h is an function based on the measurement process and $\boldsymbol{e}_t$ is an error term. For most weather forecasts the primary observations are vertical profiles of the atmosphere obtained from weather balloons (rawinsondes) but other information can come from satellite instruments, Doppler radar or other surface measurements. Making a forecast has two steps. The first *assimilates* the data to give an updated, and we hope better, estimate of the current state. In Bayes language, given a prior for $\boldsymbol{x}_t$ and a model for $\boldsymbol{y}_t$ given $\boldsymbol{x}_t$ one finds a posterior distribution of $\boldsymbol{x}_t$. The next step is to propagate the state forward using g to get a forecast. Formally this is just a change of variables: given the posterior distribution for $\boldsymbol{x}_t$ find the distribution of $g(\boldsymbol{x}_t)$.

The atmospheric sciences have a mature literature on this basic problem and solutions when distributions are Gaussian and g and h are approximately linear. Even with these assumptions the calculations are formidable: for operational forecasts $\boldsymbol{x}$ has a dimension on the order of 10^6 and $\boldsymbol{y}$ is on the order of 10^5. Gaussian prior distributions on $\boldsymbol{x}$ depend on covariance matrices that have complex, state dependent structure. These covariance models must never-the-less have a computationally efficient representation to make the assimilation with data possible. Statistics can benefit from the computational tools developed to handle large assimilation problems and can make contributions to extending the distributional models to non Gaussian data, such as precipitation measurements.

The posterior distribution for $g(\boldsymbol{x}_t)$, essentially a probabilistic forecast, usually does not have a closed form. Current practice is to represent the distribution by a sample of forecasts termed an *ensemble*. If the ensemble is constructed correctly then the variability among ensemble members should reflect the actual uncertainty in the forecast. For a (Bayesian) statistician the simplest way to construct the ensemble is to draw a random sample from the posterior of $\boldsymbol{x}_t$ and then propagate each member of the sample forward by g. There are also more dynamically based strategies that also involve the form of g. In any case, the key issue is whether measures of uncertainty for the forecasts stand up to a frequentist criterion.

An important component of the forecast problem is that today's posterior becomes tomorrow's prior. Thus, ensemble of forecasts should be constructed so that they can function as a prior distribution for $\boldsymbol{x}_{t+1}$ during the next forecast period. In general the repetition of the assimilation/forecast cycle is a fertile ground for studying adaptive estimation and raises theoretical questions of the convergence and the limiting properties of approximate Bayes procedures.

At a deeper level the assimilation and forecast problems in meteorology raise issues of dimension reduction. The folklore in atmospheric dynamics is that the high dimensional state vector used to describe the atmosphere at a grid of points actually lives in much lower dimensional manifold. Finding representations of this attractor would help in creating ensembles and guide the choice of ensemble size.

4 Predicting Clear Air Turbulence

Turbulence in the atmosphere and ocean is an area where there are still many open questions. In most cases it can be readily identified but a simple and rigorous definition of turbulent flow is still elusive. High altitude, clear air turbulence (CAT) is a serious problem for aviation. The challenge is to forecast CAT accurately so that aircraft can be diverted or at least crews can be warned of turbulent regions. CAT is a small scale phenomenon that is often linked to vertical wind shear at the boundaries of the jet stream. But its exact causes are still uncertain. In order to make forecasts of CAT atmospheric scientists have proposed a number of indices derived from the output of numerical models. For example Ellrods number is a product of the vertical wind shear and a measure of the horizontal shearing and stretching of the wind field. This index and others have been found to have some correlation with observed instances of CAT but also have limitations because of a lack of a physical theory and because they are calculated from numerical model output run at coarser grids than spatial scale of CAT.

A statistical approach to this problem is to use the indices derived on physical grounds and improved by empirical studies to build a forecast model. The basic form is a discriminant model where the indices serve as potential covariates and the training data for model estimation is a heterogeneous mix of pilot reports and accelerometer instruments on selected aircraft. Figure 2 is an example of a 3 hour period of data indicating locations of observations and the forecast grid used by the numerical model. The details of an approach that uses flexible discriminant analysis (linear discriminant analysis combined with multivariate adaptive regression splines) is described in [8]. This study indicates that a statistical model achieves levels of detection that approach acceptable thresh-holds for its use as an operational forecast.

Although we have focused on a particular study involving CAT there are many features at small scales that are difficult or expensive to resolve with traditional numerical models. Model selection and validation, dimension reduction, departures from Gaussian distributions and nonparametric function estimation all are relevant tools for building forecast models and are also areas of active statistical research. It should be noted that unlike more generic forecasting or classification problems, those in the atmospheric sciences often have a physical underpinning that guides selection of covariates and model building. Similar to weather forecasting there is also a sequential aspect to these problems and the opportunity to use adaptive methods that improve over time.

5 Closing Remarks

Popular views of statistics in the geophysical sciences often focus on spatial methods (e.g. Kriging) and time series. Of course these remain standard methods for data analysis and their extension to nonstationary and non Gaussian processes pose new research problems. Recently, Bayesian hierarchical models, coupled with Monte Carlo Markov chains for sampling posteriors, have gained prominence as an important modeling tool (e.g. [9]). One strength of hierarchical models is the ease that they incorporate the physical constraints of geophysical processes and their clarity of interpretation. I hope that this vignette balances some conventional views of geostatistics with an emphasis on the emerging areas in atmospheric science.

Due to limited space I have focused solely on the atmosphere. But a complete understanding of the earth's environment must include chemical and biological processes along with study of the Sun. Overlaid on this physical framework is the impact on society and the influence of physical systems by human activities. These areas challenge statisticians by the need to work closely with substantive numerical models and large, complicated observational data sets. I hope that it is also clear that no single discipline can approach these problems alone: the easy stuff has been done! Progress in understanding and forecasting the earth's systems requires collaborative effort among teams of scientists including statisticians.

References

[1] Bailey, B.A., L.M. Berliner, W. Collins, and J.T. Kiehl, 1999: Neural Networks: Cloud Parameterizations. In *Studies in the Atmospheric Sciences*, L. M. Berliner, D. W. Nychka T. Hoar eds., Springer Verlag, New York (to appear).

[2] Berliner, L.M. (1992) Probability, statistics, and chaos. *Statistical Science*, 7, 69–122.

[3] Berliner,L.M., Levine, R. and Shea, D. (1999) Bayesian detection of climate change. manuscript.

[4] Berliner,L.M, Nychka, D. W., Hoar, T. (1999) *Studies in the Atmospheric Sciences*, Springer Verlag, New York (to appear).

[5] Keihl, J. T., Hack, J.J., Bonan, G.B., Boville, B. A., Williamson, D.L. and Rasch, P.J. (1998). The National Center for Atmospheric Research Community Climate Model: CCM3. *Journal of Climate*, **11** 1131-1149.

[6] Boville, B. A. and Gent, P. (1998). The NCAR Climate System Model, Version One. *Journal of Climate*, **11** 1115-1130.

[7] Salby, M. L. (1996). *Fundamentals of Atmospheric Physics*, Academic Press, San Diego.

[8] Tebaldi,C. Nychka,D. and Brown, B. (1999) Predicting Clear-Air Turbulence In *Studies in the Atmospheric Sciences*, L. M. Berliner, D. W. Nychka T. Hoar eds., Springer Verlag, New York (to appear).

[9] Wikle, C.K., Milliff, R.F., Nychka, D., and L.M. Berliner, (1999). Spatio-temporal hierarchical Bayesian Modeling: Tropical ocean surface wind data., in review for *Journal of the American Statistical Association*

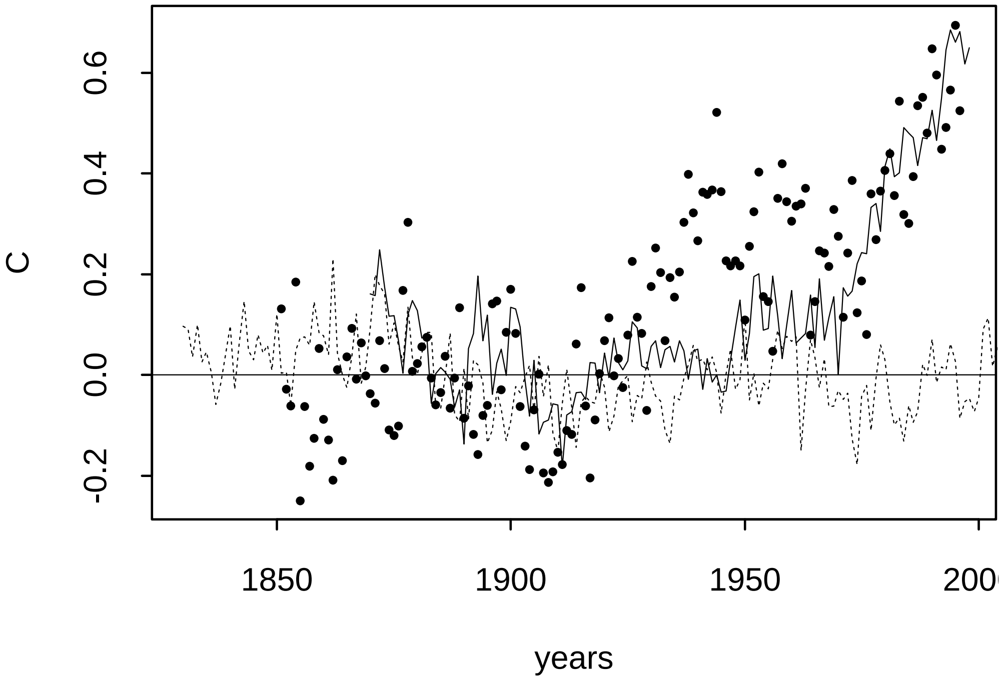

Figure 1: A comparison of observation and climate model results for annual, global average temperature anomalies. Deviations in annual global temperature relative to the observed temperatures over the period 1861-1900. Points are observed avergage temperatures. Dashed and solid lines are experimental results from Climate System Model (CSM), National Center for Atmospheric Research. Dashed: 1870 constant levels of greenhouse gases aerosols. Solid line: increasing greenhouse gases and aerosols.

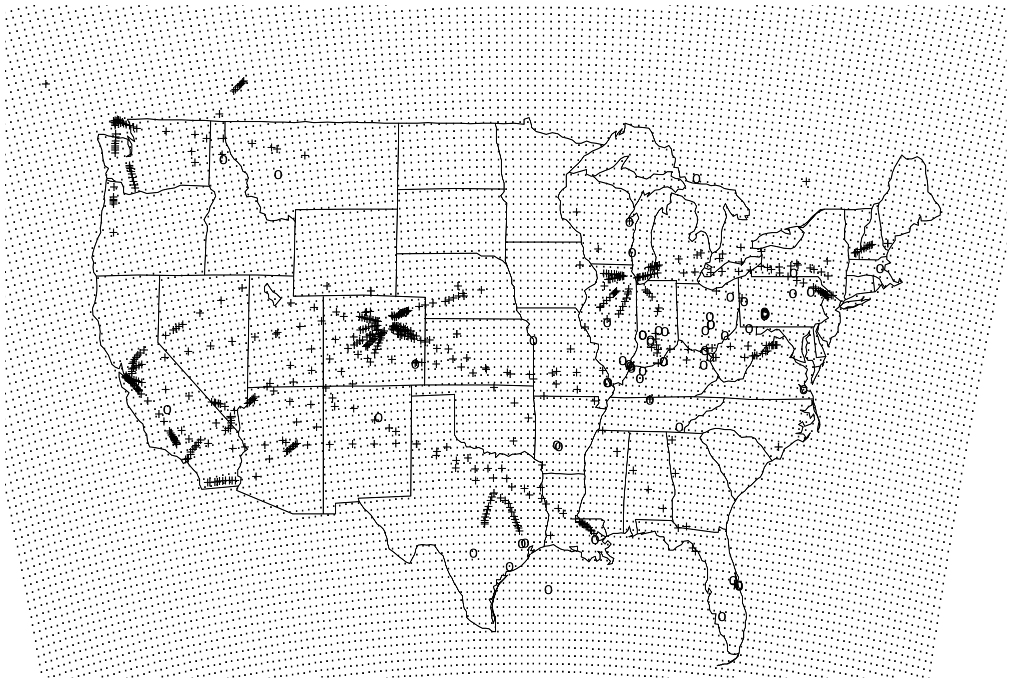

Figure 2: Pilot reports of turbulence (o) and accelerometer readings of no turbulence (+) for December 12, 1997 and above 20,000 ft. The regular grid is the resolution of the Rapid Update Cycle-60 numerical model used to construct indicators of clear air turbulence.