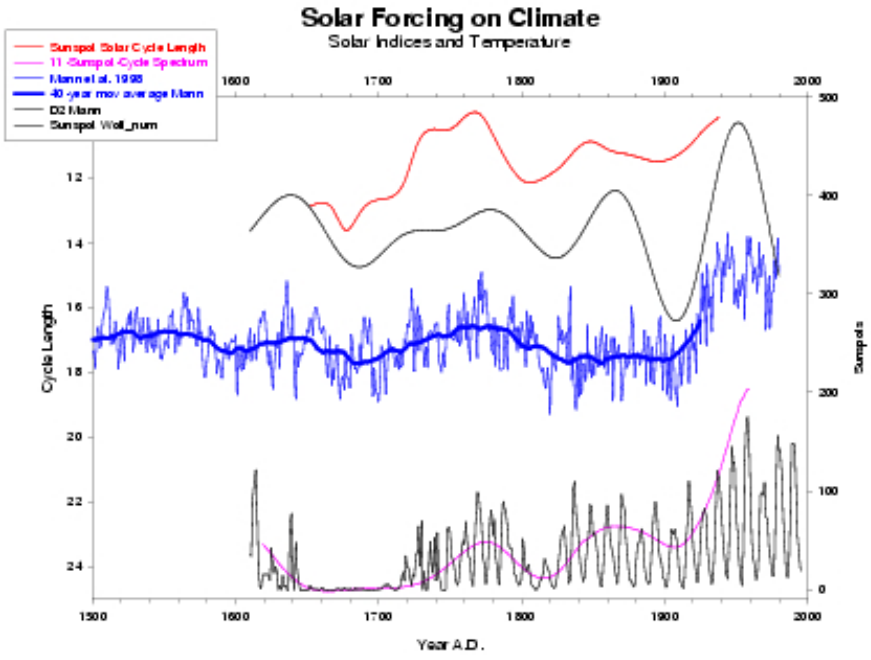
FY01 Geophysical Statistics Project (GSP)

The mission of the Geophysical Statistics Project (GSP) is to encourage the application of statistical analysis and the development of new statistical methods in the geophysical sciences. To fulfill this goal, GSP must also be engaged in basic statistical research including mathematical statistics and probability theory. Although GSP is administered through the Climate and Global Dynamics (CGD) Division, this project is an NCAR-wide effort and the research reported in the following sections reflects this breadth of scientific collaboration across the center's divisions. A large part of GSP's impact is through multi-year projects with substantial scientific involvement. For this reason many of the research results below are based on efforts continuing from last year's report. Project leadership is provided by Douglas Nychka. Co-Principal investigators for the project are Richard Katz (Environmental and Societal Impacts Group, ESIG), Joseph Tribbia (Global Dynamics Section, GDS), and Nychka (GSP).

Time Series

Multiresolution Analysis of Potential Solar Forcing on Climate

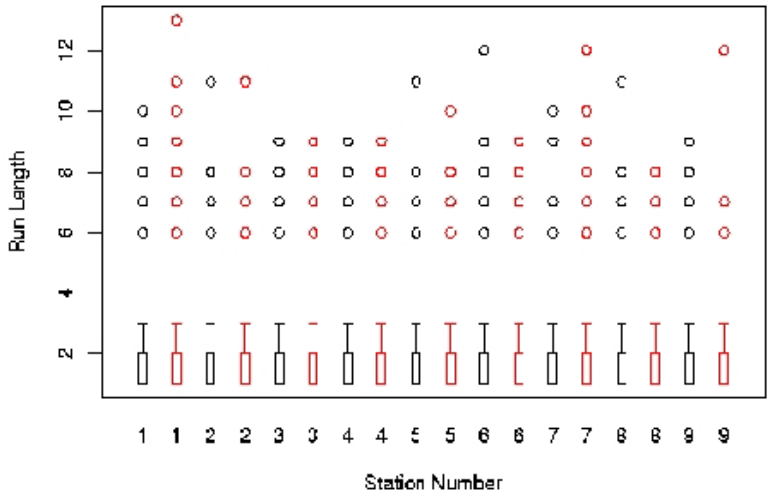
This project is a collaboration between Hee-Seok Oh (GSP) and Caspar Ammann (Climate Change Research Section). A good handle on natural climate variability is crucial for global climate change studies and the evaluation of the sensitivity of the climate system to perturbations. Formally, the derivative of the global temperature with respect to solar flux is termed the climate sensitivity. Deriving this quantity from observational data is an important contribution. Although external forcing factors might contribute substantially to both low and high frequency variations in climate, a clear separation of their impact from internally forced fluctuations is difficult. To characterize major components of solar and temperature variability in the different proxy records, and to decompose the different types of climate records into their dominant modes, a time scale decomposition is used based on a non-decimated wavelet transform. This approach gives a much better temporal resolution at coarser scales than that of the ordinary wavelet transform. The decomposition into short- and long-term components offers tools to further improve the reconstruction of solar flux based on proxy data, as well as a new way to compute previously published indices of solar irradiance.



This figure shows the long-term (D2) and short-term (D5) variations of solar forcing and temperature record obtained by time scale decomposition. (Here Dx refers to the x level of resolution in the wavelet decomposition.) In the top panel, the D2 signal represents low frequency of solar irradiance by Lean et al. (1995), and the D5 signal captures group-sunspot number (gray line) well. The bottom panel illustrates the solar forcing on climate by using long-term variations and some new solar indices. The blue line and thick blue line are temperature reconstruction by Mann et al. (1999) and 40-year moving average with Gaussian window. The solid (black) line indicates long-term variation of temperature. The red line is solar cycle length computed with the D5 signal, and the pink denotes the wavelet spectrum of the D5 signal, which will be new solar indices.

Observation Driven Models for Rainfall Occurrence

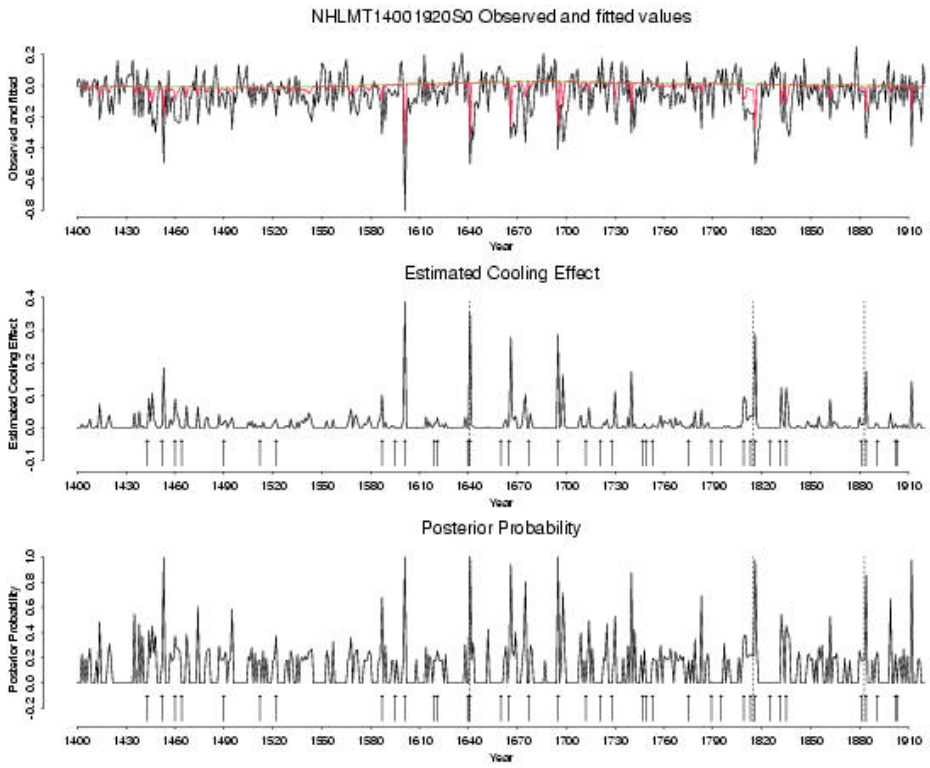
This project is a collaborative effort among Sarah Streett (GSP), Linda Mearns (ESIG), and Richard Davis (Colorado State University). The goal is to provide a realistic model for precipitation occurrence as one component of a statistical description of daily meteorology at a point location. The observation model consists of a Bernoulli {0,1} process where the probability is the logistic transformation of an autoregressive stochastic process. The key feature of this stochastic process is its dependence on past occurrences and seasonality, and a special case of this model is a Markov process for rainfall where transition probabilities depend on the past history. The observation driven extension exhibits more variability than a simple Markov chain and is easier to fit to data than hidden state models.



This figure shows the number of wet runs (number of consecutive days with precipitation) for both the measured data and data simulated from the observation driven model for 9 locations in North Carolina and South Carolina. The simulated data is represented by the black boxplots, the true by the red, and is compiled from 20 years of daily data. As is evidenced by the plot, the model does quite well in predicting the persistence of rainfall.

Extracting Volcanic Signal in Climatic Time Series

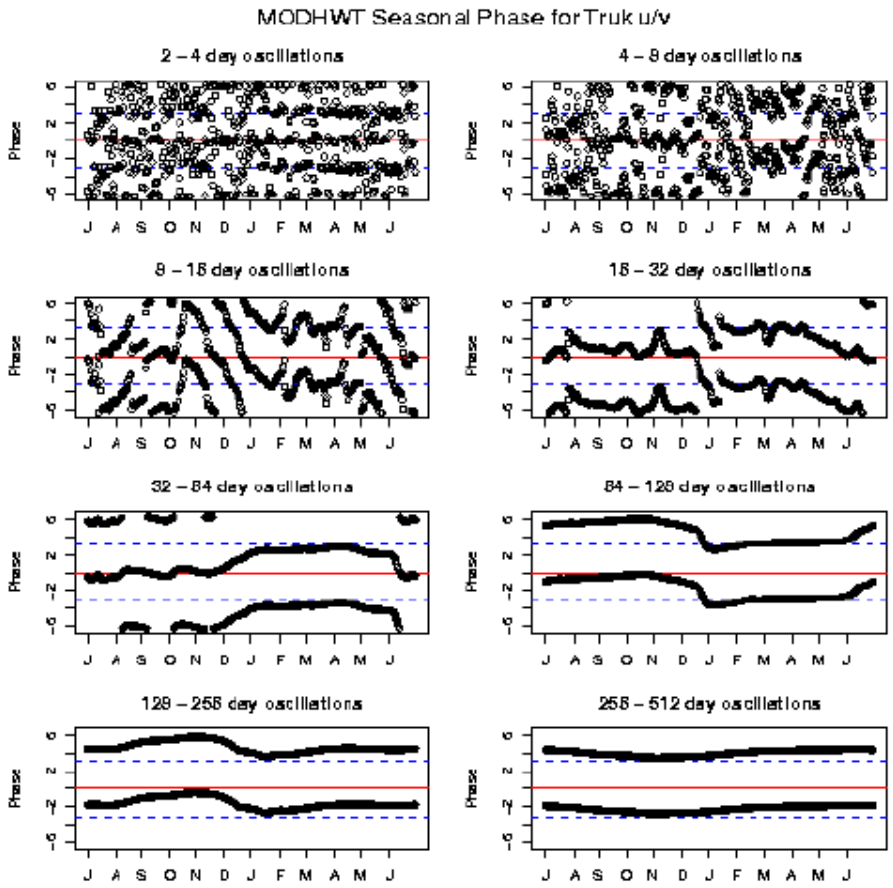
This project is led by Philippe Naveau (Ecole Polytechnique, France) in collaboration with Ammann and Oh. To understand climate variability and to attribute past climate variations to particular external forcing factors, it is necessary to implement statistical tools that are well adapted to the nature of the forcing under study. The main focus of this project is to estimate the impact of strong but short-lived perturbations from large volcanic eruptions on climate. Such eruptions inject large amounts of aerosols into the atmosphere and, on a time scale of years, influence its radiative properties. An extraction method to model simultaneously the slowly changing background climate component and the superimposed volcanic pulse-like events is presented and applied to a variety of datasets, such as temperature reconstructions. A smoothing spline is used to model the overall trend, a mixture of distributions characterizes the pulse-like event, and a multiprocess state space model offers a global mathematical framework to combine these different statistical elements. This approach provides a very good estimator of the timing of an eruption and allows a more objective estimation of its associated amplitude. In addition, a posterior probability for each cooling event is derived.



This figure illustrates the statistical extraction of a volcanic signal from the temperature reconstruction by Briffa et al. (1998). In the top panel, temperature observations (black line) and fitted values (red line) obtained by global trend and estimated volcanic signal are plotted. The next two panels indicate estimated cooling events and their posterior probability. The arrows at the bottom of the figures denote the timing of volcanic eruptions.

Seasonal Spectral Analysis of Atmospheric Time Series Using Hilbert Wavelet Pairs

This project is a collaboration between Brandon Whitcher (GSP) and Roland Madden (CGD). Time-varying spectral analysis of atmospheric time series is performed using a new class of wavelet filters, known as Hilbert wavelet pairs. Two wavelet filters, one the approximate Hilbert transform of the other, are used to band-pass filter a given time series. Because the wavelet filters have compact support, traditional statistics in spectral analysis (energy spectrum, coherence, phase) are associated with both time and scale. Thus, non-stationary features in the observed processes may be exposed.



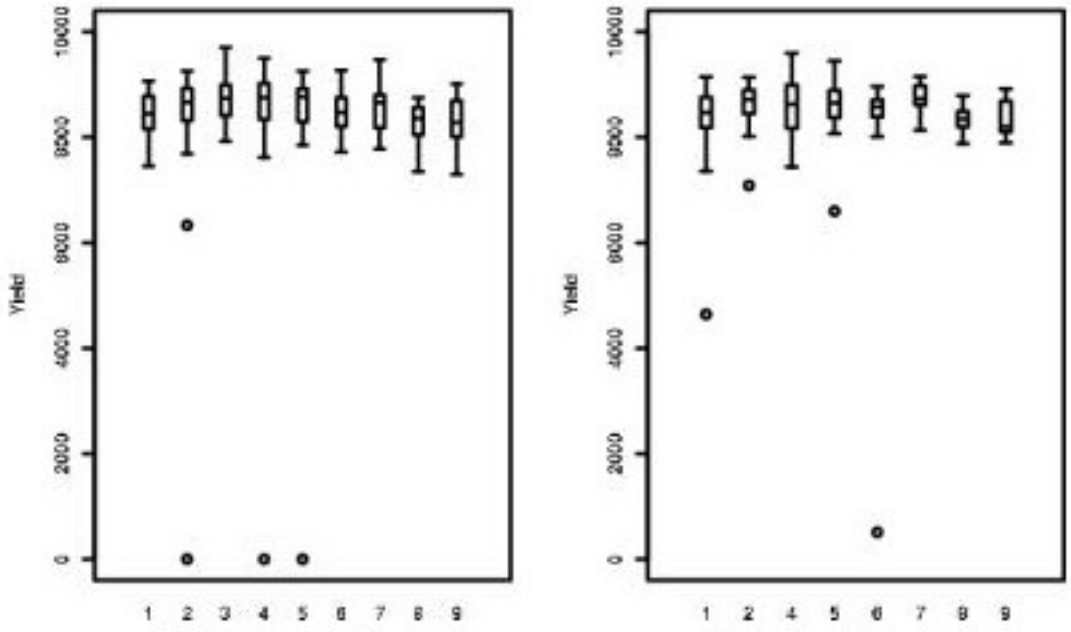
This figure shows the seasonal phase spectrum between Truk Island u-150 mb and v-150 mb winds. The band-pass filtering scheme of the discrete wavelet transform has been used. The sub-plot of interest is for 32-64 day oscillations, roughly corresponding to the Madden-Julian Oscillation. We have found a distinct phase reversal between the winds during January to June versus July to December. This feature would be lost in traditional spectral analysis that assumes both observed series to be stationary.

Spatial and Spatio-temporal Processes

Weather Generators as Tools for Assessing Climate Change on Agriculture

This project is a collaborative effort among Streett, Mearns, Nychka, Jim Jones (University of Florida), Ken Boote (University of Florida), and Tim Kittel (CGD). The overall goal is to quantify the uncertainties for predicted changes in the agricultural industry due to climate change. Typically numerical experiments using global climate models do not provide the resolution necessary to study the direct response of crops to a changing climate. An important tool for studying the response is a statistical weather generator (SWG). A SWG is a multivariate stochastic model for daily meteorology (rainfall, min/max temperature and solar radiation) at a point location and is used to generate the weather inputs for a crop

model. The basic strategy is to modify the parameters of the SWG to reflect a different climate but also to retain much of the complicated multivariate structure among the weather variables estimated from observations. In this work we have used nonparametric spline transformations to model the non-Gaussian distributions of the weather variables and incorporated a observation driven model for precipitation occurrence. These models have been extended to a spatial domain using a latent Gaussian field for precipitation and so provide the capability of generating coherent weather across a domain that is not tied to the irregular spacing of observation stations.

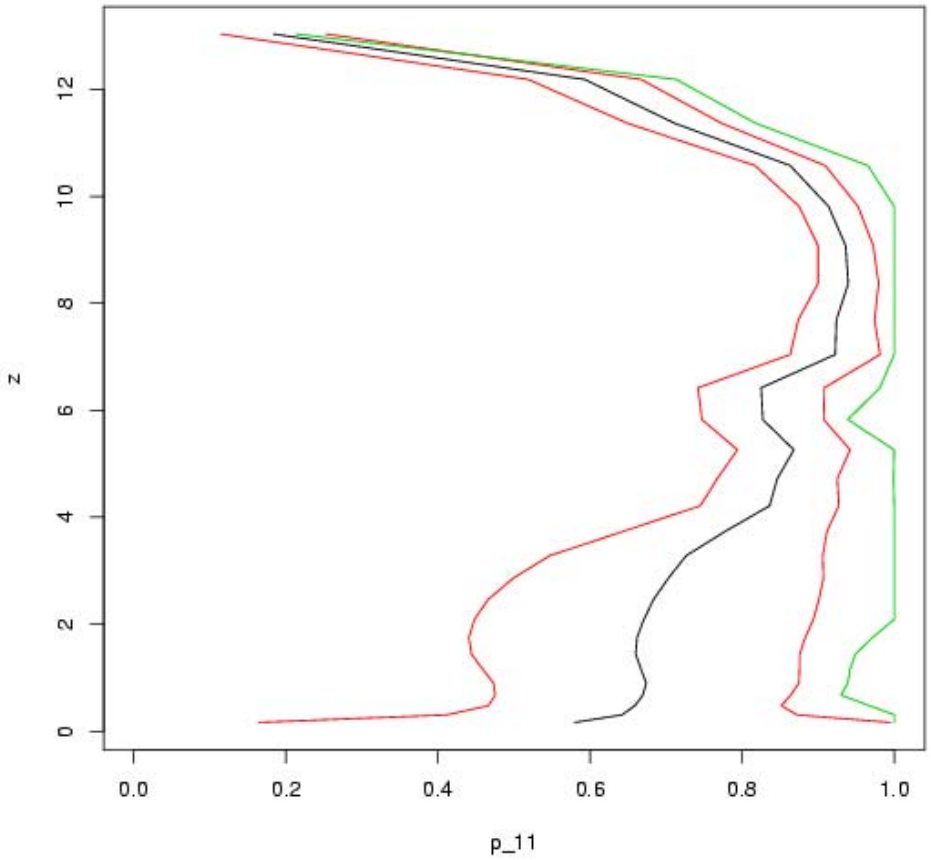


This figure gives results of evaluating a stochastic weather generator based on crop model yields for corn. Plotted are the yields over 20 years (1965-1995) for a set of nine stations from the Piedmont of North and South Carolina. On the observed daily weather is used, on the right are 20 years of simulated weather from the stochastic generator. The variability from the simulated yields is mostly within the variability due to yearly weather fluctuations and confirms the adequacy of the generator for reproducing seasonal corn yields in this region.

Spatial Distribution of Clouds and Model Parameterizations

This project is led by Enrica Bellone (GSP) in collaboration with William Collins (CGD), Xiaoqing Wu (Mesoscale and Microscale Meteorology, MMM), and Mitch Moncrieff (MMM). Clouds play an important role in the energy budget of the Earth by affecting radiation fluxes. The need to improve some aspects of the cloud system parameterizations in general circulation models (GCMs) may be addressed by studying the output of higher spatial resolution Cloud Resolving Models (CRMs). In particular, two

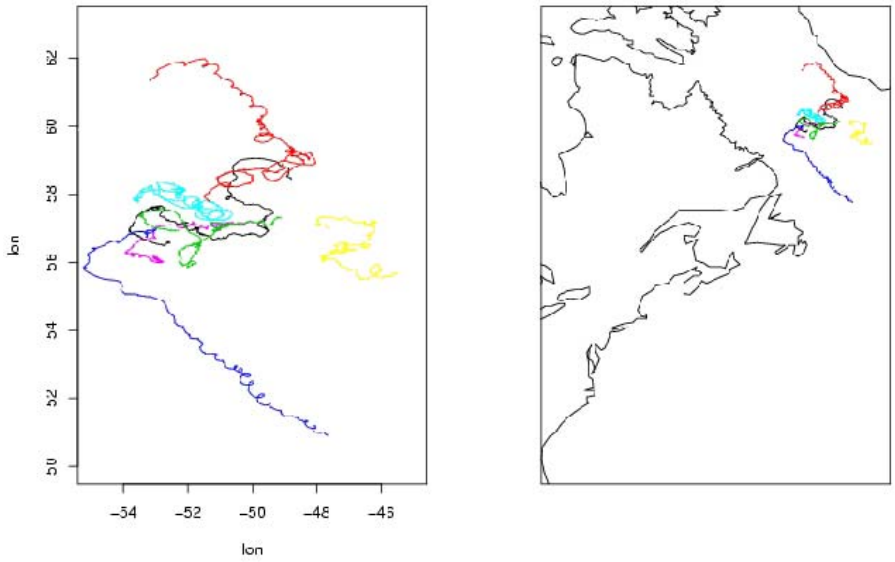
issues are crucial to the radiation calculations: 1) the vertical overlap of clouds, and 2) the horizontal inhomogeneity of condensate amounts. The first step in addressing the overlap problem is to treat the presence or absence of clouds at different vertical levels, such as a Markov chain. Specifically, the conditional distribution of cloud presence at a particular level given all other levels only depends on the presence or absence of cloud at the level immediately below. The conditional probabilities for each level are estimated and are found to be different from those implied by the maximum over lap model used in the NCAR CSM. Simulations of binary clouds can then be used to determine the impact of different overlap assumptions on the radiation fluxes. Future work includes the development of a more complex model for presence/absence of clouds with a spatial component, as well as the study of the horizontal distribution of condensate amounts.



This figure graphs conditional probabilities of cloud presence estimated from a CRM experiment. Plotted are the conditional probabilities of a cloud at a given vertical level given cloud presence at the level immediately below it. The black line represents the Markov chain probabilities estimated from the CRM data. The red lines form a simultaneous confidence band around it. The green line results from estimating the Markov chain probabilities under the assumption of maximum overlap of the clouds at each level with respect to the clouds at the level immediately below. Because the green line is outside the confidence band, the maximum overlap assumption does not fit the behavior of the CRM cloud simulations.

Reconstructing Labrador Sea Surface Currents from Drifting Buoys

Richard Jones (GSP, University of Colorado-Denver), Thomas Bengtsson (GSP), and Ralph Milliff (Colorado Research Associates) are estimating basic spatial and temporal correlation scales for the Labrador Sea based on several months of telemetry data from 21 Minimet drifting buoys. A state space Kalman filter approach is used where the observed buoy's position is related to an underlying state equation based on the ocean current and surface wind vector. Parameter estimates in the state equation are consistent with the damping (~1 hour) and the inertial (~15 hours) time scales of the ocean and spatial scales of the Labrador Sea basin.



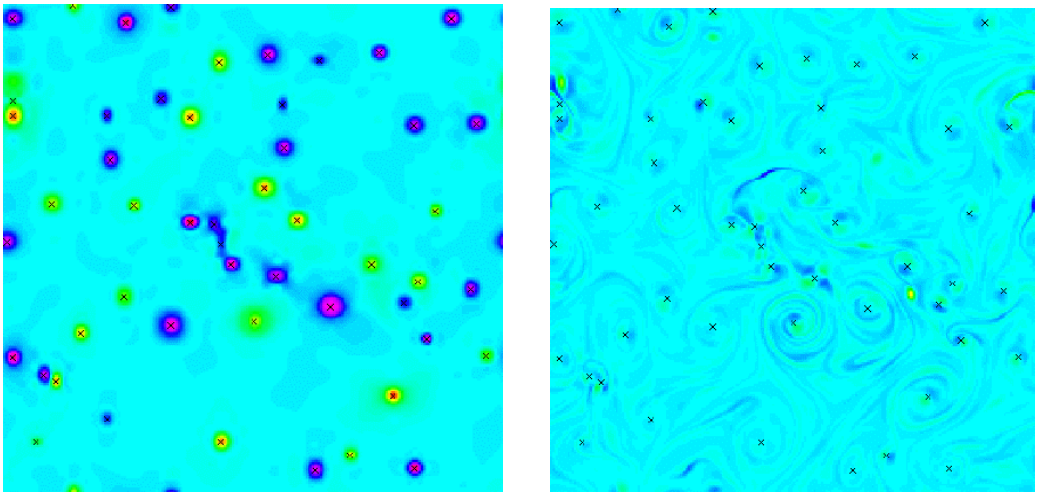
This figure shows the buoy tracks over the period October 1996 to January 1997 used to estimate the ocean current dynamics and the coupling to surface winds. Curls in the paths are due to the Coriolis force.

Regression and Classification

Stochastic Multiresolution Models for Turbulence

This project is a collaboration among Whitcher, Jeff Weiss (University of Colorado at Boulder), and Nychka. With the current state of computing technology, it is relatively straightforward to generate high-resolution numerical integrations of the Naiver-Stokes (N-S) equations, the fundamental physical equations that describe (nonreactive) fluid motion. One feature in the simulation observations of geophysical type fluids is the presence of eddies and vortices. It has become of interest in the past five years to quantitatively describe these coherent structures, (i.e., the number, size, shape, amplitude, of vorticies), and our goal is to formulate a statistical model for these structures without the necessity of simulating them from the N-S equations. As a preliminary task one needs to be able to identify individual vortices from the numerical model output of a simulation, and we

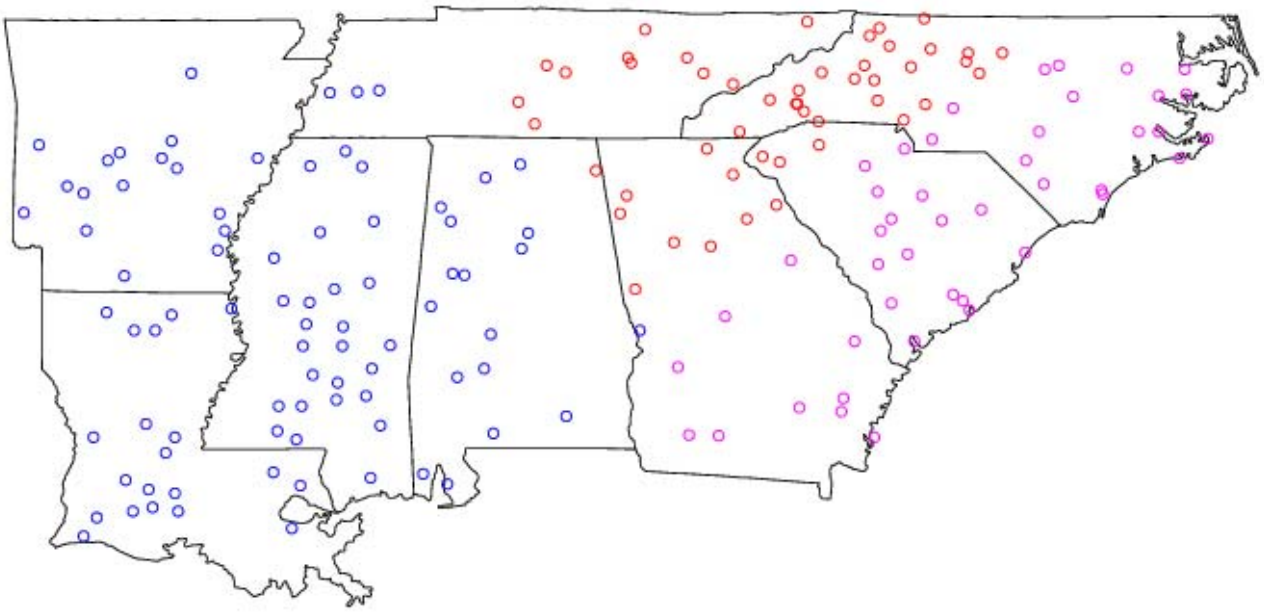
describe the results for this image segmentation problem. For a 2-d simulation of turbulence, the observed vorticity fields (vorticity is the curl of velocity) are decomposed via a non-decimated discrete wavelet transform. Our model for the field is formulated as a multivariate multiple linear regression in the wavelet domain, and individual vortex models are isolated through a model selection strategy. The identification procedure appears to be reliable and is currently being implemented on a parallel architecture to make analysis of a reasonable numerical experiment feasible. Extensions of the proposed methodology include the analysis of more realistic simulations (e.g., forcings that mimic behavior of surface currents in the Pacific Ocean) and three-dimensional datasets.



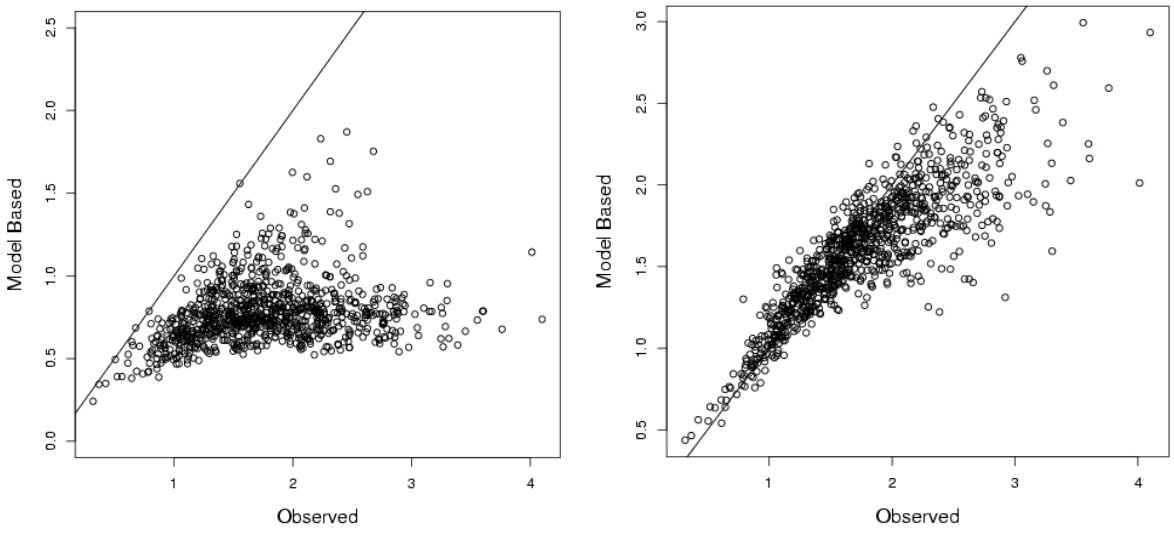
This figure shows the result of the multivariate multiple linear regression (right) and residual (left) fields from a realization of 2-d turbulence. The crosses denote where the algorithm determined there was a coherent structure (a vortex) in the vorticity field.

Precipitation Regimes for the Southeast U.S. Based on Hidden Markov Models

This project is led by Bellone in collaboration with James Hughes (University of Washington) and Peter Guttorp (University of Washington). Precipitation is an important input for crop growth models. Assessing the effect of climate variability and climate change on crop yields requires simulations of realistic precipitation fields under current and altered scenarios. Nonhomogeneous hidden Markov models (NHMMs) provide a relatively simple framework for simulating precipitation incidences and amounts at multiple rain gauge stations, conditional on synoptic atmospheric patterns. This project's case study involves 175 rain gauges that cover the Southeast region of the U.S. Given the size of the network, direct application of the NHMM methodology is computationally intensive and beyond the resources of most statisticians. To overcome this problem, model fitting is done in two stages. The first step consists of identifying different geographical regions, and the second step involves fitting a separate NHMM to all stations in each region. Common atmospheric variables should provide a link between the different regions, allowing the simulations from the separate models to reproduce the spatial dependence observed in the original data.



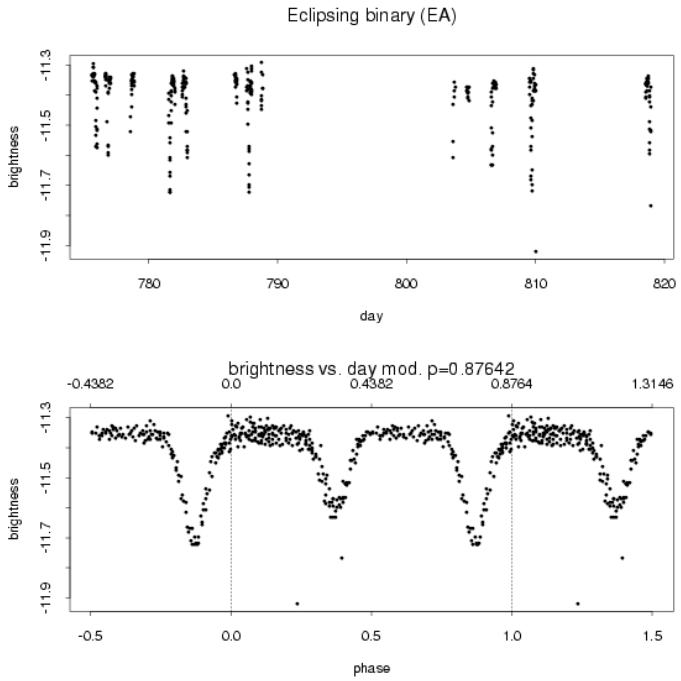
This figure shows the 175 rain gauges that were selected for the study. Only stations that have good quality data daily precipitation for the April-June period between 1965 and 1987 were included. The color scheme reflects the partition into three separate regions. The regions were found by fitting a hidden Markov model to a sample of gauges throughout the study region. The model fitting procedure identifies "hidden" states of the weather that are associated to spatial precipitation patterns. This results in a natural partition of the Southeast U.S. into several smaller areas with coherent precipitation patterns.



The left figure shows the model based versus observed logodds ratios between precipitation occurrences at the gauges in the Atlantic coast sub-region. To obtain the logodds in both plots, a Hidden Markov Model was fit to a sample of 30 stations throughout the Southeast. A final M-step in the EM algorithm used to maximize the likelihood provided the parameter estimates for the rest of the gauges. In this "global" model the observed dependence between stations is largely underestimated by the model. In the right figure, the model-based logodds result from the "regional" NHMM for the Atlantic Coast. Most of the observed dependence is reproduced by the regional model.

Period Analysis of Variable Stars

This is a collaborative effort among Oh, Nychka, Tim Brown (High Altitude Observatory, HAO), Paul Charbonneau (HAO) and John Rice (University of California-Berkeley). Variable stars show variations (or periodicity) as a function of time in brightness. The objective of this project is to estimate the period and the light curve (or periodic function) of a variable star with the goal of classifying the stars according to different types. A large census of variable stars would be a useful quantification of stellar types and could be used for testing theories of stellar formation and evolution. A smoothing spline regression is used to estimate the light curve given the period, and the method is to find the period that minimizes the generalized cross-validation (GCV). This approach works well, matching an intensive visual examination of a few hundred variable stars. However, because the GCV function is sensitive to outliers, a robust spline regression method was developed as a robust modification for finding the period. Once the period is determined, spline or wavelet regression is used to estimate the light curve. In addition, to estimate multi-periodicity in variable stars, a backfitting algorithm is employed with robust GCV method. Future work is to use the Stellar Astrophysics & Research on Exoplanets Project (STARE) database to classify stars into different groups. This will be based on functional data methods applied to the estimated light curves and prior knowledge of variable stars.

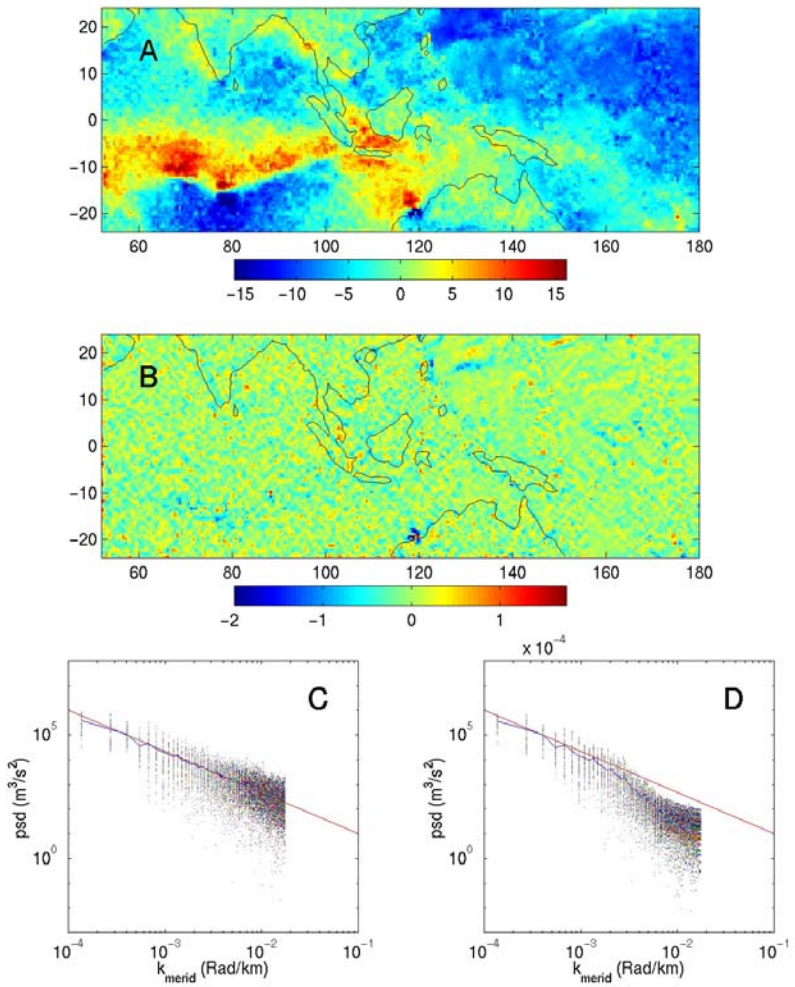


This figure shows the period and light curve estimated by robust GCV method. The brightness of an eclipsing binary star is plotted in top panel. The bottom panel shows the estimated period and the light curve with the period.

Statistical Computing

Statistical Supercomputing with the Gibbs Sampler

This project is a collaboration between Tim Hoar (GSP), Chris Wikle (University of Missouri), and Ralph Milliff (Colorado Research Associates). Gibbs sampling for a Bayesian hierarchical spatial model was implemented using portable FORTRAN 90 code and ported to a massively parallel architecture, such as that of the NCAR IBM SP RS/6000 machines. This software was used to blend scatterometer wind observations (QuikSCAT) and analysis fields from the National Center for Environmental Prediction (NCEP) to produce high-resolution surface wind fields for the tropical Pacific. The current work has been able to efficiently blend QuikSCAT surface winds with analysis winds over a 48° x 128° (at 0.5° resolution) domain in the Pacific and produce posterior samples of the wind vector field.

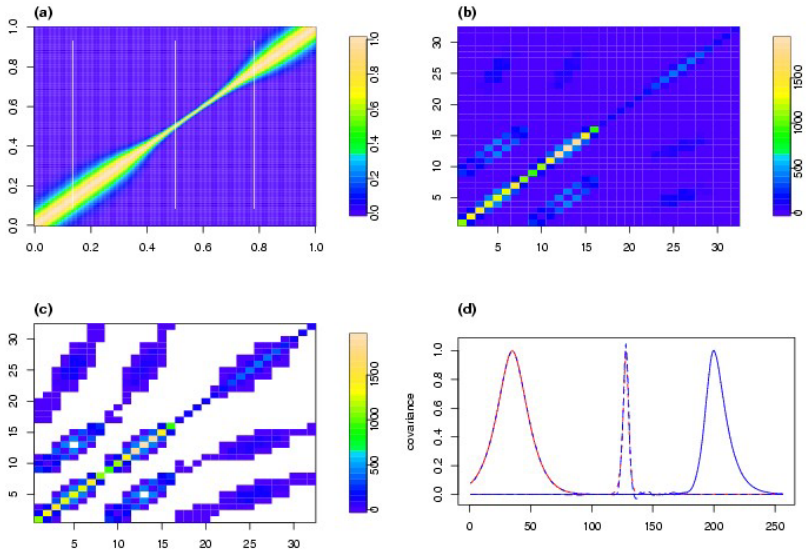


This figure shows A) One of 50 equally probable U wind fields for 30-Jan-2001 18Z. Not shown are the 50 V wind fields for this time. B) The divergence of the above field. C) The periodograms of each of the 96 latitudes. The reference line depicts a $k^{-5/3}$ power law relationship. D) The periodograms of the NCEP winds interpolated to the same prediction gridpoints as in A. Note that the NCEP winds do not have enough energy at the higher wavenumbers.

Nonstationary Fields Using Multiresolution Bases

Nychka and Whitcher, in collaboration with Wikle and Andrew Royle (U.S. Fish and Wildlife Services), have developed a wavelet model for nonstationary fields. The basic idea is to expand the field in a computationally efficient wavelet basis with the intent that the covariance structure among basis function coefficients will be both simple and sparse. This is a continuation of work from the previous period, but new developments include a deliberate exploration of the approximation properties of this model for the Matern class of covariances. Also, analytic formulas have been derived using *Mathematica* that give

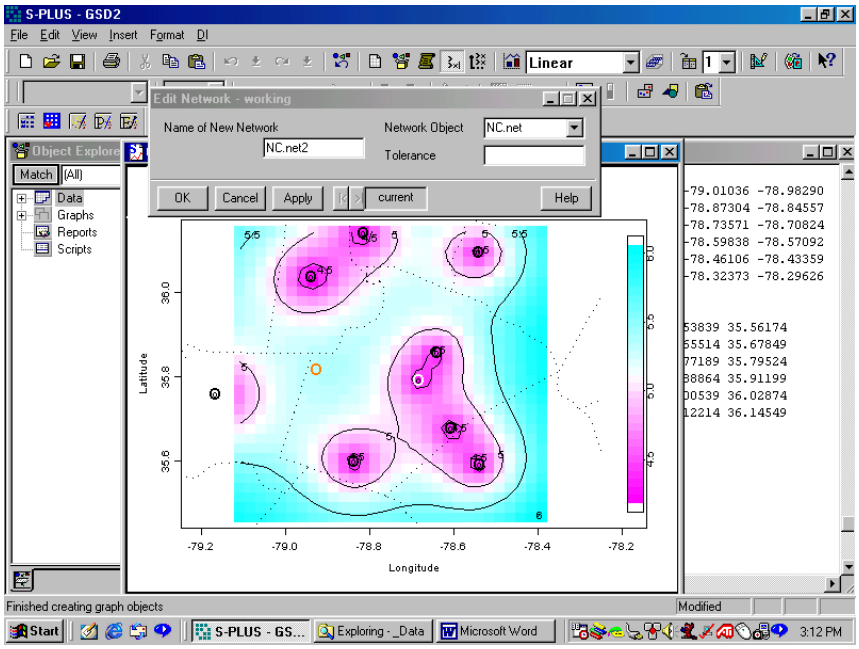
explicit forms for the covariances among wavelet coefficients. These formulas are important for proving general results about the sparsity of the wavelet representation.



This figure considers a one-dimensional nonstationary covariance constructed by a spatial deformation. (a) is the covariance matrix for the fields at 128 equally spaced points. Note the reduced correlation range for locations in the center of the interval. (b) is the portion of the square root of the covariance matrix among wavelet coefficients that will exactly reproduce the covariance in (b). (c) is the 98% decimated version of the matrix from (b). (d) Compares the approximate version of the covariance based on decimation to the exact covariance. Plotted are three columns of the covariance matrices where the locations for these columns are indicated as white lines in (a). The agreement between exact and approximate is very good, although note for low values of the covariance in the middle of the interval, there is some ringing in the approximation.

Design Interface (DI): Interactive Statistical Tools for Evaluating Spatial Networks

Eric Gilleland (GSP, Colorado State University), Nychka, and William Cox (U.S. Environmental Protection Agency, EPA) have developed a graphical user interface (GUI) for evaluating and modifying spatial designs. This current development version is targeted to support the EPA in determining the sensitivity and coverage of air quality monitoring networks for different pollutants. The goal is to give a nonstatistician insight into the influence of monitoring site locations on the accuracy of the spatial prediction of pollutants at locations where data is not collected. A practical question that motivated this work is: if a monitoring site violates the EPA standards how far can this nonattainment be extended to an area surrounding a station? DI is written in the S language with S-PLUS® GUI extensions and includes a web-based manual (see http://www.cgd.ucar.edu/stats/DI). Although top level functions are "point and click" based, DI also supports more deliberate spatial analysis through a suite of command-line S functions from the package FIELDS also developed by the GSP at NCAR.

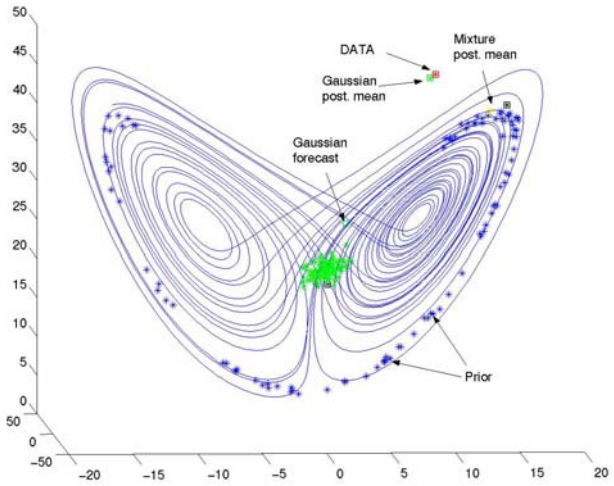


This figure is a screen shot of an Splus session using the DI module. Here the function for editing a design is being used. The GUI box controls the network being edited and the new network that will be created with modified locations. The plot is a network centered over Research Triangle Park, North Carolina, and for reference plots the spatial prediction standard based on the (old) network. The white (orange) circle indicates a station that has been interactively deleted (added). Once this function is exited, a new network object is created that can be subsequently analyzed with other DI spatial functions.

Dynamical Systems

Forecasting and Data Assimilation Adapting Particle Filter Methodology

This project is lead by Bengtsson in collaboration with Nychka and Chris Snyder (MMM). Several aspects of numerical weather prediction (NWP) make forecasting and data assimilation particularly challenging: very high-dimensional systems, strongly non-linear (possibly chaotic) dynamics, and real-time requirements for assimilating data and physical models. In practice one must: address multi-modal forecast distributions, specify spatial covariance structures, use severely rank-deficient matrices, devise sampling schemes, and understand the properties of sample based filtering algorithms. In this project we implement Bayes theorem through an approximation based on a discrete sample from a mixture of Gaussian distributions. To handle non-linear systems and still have a stable filtering method, we have found that it is important to use nearest neighborhoods of states to derive updates.



This figure illustrates how the usual Ensemble Kalman filter based on the Gaussian distribution goes wrong. Given an initial distribution in the saddle of the Lorenz attractor, the system is forecast forward one time step and a new data point is assimilated. Because the ensemble is spread widely on the attractor, the covariance is large and the update largely reverts to the observed value. The mixture model only uses local covariance structure around state vectors consistent with the observation and so the update is clearly superior.

To get a feel for the physics involved in this example, watch the MPEG movie of the Lorenz '63 system (www.cgd.ucar.edu/stats/pub/lorenz.mpg). The Lorenz (Lorenz, 1963) attractor is loosely modeled on the convective flow of a fluid heated from below in a gravitational field. Although a simple non-linear model, the system illustrates the difficulties involved in making predictions in highly non-linear systems. In particular, as the system passes through its saddle point, a small perturbation will determine which "wing" the state passes through next. This exemplifies the difficulty of making accurate weather forecasts when the atmosphere is currently in (or near) an "unstable" state (in this case a saddle point). To make an accurate (weather) forecast under such sensitive conditions, one must have a very accurate estimate of the current atmospheric state.