Annual report for DMS-9815344
July 1, 2002 - June 30, 2003
A statistics program at the National Center for Atmospheric Research.
(revised for external advisory panel 8/18/2003)

August 18, 2003

1 Overview

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 projects outlined below reflect a breadth of scientific collaboration across the center's divisions. [1]

2 GSP Members

For this annual period GSP supported the following long term visiting scientists and research assistants. Research assistants are enrolled in the graduate statistics program at Colorado State University (CSU). Other sources of support that leverage GSP funds are listed along with term dates.

Post docs

Enrica Bellone 9/2000 - 8/2002
Thomas Bengtsson 7/2000 - 6/2003

[1]NCAR research groups: Atmospheric Chemistry (ACD), Climate and Global Dynamics (CGD), High Altitude Observatory (HAO), Mesoscale and Microscale Meteorology (MMM), Research Applications Program (RAP), Environmental and Societal Impacts Group (ESIG), Atmospheric Technology (ATD), Scientific Computing (SCD), Advanced Study Program (ASP)

Heeseok Oh 7/2000 - 8/2002
Sarah Street 9/2000 - (50% ESIG)
(1/2 time 5/03-12/03, leave 1/03-5/03)
Brandon Whitcher 9/2000 - 8/2002

Graduate Students

Eric Gilleland (7/2002 - 3/2003) Research Assistant (PhD program CSU)
Dorin Drignei (9/02- 12/02) Statistics Department, Iowa State University
Chris Paciorek, (7/02) Statistics Department, Carnegie Mellon University
Uli Schneider, (7/02-) Applied Mathematics Department, CU-Boulder
(Ph D. adviser is Jem Corcoran)
Daniel Coley, (5/03-) Applied Mathematics Department CU-Boulder
(Ph D. adviser is Philippe Naveau)

Visiting Faculty

Thomas Lee, Colorado State University 20% GSP 1/2003-
Craig Johns, University of Colorado-Denver 7/2002
Steve Sain, University of Colorado-Denver 20% GSP 7/2002-
Philippe Naveau, University of Colorado-Boulder 5/2002

3 Research

Some of the major research accomplishments and progress over this period are reported in this section. The work is organized by statistical topics and has GSP members highlighted. NCAR scientists names are followed by the initials for their science division (see footnote page 1). Another version of this material with figures is disseminated to the statistics and atmospheric community community as part of the NCAR Annual Science Report 2002 `www.cgd.ucar.edu/stats/asr02`. Finally it should be noted that many of the projects reported below are continuations from the previous funding year reflecting the continuation of the GSP post docs from the previous year. Thus, the reported progress is often incremental rather than initiating a new area of application.

Time Series

Multivariate Observation Driven Models for Rainfall Occurrence This project is a collaborative effort among Sarah Streett (GSP), Linda Mearns (ESIG), and Steve Sain (University of Colorado at Denver). The goal is to provide a realistic model for precipitation occurrence as one component of a statistical description of daily meteorology. 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 past history. The main accomplishment for this period is the extension of these models to produce coherent multivariate sequences of binary process. Preliminary simulation studies have shown that such coupled processes are stable.

Spatial and Spatio-temporal Processes

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). The presence or absence of clouds at different vertical levels is modeled 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 overlap model used in the NCAR Climate System Model. Different cloud overlap distributions can be evaluated by computing the resulting radiation fluxes based on the different overlap schemes. The fluxes determined for the (statistical) Markov overlap model are more accurate than maximal or random overlap for the CRM model experiments. A second part of the model is determining the cloud amount (either water or ice) conditioned on the presence of a cloud. Here the CRM is used to identify a spatial covariance for amount and radiative flux calculations are found to improve over using just the average amount of cloud water and ice for each vertical layer.

Extracting Statistical Properties of Sea Surface Currents from Drifting Buoys Thomas Bengtsson (GSP), Richard Jones (University of Colorado at Denver), 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 continuous time state equation based on the ocean current and surface wind vector. Parameter estimates in the state equation indicate a substantial departure in the Labrador Sea from the usual damping and the inertial time scales seen in other areas of the ocean.

Determining Extremes in Daily Ozone Eric Gilleland (GSP and Colorado State University), Nychka in collaboration with William Cox (U.S. Environmental Protection Agency (USEPA), and Shelly Eberly (USEPA) have studied the problem of estimating the spatial field comprising the ozone standard proposed by the USEPA. The standard for an ozone monitoring station is an extreme order statistic (fourth highest) of the daily ozone averages observed over the summer season. This is a nonlinear statistic that has different spatial structure from daily ozone measurements and is possibly not normally distributed. Despite these difficulties the goal is to estimate the spatial field of the standard at unmonitored locations. The approach is to create a space-time model for daily ozone and sample the conditional distribution of daily ozone for the entire summer season at arbitrary locations given the record of observed data. The conditional simulation of the standard is then the order statistic applied to the conditional sample of daily values. This model has been used to quantify how well the ozone standard can be predicted from a given monitoring network and also confirmed the reliability of a fast spatial statistics approximation to the conditional distribution of the standard. The analysis has been extended to the Eastern US using a nonstationary covariance model.

Regression and Classification

Meta Analysis of Climate Simulations Claudia Tebaldi (ESIG, and Research Applications Program), in collaboration with Linda Mearns (ESIG) and Richard Smith (University of North Carolina), has developed a Bayesian, random effects model to combine the results of different climate model simulations and quantify the uncertainty. Different atmosphere-ocean general circulation models (AOGCM) produce different climate change projections, these differences being especially striking when trying to evaluate climate change on a regional (subcontinental) scale. Studies of multi-model ensemble predictions try to reconcile the AOGCM's responses by combining them into (weighted) means and to use the ensemble spread as an envelope of uncertainty around them. It is natural to embed this analysis in a statistical context in order to formalize these otherwise heuristic procedures and a Bayesian random effects model and been developed for this problem. Unlike more ad hoc approaches for combining climate model results, use of a statistical model delineates the assumptions made in synthesizing numerical results. As a specific project, we look at a data set studied by Giorgi and Mearns (Journal of Climate, May 2002) consisting of nine AOGCM experiments run under two emission scenarios with simulations of climate for present and future. The statistical method produces posterior distributions for regional temperature changes and has been found to give multimodal results for some regions suggesting model disagreement.

Scaling Climate Model Experiments Dorin Drignei (GSP, Iowa State University) and Claudia Tebaldi (ESIG, RAP) in collaboration with Linda Mearns (ESIG) and Thomas Wigley (CGD) have explored methods for inferring patterns of climate change by scaling the results of one climate experiment to infer results under other conditions. Temperature patterns were found to scale accurately between two different emissions scenarios (A2 and B2) but this accuracy degraded as one moves from large regional averages to individual model grid points. In errors in the scaled field can be modeled as a Gaussian nonstationary field and so quantify the uncertainty in the scaled estimate.

Uncertainty in characterizing a soil pedotransfer functions Steve Sain (GSP, CU-Denver) in collaboration with Linda Mearns (ESIG) and Shrikant Jagtap (University of Florida) has estimated and also characterized the uncertainty in the relationship between soil composition and water drainage properties that influence plant growth. The pedotransfer functions for the drained upper and lower limits (DUL, DLL) were determined using nonparametric regression. Thin-plate splines were used to predict DUL and DLL based on soil composition. The performance is comparable to the nearest neighbor methods described in the preceding section but have the advantage in quantifying the uncertainty in the estimated pedotransfer functions. From a statistical point of view this application is interesting as a model for a multivariate response based on compositional covariates (e.g percentages of sand, silt and clay). The statistical model fit to the Radcliffe and Richie soil data base was used to generate a posterior distribution of DUL and DLL for a several soil compositions. The uncertainties in these parameters were examined by the resulting variability in yield using the CERES (3.0) crop model for corn. Preliminary results for a Southeastern climate suggest that the natural inter-annual variability of weather dominates the uncertainty in soil parameters.

Stochastic Multi-resolution Models for Turbulence This project is a collaboration among Brandon Whitcher (GSP), 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 describe these coherent structures (i.e., the number, size, shape, amplitude of vortices), and our goal is to formulate a statistical model for these structures. The initial statistical problem is to segment the time slices from a numerical simulation of turbulence to reliably identify vortices. The observed vorticity fields (vorticity is the curl of velocity) are decomposed via a non-decimated discrete wavelet transform. This 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. An important component of the selection is a hard thresh-holding step to limit the number of variables entering each regression. The identification procedure appears to be reliable and has been implemented on a parallel architecture to make analysis of a large and scientifically interesting numerical experiment feasible.

Robust Smoothing This is theoretical work by Hee-Seok Oh (GSP) that is motivated by smoothing the light curves for variable stars (see the entry from the 2001 ASR). In that application occasional outliers not only influence the spline fit but more fundamentally effected the determination of a data-driven smoothing parameter. The practical need for a less sensitive version of cross-validation leads to the development of a general principle for robust smoothing. The basic idea is to use the least squares spline based on pseudo data to study the properties of the nonlinear robust estimate. Also, the pseudo data concept suggests a general algorithm to find robust estimators. This has been productively applied to wavelet shrinkage type smoothers and is substantially simpler than other approaches for robust wavelet estimators.

Statistical Computing Statistical Supercomputing with the Gibbs Sampler This project is a collaboration among Tim Hoar (GSP), Chris Wikle (University of Missouri), and Milliff. 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 Centers 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. Statistical analysis of the Gibbs iterates suggests that the Markov chain is close to stationary after 50 iterates and approximately independent when separated by five iteration cycles. These results guided the Markov chain Monte Carlo (MCMC) generation of posterior samples. Posterior means from this analysis along with 50 posterior samples are being made available on a four DVD set in netCDF format. To our knowledge this spatio-temporal data product is unique in the geophysical community for providing an ensembles of the fields. In the statistics community, to our knowledge, this the largest MCMC application yielding a substantive product attempted by a statistical research group.

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. Previous results suggested an estimator where the square root matrix of empirical covariances among wavelet coefficients is decimated. This approach was successfully applied to the daily ozone field simulated by a pollution model. Future work includes extending this approach to irregularly observed data.

Dynamical Systems

Forecasting and Data Assimilation: Adapting Particle Filter Methodology This project is led 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. Building on the success of this non-Gaussian filter in Lorenz '63 3-dimensional system an extension was formulated for higher dimensional systems. The basic idea is to consider mixture approximations that are local both in the state space and also in the physical space. As a starting point it is demonstrated that this approach can be effective in handling distinctly Non-Gaussian distributions produced by the 40-dimensional Lorenz '96 system.

4 Placement of GSP members

GSP placement

- *Enrica Bellone*: Research Associate, Department of Civil and Environmental Engineering, Imperial College, London
- *Brandon Whitcher*: Senior Researcher, Research Statistics Unit, GlaxoSmithKline
- *Heeseok Oh*: Assistant Professor, University of Alberta
- *Thomas Bengtsson*: Jerzy Neyman (visiting) Assistant Professor, University of California at Berkeley.
- *Eric Gilleland*: Associate Scientist, ESIG/RAP NCAR

5 Future Plans

New research directions

Several promising new collaborations have been identified and are listed below:

- Continue our work on statistical models for cloud overlap parameterizations and we will do this collaborating with James Hack, David Bader and Mitch Moncreiff.
- We have outlined a new project to collaborate with Peter Hess and Natalie Mahowald on inter-annual variability of atmospheric constituents due to transport. This will give GSP a new project in ACD.

- In collaboration with the Terrestrial Sciences Section (CGD) a study of subgrid effects of precipitation uses fine scale statistical models for the Southeastern US. The goal is to understand how different levels of resolution in the land component of the Community Climate System Model respond to realistic meteorological forcing.
- In collaboration with Arthur Richmond HAO, transfer of ensemble and Monte Carlo data assimilation methods to spatial estimates of the electric fields in the ionosphere.

GSP members and visitors for project Year 5

New post docs
Uli Schneider (CU-Boulder Applied Math) starting 7/03
Tomoko Matsuo (UCAR/SUNY-Stonybrook) starting 7/03
Dorin Drignei (Iowa State University) starting 1/04
Continuing post docs:
Reinhard Furrer
Sarah Street (1/2 time)
Jarrett Barber (1/4 time)

Sarah Street has been on leave to teach two courses at CSU. She will return in June to finish out 6 months at half time. One criticism we have from external panels is that the post doc appointments should be staggered. So we will consider filling another position to start 1/04. A promising candidate already identified is Dorin Drignei (Iowa State) who has an interest in the statistical analysis and scaling of climate models. Dorin is described by his advisor as his best student.

GSP Visiting faculty:
Thomas Lee, Colorado State University 20% GSP 1/2003-
Craig Johns, University of Colorado-Denver 7/2002
Steve Sain, University of Colorado-Denver 20% GSP 7/2002-
Philippe Naveau, University of Colorado-Boulder 5/2002

GSP sponsored workshop

A joint SAMSI/GSP Workshop on Spatio-Temporal Modeling will be held 1-6 June 2003 Boulder, CO. This will involve approximately 80 participants with more than 25 young researchers being supported to attend. This is last workshop in SAMSI's 6 month program on Large-Scale Computer Models for Environmental Systems and is the only SAMSI workshop to be held outside North Carolina. SAMSI is contributing 40% of the costs.

6 Visibility

Presentations

Besides numerous individual seminars, a session featuring GSP members was organized for TIES Annual Conference of The International Environmetrics Society in Genoa, Italy,. June 2002. GSP members also participated in the workshop on Multi-scale models for the environment, Statistics and Applied Mathematical Sciences Institute (SAMSI), RTP, NC, March 2003.

GSP hosted more than ??? short term visitors mainly faculty members from academic institutions and a list is posted on the project webpage.

Software

GSP maintained an extensive web page (`http://www.cgd.ucar.edu/stats`) and along with several contributed packages in the R/S language for statistical analysis. These are publically available at: `http://www.stat.math.ethz.ch/CRAN`

- **fields** Statistical tools for spatial data.
- **RadioSonde** Skew-t and related plots for radiosonde profile data.
- **waveslim** Wavelet analysis and algorithms for 1 and 2 dimensions.

7 Bibliography

Published papers and manuscripts submitted

Ammann, C. and Naveau, P. (2001). Multi-Decadal periodicity in tropical explosive volcanism and its influence on climate. *Geophysical Research Letters.*

Bengtsson, T., Snyder, C. Nychka, D.(2003) A nonlinear filter that extends to high dimensional systems. *Journal of Geophysical Research-Atmosphere* (in revision)

Holland, D. Cox, W.M., Scheffe, R., Cimorelli, A., Nychka, D. and Hopke, P. (2003) Spatial Prediction of Air Quality Data. *Environmental Manager* (to appear)

Furrer, R.(2003) Covariance Estimation under Spatial Dependence, submitted to *Journal of Multivariate Analysis.*

Kupper, T., Berset, J.D., Etter-Holzer, R. Furrer, R. and Tarradellas, J. (2003) Concentrations and Specific Loads of Polycyclic Musks in Sewage Sludge Originating from a Monitoring Network in Switzerland. submitted to *Chemosphere.*

Nychka, D., Wikle, C.K. and Royle, J.A. (2002). Multi-resolution models for non-stationary spatial covariance functions. *Statistical Modeling* 2 315-331.

Nychka, D. and C. Tebaldi, (2002), Comment on 'Calculation of Average, Uncertainty Range and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability Ensemble Averaging" (REA) method' *Journal of Climate*, 16, 883–884.

Oh, H-S, Nychka, D. Brown, T. and Charbouneau, P. (2003) Periodic Analysis of Variable Stars by Robust Smoothing. *Journal of the Royal Statistical Society Series C* (accepted)

Oh H.S., Ammann, C., Naveau P., D. Nychka and B. Otto-Bliesner. (2003) Solar Forcing and Climate: A multi-resolution analysis. *Journal of Atmospheric and Solar-terrestrial Physics* 65, 191-201

Johns, C. Nychka, D. Kittel,T., Daly C.(2003) Infilling Sparse Records of Precipitation Fields. *Journal of the American Statistical Association* (accepted).

Whitcher, B., S.D. Byers, P. Guttorp, and D.B. Percival (2002). Testing for Homogeneity of Variance in Time Series: Long Memory, Wavelets and the Nile River. *Water Resources Research*, 38 (5)

Whitcher, B., J.B. Weiss, D.W. Nychka and T.J. Hoar (2003). Stochastic Multiresolution Models for Turbulence. In M. Akritis and Talks D. Politis (Eds.) Recent Advances and Trends in Nonparametric Statistics, forthcoming.

Significant Manuscripts

Bellone, E. Wu, X-Q, Collins, W., Nychka, D. Moncreiff M. (2003) Statistical Models for Cloud Overlap from a Cloud Resolving Model

Bengtsson, T. and Cavanaugh, J. (2001). State-space discrimination using Kullback information measures. *Manuscript.* Sain, S. and Nychka D. (2003) Nonparametric models for soil pedotransfer functions.

Furrer, R. (2003) Covariograms of Nonstationary Spatio-Temporal Processes.

Furrer, R. and Bengtsson, T. (2003) Evaluation of forecast and Analysis Covariances in Kalman Filter Variants.

Furrer, R. and Naveau, P. (2003) Robust Estimation of the GPD's Parameters for some Dependent Cases.

Naveau, P., Furrer, R., Keckhut, P., Hauchecorne, A. and Goldfarb, L. (2003) Long-term variability analysis of stratospheric ozone over the Arctic region.

Streett, S. (2003) Observation driven temporal models for daily precipitation

Streett, S., Sain, S. and Nychka D. (2003) A latent field model for precipitation occurrence.

Tebaldi, C. Smith, R. and Nychka, D. (2002) A Bayesian analysis of regional climate change projections, based on a multi-model ensemble.

Whitcher, B., T. Lee, J.B. Weiss, D.W. Nychka and T.J. Hoar (2002). A Multiresolution Census Algorithm for Calculating Vortex Statistics.