Annual report for *A statistics program at the National Center for Atmospheric Research.* DMS-0355474 July 1, 2004 - June 30, 2005 (DRAFT)

April 28, 2005

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. During this period as part of the NCAR reorganization, GSP administration was transfered from the the Climate and Global Dynamics (CGD) Division to the Institute for Mathematics Applied to Geosciences (IMAGe). This new home for GSP facilitates its perspective as 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. Other sources of support that leverage GSP funds are listed along with term dates.

Post docs

- Reinhard Furrer 9/03- 6/05 (19% Weather and Climate Impacts Assessment Initiative, NCAR)

[1] NCAR research groups: Atmospheric Chemistry Division (ACD), Climate and Global Dynamics Division (CGD), High Altitude Observatory (HAO), Mesoscale and Microscale Meteorology (MMM), Institute for Study of Society and Environment (ISSE), Scientific Computing Division (SCD), Advanced Study Program (ASP), The Institute for Multidisciplinary Earth Studies (TIMES), Earth Observing Laboratory (EOL), Research Applications Laboratory (RAL)

- Tomoko Matsuo 7/03- (25% Data Assimilation Initiative, NCAR)
- Dorin Drignei 6/04- (50% CGD, NCAR)
- Uli Schneider 6/04- 9/04 (19% Weather and Climate Impacts Assessment Initiative, NCAR) on leave 10/04-6/05
- Jarrett Barber 5/04-7/04

Graduate Students

- Daniel Cooley, (5/03-) Applied Mathematics Department, CU-Boulder
  (Philippe Naveau, advisor)
- Curtis Storlie, (10/03 - 6/05) Statistics, CSU
  (Thomas Lee, advisor)

Visiting Faculty

Thomas Lee, Colorado State University 20% GSP 1/2003-
Steve Sain, University of Colorado-Denver 20% GSP 7/2002-

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 as part of the NCAR Annual Science Report 2004 `www.asr.ucar.edu/2004/CGD/gsp/narrative_gsp.html`. Finally, it should be noted that some of the projects reported below are continuations from the previous year reflecting the continuation of the GSP post docs from the previous year. Thus, the reported progress can be incremental rather than initiating a new area of application.

Spatial and Spatio-temporal Processes

Extremes for atmospheric variables This project considers models to describe the variation of extreme meteorological variables over space. Two datasets are used as testbeds for this research.

A hierarchical extremes model has been developed for extreme precipitation data from Colorado's Front Range with the goal of improving on methods currently in use by the United States National Weather Service in the construction of regional

precipitation atlases. A Bayesian hierarchical model is proposed in which the parameters of the generalized Pareto distribution at each location are spatially dependent and possibly dependent on other covariates (such as the annual precipitation, slope and elevation). From this model one obtains a spatial prediction of precipitation return levels, along with ensemble-based measures of uncertainty. One interesting result for this analysis is the lack of strong dependence of the parameters on the annual average precipitation suggesting some decoupling between the mean and tail properties.

A space/time model for surface ozone pollution was developed under previous NSF funding. As a companion to this project the distribution of high ozone values was considered from an extremes approach. The regulatory standard for an ozone monitoring station is an extreme order statistic (fourth highest) of the daily ozone averages observed over the summer season. A practical statistical problem is to draw inferences about the value of this standard at locations that are not monitored. Similar to modeling extreme precipitation, a generalized Pareto distribution (GPD) was used to model ozone extremes as a function of space. Here the region had a common shape parameter, and the scale parameter varied according to a spatial model. The probability surface estimated through this model is very similar in structure to a spatial analysis obtained by a spatial autoregressive model applied to the daily ozone measurements. The fact that an extremes approach agrees with a more conventional space time model is surprising and reinforces the credibility of each method

Some theoretical questions related to spatial dependence extremes have also been investigated. Spatial dependencies are often modeled with the variogram; however, use of the variogram requires that second moments be finite, which is often not the case with extreme value distributions. However, the madogram (an L1 version of the variogram) only requires that first moments be finite, and it also has convenient relationships with maxima. Preliminary results suggest that the madogram can be used to estimate spatial dependence parameters for fields of stationary max-stable random fields.

Multiresolution covariance models Flexible classes of nonstationary covariance functions are important for the analysis of historical climate data or atmospheric soundings, sparse measurements of the upper atmosphere, and as a background (or prior) covariance in data assimilation. The basic idea behind a wavelet representation is that the random field is a composition of basis functions with localized support multiplied by random coefficients. The main contribution of this modeling project is the construction of a covariance matrix among the wavelet coefficients that is much simpler than the covariance of the original field, can be adaptive over space, and has many zero elements. One result of this project is a statistical parameterization of the square root of the coefficient covariance matrix (H) that can accurately approximate the shape/scale Matern family of one-dimensional covariances. Preliminary results also suggest that this parameterization can be extended to a two-dimensional family

including anisotropy. The sparsity of the H matrix makes it possible to compute the likelihood of the covariance efficiently for large spatial datasets using sparse matrix algorithms and a Monte Carlo Expectation/Maximization (MCEM) algorithm has also be shown to work for irregularly sampled data. Here the E step is accomplished using conditional simulation from the spatial model to estimate the missing values of the field on a regular grid. The M step takes advantage of fast wavelet algorithms for regular grids, and unlike Fourier techniques, does not require tapering at the boundaries. As an illustration of the MCEM approach, a nonstationary covariance function can be found for daily surface ozone measurements for 500+ stations in the Eastern US. This covariance representation will be particularly useful for extending the extremes model for ozone to this large region.

Regression and Classification

Parameter estimation for intermediate climate models This project is collaborative research among Chris Forest (Massachusetts Institute of Technology, MIT), and Bruno Sanso (University of California at Santa Cruz) and is intended to revise the method for estimating the uncertainty in climate system properties from Forest, et al. (2002). From a statistical point of view, the goal is to estimate the parameters of a nonlinear regression model where the regression function is expensive to evaluate and the response has correlated errors. Some of this work draws on the design and analysis of computer experiments, but has the additional complexity that a run of the climate model only yields a realization of the climate and so contains some sampling error. To apply a fully Bayesian approach, we first approximate the response of the MIT intermediate climate model with a statistical model that provides a response surface in the uncertain parameter space. The three-dimensional parameter space is defined as climate sensitivity, rate of deep-ocean heat uptake, and the net aerosol forcing, and covers the three major uncertain quantities that affect the ability to simulate accurately the 20th century climate record. The availability of this response surface gives one an approximate likelihood surface for a continuous range of parameters and allows for sampling the joint posterior distribution of the parameters using a MCMC techniques. Preliminary results have shown that the uncertainty of the parameter estimates taking into account the uncertainty in the likelihood are comparable to those obtained by Forest and collaborators using more ad hoc statistics.

Along a another line of development is the use of more adaptive covariance models when fitting the model output to observations. Previous work has used a single error covariance based on a long control run with a fixed set of physical parameters. It is possible that the error covariance is dependent on these parameters and a better approach would be to construct a global statistical model that allows the error covariance to vary as a function of the climate model parameters.

Stochastic Multiresolution Models for Turbulence With the current state of computing technology, it is relatively straightforward to generate high-resolution numer-

ical integrations of the Naiver-Stokes equations, the fundamental physical equations that describe (nonreactive) fluid motion. One feature in the simulation of geophysical type fluids is the presence of eddies and vortices – and more complex structures for three-dimensional turbulent flows. It is of interest in the study of turbulence to describe these coherent structures (i.e., the number, size, shape, amplitude of vortices), with the goal of formulating a statistical model for these structures. We have been successful in identifying vortices in 2-d turbulent flow using a template that can been modified in its shape and scale by a basis derived from wavelets. The key to the method is to select basis functions using generalized cross-validation, in which the number of effective parameters is obtained via a thresholding rule. This method accurately identifies vortices and reproduces the scaling relationships among different properties of the vortices (e.g. enstrophy or size) and time.

Meta Analysis of Climate Simulations A statistical approach to combine climate projections for different climate model is an important tool for synthesizing the results from different modeling centers. The objective is to develop a Bayesian mixed effects model to combine the results of different climate models simulations and quantify the uncertainty. Here the random effects are the model biases from the true climate. Although these statistical models have been developed for model output for a particular point or region, the goal here is to handle the complete spatial fields produced by each model. Thus in this case the random effects are actually spatial fields and the modeling needs to consider the spatial structure among these random variables. One successful approach is to represent the climate change signal as a heterogeneous expansion of basis functions that involve spherical harmonics, indicators for significant land masses and covariates based on the pattern of current climate. The departures from this signal are assumed to be isotropic and Gaussian spatial processes. Using largely uninformative priors for the models parameters and MCMC techniques for computing, it is possible to derive estimates of the posterior distribution for climate change that includes the uncertainty among models. This analysis has been applied to the climate model experiments completed for the Fourth Intergovernmental Panel on Climate Change.

Robust Smoothing This is a theoretical project in collaboration with Hee-Seok Oh (Seoul National Univ., Korea) that is motivated by smoothing the light curves for variable stars. 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.

Dynamical Systems

Tracking vortices and storms In order to study many geophysical processes it is useful to be able to track coherent features through a sequence of images. The motivating examples used in this project are vortices in 2-dimensional turbulent flow and propagating, convective storms imaged from Doppler radar reflectivity. The basic model assumes the structures for tracking have centers, sizes and aspect, can be born, die, merge and split and move as independent random walks. Based on images sampled in time the paths are estimated by maximizing a likelihood. The approach also has been successful when the image is corrupted by noise (clutter) although maximizing the likelihood requires longer computing times. Some theory associated with this problem gives results on consistency of the method as the sampling times decrease.

Ensemble filters applied to a mesosphere model for the atmosphere. Ensemble Kalman filters (EnKF) are a useful Monte Carlo approximation to the computation of the Bayesian posterior distribution when combining observations with a numerical model. Most work using geophysical models and EnKF techniques has focused on deterministic models that have significant internal variability. This variability helps to maintain the ensemble spread making it a viable filter. In this work we consider a geophysical model where additional adjustments may need to be made to insure adequate ensemble spread.

The mesosphere and lower thermosphere (MLT) region is a transition region where wave energy originating from the troposphere and stratosphere dissipates and is absorbed via dynamical, chemical, and radiative processes. This project uses a 3-d chemical and dynamical middle atmosphere model (ROSE) and observations based on the TIMED spacecraft and ground based data. This numerical model and observation operators have been successfully combined in a data assimilation software environment that facilitates experiments for large problems. Preliminary results indicate that the EnKF is stable.

Statistical Computing

Improvements to the Gibbs Sampler for large spatio-temporal data sets. Previously funded work has developed a Bayesian hierarchical spatial model to combine 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. Due to computational and storage constraints a production data product based this method was limited to two week time intervals. Thus there is the possibility of introducing some discontinuity between adjacent two week blocks of analyzed winds fields. This project proposed a solution to this problem by proposing an iteration in the Gibbs sampler where the posterior samples for a two week segment are updated by conditioning on the results for adjacent blocks. Although conceptually simple, the challenge is to effectively manage the

large data streams for these computations and take full advantage of the modular code developed for the disjoint case.

KriSp package for large spatial datasets A package has been created for spatial statistics using sparse matrices for the high-level language R. The functions are based on similar methods and classes of the well-established GSP package. The KriSp package allows one to handle large spatial prediction problems where observations occur at irregular locations. The approach is to introduce sparsity in the covariance matrix by tapering with a compactly supported (positive definite) kernel (e.g., one finds the direct product of the covariance matrix and the tapering matrix). Although this technique is common in atmospheric data assimilation, it is not well used in general. The theoretical background was established in the previous year stating that, given the correct choice of a taper (e.g., matching the tail behavior of the spectral density of the covariance function), the approximation is asymptotically optimal. The potential speedup, based a sparse approximations and implemented KriSp, is striking suggesting that a spatial prediction (Kriging) can be computed for several thousand points interactively in the R environment. Also, the tapering approximation is quite accurate sacrificing little statistical efficiency compared to the exact calculation.

4 Placement of GSP members

- Reinhard Furrer, Assistant Professor, Mathematics Department, Colorado School of Mines.
- Jarrett Barber, Assistant Professor, Statistics Department, Montana State University.
- Curtis Storlie, Postdoctoral visitor, Statistics and Applied Mathematical Sciences Institute and North Carolina State University.
- Daniel Cooley, Postdoctoral visitor, Statistics, Colorado State University and NCAR.

5 Future Plans

New research directions

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

- In collaboration with the Data Assimilation Section (IMAGe) develop flexible statistical models for the model error component in an Ensemble Kalman Filtering framework.

- In collaboration with the Computational Sciences Section (SCD) develop fast interpolation algorithms based on radial basis functions that are adaptive to structure in the interpolated function.
- In collaboration with the Terrestrial Sciences and Climate Modeling Sections (CGD) improve statistical methods for testing and comparing two spatial fields. This will extend recent work on controlling false discovery rate and wavelet based representations for spatial processes.
- In collaboration with the Weather and Climate Impacts Assessment Initiative (ISSE) extend the Bayesian model for combining climate model experiments to a multivariate version that can incorporate both seasonal temperature and precipitation.
- In collaboration with ISSE, University of Colorado-Denver and Colorado State University use spatial models for extremes to characterize the climate response of regional climate models to different future scenarios of greenhouse gas emissions and also to changing landuse.
- In collaboration with Global and Stratospheric Studies (ACD) identify statistical models for the (atmospheric) tidal forcing on the mesosphere and incorporate this into the ROSE numerical model for the mesosphere.

GSP members and visitors for project Year 2

Post docs
Uli Schneider 7/05- 7/06 (returning from leave)
Mikyoung Jun 9/05 - 12/05
Dan Cooley 9/05 - 6/06 (50% CSU)
Shree Khare 7/05 - (75% IMAGe)
Dorin Drignei 7/05 - 6/05 (50% with NCAR Climate Division)
Tomoko Matsuo 7/05 - 6/05 (25% with IMAGe)

6 Visibility

GSP sponsored workshops

GSP will cosponsor a summer school (June 13-17, 2005) with SAMSI and the NCAR Data Assimilation Section on *Fusing Models and Data: from practice to theory to practice.* This is being held as part of the SAMSI six month program on Data assimilation in the geophysical sciences. The total participation is expected to be more than 30 students and researchers with computer lab space for 20 for hands-on afternoon practicums.

Presentations

GSP Organized a Topic Contributed Session at the 2004 Joint Statistical Meetings: *Spatial Statistics for Geophysical and Environmental Process*

GSP members have given more than 30 invited seminars and posters.

Visitors

GSP hosted four statistics graduate students as summer interns. These students were given short projects that had a scientific component mentored by NCAR staff. This is a experimental program for GSP and IMAGe to engage statistics students during their Ph D. Given the success of this program, a similar summer program will be supported by IMAGe visitor funds for summer '05.

- *Angela Pignotti*, UC Santa Cruz, Climate model diagnostics related to forcing regional models.

- *Tingting Shu*, Univ. Alberta, Nonlinear, multiresolution decomposition of time series.

- *Bin Shi*, Georgia Tech, Statistical models for 3-d turbulence experiments.

- *Cari Kaufman*, Carnegie Mellon, Estimating covariance parameters for large data sets and temperature lapse rates for atmospheric soundings in the Front Range of Colorado.

GSP hosted more than 8 short term visitors mainly faculty members from academic institutions and a list is posted on the project web page.

Software

GSP maintained an extensive web page (`http://www.image.ucar.edu/GSP/`) 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` GSP R packages modified or created during this period.

- **fields** Statistical tools for spatial data.

- **KriSp** Kriging for large spatial data sets using sparse matrix methods.

7 Bibliography

Published and accepted papers

Bengtsson, T., R. F. Milliff, R. Jones, D. Nychka, and P. Niiler (2005). A state space model for ocean drifter motions dominated by inertial oscillations. To appear *Journal of Geophysical Research- Oceans*.

Drignei, D. (2005). Empirical Bayesian Analysis for High-Dimensional Computer Output, To appear in *Technometrics.*

Fournier, B., and R. Furrer (2004). Automatic Mapping in the Presence of Substitutive Errors: A Robust Kriging Approach. To appear *Applied GIS.*

Fuentes, M., T. G. F. Kittel, and D. Nychka (2004). Sensitivity of ecological models to spatial-temporal estimation of their climate drivers: Statistical ensembles for forcing. To appear *Ecological Applications*

Furrer, R. (2004). Covariance Estimation under Spatial Dependence. To appear *Journal of Multivariate Analysis*.

Gilleland, E., and D. Nychka (2004). Statistical Models for Monitoring and Regulating Ground-level Ozone. To appear *Environmetrics*

Gilleland E., D. Nychka, and U. Schneider (2005). Spatial models for the distribution of extremes. In *Applications of Computational Statistics in the Environmental Sciences: Hierarchical Bayes and MCMC Methods* ed. J.S. Clark and A. Gelfand, Oxford University Press. To appear.

Katz, R. W. (2005). Environmental sciences. In *Statistical Analysis of Extreme Values with Applications to Insurance, Finance, Hydrology and Other Fields* (Third Edition), R.-D. Reiss and M. Thomas, Birkhauser, Basel, Switzerland. To appear.

Katz, R. W., G. S. Brush, and M. B. Parlange (2005). Statistics of extremes: Modeling ecological disturbances. To appear *Ecology*.

Poncet P., D. Cooley, and P. Naveau (2005). Variograms for max-stable random fields. Book chapter in *Springer Lecture Notes in Statistics on Statistics for Dependent Data.*

Tebaldi, C., R. L. Smith, D. Nychka, and L. Mearns (2005). Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles. To appear *Journal of Climate*.

Papers in Review

Feingold, G., R. Furrer, P. Pilewskie, L. A. Remer, Q. Min, and H. Jonsson (2004). Aerosol Indirect Effect Studies at Southern Great Plains during the May 2003 Intensive Operation Period. Under revision, *Journal of Geophysical Research.*

Furrer, R., and T. Bengtsson (2004). Estimation of High-dimensional Prior and Posteriori Covariance Matrices in Kalman Filter Variants. Submitted to *Journal of Multivariate Analysis*.

Furrer, R., S. Sain, D. Nychka, C. Tebaldi, and G. A. Meehl (2005). Multivariate Analysis of Atmosphere Ocean General Circulation Models. Submitted to review *Environmental and Ecological Science*

Furrer, R., M. G. Genton, and D. Nychka (2004). Covariance Tapering for Interpolation of Large Spatial Datasets. Submitted to *Journal of Computational and Graphical Statistics*.

Katz, R. W., and M. Ehrendorfer (2005). Bayesian approach to decision making using ensemble weather forecasts. In review *Weather and Forecasting*.

Naveau P., J. M. Nogaj, C. Ammann, P. Yiou, D. Cooley, and V. Jomelli (2004). Statistical Analysis of Climate Extremes. Submitted to the *Comptes Rendus de l'Academie des Sciences*.

Naveau P., P. Poncet, and D. Cooley (2004). First-order variograms for extreme bivariate random vectors. Submitted to *Mathematical Geology*.

Sain, S., and D. Nychka (2005). A multivariate spatial model for soil water profiles. In review *Journal of Agricultural Biological and Environmental Statistics*

Significant Manuscripts and Conference Proceedings

Lee, T. C. M., D. Nychka, B. Whitcher, C. Davis, and J. Weiss (2004). Identifying and Tracking Turbulence Structures. *Proceedings of the 38th Asilomar Conference on Signals, Systems, and Computers.*

Oh, H-S., T.C.M. Lee, and D. Nychka (2005). Fast Robust Wavelet Regression.

Oh H-S., and D. Nychka (2005). Smoothing Spline Regression by Robust Cross-Validation.

Sain, S. R., and R. Furrer (2004). Fitting Large-Scale Spatial Models with Applications to Microarray Data Analysis. *Proceedings of Interface 2004: Computational Biology and Bioinformatics*, Baltimore, Maryland.

Tebaldi, C., D. Nychka, and L. O. Mearns (2005). Inferring climate from climate models. *Proceedings of the 55th Session of the International Statistical Institute.*