Research Statement

Gabriel Huerta

My main contributions to statistical science are based on Bayesian methods that had raised from multiple collaborations. I had performed research in the areas of Bayesian time series, space-time modeling, parameter estimation in climate modeling and extreme value analysis, among other areas.

Time series modeling

In my paper "'Priors and Component Structures in Autoregressive Time Series Models" which was published in the *Journal of the Royal Statistical Society*, (JRSS) Series B, I developed a new class of prior distributions for autoregressive models based on novel results for time series decompositions and the characteristic root parametrization. This class of prior distributions addresses issues about model order uncertainty, inference on latent structure, initial values and unitary roots. With this approach, I analyzed the *Southern Oscillation Index*, a time series that has been central in high-profile debates in the atmospheric sciences about apparent trends in climatological indicators. In a follow up paper, "Bayesian Inference on Periodicities and Component Spectral Structure in Time Series" published in the *Journal of Time Series Analysis*, I studied the implications of these structured priors on the spectral domain with observations from astronomical data from the *S. Carinae* star, a time series that has been considered for frequency analysis and regularities of movement of objects in space. Since publication these papers had been regularly cited in the specialized literature of time series analysis and had been extensively considered in the study of Electro-encephalogram (EEG) data in the context of sequential Monte Carlo known as *Particle Filters*. Furthermore, in my paper 'Structured priors for multiple time series" which appeared in *Journal of Statistical Planning and Inference*, I studied a multivariate extension of these structured prior distributions for multiple time series in the form of vector autoregressions and with the goal of identifying common latent structure across multivariate series. Particularly, I illustrated this approach with an analysis of seasonally adjusted US Housing data consisting of housing starts and housing sales values over the period 1965 to 1975. An open research problem in this area is to assess how well these structured AR models can be used within a multilayer framework to represent, for example, changes in stochastic volatility across time. Another open problem is to extend these structured priors into MA terms to permit for a time series model that includes model order uncertainty for both AR and MA terms and considers for a stationarity and invertibility within the ARMA process context. It is also plausible to extend these approaches into models for unequally-spaced time series. Due to the complexities in the modeling of autocorrelations and frequency characteristic of data that is measured at irregular time intervals, this area of research does not have much development yet. However, plenty of data that arises from climatology, astronomy and biology is unequally-spaced. I consider that any new developments along this line of research would be a milestone for time series analysis. In particular, I had researched a space-state framework where the eigenvalue/eigenvector decomposition of the evolution matrix is modified to include time gaps between observations. In this way, one can avoid the technical difficulties of continuous time stochastic differential equations and its corresponding discrete process approximations.

Spatial and space-time modeling

Another important aspect of my research program has been spatio-temporal analysis with environmental applications. In my paper, "A spatio-temporal analysis for Mexico City ozone levels" that appeared in the *Journal of the Royal Statistical Society*, Series C (Applied Statistics), I created a spatio-temporal model for hourly ozone levels in Mexico City that allows spatial smoothing and tem-

poral prediction. The model incorporates the daily cyclical patterns of ozone and is dependent on covariates such as temperature and spatial coordinates which are incorporated through time-varying regressions. Such a regression requires interpolated values of temperature at locations and times where readings are not available. The main product of this research is retrospective/prospective analysis of the readings and spatio-temporal maps of ozone. Since this model is empirical, a possible extension is to incorporate the influence of wind fields on ozone and use transport theory to include a theoretical representation of the chemical process that generates the pollutant. This paper was one of the first ones to consider a real data application of Dynamic models to study information from a monitoring environmental network. This paper has been highly cited, and has produced wide interest in the modeling of pollution concentration fields for a large variety of data sets along with many studies around univariate/multivariate spatial models and space time dynamic models.

Extreme value models

In my paper "Time-varying models for extreme values" that appeared in *Environmental and Ecological Statistics*, I developed a new approach to model extreme values that are measured in time and space . The main motivation for these models was to provide a statistical description of block-maximum ozone values in Mexico City. The key idea of this work is to use a hierarchical Bayes approach to impose a spatial and/or temporal latent process formulated over the model parameters of the well known Generalized Extreme Value (GEV) distribution. Combining Dynamic Linear models and MCMC methods, the framework provides a smooth version of extremes in space and time in the form of quantile probability estimation and trends. Currently, I am applying these models to study relationships between extreme rainfall data in Venezuela and climatological indexes such as the Atlantic-Pacific index. My former Ph.D. student Wenxia Ying extended these models to allow for an AR component in the parameter structure of the GEV distribution. An important aspect of this work is to assess if a latent stochastic representation is comparatively better to describe trends of extremes compared to deterministic representations given by simple linear regression or polynomial trends. Furthermore, there is an increased interest to consider extreme values that are generated from output of climate models and treat this output as data that could be analyzed with Bayesian hierarchical analysis and GEV distributions to produce quantile estimates. The purpose of this approach is to guide climate modelers in assessing the validity of regional climate models and to test theories about climate change. In the book chapter "Dynamic and Spatial Modeling of Block Maxima Extremes" that will appear in the volume *Bayesian Inference and Markov Chain Monte Carlo: In Honor of Adrian Smith*, an written with my Ph. D. student, Glenn Stark, I had illustrated the use of such hierarchical models by combining GEV distributions with methods from *Gauss Markov Random Fields*. In particular, we are re-analyzing extreme precipitation that was produced by a Penn State/NCAR regional climate model and that covers the southwestern United States. In this modeling framework, the location and scale parameters of the GEV distribution have a spatial component that is defined through neighbors and that allows to capture local phenomena as storms. In addition, the location parameter also includes an additive trend component to capture for time changes. One of the main products of this research are maps of predictive quantiles of extreme rainfall across the southwestern United States. In particular, these maps show that the higher values of quantiles are observed along the Pacific coast from British Columbia through northern California, through northern and central California, and in central Arizona.

Estimation problems in Climate Models

In collaboration with Dr. Charles Jackson a climate scientist at the Institute of Geophysics at the University of Texas-Austin, I had developed statistical methods to study parameter estimation in

climate models. My paper "Error Reduction and Convergence in Climate Prediction" that appeared in *Journal of Climate* describes the general Bayesian framework to the problem of parameter calibration via stochastic sampling and with likelihood functions formulated as metrics of model skill. The paper highlights the estimation of posterior density functions for parameters of a general atmospheric and circulation model known as CAM 3.1 and offers improvements over previous methods used in the climate literature to deal with uncertainty quantification in this context. Furthermore, my collaborations with Dr. Jackson were originally funded by a National Science Foundation grant that allowed me to support the studies of a former Ph.D. student, Alejandro Villagran who took a faculty position at the University of Connecticut. The original goal of this NSF project was to study in more detail the sampling/optimization based methods (Multiple Very Fast Simulated Annealing) of the *Journal of Climate* paper and to establish comparisons for estimation of posterior distributions with efficient MCMC type methods (Adaptive Metropolis-Hastings). The results published in the paper "Computational Methods for Parameter Estimation in Climate Models" that appeared in *Bayesian Analysis* are based on a 3-D surrogate climate model where the response is surface air temperature in relation to the Earth's orbital parameters. The results of this research indicates that although MVFSA can converge rapidly, this technique may have serious biases at the level of posterior quantile estimation when compared to Adaptive Metropolis. A serious challenge with these climate models, particularly with the large atmospheric circulation models, is that the computational burden to produce single climate runs is enormous, which limits the capability of generating thousands of posterior samples from traditional MCMC methods. Therefore, there is still a need to study surface approximation methods for climate models based on Gaussian process with a large number of dimensions or inputs. This is inside the frontiers of current research in the area of analysis of computer experiments. In collaboration with the Los Alamos National Laboratories, we are considering Principal Component Analysis and Gaussian Process emulators to analyze output from an NCAR-Community Atmospheric Model (CAM 3.1). Representation and dimension reduction for climate response fields is possible with these approaches, but attention has to be paid to sufficient averaging of the climate model runs. In addition, my current funded grant by the Climate and Environmental Science Division of the Department of Energy addresses the problem of creating a multi-variate metric that takes into account spatial and field dependencies that avoid many of the mathematical and scientific limitations imposed by more traditional skill metrics based on singular value decomposition and empirical orthogonal functions as in our *Journal of Climate* paper. This project provides funding for a Postdoctoral Fellow, Dr. Alvaro Nosedal-Sanchez, who is currently implementing these new methodologies for CAM 3.1 experiments that were performed base on a factorial design in parameter space. The new proposed measures take into account spatial variability between grid locations modeled with Gaussian Markov Random Fields, and variability between fields through data estimated covariance matrices. The first results of this research show that the new approaches are able to pinpoint parameter combinations in a more concise manner compared to previous approaches based on a simplistic mean square error metric.

Other contributions to statistical research

My other research in time series includes the modeling approach known as Hierarchical Mixture-of-Experts (HME). This approach is very flexible and allows comparison of arbitrary models, nonlinear modeling and incorporation of covariates through weights that have a specific parametric form (multinomial logistic). The applications for this approach stem from issues about economic time series such as "trend-stationary" versus "difference-stationary" (papers published in *Statistica Sinica* and *Journal of Computational and Graphical Statistics*). More recently and supervising

a graduate student, I considered a simplified version of HME, Mixture-of-Experts models, with the goal of estimating and predicting stochastic volatility for exchange rates and stock market data. Particularly, we found some interesting relationships between an emerging stock market and the Dow Jones Index (paper published in *Advances in Econometrics*). I had also considered estimation of HME for vector autoregressive (VAR) processes, models that had been useful to describe structure in multivariate data from electroencephalogram (EEG) studies (paper published in *Computational Statistics and Data Analysis*). In the area of Bayes Wavelet shrinkage, I had developed statistical models for data denoising problems (papers in *Journal of Applied Statistics* and *Wiley Interdisciplinary Reviews*). Also, I had developed a Bayesian model to assess the impact of managed-care strategies of length-of-stay in hospitals of children with a psychiatric condition (paper in *Lifetime Data Analysis*). I have also experience with data analysis of problems that arise in Biology, by studying methods to describe the spatial and temporal distribution of nano-particles across a cell membrane (paper in *Bulletin of Mathematical Biology*). More recently and jointly with the Ph.D student Glenn Stark, I had been developing Bayesian models to study elevated disease risk due to exposure of uranium mining and mill waste across the Navajo Nation. Our models consider logistic type distributions that are combined with Bayesian model averaging to determine which are the main factors that produce elevated disease risk. Another aspect of my research is the analysis of Adaptive Metropolis methods and of novel adaptation schemes with the right ergodic properties. These methods are not only attractive for climate modeling estimation problems but also for situations with difficult-to-evaluate likelihood functions where traditional MCMC methods have slow convergence properties or fail to converge. My results with these adaptive algorithms are very promising since for complicated situations (mixture modeling, Poisson Regressions, non-parametric methods), the convergence of this new method is faster than with standard MCMC simulations.