Research Statement

My research has centered on the following areas: (i) statistical computing involving Markov chain Monte Carlo (MCMC) algorithms, (ii) Gaussian random field models for spatial data and computer experiments with applications to interdisciplinary research, and (iii) statistical approaches for problems in software engineering. My interdisciplinary work has involved collaborations with researchers in geosciences, geography, ecology and infectious disease modeling.

MCMC algorithms are general algorithms for simulating from probability distributions. Much of my research has focused on developing MCMC algorithms that are efficient, so accurate answers are obtained quickly. I have developed MCMC algorithms that work well for very popular and flexible spatial models, which often pose serious challenges to existing MCMC methods. I have worked on approaches for producing rigorous estimates of the accuracy of MCMC-based estimates, and have shown how these accuracy estimates can be used to determine when to stop the algorithm. I am thus working towards creating automated MCMC algorithms that are efficient but where little user intervention is required, accuracy estimates are easily obtained, and simple but theoretically justified approaches exist for knowing when to stop the algorithm.

I have developed heavy-tailed approximations to the spatial model posterior distributions. in conjunction with and theoretical work on stopping rules for MCMC to construct a virtually completely automated MCMC algorithm for a class of spatial models. This work is described in a paper that I have revised for the *Journal of Computational and Graphical Statistics (JCGS)*. My work on rigorous, theoretically justified rules for stopping MCMC, based on a consistent estimate of Monte Carlo standard errors, is summarized in a paper published in the *Journal of the American Statistical Association (JASA)*. A related paper where I compare our approach with other standard ways of stopping MCMC runs, has appeared in *Statistical Science*. I have also developed a host of efficient MCMC algorithms based on block updating parameters for spatial models, a comparative study of which has led to a publication in *JCGS*. I have more recently worked on slice sampling algorithms in order to automate MCMC. Our approach produces better mixing Markov chains but is computationally very expensive. In order to resolve the computational burden, our work takes advantage of the latest in parallel computing technology, thereby producing a fast mixing *and* computationally

efficient algorithm. This work has been submitted to *Statistics and Computing*. I have summarized the state of the art in modeling and computation for Gaussian random field models in an invited (peer-reviewed) chapter on "Gaussian random field models for spatial models" for the edited volume *The MCMC Handbook*, an updated version of the influential *Markov chain Monte Carlo in Practice* volume published in 1996.

My collaborative work began with graduate research in modeling air pollution data, leading to a publication in *Case Studies in Bayesian Statistics*. I have developed collaborations at Penn State with researchers from geography, geosciences, plant pathology, ecology, and infectious disease dynamics. My crop epidemiology research with plant pathologists won an *Outstanding Poster* award at a Bayesian Statistics conference, and has led to a publications in the *Journal of Agricultural, Biological and Environmental Statistics*. I have worked closely with geoscientists on climate change research. Our group has developed statistical approaches for learning about climate system characteristics combining information from multiple large spatial data sets with output from complex (computer) climate models. One of our areas of focus has been making predictions about the future of the global ocean circulation system, which plays a key role in global climate. We have submitted one approach that builds a flexible model for the multivariate spatial data, taking advantage of special matrix structures to make computations tractable, to *Annals of Applied Statistics*; a somewhat less flexible but very fast approach to solve this problem has been submitted to *Journal of Geophysical Research.*

My collaboration with geography is on space-time models for the spread of invasive plant species — this involves modeling binary and zero-inflated spatial data. I have one paper submitted to *Environmetrics*, and am currently working with a Ph.D. student on studying modeling and computation for binary data on a lattice. I recently had a grant funded with a geographer to pursue invasive plant species modeling. While it has taken considerable time and effort to build these first time collaborations — in each case I have been the sole statistician (besides graduate students) — my collaborative research has started to produce several manuscripts, and promises to develop new statistical methods for important scientific challenges, and provide training and financial support for graduate students. I have been succesful with obtaining grant funding from several agencies, including the *National Science Foundation (NSF)* and the *U.S. Geological Survey* on grants related to my work with geoscientists, the *National Oceanic and Atmospheric Administration*

(NOAA) for work with geographers/ecologists, and *The Gates Foundation* for work on infectious disease dynamics. This year, for instance, I will be supporting two Ph.D. students (K.S. Bhat and R. Jandarov) through my US Geological Survey and Gates Foundation grants for climate change and disease dynamics research respectively. I also currently have four pending proposals.

In addition to my collaboration with scientists across Penn State, I have worked with computer scientists and other statisticians on statistical approaches called random forests for predicting failures in large software systems. Two papers based on this work have been published: one in the proceedings of *ICSE*, the premier international conference for software engineering research, and the other in *IEEE Transactions in Software Engineering*, the flagship journal of the field. Both have acceptance rates of under 20%.

I have worked with several Ph.D. students: V.Recta (joint with J.L.Rosenberger) on spatial generalized linear models, Ph.D. completed February 2009; K.S. Bhat on computer experiments, multivariate spatial data in the context of climate science, Ph.D. expected summer 2010; M.M. Tibbits (joint with J.L. Liechty) on parallel computing and automated MCMC algorithms; J.P. Hughes (joint with J. Fricks) on computing . Three students have completed M.S. theses with me. Two other Ph.D. students, R. Jandarov and J. Hughes, have recently started working with me on infectious disease dynamics and computation for exponential family ('auto-exponential') models respectively. I am working with C. Congdon, an M.S. student, on modeling computer experiments where the output is multivariate and has an equality constraint. In addition, I have served or am currently serving on a total of 9 Ph.D. thesis committees in statistics and 10 Ph.D. thesis committees in other departments, frequently with students of collaborators or other researchers who need expertise with Monte Carlo methods, statistical computing or spatial modeling.