Challenges in the Statistical Modeling of Stochastic Processes for the Natural Sciences (17w5107)

Peter F. Craigmile (The Ohio State University),
Richard Lockhart (Simon Fraser University),
Wendy Meiring (University of California, Santa Barbara),
Vladimir Minin (University of California, Irvine),
Debashis Mondal (Oregon State University),
Paul D. Sampson (University of Washington)

Sunday July 9–Friday July 14 2017

1 Overview of the Field

Stochastic processes are constructed to explore and answer key scientific questions about varied scientific phenomena (Guttorp, 1995). These stochastic processes can capture relationships between the different physical processes of interest on spatially and temporally varying scales, while accounting for the uncertainty inherent in complex data sources or in specifying the stochastic model itself. As the science and the corresponding stochastic process models become more complex, and are informed by richer data sources, there is a greater need to develop robust tools for statistical inference for statistical models based on stochastic processes in the natural sciences.

2 Recent Developments and Open Problems

Traditional statistical models for natural science phenomena have largely been either linear models or black boxes. It is now recognized that stochastic models whose behaviour more closely mirrors the scientific structure of the system under study provide a more interpretable framework for analyzing data. These ideas are now being used in such disparate fields as atmospheric science, biology, climate, environmental sciences, geophysics, and hydrology. This diversity of applications leads, of course, to a diversity of modeling approaches. There are, however, a number of challenges common to quantifying uncertainty in data analysis using these stochastic processes which cut across the range of applications. One example is in building stochastic processes that respect the current known science, but also incorporate believable uncertainty quantification. Another is building statistical inference tools for understanding complicated stochastic processes, observed on potentially massive or complex datasets. Solving these problems will help scientists, for example, understand uncertainty in forecasts of the effects of climate change, in prediction of flood risk, and in the study of blood cell differentiation in hematology.

This workshop was dedicated to understanding what is common to the problems in all these areas, to encouraging transmission of ideas from one application area to others, and to working towards solutions to the problems we identify.

3 Presentation Highlights

The workshop was constructed around a set of diverse areas in statistical science involving the use of stochastic processes to analyze different areas of applications in the natural sciences. The majority of the talks at the workshop were intended to motivate 12 breakout sessions that were held on the Tuesday and Thursday of the week (sessions A and B were later combined). Summaries of these breakouts are provided below. Breakout sessions were held in parallel. This guaranteed that each breakout session was of reasonable size (i.e., not too large, but not too small) to facilitate discussion. Participants were encouraged to discuss the statistical methods used in each topic area and also to outline open challenges as a way to motivate more general cross-area discussion on issues in building statistical models for stochastic processes. Students were active in all the breakout sessions. There was a report back period where all members of the workshop could comment on what each group had discussed.

Early in the workshop, 8 poster presentations by Ph.D. students and postdoctoral researchers outlined new research in the statistical modeling of stochastic processes, applied to problems in seismology, epidemiology, and various aspects of the climate system (land-based, in the atmosphere, and in the oceans).

4 Scientific Progress Made

A description and summary of the scientific findings for each breakout session is presented below. Many additional references for each breakout session can be found in the attached supplement, or at the BIRS website:

http://www.birs.ca/workshops/2017/17w5107/files/

Breakouts A. and B. Statistical parameter estimation and inference for dynamical models and Statistical and computational challenges posed by partially observed compartmental models

Moderators: Jennifer Hoeting, Colorado State University and Vladimir Minin, University of California, Irvine

Description of the session: In the study of biological, ecological, or environmental dynamical processes, many theoretical models have been developed but it is not common practice to estimate model parameters using statistical functions of observed data. Often these (compartmental) models describe dynamics of populations, where individuals can be assigned types, and individuals are allowed to switch types as time goes by. A challenge for statisticians is to develop methods to address the issue of the computationally intensive or intractable likelihoods required for these problems. This session reviewed statistical challenges that arise when analyzing such models and highlighted how different research communities propose to tackle these challenges. The main discussion revolved around statistical procedures in compartmental models under realistically complicated observation schemes with noisy observations and large fractions of missing data.

Outcomes:

1. Participants discussed various ways to make stochastic compartmental models statistically and computationally tractable. Many approaches exist that approximate the state of counts with a vector of non-negative real numbers using diffusion limit arguments. Such approximations are questionable when some components of the vector are close to zero. It was noted that hybrid models in which only some components of the vector are subject to the diffusion limit approximation are necessary. Sometimes it is easier to introduce real valued latent variables for computational tractability. Other computational methods, such as particle filters and Markov chain Monte Carlo (MCMC), were discussed.

2. Often parameters of stochastic compartmental models are only weakly identifiable from sparse data. Bayesian statisticians should encourage practitioners to superimpose priors and posteriors for all model parameters to catch identifiability issues. Remedies for weak identifiability offered during the discussion included:

(a) Helping scientists design better experiments or better data collection protocols;

(b) Integrating multiple sources of information (e.g., epidemiological data and sequence data during analysis of infectious diseases outbreaks);

(c) Changing the model;

(d) Checking whether interpretable transformations of parameters make physical sense;

(e) Admitting model identifiability may be a problem and investing time into developing very informative priors for some model parameters, for which prior information exists.

3. Participants pointed out that mechanistic models may be better for inference and interpretability, but often phenomenological models are good enough for prediction. For example, understanding dynamics of cholera spread may require mechanistic models, but possibly Poisson regression may be good enough for predicting cholera case counts.

Breakout C. Spatio-temporal modeling 1

Moderators: Michael Stein, University of Chicago and Paul Sampson, University of Washington

Description of the session: This session targeted aspects of space-time statistical modeling that have received little attention from the statistical community. A second theme was how recent developments in measuring processes in space-time may change the kinds of questions we might want to ask or can answer about space-time statistical models. Michael Stein delivered the motivating presentation beginning with consideration of the specification of mean and covariance function for challenging spatio-temporal datasets, and specifically the ocean based Argo float data or anomaly field. This and a second example, the Harris, TX, ozone dataset presented in an early JASA discussion paper on space-time modeling led to two lessons. Lesson 1: One must consider models with space and time trend components. Lesson 2: What matters in a model depends on the problem.

Another issue involved the estimation of a "mean" of a stationary geostatistical process. Two possible types of "mean" arise:

(i) The expectation of the process, which is constant since the process is stationary;

(ii) The average of the process over the spatial domain (a spatial integral).

The characteristics of the process that matter differ for each problem: in case (i) the low frequency properties of the process matter, whereas for (ii) the high frequency characteristics matter.

Later discussions involved the interaction of sources of variation in space-time data, leading to consideration of the frequency or spatio-temporal resolution of observations in space and time, particularly for satellite instruments like data from the GOES-16 geostationary satellite (https://www.nasa.gov/press-release/nasa-successfully-launches-noaa-advanced- geostationary-weather-satellite) and then to the challenge of space-time processes that include a vertical dimension. In both oceans and the atmosphere, the nature of circulation patterns is fundamentally different near the Earth's surface than away from it, so just adding the vertical as a third spatial component to existing models will not be adequate.

Outcomes: The group discussion touched on most of the issues above and raised some new issues needing further attention, including (1) the need to incorporate inequality constraints in models, (2) strategies for modeling nonstationarity, and (3) statistical modeling of means versus variability in extremes for certain problems. We also learned that (4) experience with waves leads to the requirement of combining properties of physical processes with statistical models, and (5) there is a need for greater involvement of statisticians in making data products that included expressions of uncertainty, possibly with the incorporation of multiple imputation. This was part of a discussion that moved outside of the initial topics of interest to discuss communication issues and uncertainty.

Breakout D. Non-Gaussian processes

Moderator: Alexandra Schmidt, McGill University

Description of the session: In the analysis of most spatio-temporal processes in environmental studies, observations present skewed distributions, with a heavy right or left tail. Usually, a single transformation

of the data is used to approximate normality, and stationary Gaussian processes are assumed to model the transformed data. The choice of a distribution for the data is key for spatial interpolation and temporal prediction. Initially we will discuss advantages and disadvantages of using a single transformation to model such processes. Then we will focus on some recent advances in the modeling of non-Gaussian spatial and spatio-temporal processes and discuss some possible avenues for research.

Outcomes: Several issues and questions were discussed and investigated during the session:

1. *Why do we need models for Non-Gaussian Processes?* It was noted that point predictions are quite robust to model misspecification, but that the correct estimation of non-Gaussian processes leads to big improvement in estimates of variability. Non-Gaussian processes are also useful in identify outlying regions and spatial outliers (e.g., Palacios and Steel, 2006)

2. *What features do you need your model to capture?* Non-Gaussian and/or non-stationary; Sharp versus smooth changes in the field over space; Low frequency versus high frequency signals; spatial interpolation versus parameter (effect) inference. A complicating factor was that non-Gaussian processes can look like Gaussian process data – this often occurs because we only have one replication.

3. *Comparing approaches: stochastic partial differential equation/process convolution versus transforming Gaussian processes.* There is a drawback with transformed Gaussian fields in that you cannot change the conditional distributions. We need non-Gaussian models to account for sharp transitions, since commonly used Gaussian models are too smooth.

4. *Distinguishing (if possible) between non-normal likelihood with a non-normal latent field.* Participants recognized there is a need to develop methodology to distinguish between non-Gaussian data (easier) and non-Gaussian latent field (harder). This is especially challenging when modeling discrete data (e.g., binary data observed over space). There is an open question of whether heavy tails are needed for binary data.

5. *Computation.* Many questions in this area. First question for any new model: can you estimate the model parameters? It was remarked that it is easy to create a new fancy model, but in practice the model parameters are very hard to estimate. How difficult are the computations for these models? (As the number of locations increases, analyses can get computationally intensive.) As data sizes increase, how do we extend to big data? (For example with Markov random fields the computational complexity jump is high as one moves move from 1 to 2 dimensions.) How do you know that you are getting the answer correctly from a computational method if there is only one way to compute it? There was a suggestion that we need multiple developments of a single model, or ways to verify that computational methods are correct.

6. *Diagnostics.* Diagnostics are extremely difficulty in this area. We need to develop statistical methods to assess whether we are modeling the complex dependencies reasonably. If the problem is high dimensional, we can try to think about conditional distributions to simplify the diagnostic problem. Also, given the issue with distinguishing between different processes, can we find something that looks and behaves similarly to having replicates, for model verification?

Breakout E. Spatial point processes

Moderator: Aila Särkkä, Chalmers University

Description of the session: In the early spatial point process literature, point patterns were typically small, observed in 2D, had quite simple interaction structures, and there were no repetitions of the point patterns available. The observed point patterns were assumed to be realizations of stationary and isotropic point processes. Furthermore, e.g. clustered patterns were typically modelled by assuming conditional independence between the cluster points given the Poisson distributed parents. Nowadays, large data sets (with repetitions) observed both in 2D and in 3D have become more and more common and it is less likely that stationarity and/or isotropy assumptions hold and that simple interaction structures are enough for realistic modelling of

the data. Adding temporal information increases the complexity, as spatio-temporal point pattern data have become common.

The theme of the session is to discuss the challenges of the more complicated point pattern data mentioned above. Some possible topics for discussion are:

1. Inhomogeneous or anisotropic?
2. Different types of anisotropy. How to test anisotropy?
3. How to model dependencies in clustered point patterns?
4. Spatio-temporal point processes.

Outcomes: We discussed two particular problems related to items 3 and 4 above, namely the Hawkes process and the Bartlett-Lewis instantaneous pulse model for rainfall. The Hawkes process is often used to model earthquakes which cause new earthquakes in the form of aftershocks. Max Schneider is working on these models and is especially interested in investigating how to include measurement error in the analysis. We discussed previous work on measurement errors in the point process analysis. We also discussed the parallels between the modeling of nerve structures in the skin, which was part of the motivating talk for the session, and modelling earthquakes.

The pulse model for rainfall incorporates layers: a process describing storms (Poisson start times, exponential durations), a process of modelling storm cells within each storm (Poisson start times, exponential durations), and a pulse process (rain falls) inside cells (marked Poisson process). The problem of producing useable forms of the likelihood was discussed; the dataset whose likelihood was discussed consists of 5 minute total rain falls at a station. All these discussions were very fruitful in the sense that ideas passed from people working on one kind of model to people working on other related models.

Breakout F. Spatio-temporal modeling 2

Moderators: Finn Lindgren, University of Edinburgh, and Wendy Meiring, University of California, Santa Barbara

Description of the session: Spatio-temporal problems often appear in contexts where the observation structure is complex at the same time as the latent or true spatio-temporal structures are non-trivial. Modern data sets frequently are massive. Statistical study of high-dimensional nonlinear scientific processes based on diverse observation structures leads to numerous modelling and computational challenges. We discussed this combination of challenges.

Outcomes:

1. We highlighted distinctions between model-development (physically based and/or statistically based), and computationally efficient methods needed to study parameters or processes based on these high-dimensional and complex models, using available observations. We discussed the need for scientifically based inspiration for statistical models, such as physics inspiration in the SPDE approach presented by Finn Lindgren.
2. We discussed multi-scale properties of the true processes and of models. Researchers need to ask what question they want to answer in each study, since different models and methods may be appropriate for different space-time scales, and for different target fields and measures (e.g., temperature or precipitation; extremes or averages).
3. In particular, space-time processes may exhibit diverse smoothness and association properties in different spatial and temporal dimensions/scales, requiring additional study in space-time statistics. An illustrative example contrasted the vertical versus horizontal directions in the atmosphere or ocean.

4. The field of space-time statistics will benefit if clear specification of the limitations of statistical models and computational methods is valued in publications, in addition to reporting cases where good performance was achieved. Researchers need to know when one model class/computational method may be preferred versus another.

5. Finn Lindgren's motivating talk concerned SPDE-based Matern fields with different operators, using mesh based computation methods and sparse precision matrices. We discussed sensitivity to the choice of mesh and basis function approximations in this approach. Finn mentioned choices he is finding valuable in studying global processes with approximately 10-15 terabytes of input data. We provided references to several other space-time methods and computational approaches.

6. We recognized the need for increased collaboration between statisticians and applied mathematicians/ computer scientists in order to benefit from many recent computational advances, in addition to collaborations with problem-specific scientists. We requested recommendations from participants on publically available resources for numerical analysis and computer science training for statisticians.

7. We mentioned the diversity of scientific areas of space-time study. Due to time our discussion concentrated on two areas: climate studies and the study of wave dynamics for risk analyses of the amount of fatigue a ship encounters on a particular journey. In the latter, waves are inherently asymmetric and difficult to measure and model.

8. We discussed potential sources for measurement biases, the need to understand data sources and potential biases, and to assess whether the data agree with the models and vice versa.

Breakout G. Ideas in visualization

Moderator: David Bolin, Chalmers University

Description of the session: Visualising estimates of low-dimensional random variables and their uncertainty is a rather straightforward problem. Also for time series and stochastic processes there are several good options for displaying uncertainty, such as simultaneous envelopes and functional boxplots. Uncertainty visualization for spatial and spatiotemporal fields is, however, not as straightforward. It is nevertheless of great importance when communicating results in spatial statistics and related subjects, or as Brodlie et al. (2012) writes:

We may encounter error bars on graphs, but we rarely see the equivalent on contour maps or isosurfaces. Indeed the very crispness of an isosurface gives an impression of confidence that is frankly often an illusion.

This is a major issue when visualizations are used in decisionmaking - such as planning evacuations based on a visualization of a predicted hurricane path. The central topics of discussion for this session are:

1. What is the best way of visualizing estimates of spatial fields and their uncertainties?

2. How should one do visualization for more complicated scenarios, such as problems in three spatial dimensions, spatio-temporal applications, and hierarchical models?

3. How can visualization effectively be used for model validation, and also for conveying data uncertainty/biases?

Outcomes: We continued discussing David Bolin's motivating talk on visualizing uncertainty in contour curves, including how many contours to include, contour map quality measures, and alternatives. Alternatives included presenting credible intervals/regions for the contours, and fog maps using color and opacity variation to indicate statistical measures of confidence. We discussed that the specific contours of primary scientific interest may be dictated by the intended audience, and that contour maps may be most valuable when the user cares about a specific level or very few levels. In two dimensions, contour maps can convey spatially explicit uncertainty estimates, although substantial challenges remain in combining uncertainty visualization from two spatial maps (such as mean and standard error), and in higher dimensions. We discussed pros and

cons of visualizing/returning many samples from the posterior distribution in a Bayesian Analysis instead of presenting summary contour maps. Samples from the posterior distribution may be especially valuable when the user is most interested in functions of the field (such as flow fields), however we recognized that substantial information needs to be provided to the user on how to interpret and use the posterior samples they are receiving. Regarding data uncertainty, we discussed ongoing work on interactive visualization (including Sam Shen's presentation earlier in the workshop). Meta-data may be a valuable way to convey data representativeness or reliability when interactively mousing over a region. We also mentioned ongoing work in 3D animation and virtual reality, including displaying neural connections from brain imaging. We are excited by rapidly developing visual analytic techniques and computational methods that are proving valuable in diverse areas such as neuroscience, chaos, environmental sciences, and media arts and technology. We recognize the need for increased collaborations with visualization experts. We also discussed the need to adjust the degree of detail depending on the target audience, and the necessity to provide all the information necessary for the target audience to understand visualizations of the uncertainty most important for their scientific use.

Breakout H. Communicating science to decision makers

Moderator: Richard Lockhart, Simon Fraser University

Description of the session: Like all scientists, statisticians are increasingly interested in seeing their work used in support of sound public policy. They want to communicate results while being honest about the unavoidable uncertainty accompanying real data and do that without finding that their advice is ignored because of that uncertainty. The session was intended to focus on the following issues:

- Who is communicating? Which expert [individual scientist or small research group, science society, advocacy group] communicating with which policy maker [boss, government bureaucrat, politician]?
- One-way or two-way communication?
- Do people know of successful communication examples? Did those communicators use specific strategies?
- How can we get trained in a hurry if need be?
- What dangers for individual scientists need to be watched out for?
- We are uncertainty quantification experts. Are we decision making experts?
- How can we elicit from policy makers what sort of questions they would really like answered? Are they the sorts of questions we can answer?
- Are we providing information or trying to influence the decision?

Outcomes: The list of questions served to get discussion going but was naturally, and intentionally, too long to permit discussion of every point. In the end the following major issues arose:

- There is a need for a taxonomy of this sort of communication to highlight the different types of communication to decision makers. Discussants had experience in such contexts as: an advisory panel to the EPA tasked with recommending a specific ozone standard; advisory committees to government agencies (the EIA, Statistics Canada, and water boards in Australia), working on a large interdisciplinary grant studying climate impacts in Norway where part of the task was regular communication with stakeholders, and meeting congressional staffers to communicate scientific understanding of issues such as those surrounding climate change.
- We identified the issue of indirect communication. It is common for the statistician to have access to staff in an agency or a political body. In turn the staffer communicates with senior administration or a congressional representative and in turn those people communicate with the final decision maker. The EPA ozone rule, 8 years in the making, provided a case study. The statistician, and her/his scientific colleagues, must communicate with those staff to whom they have access in such a way that the information finally communicated to the ultimate authority is clear, compelling, and correct.

- We discussed the process of becoming involved in this sort of communication. It was apparent that this is a slow, difficult process. Statisticians must first succeed in becoming involved with other scientists who have an interest in influencing decision making. Then they must demonstrate their utility to those other scientists. The process of establishing this reputation is long.
- There was some discussion of the issue of one-way versus two-way communication. The Norwegian example provided a case study of the benefits of involving stakeholders. The EPA ozone example illustrated the difficulty when communication flows through a long chain.
- We discussed at some length the danger that communicating accurately the uncertainty in scientific conclusions can lead to policy makers ignoring advice. We also noted that advice can be ignored even in the absence of any significant uncertainty.

Breakout I. Popular Science

Moderator: Peter Guttorp, University of Washington and Norwegian Computing Center

Description of the session: There are many outlets for statisticians to publish their work in popular science:

1. Journals for people with some statistics background (e.g., Chance, http://chance.amstat.org; Significance; https://www.significancemagazine.com)
2. General science magazines (e.g., Popular Science; http://www.popsci.com;
   Science News, https://www.sciencenews.org)
3. Blogs and online magazines (e.g., Hakai Magazine, https://www.hakaimagazine.com; PLOS Blog, http://blogs.plos.org)
4. Tweets

This session will discuss the different ways that statisticians can communicate their research, including building links with journals and science writers.

Outcomes: A number of different topics were covered:

1. *Identifying and differentiating between the roles of statisticians and journalists.* It was agreed that working together or providing joint training was beneficial in communicating statistics to a wider audience. We noted that this may be harder for junior statisticians. We also discussed differences between communicating a single article and a body of work (the latter is less common).
2. *Where do we start?* Possible outlets identified and discussed included: University/research organization public relations office; Personal relations with journalists; Online popular science journals/blogs; Going out to schools.
3. *Misstatements and oversimplifications.* It is very common to overstate or oversimplify the message of scientific studies (e.g., The science is settled on climate change). The session identified that we were all responsible to communicate our ideas as simply as possible, while keeping the message accurate. There are still open questions on how to communicate uncertainty in popular science. Often this is oversimplified.

Breakout J. Modeling in Transformed Domains: Future Directions

Moderators: Debashis Mondal, Oregon State University, and Donald Percival, University of Washington

Description of the session: Transformations are commonly used for many different reasons in statistical analysis. In traditional regression courses students are taught to use transformations to force non-Gaussian data into a Gaussian shoe. Transformations can make distributions more Gaussian, symmetric, can stabilize

variance, and can be used to guard against multicollinearity in regression problems. Sometimes transformations (e.g., ranks) are used to make the inference robust or resistant to outliers, and are highly efficient for many distributions. For time series and spatial analysis, transformations facilitate or simplify modeling, characterize or compensate for correlation, help extract signals in the presence of noise and handle nonstationarities, by forcing data into a stationary shoe (Sampson and Guttorp, 1992). The session had a number of aims:

1. How can we adapt existing transforms (Fourier, wavelet etc.) to handle new problems arising in environmental data analysis: what are interesting directions to pursue?
2. Are there other transforms of interest for analysis of environmental data that have yet to be fully explored, especially for non-standard domains such as networks?
3. What is the best way to deal with transforms that help satisfy distributional problems, but at the cost of messing up the correlation structure?
4. For transforms that do not preserve variance, are there ways in which we can do at least a quasi-ANOVA?

Outcomes: We discussed the following:

1. For correlated spatio-temporal data, we typically use transformation to find patterns and scales on which correlation makes sense. While certain machine learning methods (e.g., random forests) may be easier for finding patterns, it was noted that spatially agnostic transformations, which are easy to implement, may be harder to interpret. We agreed that there is a need to have transforms that are spatially and temporally aware. This is a problem with empirical orthogonal functions, where it can be hard (impossible) to compare principal components for different datasets.
2. There are many transformations for gridded data (e.g., wavelets, Fourier, and needlets on the sphere). There are methods for wavelet-like transforms for nearest neighbors, although these cannot be interpreted as being on a fixed scale. For methods on graphs we can use wavelets (e.g., Sharpnack, et al. 2013) or Laplacians.
3. Often transformations of parameters are used instead of transformations of data. This can pose its own difficulties as parameters may need to be constrained (e.g., constraining the direction of flow of a river using directed graphs). In time series analysis reparametrization of model parameters using the partial autocorrelation function can help immensely (similar ideas apply to imposing sparsity on spatial and spatio-temporal data using Gaussian Markov random fields).
4. There is a need to understand how transformations are useful in modeling nonlinear dynamics (e.g., in climate how do we model El Nino year dependencies versus La Nina year dependencies?) This is further complicated with teleconnections, which are easier to represent with certain transformations (e.g., empirical orthogonal functions) compared to others. Multiscale nonstationary statistical models can help.
5. Often transformations impose problems themselves (e.g., there is no exact analysis of variance decomposition), or the transformation induces correlation in observations (e.g., satellite monitoring, tomography and other medical imaging data). While there are inverse modeling techniques available, more work is needed to build reliable statistical models for these transformation-based data products.
6. Many transformations are not explored well in the statistical literature and need further study: e.g., independent component analysis, random projections, empirical mode decomposition, dynamic mode decomposition.

Breakout K. Extremes

Moderators: Holger Rootzén, Chalmers, and Thordis Thorarinsdottir, Norwegian Computing Centre

Description of the session: In this breakout session, we will focus on the topic of multivariate extremes and the associated risks. Atmospheric, hydrological and earth scientists spend important efforts on estimating risk of catastrophes at a very large set of locations all over the world. Typically, this is done separately at each single location so that one, for example, tries to estimate risks of flooding at an individual dam, and then uses the result for design and operation at this dam. However, few tools seem to be available for estimating joint risks, say the risk that at least one of the perhaps several thousand dams in a country is flooded during the coming year. An accurate assessment of such risks would, in particular, provide valuable information for national and regional authorities when planning their disaster risk management strategies. Are there methods and data available which make it possible to make such risk estimates? In particular, how do joint risks change if, e.g., design criteria at individual dams are changed? And, if not, could such methods be developed? These problems seem both important and exciting, with climate change adding to their importance.

Outcomes: We identified a wide variety of problems where there is great need for understanding of potentially dependent extremes. These are problems where you have many structures and are concerned about failures at any one of them. In particular we discussed bridges, wooden structures, dams, flooding (urban, causing landslides, destroying roads), dam and dike breaches; windstorms; extreme precipitation/temperatures and storm floods. We are concerned both about extreme climate and weather events and other natural disasters, and about extreme consequences in dollars, lives, or human suffering. We discussed examples where extremes at different locations (worst cases over a period of time) would likely be independent or nearly so and other examples (extreme cold in northern Sweden, the Calgary flood, etc) where there would clearly be significant spatial correlation because many of the extremes would be simultaneous, or essentially so. We discussed the role of extreme value theory in designing large structures and in setting design codes. We contrasted the complexity of modern theoretical analyses with a desire for extreme simplicity in codes. We asked "Who might care about joint risks?" They should be important for emergency plans at a national/regional level, for design codes and regulations, and for re-insurers. We discussed the role of statisticians in these processes, the extent to which money was the driving consideration and the need to participate in discussions with the civil engineering community. Finally we identified a number of areas needing statistical research: we need to understand individual events better before we can think about multivariate modeling; we observed that multivariate extreme value theory is under rapid development but is not very high dimensional yet (as an example of the issue there are 2300 dams in Norway to be considered and assessing the risk of the worst case over a period of time and over 2300 spatially dependent structures is hard); we likely need spatial model development on a case-by-case basis; when we try to model joint extremes we need to cope with what is happening at non-extreme locations (conditionally on an extreme at some location or locations); we need to know when we can trust our models, for the questions of interest, and we need to distinguish non-human and human causes.

Breakout L. Modern approaches to climate and hydrological data analysis and modeling

Moderators: Efi Foufoula-Georgiou, University of California, Irvine, and Sam Shen, San Diego State University

Description of the session: It is well established by now that the earth system is a highly interconnected system across processes (atmospheric, oceanic, and eco-hydrologic) and across scales (cloud microphysics to large scale dynamics). Improving modeling and prediction relies on using state-of-the-art methodologies for extracting from available data relevant information and hidden relationships, merging observations at different scales, advancing data assimilation methodologies, and identifying trends and patterns. Listed below are a few references on topics: (i) Modern spatial data analysis and reconstruction methods: from EOF regression to random forests, and dictionary learning; and (ii) Stochastic modeling approaches for the next generation of climate models: from self-similarity to connections between microscopic scales and global circulations scales.

In the first 30-minute session, Efi Foufoula-Georgiou used the blackboard to illustrated a series of open problems on statistical modeling and hydrology. Sam Shen supplemented three more problems on climate data analysis, visualization and modeling.

Outcomes: This breakout parallel session was not formally held due to very few participants. However, people discussed the topics of the session at other occasions. Efi Foufoula-Georgiou and Sam Shen invited the participants to comment, contribute and collaborate on their proposed topics on modern approaches to climate and hydrological data and modeling. In addition, they provided the following four concrete research directions on these topics:

1. How to diagnose/use climate models to construct regionally relevant climate projections?
2. Using big data technology and computer visualization to create movies for the past climate.
3. Sub-seasonal-to-seasonal (S2S) forecasting needs better statistical models: This is one of the current NOAA research foci with a goal of breaking the 4-day barrier of numerical weather forecasts by using both statistical models and numerical weather forecast models.
4. Next generation GCMs linking 10^{-6} meter to 10^{6} meter scales need nonlinear stochastic models: The main question is whether one can use nonlinear stochastic models to link the microscopic scale dynamics of atmosphere, such as the condensation process and cloud ice crystal formation, with mesoscale and large-scale precipitation.

5 Outcome of the Meeting

There were a number of common themes that emerged from this workshop that participants of the workshop will continue to work on:

1. In all areas of statistical modeling of stochastic processes in the natural sciences there is a need to develop more methods for model diagnostics, verification, and validation. This is especially important in areas where the science is not fully worked out, and statistical models are being used to assess and validate new scientific hypotheses or statistical models are being used for out-of-sample prediction and risk assessment. Additionally there is a need to have ways to test that complex statistical models are being fit correctly.
2. Problems are becoming more complicated with richer data sources; this necessitates the construction of more involved, but realistic, statistical models that capture the necessary interactions between the scientific processes of interest. This is difficult in non-Gaussian settings (e.g., modeling precipitation, characterizing spatio-temporal extremes, and in studying biological phenomena using complex hierarchical representations).
3. Uncertainty quantification is still being under-utilized in the sciences. Using statistical models is key to making more reliable inferences using data, with realistic uncertainty assessment.
4. Clear and effective communication of our statistical results as well as the science is key. There is a need for statisticians to collaborate more with scientists and journalists to make this happen.

Further discussions of these topics will occur in other workshops this year. For example, the Program on Mathematical and Statistical Methods for Climate and the Earth System (CLIM), at the Statistical and Applied Mathematical Sciences Institute.

References

[1] K. Brodlie, R.A. Osorio, and A. Lopes (2012), A Review of Uncertainty in Data Visualization, in Expanding the Frontiers of Visual Analytics and Visualization, 81–109, Springer, London.

[2] P. Guttorp (1995), Stochastic Modeling of Scientific Data, Chapman & Hall, London.

[3] M. B. Palacios and M. F. J. Steel (2006). Non-Gaussian Bayesian geostatistical modeling. *Journal of the American Statistical Association*, **101**, 604-618.

[4] P.D. Sampson and P. Guttorp (1992), Nonparametric estimation of nonstationary spatial covariance structure, *Journal of the American Statistical Association*, **87**, 108-119.

[5] J. Sharpnack, A. Krishnamurthy and A. Singh (2013), Detecting Activations over Graphs using Spanning Tree Wavelet Bases, Proceedings of the 16th International Conference on Artificial Intelligence and Statistics (AISTATS), Scottsdale, AZ, USA.

Supplement to: Challenges in the Statistical Modeling of Stochastic Processes for the Natural Sciences (17w5107)

Additional references from the breakout sessions

[1] J. C. Aerts and W. W. Botzen. Climate change impacts on pricing long-term flood insurance: A comprehensive study for the Netherlands. *Global Environmental Change*, 21:1045–1060, 2011.

[2] S. Ambikasaran, D. Foreman-Mackey, L. Greengard, D. W. Hogg, and M. O'Neil. Fast direct methods for Gaussian processes. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 38:252–265, 2016.

[3] I. T. Andersen and U. Hahn. Matérn thinned Cox processes. *Spatial Statistics*, 15:1–21, 2016.

[4] C. Andersson, P. Guttorp, and A. Särkkä. Discovering early diabetic neuropathy from epidermal nerve fiber patterns. *Statistics in Medicine*, 35:4427–4442, 2016.

[5] C. Andrieu, A. Doucet, and R. Holenstein. Particle Markov chain Monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72:269–342, 2010.

[6] A. Bárdossy. Copula-based geostatistical models for groundwater quality parameters. *Water Resources Research*, 42, 2006.

[7] D. Bolin and F. Lindgren. Excursion and contour uncertainty regions for latent Gaussian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:85–106, 2015.

[8] D. Bolin and F. Lindgren. Quantifying the uncertainty of contour maps. *Journal of Computational and Graphical Statistics*, 2016.

[9] A. Bostrom, L. Anselin, and J. Farris. Visualizing seismic risk and uncertainty. *Annals of the New York Academy of Sciences*, 1128:29–40, 2008.

[10] K. Brodlie, R. A. Osorio, and A. Lopes. A review of uncertainty in data visualization. In *Expanding the frontiers of visual analytics and visualization*, pages 81–109. Springer, 2012.

[11] N. Chen and A. J. Majda. Simple stochastic dynamical models capturing the statistical diversity of El Niño southern oscillation. *Proceedings of the National Academy of Sciences*, pages 1468–1473, 2017.

[12] G. S. Chiu and J. M. Gould. Statistical inference for food webs with emphasis on ecological networks via Bayesian melding. *Environmetrics*, 21:728–740, 2009.

[13] R. de Fondeville and A. C. Davison. High-dimensional peaks-over-threshold inference for the Brown-Resnick process. *arXiv preprint arXiv:1605.08558*, 2016.

[14] V. Dukic, H. F. Lopes, and N. G. Polson. Tracking epidemics with google flu trends data and a state-space seir model. *Journal of the American Statistical Association*, 107:1410–1426, 2012.

[15] A. V. Dyrrdal, A. Lenkoski, T. L. Thorarinsdottir, and F. Stordal. Bayesian hierarchical modeling of extreme hourly precipitation in norway. *Environmetrics*, 26:89–106, 2014.

[16] A. M. Ebtehaj, E. Foufoula-Georgiou, and G. Lerman. Sparse regularization for precipitation down-scaling. *Journal of Geophysical Research: Atmospheres*, 117, 2012.

[17] G. Faÿ, E. Moulines, F. Roueff, and M. S. Taqqu. Estimators of long-memory: Fourier versus wavelets. *Journal of Econometrics*, 151:159–177, 2009.

[18] P. Fearnhead, V. Giagos, and C. Sherlock. Inference for reaction networks using the linear noise approximation. *Biometrics*, 70:457–466, 2014.

[19] B. Fischhoff. The sciences of science communication. *Proceedings of the National Academy of Sciences*, 110(Supplement 3):14033–14039, 2013.

[20] T. C. O. Fonseca and M. F. J. Steel. A general class of nonseparable space-time covariance models. *Environmetrics*, 22:224–242, 2011.

[21] T. C. O. Fonseca and M. F. J. Steel. Non-Gaussian spatiotemporal modelling through scale mixing. *Biometrika*, 98:761–774, 2011.

[22] E. Foufoula-Georgiou, A. M. Ebtehaj, S. Q. Zhang, and A. Y. Hou. Downscaling satellite precipitation with emphasis on extremes: A variational 1-norm regularization in the derivative domain. *Surveys in Geophysics*, 35:765–783, 2013.

[23] G.-A. Fuglstad, F. Lindgren, D. Simpson, and H. Rue. Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statistica Sinica*, 2014.

[24] A. E. Gelfand and S. Banerjee. Bayesian modeling and analysis of geostatistical data. *Annual Review of Statistics and Its Application*, 4:245–266, 2017.

[25] T. Gneiting, M. G. Genton, and P. Guttorp. Geostatistical space-time models, stationarity, separability, and full symmetry. *Monographs On Statistics and Applied Probability*, 107:151, 2006.

[26] J. A. González, F. J. Rodríguez-Cortés, O. Cronie, and J. Mateu. Spatio-temporal point process statistics: A review. *Spatial Statistics*, 18:505–544, 2016.

[27] B. Gräler. Modelling skewed spatial random fields through the spatial vine copula. *Spatial Statistics*, 10:87–102, 2014.

[28] P. Guttorp and A. M. Schmidt. Covariance structure of spatial and spatiotemporal processes. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5:279–287, 2013.

[29] H. Häbel, T. Rajala, M. Marucci, C. Boissier, K. Schladitz, C. Redenbach, and A. Särkkä. A three-dimensional anisotropic point process characterization for pharmaceutical coatings. *Spatial Statistics*, 2017.

[30] S. Hallegatte, C. Green, R. J. Nicholls, and J. Corfee-Morlot. Future flood losses in major coastal cities. *Nature Climate Change*, 3:802–806, 2013.

[31] M. B. Hooten and C. K. Wikle. A hierarchical Bayesian non-linear spatio-temporal model for the spread of invasive species with application to the eurasian collared-dove. *Environmental and Ecological Statistics*, 15:59–70, 2007.

[32] C. Hutengs and M. Vohland. Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sensing of Environment*, 178:127–141, 2016.

[33] E. L. Ionides, D. Nguyen, Y. Atchadé, S. Stoev, and A. A. King. Inference for dynamic and latent variable models via iterated, perturbed Bayes maps. *Proceedings of the National Academy of Sciences*, 112:719–724, 2015.

[34] M. Jun and M. L. Stein. An approach to producing space-time covariance functions on spheres. *Technometrics*, 49:468–479, 2007.

[35] H.-M. Kim and B. K. Mallick. A Bayesian prediction using the skew Gaussian distribution. *Journal of Statistical Planning and Inference*, 120:85–101, 2004.

[36] P. Krupskii, R. Huser, and M. G. Genton. Factor copula models for replicated spatial data. *Journal of the American Statistical Association*, 2016.

[37] E. Lakatos, A. Ale, P. D. W. Kirk, and M. P. H. Stumpf. Multivariate moment closure techniques for stochastic kinetic models. *The Journal of Chemical Physics*, 143:094107, 2015.

[38] H. Liang and H. Wu. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103:1570–1583, 2008.

[39] J. Liepe, P. Kirk, S. Filippi, T. Toni, C. P. Barnes, and M. P. H. Stumpf. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nature Protocols*, 9:439–456, 2014.

[40] F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73:423–498, 2011.

[41] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32:139–160, 2005.

[42] R. E. Morss, O. V. Wilhelmi, M. W. Downton, and E. Gruntfest. Flood risk, uncertainty, and scientific information for decision making: Lessons from an interdisciplinary project. *Bulletin of the American Meteorological Society*, 86:1593–1601, 2005.

[43] V. D. Oliveira, B. Kedem, and D. A. Short. Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association*, 92:1422, 1997.

[44] M. B. Palacios and M. F. J. Steel. Non-Gaussian Bayesian geostatistical modeling. *Journal of the American Statistical Association*, 101:604–618, 2006.

[45] J. Parslow, N. Cressie, E. P. Campbell, E. Jones, and L. Murray. Bayesian learning and predictability in a stochastic nonlinear dynamical model. *Ecological Applications*, 23:679–698, 2013.

[46] D. B. Percival, S. M. Lennox, Y.-G. Wang, and R. E. Darnell. Wavelet-based multiresolution analysis of wivenhoe dam water temperatures. *Water Resources Research*, 47, 2011.

[47] N. Pidgeon and B. Fischhoff. The role of social and decision sciences in communicating uncertain climate risks. *Nature Climate Change*, 1:35–41, 2011.

[48] T. A. Rajala, A. Särkkä, C. Redenbach, and M. Sormani. Estimating geometric anisotropy in spatial point patterns. *Spatial Statistics*, 15:100–114, 2016.

[49] J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69:741–796, 2007.

[50] H. Rootzén, J. Segers, and J. L. Wadsworth. Multivariate peaks over thresholds models. *Extremes*, 2017.

[51] H. Rue, A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren. Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application*, 4:395–421, 2017.

[52] B. T. Russell, D. S. Cooley, W. C. Porter, B. J. Reich, C. L. Heald, et al. Data mining to investigate the meteorological drivers for extreme ground level ozone events. *The Annals of Applied Statistics*, 10:1673–1698, 2016.

[53] J. Sharpnack, A. Singh, and A. Krishnamurthy. Detecting activations over graphs using spanning tree wavelet bases. In *Artificial Intelligence and Statistics*, pages 536–544, 2013.

[54] S. S. P. Shen, N. Tafolla, T. M. Smith, and P. A. Arkin. Multivariate regression reconstruction and its sampling error for the quasi-global annual precipitation from 1900 to 2011. *Journal of the Atmospheric Sciences*, 71:3250–3268, 2014.

[55] D. Spiegelhalter, M. Pearson, and I. Short. Visualizing uncertainty about the future. *Science*, 333:1393–1400, 2011.

[56] S. N. Stechmann and J. D. Neelin. A stochastic model for the transition to strong convection. *Journal of the Atmospheric Sciences*, 68:2955–2970, 2011.

[57] M. L. Stein. Space-time covariance functions. *Journal of the American Statistical Association*, 100:310–321, 2005.

[58] L. Sun, C. Lee, and J. A. Hoeting. Parameter inference and model selection in deterministic and stochastic dynamical models via approximate Bayesian computation: modeling a wildlife epidemic. *Environmetrics*, 26:451–462, 2015.

[59] L. Sun, C. Lee, and J. A. Hoeting. A penalized simulated maximum likelihood approach in parameter estimation for stochastic differential equations. *Computational Statistics & Data Analysis*, 84:54–67, 2015.

[60] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of The Royal Society Interface*, 6:187–202, 2009.

[61] A. A. Tsonis and K. L. Swanson. Topology and predictability of El Niño and La Niña networks. *Physical Review Letters*, 100, 2008.

[62] J. Wallin and D. Bolin. Geostatistical modelling using non-Gaussian Matérn fields. *Scandinavian Journal of Statistics*, 42:872–890, 2015.

[63] Y.-X. Wang, J. Sharpnack, A. Smola, and R. J. Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17:1–41, 2016.

[64] G. Xu and M. G. Genton. Tukey g-and-h random fields. *Journal of the American Statistical Association*, pages 1–14, 2016.

[65] H. Zareifard and M. Jafari Khaledi. Non-Gaussian modeling of spatial data using scale mixing of a unified skew Gaussian process. *Journal of Multivariate Analysis*, 114:16–28, 2013.

[66] H. Zhang and A. El-Shaarawi. On spatial skew-Gaussian processes and applications. *Environmetrics*, 21:33–47, 2009.