# SAMSI Climate Program Report: Table of Contents

# Mathematical and Statistical Methods for Climate and the Earth System

# Final Report on SAMSI Program 2017-18

# Coordinated by Richard L. Smith

## 1. Summary of Program

The objective of this program was to develop various mathematical and statistical methodologies that are relevant to modern research on climate and earth systems science. On the mathematical side, the main themes were data assimilation, stochastic parametrizations in computer models, and applications to sea ice and food systems, among others. On the statistical side, there were too many themes to list all of them here, but major emphases were remote sensing, extremes and risk, the use of machine learning methods for climate applications, and environmental health. Activities included a preceding summer school at the National Center for Atmospheric Research (NCAR), 6 program workshops including one organized at the Alan Turing Institute (London, U.K.), fall and spring graduate courses, 2 education and outreach workshops, a public lecture by the distinguished hurricane scientist Kerry Emanuel (MIT), 6 postdocs, 14 research visitors and 12 active working groups. Products of the program (as of March 2019) include 19 papers published or in press; 14 under review; 37 in preparation; 5 software products; 14 grant applications submitted based at least in part on work done during the program, 66 reported conference presentations, and a conference on climate risk and the insurance industry that grew out of discussions in the program.

## 2. Summer School

Before any of the regularly scheduled activities of the program, a summer school took place at NCAR on Climate Datasets. This was jointly organized by SAMSI and by STATMOS, an NSF-sponsored statistics in atmospheric research network. A novel feature of this workshop was that the activity was entirely focused on nine subgroups, each of which combined expertise from both mathematical/statistical sciences and from atmosphere/ocean sciences and was devoted to analysis of a specific dataset. A report of this summer school is included in the Appendixes of this report.

## 3. Opening Workshop

The Opening Workshop took place from August 21-25 at the North Carolina Biotechnology Center and attracted around 100 participants. The opening day featured a set of five tutorial lectures on a wide range of topics including climate informatics, climate and health, climate extremes, climate prediction and climate datasets. There followed five invited research paper sessions on Remote Sensing, Ice Dynamics, Informatics, Climate and Health and Stochastic Parameterizations. There were two panel discussions, one on climate extremes and the other on climate risk. As is traditional in SAMSI opening workshops, there was a half day "working group formation" sessions and another half day devoted to initial meetings of the thirteen

working groups. The final half day session consistent of a very well attended tutorial on the use of high-performance computing in climate research, led by Doug Nychka and Dorit Hammerling of NCAR.

The complete program of the Opening Workshop is appended to this report.

## 4. Postdocs

For this program we were able to appoint six SAMSI postdocs, a record number for a single SAMSI program. The reason we were able to do this was our success in obtaining externally supported second-year placements for all six (except for Mikael Kuusela, who came from a STATMOS postdoc at the University of Chicago and spent his second postdoc year with SAMSI). The full list is:

- Kuusela, statistician with a PhD from EPFL (Lausanne, Switzerland)
- Yawen Guan, statistician with PhD from Penn State; second year placement with Brian Reich (NCSU);
- Christian Sampson, applied mathematician with PhD from the University of Utah; second year placement with Chris Jones (UNC);
- Whitney Huang, statistician with PhD from Purdue, second year placement joint with CANSSI and the University of Victoria;
- Huang Huang, statistician with PhD from KAUST (Saudi Arabia); second year placement with NCAR
- Maggie Johnson, statistician with PhD from Iowa State; second year placement still to be confirmed but tentatively with JPL.

## 5. Visitors

The following spent periods ranging from several days through to the full academic year at SAMSI:

- Amit Apte, former SAMSI postdoc visiting from the International Center for Theoretical Sciences (ICTS) in India (full-year visitor);
- Veronica Berrocal, former SAMSI postdoc and a biostatistician at the University of Michigan (full-year visitor);
- Asim Dey, PhD student from the University of Texas at Dallas, working with Lyubchich jointly with Yulia Gel in Dallas (fall);
- Dorit Hammerling, former SAMSI postdoc and a statistician from the Institute for Mathematics Applied to Geosciences, NCAR (now at Colorado School of Mines) (fall);
- Jong-Joon Jeon, statistician from the University of Seoul, South Korea (full year);
- Emily Kang, former SAMSI postdoc and a statistician at the University of Cincinnati (fall);
- Bo Li, statistician from the University of Illinois (fall);

- Slava Lyubchich from the University of Maryland Center for Environmental Sciences (fall);
- Pulong Ma, PhD student with Emily Kang (fall);
- Adway Mitra, visiting postdoc from ICTS (full year);
- Adam Monahan, applied mathematician at the University of Victoria in Canada (fall);
- Doug Nychka, statistician from the Institute for Mathematics Applied to Geosciences, NCAR (now at Colorado School of Mines) (fall and spring);
- Mark Risser, statistician recently appointed to a permanent research position at the Lawrence Berkeley National Laboratory (LBNL) (fall);
- Erik Van Vleck, applied mathematician from University of Kansas (spring);
- Michael Wehner, atmospheric scientist from LBNL (fall).

## 6. Graduate Courses

The Fall Semester featured a graduate class on "Statistics for Climate Research", coordinated by Richard Smith (UNC/SAMSI), Brian Reich (NCSU) and Doug Nychka (NCAR), with guest lectures from Bo Li (Illinois), Veronica Berrocal (Michigan), Surya Tokdar (Duke) and Murali Haram (Penn State). Over 20 students registered for the course (plus several auditors in regular attendance) with students from departments such as Marine Sciences at UNC and Economics at NCSU, in addition to the local mathematics and statistics departments.

The Spring 2018 graduate course was on "Data Assimilation for Dynamical Systems", jointly taught by Chris Jones (Mathematics, UNC), Amit Apte (ICTS) and Erick Van Vleck (visiting SAMSI from the Mathematics Department, University of Kansas).

The schedules for both courses are included in the Appendices.

## 7. Kerry Emanuel public lecture and workshop

A public lecture was given by Kerry Emanuel, of the Department of Atmospheric Sciences of MIT, on "The Storm Next Time: Hurricanes and Climate Change." This took place in the Genome Sciences Auditorium at UNC and attracted an audience of around 200.

In conjunction with Kerry Emanuel's visit, a short workshop took place at RENCI (the Renaissance Computing Institute, at UNC) on "Mathematical Problems in Radiative-Convective Balance of the Atmosphere," and featured Emanuel and five other external invited participants as well as several local researchers. The program for this workshop is included in the appendices.

## 8. Faculty and Graduate Fellows

Faculty Fellows were Chris Jones (Mathematics, UNC), Brian Reich (Statistics, NCSU) and Surya Tokdar (Statistical Science, Duke).

The Graduate Fellows for this program were Suman Majumder (Department of Statistics, NC State University), Amanda Muyskens (Department of Statistics, NC State University), Colin Guider (Department of Mathematics, UNC) and Sheng Jiang (Department of Statistical Science, Duke University).

## 9. Workshops

Apart from the opening workshop and the one organized in connection with Kerry Emanuel's visit, the following program workshops were organized:

- Remote Sensing, Uncertainty and the Theory of Data Systems, at Caltech, jointly funded by SAMSI and JPL, February 12-14 2018. Organizers: Amy Braverman (JPL/Caltech) and Jessica Matthews (CICS);
- Data Sciences for Climate and Environment, a one-day workshop organized jointly by SAMSI and the Alan Turing Institute, London, UK, March 26 2018;
- The Climate Transition Workshop, set for May 14-16, at SAMSI;
- Climate Extremes, at SAMSI, May 16-17;

## 10. Working Groups

Twelve Working Groups remained active throughout the program and presented their results at the Transition Workshop in May. The titles and organizers were as follows:

- Remote Sensing: Amy Braverman (JPL) and Jessica Matthews (NCEI/CICS);
- Parameter Estimation: Ben Timmermans (LBNL);
- Data Assimilation: Chris Jones (UNC) and Erik Van Vleck (Kansas);
- Climate Informatics: Doug Nychka (NCAR);
- Extremes: Dan Cooley (Colorado State) and Richard Smith (UNC);
- Stochastic Parameterizations: Adam Monahan (Victoria);
- Environmental Health: Brian Reich (NCSU);
- Food Systems: Hans Kaper and Hens Engler (Georgetown);
- Ice Dynamics: Chris Jones (UNC) , Alberto Carrassi (NERSC), Murali Haran (Penn State);
- Detection and Attribution: Dorit Hammerling (NCAR) and Matthias Katzfuss (Texas A&M);
- Risk and Coastal Hazards: Brian Blanton (RENCI), Slava Lyubchich (Maryland), Richard Smith (UNC);
- Statistical Oceanography: Michael Stein (University of Chicago) and Mikael Kuusela (SAMSI).

Final reports from the Working Groups are included as an appendix to this report. The Stochastic Parameterization group eventually merged with the Climate Informatics group and their results are combined as a joint report.

## 11. Education and Outreach Activities

A very successful Fall workshop on Climate Extremes took place on October 23-24, organized by SAMSI Deputy Director Elvan Ceyhan and featuring SAMSI visitors Michael Wehner, Doug Nychka and Bo Li, as well as a very lively session on data reconstruction by Scott Stevens and Jared Rennie, Cooperative Institute of Climate and Satellites – North Carolina (CICS-NC) and

National Centers for Environmental Information (NCEI). The workshop also featured talks and demos by SAMSI postdocs Huang Huang, Whitney Huang and Maggie Johnson.

This year's Undergraduate Modeling workshop (May 21-25, 2018) was led by Chris Jones (Mathematics, UNC) and Doug Nychka (NCAR), based on their joint project which uses climate datasets to teach undergraduate-level concepts in calculus and statistics. The schedule of projects is included in the Appendix.

**12. Program products**

a. Papers published or in press

Bessac, J., Monahan, A.H. , Christensen, H.M. and Weitzel N. (2019), Stochastic parameterization of subgrid-scale velocity enhancement of sea surface fluxes. In press, Monthly Weather Review.

Ebert-Uphoff, I., Huang, W.K, Mitra, A, Cooley, D.S., Chatterjee, S.B., Chen, C., and Wang, Z. Studying extremal dependence in climate using complex networks. Proceedings of the 8th International Workshop on Climate Informatics (CI 2018), Boulder, CO, 2018. NCAR Technical Note NCAR/TN-550+PROC, doi:10.5065/D6BZ64XQ, p. 37-40, Sept 2018.

Giglio, D., Lyubchich, V., and Mazloff, M. R. (2018). Estimating oxygen in the Southern Ocean using Argo temperature and salinity. Journal of Geophysical Research: Oceans 123: 4280–4297. DOI: 10.1029/2017JC013404

Yawen, G. and Haran, M. (2018), A Computationally Efficient Projection-Based Approach for Spatial Generalized Linear Mixed Models, Journal of Computational and Graphical Statistics 27, 701-714.

Guan, Y., Sampson, C., Tucker, J.D., Chang, W., Mondal, A., Haran, A. and Sulsky, D. (2019), Computer model calibration based on image warping metrics: an application for sea ice deformation. In press, Journal of Agricultural, Biological, and Environmental Statistics.

Huang, W. K., Cooley, D. S., Ebert-Uphoff, I., Chen, C., Chatterjee, S.B. New Exploratory Tools for Extremal Dependence: Networks and Annual Extremal Networks (2019). In press, Journal of Agricultural, Biological, and Environmental Statistics, Special Issue on Climate and Earth System. arXiv: https://arxiv.org/abs/1901.08169

Huang, W. K., Nychka, D. W., and Zhang, H. (2018). Estimating precipitation extremes using the log-histospline. Environmetrics, e2543.

Johnson, M., Caragea, P., Meiring, W., Jeganathan C. and Atkinson, P. (2019), Bayesian Dynamic Linear Models for Estimation of Phenological Events from Remote Sensing Data. In press, Journal of Agricultural, Biological and Environmental Statistics.

Kaper, H. and Engler, H. (2019), Modeling Food Systems. To appear in *Mathematics of Planet Earth: Protecting Our Planet, Learning from the Past, Safeguarding the Future*, H. Kaper and F. Roberts (eds), Ch. 2 (30 pp), Springer Verlag.

Kuusela, M. and Stein, M. L. (2018), Locally stationary spatio-temporal interpolation of Argo profiling float data. Proceedings of the Royal Society A 474: 20180400.

Li, A., Chen, S., Zhang, X. and Liu, Z. (2017), Political Pressures Increased Vulnerability to Climate Extremes for Nomadic People and Livestock in Inner Mongolia, China. Scientific Reports, 7: 8256.

Liang, Y., Mazloff, M.R., Rosso, I. Fang, S. and Yu, J. (2018), A Multivariate Empirical Orthogonal Function Method to Construct Nitrate Maps in the Southern Ocean. J. Atmos. Oceanic Technol., 35, 1505–1519, https://doi.org/10.1175/JTECH-D-18-0018.1

Ma, P., Kang, E. L., Braverman, A., and Nguyen, H. (2018), Spatial statistical downscaling for constructing high-resolution nature runs in global observing system simulation experiments, Technometrics, https://doi.org/ 10.1080/00401706.2018.1524791.

Mitra, A., Apte, A., Govindarajan, R., Vasan, V. and Vadlamani, S. (2019), A Discrete View of the Indian Monsoon to Identify Spatial Patterns of Rainfall. Dynamics and Statistics of the Climate System, https://doi.org/10.1093/climsys/dzy009, https://arxiv.org/abs/1805.00414

Mitra, A., Apte, A., Govindarajan, R., Vasan, V. and Vadlamani, S. (2019), Spatio-temporal Patterns of Indian Monsoon Rainfall, Dynamics and Statistics of the Climate System, https://doi.org/10.1093/climsys/dzy010, https://arxiv.org/abs/1805.00420.

Monahan, A.H., "Temporal filtering enhances the skewness of sea surface winds" (2018). Journal of Climate, 31, 5695-5706. (Work done while Research Fellow at SAMSI)

Nychka, D., Hammerling, D., Krock, M., & Wiens, A. (2018). Modeling and emulation of nonstationary Gaussian fields. Spatial statistics, 28, 21-38.

Risser, M.D., Paciorek, C.J., O'Brien, T. A., Wehner, M. F. and Collins, W. D. (2019), A probabilistic gridded product for daily precipitation extremes over the United States. Climate Dynamics. DOI: 10.1007/s00382-019-04636-0

Vakulenko S.A., Sudakov I., Mander L. (2018), The influence of environmental forcing on biodiversity and extinction in a resource competition model. Chaos, 28(3):031101. doi: 10.1063/1.5017233.

b. Papers currently under review

Berrocal, V.J., Guan, Y., Muyskens, A., Wang, H, Reich, B.J. and Chang, H.H. A comparison of statistical and machine learning methods for creating national daily maps of ambient PM concentration. Submitted, Environmental Health Perspectives.

Castro-Camilo, D., Huser, R. and Rue, H., A spliced Gamma-generalized Pareto model for short-term extreme wind speed probabilistic forecasting, Submitted to the JABES Special Issue related to the SAMSI climate program.

Chen, S., Whiteman, A., Li, A., Rapp, T., Delmelle, E., Chen, G., Brown, C.J., Robinson, P., Coffman, M.J., Janies, D., and Dulin, M., Effects of Socioeconomic and Landscape Patterns on

Mosquito Abundance in an Urban Landscape: a Machine Learning Approach. Landscape Ecology, under review.

Erhardt, R., Bell, J., Blanton, B., Nutter, F., Robinson, M., Smith, R. Nexus of Climate Data, Insurance, and Adaptive Capacity. Submitted to BAMS.

Guan, Y., Johnson, M., Katzfuss, M., Mannshardt, E., Messier, K.P., Reich, B.J. and Song, J.J. Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google Street View vehicles. In revision, Journal of the American Statistical Association.

Konomi, B. A., Hanandel, A. A., Ma, P. , and Kang, E. L., Computationally efficient nonstationary nearest neighbors Gaussian processes using data-driven techniques, requested revision by Environmetrics.

Ma, P. and Kang, E. L., Fused Gaussian process for very large spatial data, requested revision submitted to Journal of Computational and Graphical Statistics.

Ma, P. and Kang, E. L., Spatial-temporal data fusion for massive sea surface temperature data from MODIS and 2AMSR-E instruments, under requested revision for Environmetrics.

J. Maclean and E. S. Van Vleck, Projected Data Assimilation. Submitted.

Raha, S. and Ghosh, S., Heatwaves: Characterizations using Probabilistic Inference. Submitted to JASA Case studies and Applications.

Reddy, A.S., Apte, A. and Vadlamani, S., Asymptotic Properties of Linear Filter for Noise Free Dynamical System. Submitted, https://arxiv.org/abs/1901.00307

Russell, B.T., Risser, M., Smith, R.L., Kunkel, K., Investigating sea surface temperature's link to US Gulf Coast precipitation extremes with focus on Hurricane Harvey. Submitted.

Stein, M.L., Some Statistical Issues in Climate Science. Submitted to Statistical Science.

Tian, Y. and Reich, B., A Bayesian semi-parametric mixture model for bivariate extreme value analysis with application to precipitation forecasting. Submitted.

c. Papers in preparation

Apte, A., Reddy, A. and Van Vleck, E. S., Dynamics and Asymptotic Behavior of Differential Riccati Equations.

Bravo de Guenni, L., Risk evaluation to extreme meteorological events in South Carolina: a retrospective approach. To be submitted to the Journal Stochastic Environmental Research and Risk Assessment.

Carrassi, A., Guider, C., Rampal, P. and Aydogdu, A., Data Assimilation using adaptive non-conservative, moving mesh models. To be submitted to Nonlinear Processes in Geophysics.

Chang, W., Konomi, B. A., Guan, Y., Haran, M., Ice Model Calibration using Semi-continuous Spatial Data.

Chang, W., Kim, B., Guan, Y., Gong, M., and Slutsky, D., Arctic Sea Ice Model Calibration using Dirichlet process-Gaussian Mixture Model-based Feature Tracking.

Y. Chung, S. Day, K. Keegan, C. Sampson, Geometric and Topological understanding of Images: Multiparameter Persistent Homology Approach, International Conference on Computer Vision 2019, In preparation.

Cunningham, E. and Tokdar, S., A data-retaining method for tail estimation.

Dey, A., Gel, Y. R., and Lyubchich, V. Predicting precipitation-related home insurance claims: building ensembles based on climate projections.

Ebert-Uphoff, I., Mitra, A, Huang, W.K, Cooley, D.S., Chatterjee, S.B., and Chen, C., Recent Developments in Climate Networks

Erhardt, R. and Nutter, F. Climate Adaptation through Insurance Markets. In preparation for Nature Climate Change.

Gagne, D.J., H.M. Christensen, A. Subramanian, and A.H. Monahan, Machine learning for stochastic parameterization: Generative Adversarial Networks in the Lorenz '96 Model.

Guan, Y., Chang, H.H., Reich, B.J., Multivariate Spectral Downscaling for Multiple Air Pollutants.

Guan, Y. and Haran, M., Fast maximum likelihood inference for spatial generalized linear mixed models.

Guan, Y. and Reich, B.J., Dimension reduction with auxiliary data for exposure measures.

Hill, K. and Wang, Y., Resilience of the global beef trade network to climate change-related disruption.

Huang, W. K., Smith, R. L., Asher, T. G., Blanton, B. O., Luettich, R. A. Estimating r-year Storm Surge Heights: A Statistical Perspective. Planned submission to *Annals of Applied Statistics*

Huser, R. and Stein, M.L., Inference for max-stable processes based on the Vecchia approximation.

Johnson, M., Braverman, A., Zhu, Z., Sanso, B., Grenier, I., Rosenberger, J. and Reich, B., A Mathematical Formulation of the Cost Tradeoff for the Design of Distributed Statistical Analyses of Massive Data.

Johnson, M., Braverman, A. and De Domenico, M., A Network Framework for Quantifying the Cost Tradeoff in Distributed Data Analysis Systems.

Johnson, M., Reich, B.J., Gray, J. and Furman, M. Multisensor Fusion of Remotely Sensed Vegetation Indices using Space-Time Dynamic Linear Models.

Johnson, M., Caragea, P. and Bramer, L., Quantifying Dynamics of Electrochemical Processes in Sub-sampled Scanning Transmission Electron Microscopy Images using Bayesian Constrained Gaussian Mixture of Regression Models.

Joon, S., Guan, Y., Li, B. and Berrocal, V., Causal inference in the presence of an unmeasured spatial confounder.

Jones, C., Kiers, C., Chen, Y. and Maclean, J., Rate-induced tipping and data assimilation in a simplified hurricane model (provisional title).

Katzfuss, M., Berrocal, V., Brynjarsdóttir, J., Hobbs, J. and Mondal, A., Spatial retrievals for satellite remote-sensing observations.

M. Kuusela, D. Giglio, A. Mondal and M. L. Stein. Statistical Methods for Ocean Heat Content Estimation with Argo Profiling Floats. In preparation.

Lu, F., Weitzel, N. and Monahan, A. H., Joint parameter-state estimation of nonlinear stochastic energy balance models.

Ma, P., Konomi, B., Hobbs, J., Mondal, A., Song, J. J. and Kang, E. L., Statistical Emulation for High-Dimensional Functional Outputs in Large-Scale Observing System Uncertainty Experiments.

Majumder, S., Guan, Y., Reich, B.J. and Rappold, A.G., Statistical downscaling with spatial misalignment: Application to wildland fire PM2.5 emissions forecasting.

Nychka, D., Ma, P. and Kuusela, M., Local Conditional Simulations for Inference of Spatial Fields.

Reich, B.J., Guan, Y., Fourches, D., Warren, J.L. and Chang, H.H., Integrative statistical methods for exposure mixtures and health.

Sampson, C., Chung, Y.M., Lubbers, D. and Golden, K., Topological and Statistical analysis of Sea Ice Microstructure.

Sampson , C., Sudakov, I., Elsheikh, A. and Gong, M., Generation of sea ice floe geometries under ponded conditions using a General Adversarial Networks.

Sampson, C., Murphy, B., Cherkaev, E. and Golden, K., Effective Rheology and Wave Propagation in the Marginal Ice Zone.

Sass, D., Li, B. and Reich, B., Return level estimation for large spatial extremes.

Smith, R.L., Johnson, M. , Hammerling, D. and Lenssen, N., A Sparse Bayesian Factor Model Approach for Climate-Change Detection and Attribution.

Smith, R.L. and Kunkel, K.E., Extreme precipitation events in the Gulf of Mexico.

Sudakov, I. and Vakulenko. S., Species extinction at the various environmental forcing in a stochastic ecosystem model. To be submitted to Physica A and arxiv.

d. Software products

Kuusela, M. Preprocessed Argo data for the Statistical Oceanography Working Group: https://github.com/mkuusela/PreprocessedArgoData

Kuusela, M. Code for the Kuusela and Stein (2018) paper: https://github.com/mkuusela/ArgoMappingPaper

Nychka D, Hammerling D, Sain S, Lenssen N (2016). "LatticeKrig: Multiresolution Kriging Based on Markov Random Fields." doi:10.5065/D6HD7T1R (URL: http://doi.org/10.5065/D6HD7T1R), R package version 7.2, <URL: www.image.ucar.edu/LatticeKrig>. Updated version of an earlier package based on SAMSI interactions.

Nychka, D. and Hammerling, D. (2018) HPC4Stats: A framework for using R on large data analysis problems. https://github.com/dnychka/HPC4Stats. HPC4Stats was created to make parallel data analysis on a supercomputer easy to use from within R and has been applied to several examples. We see nearly linear speedup in this kind of analysis and queue wait times for even several hundred cores is not appreciable. Surprisingly the time to initiate even several hundred R worker sessions on independent cores is not appreciable. Here we see nearly linear scaling up to 1000 cores and thus can expect nearly a factor of 1000 speedup when using the HPC4Stats package on the NCAR supercomputer (Cheyenne). To our knowledge creating these parallel tools for the R data analysis community and on an NSF supercomputing facility is path breaking.

Tokdar, S. and Cunningham, E., sbde: Semiparametric Bayesian Density Estimation. R-package. In preparation.

e. Grant applications resulting from the program

Amy Braverman and Jonathan Hobbs: The Orbiting Carbon Observatory (OCO-2) project at the Jet Propulsion Laboratory (JPL) provided research subcontracts for M. Katzfuss and J. Brynjarsdóttir to work spatial retrieval methodology for estimating atmospheric carbon dioxide.

Mark Borsuk (Associate Professor of Civil and Environmental Engineering, Duke University) was Lead Principal Investigator on the submitted proposal DoD SERDP SON #: RCSON-18-L2, Detecting Shifts in Climate Classification Using Bayesian Reasoning on Extremes (DESCUBRE).

Robert Erhardt, NSF #1824394 WORKSHOP: Nexus of Climate Data, Insurance, and Adaptive Capacity, $44,190, https://nsf.gov/awardsearch/showAward?AWD_ID=1824394&HistoricalAwards=false

A. R. Gray (PI) and M. Kuusela (Institutional PI). AI for Monitoring Global Ocean Health. Grant proposal submitted to the Google.Org AI Impact Challenge. Proposed budget $1.28 million.

Christopher Jones (UNC-CH, PI) with Co-PIs: Michael Ghil (UCLA), Mickael Chekroun (UCLA), Deborah Sulsky (UNM), Hayley Shen (Clarkson), Erik Bollt (Clarkson). MURI: A

Mathematical Framework for a New Generation of Arctic Sea-Ice Models. FOA Number: N00014-18-S-F006. MURI Topic 7, Advanced Analytical and Computational Modeling of Arctic Sea Ice.

Chris Jones and Christian Sampson, ONR travel funding for Minisymposium: Wave-Ice Interactions: Nonlinearity, Paradigms, and Modelling Approaches. In, SIAM nonlinear waves and coherent structures conference June 2018, Anaheim, CA

Emily Kang (co-investigator): Likely refugia from projections of coral reef health from observationally weighted climate model ensembles. NASA NNH18ZFS001N-SLSCVC (Kalmus) 01/2019 – 09/2021

Emily Kang (PI): Collaborative Research: Inference and Uncertainty Quantification for High Dimensional Systems: Methods, Computation, and Applications. Pending, NSF DMS-CDS&E-MSS, 09/01/2019 – 08/31/2021.

M. Katzfuss has been added as a co-investigator on a JPL Research and Technology Development grant on "Coupled Atmosphere-Surface Retrievals for Visible/Shortwave Infrared Imaging Spectroscopy". PI is Vijay Natraj; Katzfuss support includes a student research assistant.

A. Lee (PI), M. Kuusela (Co-I), K. Wood (Co-I/Institutional PI), D. Giglio (Co-I/Institutional PI). Statistical Modeling of Tropical Cyclone Intensity Changes Using Observational Data from Multiple Sources. Notice of Intent submitted to the NASA AIST-18 call. Full proposal to be submitted in April 2019.

A.H. Monahan, Discovery Grant proposal to the Natural Sciences and Engineering Research Council of Canada (proposal motivated by two papers of Monahan during the program – decision to be announced in April-May 2019).

Fred S. Roberts (Rutgers U) is PI and Midge Cozzens (Rutgers U), Hans G Kaper (Georgetown U), Amy Luers (Future Earth), and Christiane Rousseau (University of Montreal, Canada) are Co-PIs on an NSF proposal submitted in response to the solicitation "Accelerating Research through International Network-to-Network Collaborations (AccelNet)" (NSF 19-501). The proposal will link the DIMACS Network (US) with Future Earth (an international network focused on sustainability). The theme of the proposal is "Resilience in the Digital Age." The proposed activities will involve three Study Groups: Urban Systems (UrbanSSG), Ecological Systems (EcoSSG) and Food Systems (FoodSSG). Coordinators of the FoodSSG are Hans Kaper and Hans Engler (Georgetown University) and Steven McGreevy (Future Earth). The FoodSSG is a direct outcome of the SAMSI Working Group on Food Systems, which has been meeting regularly since the conclusion of the 2017-18 Climate program.

Erik Van Vleck (PI) and N. Brunsell (co-PI), Computational Dynamics for Complex Land-Atmosphere Ecosystems. Submitted to NSF-DMS Mathematical Biology Program

Yimin Xiao (Michigan State) submitted a grant proposal to NSF based on his participation in the Climate Program Opening Workshop, August 21 - 25, 2017, and working groups on Extremes

and Statistical Oceanography. He is also directing a Ph.D. student whose dissertation research is on statistical inference of multivariate Gaussian random fields.

f. Presentations at major conferences

Amit Apte, Markov random field model for the Indian monsoon rainfall. Invited talk at the IUTAM workshop on "Stochastic approaches to transitions in fluid flows;" 12-14 September 2018, Ithaca, NY, USA; Invited talk at the International Workshop on "Cloud Dynamics, Micro physics, and Small-Scale Simulation;" 13-17 August 2018, Indian Institute of Tropical Meteorology, Pune, India.

Lelys Bravo de Guenni, Risk estimation to weather hazards using a Bayesian approach. Poster presentation at the NEXUS workshop in North Carolina.

Snigdhansu Chatterjee (organizer), Session on Climate networks and extremes, Joint Statistical Meeting, Denver, CO, Tuesday, July, 30, 2019. 2-4pm. (This session includes talks by Dan Cooley and Imme Ebert-Uphoff who were both active in the SAMSI program.)

Erika Cunningham, A Data-retaining Method for Tail Estimation. Presented in the topic-contributed session, The Climate Extremes Program at SAMSI, Joint Statistical Meetings, Vancouver, Aug 2018.

Yawen Guan and Brian Reich (2019). Fine-scale spatiotemporal air pollution analysis using mobile monitors on Google Street View vehicles. Environmental Defense Fund.

Yawen Guan, Spatiotemporal air pollution analysis using mobile monitors on Google Street View vehicles. International Conference on Advances in Interdisciplinary Statistics and Combinatorics (AISC 2018), Greensboro, NC, October 2018

Yawen Guan, Multivariate Spectral Downscaling for Multiple Air Pollutants. Presented at: Statistics for the Environment: Research, Practice & Policy (ENVR 2018), Asheville, NC, October 2018; The 28th Annual Conference of the International Environmetrics Society (TIES 2018), Guanajuato, Mexico, July 2018; Symposium on Data Science & Statistics (SDSS 2018), Reston, VA, May 2018; plus multiple job interview talks.

Jonathan Hobbs (13 Sep 2018). Simultaneous Retrieval of Spatial Fields of Atmospheric Carbon Dioxide from High-Resolution Satellite Data. SIAM Conference on Mathematics of Planet Earth; Philadelphia, PA.

Jonathan Hobbs. (10 Oct 2018). Incorporating Spatial Dependence in Atmospheric Carbon Dioxide Retrievals. Workshop on Inverse Problems and Uncertainty Quantification in Satellite Remote Sensing; Helsinki, Finland.

Whitney Huang, Modeling Compound Wind and Precipitation Extremes using a Large Climate Model Ensemble. Invited session on CANSSI Postdoctoral Showcase, 2019 Statistical Society of Canada (SSC) Annual Meeting, Calgary, AB, Canada May 2019

Whitney Huang, Estimating Extreme Storm Surge Levels: A Statistical Perspective. Minisymposium on Statistics of Extreme Weather and Climate Events, Society for Industrial and Applied Mathematics (SIAM) Conference on Mathematics of Planet Earth (MPE18), Philadelphia, PA Sept. 2018.(Co-organized with Richard Smith.)

Maggie Johnson, A Notional Framework for A Theory of Data Analysis Systems, SIAM Conference on Mathematics of Planet Earth (MPE18), Sept. 13-15, 2018, Philadelphia, PA

Maggie Johnson, A Sparse Bayesian Factor Model Approach for Climate-Change Detection and Attribution, 2018 ENVR Workshop - Statistics for the Environment: Research, Practice and Policy, Asheville, NC. October 11-13, 2018.

Chris Jones, Plenary Lecture, Dynamics Days, Denver, CO, Jan 2018

Chris Jones, Plenary Lecture, Symposium on Dynamical Systems and Related Fields, Waterloo, Canada, May 2018

Chris Jones, Rudy Horne Applied Mathematics Lecture, CAARMS, Princeton, NJ, July 2018

Chris Jones, Plenary Lecture, CRITICS Workshop on Tipping Points, Cork, Ireland, Aug 2018

Chris Jones, Invited Lecture, Midwest Dynamical Systems Seminar, Minnesapolis, MN, Nov 2018

Chris Jones, Invited Lecture, ACMS30: Mathematical Modeling and Computational Methods for Multiscale Problems in Science \& Engineering, Tucson, AZ, Nov 2018

Chris Jones, Focus on Math Lecture at BYU, Provo, UT Jan 2019

Emily Kang: Invited talk at the ASA-ENVR Workshop at Asheville 2018

Emily Kang: Topic-contributed talk at JSM 2018

Emily Kang: Minisymposium talk at SIAM  Conference on Mathematics of Planet Earth (MPE18)

Matthias Katzfuss (Jul 2019). Spatial retrievals of carbon dioxide from the OCO-2 satellite. Joint Statistical Meetings; Denver, CO.

Ken Kunkel, Historical Perspective on Hurricane Harvey Rainfall. Oral presentation, Annual Meeting of the American Meteorological Society, Phoenix, AZ, January 9, 2019.

Mikael Kuusela. Statistics for Ocean Heat Content Estimation with Argo Profiling Floats, contributed talk, 2019 Joint Statistical Meetings, Denver, CO, USA, July-August, 2019 upcoming)

Mikael Kuusela. Locally stationary interpolation of Argo float data for improved estimates of ocean climate, opening invited poster, 2019 Joint Statistical Meetings, Denver, CO, USA, July-August, 2019 (upcoming)

Mikael Kuusela. Statistics for Ocean Heat Content Estimation with Argo Profiling Floats, invited talk, 14th International Meeting on Statistical Climatology, Toulouse, France, June 2019 (upcoming)

Mikael Kuusela. Statistics for Ocean Heat Content Estimation with Argo Profiling Floats, invited talk, ICSA Applied Statistics Symposium, Raleigh, NC, USA, June 2019 (upcoming)

Mikael Kuusela. Statistics for Ocean Heat Content Estimation with Argo Profiling Floats, contributed talk, 2019 IMS/ASA Spring Research Conference, Blacksburg, VA, USA, May 2019 (upcoming)

Mikael Kuusela. Locally stationary spatio-temporal interpolation of Argo profiling float data, invited talk, Workshop on Data Analytics for Climate and Earth, Arrowhead, CA, USA, March 2019 (upcoming)

Mikael Kuusela. Locally stationary spatio-temporal interpolation of Argo profiling float data, invited talk, Statistics and Data Science Workshop, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, November 12, 2018

Mikael Kuusela. Locally stationary spatio-temporal interpolation of Argo profiling float data, contributed talk, SIAM Conference on Mathematics of Planet Earth, MPE18, Philadelphia, PA, USA, September 13, 2018

Mikael Kuusela. Recent progress on statistical analysis of oceanographic data from Argo profiling floats, invited talk, 28th Annual Conference of the International Environmetrics Society, TIES 2018, Guanajuato, Mexico, July 16, 2018

Pulong Ma: Contributed talk at 2018 SIAM Conference on Uncertainty Quantification

John Maclean, Feature Data Assimilation in the Unstable Subspace. Talk at SIAM UQ (Anaheim, CA)

Suman Majumder, Yawen Guan, Brian J. Reich, Ana G. Rappold, Statistical downscaling with spatial misalignment: Application to wildland fire PM2.5 emissions forecasting. Poster presentation at 2018 IISA conference in Gainesville, Florida; also planned presentations at 2019 IMS/ASA Spring Research Conference at Blacksburg, Virginia (May 22- 24, 2019) and in Joint Statistical Meetings 2019 at Denver, Colorado (July 27 - August 1, 2019).

A.H. Monahan, Stochastic parameterization of subgrid-scale velocity enhancement of sea surface fluxes. Joint Statistical Meeting in Vancouver in 2018, the Stochastic Weather Generators conference in Boulder in 2018, the Statistics and Data Science Workshop at KAUST in 2018, and the International Union of Geodesy and Geophysics meeting in Montreal in 2019.

A.H. Monahan, Machine learning for stochastic parameterization: Generative Adversarial Networks in the Lorenz '96 Model. The American Geophysical Union Fall Meeting in

Washington DC in 2018 and the American Meteorological Society Annual Meeting in Phoenix in 2019. It will be presented at the European Geosciences Union annual meeting in Vienna in 2019 and the International Union of Geodesy and Geophysics meeting in Montreal in 2019.

A.H. Monahan, Joint parameter-state estimation of nonlinear stochastic energy balance models. American Geophysical Union Fall Meeting in Washington DC in 2018 and the Joint Mathematical Meeting in Baltimore in 2019.

A.H. Monahan, Temporal filtering enhances the skewness of sea surface winds. BIRS Workshop on Nonlinear and Stochastic Problems in Atmospheric and Oceanic Prediction in 2017 and the Canadian Meteorological and Oceanographic Society Congress in 2018 in Halifax.

Doug Nychka, Statistical methods for nonstationary spatial data. July 2018, Invited talk at International Environmetrics Society, Guanajuato, MX and August 2018, Joint Statistical Meetings, Vancouver, BC.

Doug Nychka, Data Science and Climate. March 2018, Invited talk at Alan Turing Institute, London, UK.

Doug Nychka, Large and non-stationary spatial fields: Quantifying uncertainty in climate models. Invited talk in: October 2017, North Carolina State University; October 2017, Argonne National Laboratory; March 2018, Hadley Center, Exeter, UK; May 2018, University of Minnesota; May 2018, Symposium on Data Science and Statistics, Reston, VA

Sohini Rana: Talk, 2018 AISC conference, Greensboro; Poster presentation, 2018 ENVR Workshop - Statistics for the Environment: Research, Practice and Policy, Asheville; Poster presentation, 2018 Georgia Statistics Day.

Brook Russell, Characterizing Precipitation Extremes in the US Gulf Coast through the use of a Multivariate Spatial Hierarchical Model. Presented in a Topic Contributed Session on "The Climate Extremes Program at SAMSI," Joint Statistical Meetings, Vancouver, BC (Aug 2018).

Brook Russell, Investigating the link between Gulf of Mexico sea surface temperatures and US Gulf Coast precipitation extremes with focus on Hurricane Harvey. Invited talk at "Statistics for the Environment: Research, Practice & Policy," ENVR 2018 Workshop, Asheville, NC (Oct 2018).

Christian Sampson (2018) Speaker, International Conference on Advances in Interdisciplinary Statistics and Combinatorics, University of North Carolina at Greensboro Talk Title: "Stochastic Surfaces for Sea Ice Modeling"

Christian Sampson (2018) Speaker, AMS Sectional Meeting AMS Special Session on Nonlinear Water Waves and Related Problems, University of Deleware Talk Title: "Bounds on The Effective Viscoelasticity of an Ice Covered Ocean"

Christian Sampson (2018) Speaker, SIAM Conference on Mathematics of Planet Earth (MPE18), Philadelphia, PA Talk Title: "Bounds on The E?ective Viscoelasticity of an Ice Covered Ocean"

Christian Sampson (2018) Speaker, SIAM Conference on Nonlinear Waves and Coherent Structures, Special Session: Waveice Interactions: Nonlinearity, Paradigms, and Modelling Approaches Talk Title: “Bounds on The E?ective Viscoelasticity of an Ice Covered Ocean”

Christian Sampson (2018) Speaker, Prediction and Data Assimilation for Nonlocal Di?usions, University of Edinburgh Talk Title: “Toward a Metric for Large Scale Sea Ice Model Fracture Forecast Skill”

Christian Sampson (2018) Speaker, CLIM Transition Workshop, Statistical and Applied Mathematical Sciences Institute (SAMSI) Talk Title: ”Understanding Sea Ice Data for Data Assimilation”

Christian Sampson (2018) Speaker, 42nd SIAM Southeastern Atlantic Sectional Conference, University of North Carolina at Chapel Hill Talk Title: ”Mathematics for Sea Ice and Climate”

Christian Sampson (2017) Speaker, Mathematics of sea ice phenomena: Multi-scale modelling of ice characteristics and behaviour, Isaac Newton Institute for Mathematical Sciences Talk title: “Effective Rheology and Wave Propagation in the Marginal Ice Zone”

Yuan Tian and Brian Reich, Bayesian bivariate extreme value analysis with application in environmental statistics. Poster presentation in 2018 JSM at Vancouver.

Richard Smith, Attribution of extreme precipitation storms in the Gulf of Mexico. Presented at the annual meeting of the International Detection and Attribution Group, Berkeley, CA, March 2018.

Richard Smith, Influence of climate change on extreme weather events. Buehler-Martin Plenary Talk a the conference Statistics and Data Science for Earth Systems, Institute for Research in Statistics and its Applications, University of Minnesota, May 3-5 2018

Richard Smith, Risk of Extreme Weather Events in a Changing Climate. Invited talk at the Bernard Harris Memorial Syposium, Risk in the 21st Century, organized by the Section on Risk Analysis by the American Statistical Association, NOrth Carolina State University, May 10, 2018

Richard Smith, Influence of climate change on extreme weather events, presented at the conference of the International Society for Bayesian Analysis (ISBA) meeting in Edinburgh, Scotland, July 25-29, 2018 (Smith was also organizer of an invited paper session on “Bayesian Methods for Detection and Attribution of Climate Change”)

Richard Smith, Discussant on two sessions at the Joint Statistical Meetings, Vancouver, July-August 2018. Invited paper session on "Intergovernmental Panel on Climate Change (IPCC) Reports: How Statisticians Can Get Involved" and topic contributed paper session on "The Climate Extremes Program at SAMSI"

Richard Smith, An overview of detection and attribution methods for climate extremes. Invited talk at the conference Statistics for the Environment: Research, Practice & Policy, Asheville, NC, October 11-13 2018. The conference was organized by the Section on Statistics for the

Environment of the American Statistical Association (ENVR); Smith was also a co-organizer of the conference and chair of a panel discussions session.

Richard Smith, Hurricane Harvey: Attributions and Future Projections of Damage. Invited talk at the American Association for the Advancement of Science (AAAS) annual meeting (Washington, DC, February 16 2019), session on "Extreme Event Attribution in the Context of Climate Change"

Erik Van Vleck, Bistable Traveling Waves Under Discretization: BDF Methods, Moving Meshes, and Applications. Presented at International Congress on Difference Equations and Applications (ICDEA), Dresden, Germany, May 2018; BIRS Workshop on Adaptive Numerical Methods for Partial Differential Equations with Applications, Banff, Canada, June 2018; SDS2018, Capitolo, Italy, June 2018.

Erik Van Vleck, Time Dependent Stability: Computation and Applications. NUMDIFF-15, Martin Luther University Halle-Wittenberg, Germany, September 2018.

Erik Van Vleck, Projected Data Assimilation. IUTAM: Stochastic Approaches to Understanding Transitions in Fluid Flows, Cornell University, Ithaca, NY, September 2018.

g. Other outputs

Robert Erhardt (Wake Forest University) was chief organizer and Richard Smith was a co-organizer of the conference "The Nexus of Climate Data, Insurance, and Adaptive Capacity" in Asheville, NC, November 8-9 2018

Kaitlin Hill: As the academic director of the 2018 IMA/ MathCEP Math Modeling Camp for high school students, I incorporated food systems as one of the four topics. Students chose to modeled the spread of mad cow disease through the global beef trade network, investigating historical strategies used to mitigate the outbreak vs. alternate strategies. After a week of working on the project, the students presented their model results to the rest of the camp. Link to CLIM: exposing students at the post-Calculus I level to higher-level math and issues relevant to the Math of Planet Earth part of the CLIM theme.

Jonathan Hobbs: OCO-2 project is sponsoring a breakout meeting on uncertainty quantification, March 11-12, 2019, in Pasadena, CA. Most of the spatial retrieval subgroup participants will attend.

Whitney Huang: Gave an introductory talk on "Estimating changes in temperature extremes from millennial scale climate simulations using generalized extreme value (GEV) distributions" to the Pacific Climate Seminar Series at the University of Victoria, Victoria, BC, Canada. Sept.2018

Whitney Huang: Served as the organizer of the Community Climate Science Seminars (CCSS), a seminar series designed to unite researchers and students who study climate related science on Vancouver Island. Oct. 2018 to present.

Whitney Huang and Taylor Asher (UNC-Chapel Hill) are co-leaders of a SAMSI working group on Storm Surge Hazard and Risk. This is part of the Model Uncertainty Mathematics and Statistics (MUMS) program, ongoing for 2018-19.

Maggie Johnson: Through the SAMSI program, Maggie was able to obtain a second-year postdoc at the Jet Propulsion Laboratory under Dr. Amy Braverman continuing the Theory of Data Systems (ToDAS) research. The Jet Propulsion Laboratory (JPL) has the intention to hire Maggie in a permanent position by the completion of her postdoc year. The ToDAS subgroup continues to have regular group meetings every other week to continue the research started in the SAMSI program year. The ToDAS research has led to a collaboration with Manlio De Domenico -- an expert in modeling complex systems using networks and Head of the Complex Mulilayer Networks Lab at the FBK Center for Information Technology in Trento, Italy. Maggie and Amy will visit Manlio in Italy at the end of February, 2019 to further develop the research and relationship.

Mikael Kuusela joined the Department of Statistics and Data Science at Carnegie Mellon University as a Tenure-Track Assistant Professor in August 2018

Mikael Kuusela received the Ian McNeill Presentation Award at the 28th Annual Conference of the International Environmetrics Society, TIES 2018, on July 20, 2018

Mikael Kuusela: The SAMSI Statistical Oceanography Working Group is still meeting actively following a bi-weekly schedule. In addition to projects started during the SAMSI program, two CMU PhD students (Beomjo Park and Addison Hu) have now started to work on topics related to these activities.

Mikael Kuusela: The Statistical Oceanography work was one of the key building blocks for the new CMU working group on Statistical Methods for the Physical Sciences: http://stat.cmu.edu/stamps/

Mikael Kuusela: NASA Jet Propulsion Laboratory is planning to provide summer support for Mikael Kuusela in summer 2019 to support Uncertainty Quantification activities for the OCO-2 program. This is a direct result of Mikael Kuusela's SAMSI supported visit to JPL on April 20, 2018. A CMU PhD student (Pratik Patil) has started to work on these topics under Mikael Kuusela's supervision.

Doug Nychka: The SAMSI program facilitated the interaction Mikael Kuusela and me and developed a new and efficient way of generating a conditional simulation of a spatial field. This completed Mikael’s the ARGO analysis without which he could not have quantified the uncertainty in the predictions.

Christian Sampson (SAMSI postdoc, now at UNC) and Chris Jones (UNC) are leading a project is ongoing on observation error in data assimilation. It involves an undergrad at UNC (Siyang Jing) whom they met through the SAMSI course. The project has been heavily influenced by what they learned from the SAMSI Working Group on Remote Sensing.

Richard Smith and Whitney Huang (SAMSI postdoc) were co-organizers of the Minisymposium "Statistics of Extreme Weather and Climate Events" at the SIAM Mathematics of Planet Earch conference, Philadelphia, September 13-15, 2018

Richard Smith was co-organizer of the one-day meeting, “Data Sciences for Climate and Environment,” organized at the Alan Turing Institute, London, as part of the SAMSI Climate Program; March 26, 2018

Several program participants, including Richard Smith, Doug Nychka and Dorit Hammerling, are members of the organizing committee for the International Meeting on Statistical Climatology, Toulouse, France, June 2019.

Erik Van Vleck and Chris Jones taught a summer school on data assimilation in 2018, to be repeated in summer 2019. This was funded by NSF and organized through AIM. The lectures were based on the course they developed at SAMSI.

## 13. Appendices

**SAMSI GRADUATE CLASS: STATISTICS FOR CLIMATE RESEARCH: FALL 2017**

Coordinated by Brian Reich (NCSU), Richard Smith (UNC-CH) and Doug Nychka (NCAR, now at Colorado School of Mines)

| **Date** | **Class** | **Speaker** |
|---|---|---|
| **29-Aug** | ***Introduction: Statistics and Computing Background*** | **Brian Reich, NCSU** |
| **5-Sep** | ***Detection and Attribution*** | **Richard Smith, UNC** |
| **12-Sep** | ***Guest lecture: Analysis of Climate Model data*** | **Dr. Murali Haran, Penn State University** |
| **19-Sep** | ***Climate Informatics*** | **Brian Reich** |
| **26-Sep** | ***Guest lecture: Data Fusion*** | **Dr. Veronica Berrocal, University of Michigan** |
| **3-Oct** | ***Estimating Curves and Surfaces*** | **Doug Nychka, NCAR** |
| **10-Oct** | ***Spatial Data: Models and Analysis*** | **Doug Nychka** |
| **17-Oct** | ***Nonstationary Covariance Modeling*** | **Bo Li, University of Illinois** |
| **24-Oct** | ***Geostats for Large Data Sets*** | **Brian Reich** |
| **31-Oct** | ***Quantile Regression*** | **Surya Tokdar, Duke University** |
| **7-Nov** | ***Extremes*** | **Richard Smith** |
| **14-Nov** | ***Climate Trends*** | **Richard Smith** |
| **21-Nov** | ***NO CLASS – Thanksgiving*** | |
| **28-Nov** | ***Trend Detection Using Time Series Methods*** | **Richard Smith** |
| **5-Dec** | ***Final Presentations*** | **Brian Reich and Richard Smith** |

**SAMSI GRADUATE CLASS: DATA ASSIMILATION IN DYNAMICAL SYSTEMS: SPRING 2018**

Coordinated by Chris Jones (UNC-CH), Amit Apte (ICTS) and Erik Van Vleck (University of Kansas)

| | | |
|---|---|---|
| **Date** | **Class** | **Speaker** |
| **16-Jan** | **Intro to DA and Basic Examples** | **Amit Apte** |
| **23-Jan** | **Dynamical Systems: Stability, Chaos, Predictability** | **Chris Jones** |
| **30-Jan** | **Computations of Dynamical Systems and Lab** | **Amit Apte** |
| **6-Feb** | **Lyapunov Exponents and Lab** | **Erik Van Vleck** |
| **13-Feb** | **Variational Methods** | **Chris Jones** |
| **20-Feb** | **Filtering Theory Introduction** | **Amit Apte** |
| **27-Feb** | **Discrete State Space Models** | **Erik Van Vleck** |
| **6-Mar** | **Ensemble Kalman Filter** | **Erik Van Vleck** |
| **13-Mar** | *****NO CLASS – Spring Break***** | |
| **20-Mar** | **Particle Filtering** | **Chris Jones** |
| **27-Mar** | **Computations and filtering (EnKF, PF)** | **Erik Van Vleck** |
| **3-Apr** | **Lagrangian DA (LaDA) and Parameter Estimation (PE)** | **Chris Jones** |
| **10-Apr** | **Computations in LaDA and PE** | **Amit Apte** |
| **Apr 17 & 24** | **Project presentations** | |

samsi
NSF·Duke·NCSU·UNC

# Climate Opening Workshop
## August 21-25, 2017
### SCHEDULE

**Monday, August 21, 2017**
Hamner Conference Center

| | |
|---|---|
| 8:30 | Registration |
| 8:40-9:00 | Introductions and Welcome |
| 9:00-10:00 | Overview lecture 1: **Vipin Kumar** (University of Minnesota): *"Big Data in Climate: Opportunities and Challenges for Machine Learning"* |
| 10:00-10:30 | Break |
| 10:30-11:30 | Overview Lecture 2: **Kristie Ebi** (University of Washington): *"Climate and Health"* |
| 11:30-12:30 | Overview Lecture 3: **Leonard Smith** (London School of Economics): *"Science, Simulation and Insight: Developing Confidence when Extrapolating Complicated Complex Systems* |
| 12:30-1:30 | Lunch |
| 1:30-2:30 | Overview Lecture 4: **Michael Wehner** (Lawrence Berkeley National Lab): *"Computational and Mathematical Challenges in Climate Modeling"* |
| 2:30-3:30 | Break |
| 3:30-4:30 | Overview Lecture 5: **Kenneth Kunkel** (NCSU): Climate Datasets *"Understanding the Physical Causes of Observed Trends in Extreme Precipitation: How Can Statistics Help?"* |
| 4:30-5:00 | General Discussion |
| 5:00 | Adjourn, Shuttle to Hotel |

**Tuesday, August 22, 2017**
Hamner Conference Center

| | |
|---|---|
| 8:45 | Registration |
| 8:45-9:00 | Introductions and Announcements<br>*Research Session 1: Remote Sensing* |
| 9:00-9:45 | **Noel Cressie** (University of Wollongong)<br>*"A Bird's-Eye View of Statistics for Remote Sensing Data"* |
| 9:45-10:30 | **Dan Crichton** (NASA/JPL)<br>*"Software Architecture Considerations in the Analysis of Highly Distributed Data and Computational Analysis"* |
| 10:30-11:00 | Break |
| 11:00-11:45 | **Matthias Katzfuss** (Texas A&M University)<br>*"A General Framework for Vecchia Approximations of Gaussian Processes"* |
| 11:45-12:05 | Discussant **Amy Braverman** (Jet Propulsion Laboratory) |
| 12:05-12:15 | General Discussion |
| 12:15-1:30 | Lunch |
| | *Research Session 2: Ice Dynamics* |
| 1:30-2:15 | **Deborah Sulsky** (University of New Mexico)<br>*"Modeling Arctic Sea Ice"* |
| 2:15-3:00 | **Murali Haran** (Pennsylvania State University)<br>*"Some Statistical Challenges in Studying the West Antarctic Ice Sheet"* |
| 3:00-3:30 | Break |
| 3:30-4:15 | **Alberto Carrassi** (Nansen Environmental and Remote Sensing Center)<br>*"Issues in Ensemble Prediction and Data Assimilation Using a Lagrangian Model of Sea-Ice"* |
| 4:15-4:35 | Discussant: **Christopher Jones** (University of North Carolina, Chapel Hill) |
| 4:35-4:45 | General Discussion |
| 5:00-7:00 | Poster Session and Reception |
| 7:00 | Shuttle to Hotel |

**Wednesday, August 23, 2017**
Hamner Conference Center

| | |
|---|---|
| 8:45 | Registration |
| 8:45-9:00 | Introductions and Announcements |
| | Research Session 3: Climate Informatics |
| 9:00-9:45 | **Prabhat** (Lawrence Berkeley National Lab)<br>*"Deep Learning for Extreme Weather Detection"* |
| 9:45-10:30 | **Imme Ebert-Uphoff** (Colorado State University)<br>*"Methods for Causality Analysis in Climate Science"* |
| 10:30-11:00 | Break |
| 11:00-11:45 | **Dorit Hammerling** (National Center for Atmospheric Research)<br>*"Compression and Conditional Emulation of Climate Model Output"* |
| 11:45-12:05 | Discussant **Doug Nychka** (National Center for Atmospheric Research) |
| 12:05-12:15 | General Discussion |
| 12:15-1:00 | Lunch |
| 1:00-2:00 | Panel Discussion 1: Climate Extremes |
| | Panelists: **Daniel Cooley** (Colorado State University); **Michael Wehner** (Lawrence Berkeley National Lab); **Michael Stein** (University of Chicago); **Adam Monahan** (University of Victoria) |
| | Research Session 4: Climate and Health |
| 2:00-2:45 | **Jason West** (University of North Carolina Chapel Hill)<br>*"The Effects of Climate Change on Human Health through Changes in Air Quality"* |
| 2:45-3:15 | Break |
| 3:15-4:00 | **Howard Chang** (Emory)<br>*"Projecting Health Impacts of Climate Change: Embracing an Uncertain Future"* |
| 4:00-4:45 | **Montse Fuentes** (Virginia Commonwealth University)<br>*"A Multivariate Dynamic Spatial Factor Model for Speciated Pollutants and Adverse Birth Outcomes"* |
| 4:45-5:05 | Discussant: **Veronica Berrocal** (University of Michigan) |
| 5:05-5:15 | General Discussion |
| 5:15 | Shuttle to Hotel |

**Thursday, August 24, 2017**
Hamner Conference Center

| | |
|---|---|
| 8:45 | Registration |
| 8:45-9:00 | Introductions and Announcements |
| 9:00-1:00 | Research Session 5: Stochastic Parametrizations |
| 9:00-9:45 | **Andrew Gettelman** (National Center for Atmospheric Research)<br>*"A Cloud of Numbers: Representing Physical Processes in the Earth System with Mathematics"* |
| 9:45-10:30 | **Charles Jackson** (University of Texas)<br>*"Ice Sheet Contribution to Sea Level Rise"* |
| 10:30-11:00 | Break |
| 11:00-11:45 | **Michal Branicki**, University of Edinburgh<br>*"An Information-Theoretic Framework for Improving Multi-Model Predictions & Data Assimilation Techniques"* |
| 11:45-12:30 | **Chris Bretherton** (University of Washington)<br>*"Developing Stochastic Parameterizations of Subgrid Variability of Clouds and Turbulence using High-Resolution Simulations"* |
| 12:30-12:50 | Discussant: **Adam Monahan** (University of Victoria) |
| 12:50-1:00 | General Discussion |
| 1:00-2:00 | Lunch |
| 2:00-3:00 | Panel Discussion 2: Climate and Risk |
| | Panelists: **Brian Blanton** (Renaissance Computing Institute); **Wei Lei** (University of North Carolina at Chapel Hill); **Slava Lyubchich** (University of Maryland); **Hans Kaper** (Georgetown University) |
| 3:00-5:30 | Working Group Formation Session |
| 5:30 | Shuttle to Hotel |

**Friday, August 25, 2017**
Hamner Conference Center

| | |
|---|---|
| 8:45-9:00 | Introductions and Announcements |
| 9:00-12:00 | Initial Meetings of Working Groups |
| 12:00-1:00 | Adjourn / Lunch / Shuttle to RDU Airport |
| *1:00-4:30*<br>*Dogwood Room* | ***Pre-Registration Required:***<br>*Short Course on Computational Methods (**Doug Nychka** and **Dorit Hammerling**, National Center for Atmospheric Research)* |
| *4:30* | *Buses to Airport and/or Hotel* |

RENCI Meeting: Monday Oct 9 2:00-5:00

**Mathematical Problems in Radiation and Convection**

2:00-3:15

Kerry Emanuel (MIT) Radiative-Convective Instability

Erik Van Vleck (KU) Model Hierarchy for Deep Convection (with Dave Mechem and Tommy Gebhardt)

Discussion

3:15-3:30 Break

3:30-5:00

Short talks:

Mary Silber (Bifurcation in pattern-forming systems)

Rachel Kuske (Stochastic effects and bifurcations)

Alejandro Aceves (Stochastic effects in time-varying systems)

Chris Jones (Rate-induced tipping)

Adam Monahan (Column models of low atmosphere)

Discussion

Workshop Report:

# Remote Sensing, Uncertainty Quantification, and a Theory of Data Systems

February 12–14, 2018
Cahill Center, California Institute of Technology

Amy Braverman
(Jet Propulsion Laboratory, California Institute of Technology)

and

Jessica Matthews
(Cooperative Institute for Climate and Satellites–North Carolina)

# 1 Summary

The workshop on Remote Sensing, Uncertainty Quantification, and a Theory of Data Systems was held at the Cahill Center, California Institute of Technology, on February 12, 13, and 14, 2018. It was sponsored financially by the Statistical and Applied Mathematical Sciences Institute (SAMSI), the Jet Propulsion Laboratory's (JPL) Science Visitor's and Colloquium Program (SVCP), and logistically by Caltech's Center for Data Driven Discovery (CD3) and its JPL partner, The Center for Data Science and Technology (CDST). The purpose of the workshop was to invite statisticians, applied mathematicians, computer scientists, data system architects, experts in remote sensing technology, and Climate and Earth System scientists to review, discuss, and plan research on issues related to large-scale, efficient analysis of distributed data using spatial statistical methods.

Our motivation in organizing this event was to catalyze interchange among experts on the fast-emerging problem of analysis of distributed data. As part of SAMSI's 2017–2018 Program on Mathematical and Statistical Methods for Climate and the Earth System, a Working Group on Remote Sensing was established to address statistical and mathematical research problems in the analysis of remote sensing data. The Working Group has five subgroups: 1) Spatial Retrieval Methodology (the so-called "Spatial-X" subgroup); 2) Spatial Analysis for Hyperspectral Data (the so-called "Spatial-Y" subgroup); 3) Emulators for Complex Forward Models; 4) Optimization for Remote Sensing Retrievals; and 5) Theory of Data Systems (ToDS). The ToDS subgroup spent the first half of this academic year formulating a framework in which to consider the joint problem of a) optimizing statistical methods for environments where data are distributed and too large to move to a central location, and b) the design of data system infrastructures within which to implement those statistical methods.

To fix ideas, the Workshop focused on spatial statistical methods. To date there are many new spatial statistical methods designed with massive data sets in mind, in the literature. However, very few have been implemented for remote sensing data, and none have been implemented in operational settings like those used by NASA and NOAA. A major impediment to their use in these cases is that the data are not only massive, but are stored in different physical locations. These data must be brought together in some way in order to estimate spatial covariance functions, but moving data to a central location for analysis is tedious at best and impossible at worst.

Some remote data reduction is almost certainly necessary, but how much? What are the consequences for inference? The fundamental issue underlying these questions is how to navigate the trade-space between costs and uncertainty in the estimates or inferences that are ultimately produced. Thus, the Workshop was organized around the following themes:

- The computational–statistical trade-off: theory and application
- Data systems and their architectures, especially at NASA and NOAA
- Approximations in statistical inference
- Multilayer networks as a tool for optimization and visualization

- Spatial statistics with distributed data
- Case study problems with uncertainty requirements and cost limitations

This Workshop was divided into six main sessions 1) context from NOAA and NASA (Jay Morris and Mike Little); 2) foundational research areas with experts in contributing fields (Venkat Chandrasekaran, Richard Smith, Dan Crichton, Maggie Johnson, and Manlio De Domenico); 3) members of the SAMSI Working Group on Remote Sensing who discuss relationships between distributed analysis and their sub-groups activities (Jessica Matthews, Emily Kang, and Jon Hobbs); 4) methodology and tools for distributed spatial statistics (Matthias Katzfuss, Raj Guhaniyogi, Zhengyuan Zhu, Dorit Hammerling, and Luca Cinquini); 5) case study research problems that may serve as potential testbeds (Veronica Berrocal, Hui Su, Carmen Boening, and Vineet Yadav); and 6) posters by other participants, especially graduate students and post-docs. The agenda, participant list, and full abstracts from all talks and posters is provided at the end of the main body of this report. Presentations are available at `https://www.samsi.info/programs-and-activities/other-workshops-and-post-doc-seminars/remote-sensing-uncertainty-quantification-and-a-theory-of-data-systems-workshop-february-12-14-2018/`.

# 2 Synopsis of Oral Sessions

## 2.1 Context: Distributed Access and Analysis at NASA and NOAA

This was the opening session, following welcomes and logistics. Mike Little (NASA Earth Science Technology Office, Advanced Information Systems Technology (AIST) Program) and Jay Morris (NOAA National Centers for Environmental Information, Mission Science Network) gave overviews of NASA's and NOAA's Earth remote sensing observation, data collection, and data processing systems, respectively. Morris went on to describe NOAA's Big Data Partnership that uses the "Cloud" to make data and analysis tools available to decision makers. He followed with an example of a research project called the Graph Database Proof-of-Concept. This is an experimental framework using graph database technology to improve granule-level (file-level) search and discovery. The work is being carried out at CICS (the Cooperative Institute for Climate and Satellites in Asheville, NC.

Little introduced the concept of an analytic center as "An environment for conducting a Science investigation" that "enables the confluence of resources for that investigation" and is "tailored to the individual study." He gave several examples of projects funded under the AIST-16 solicitation for creating supporting technologies, and discussed other interests and needs of the AIST Program, with an eye towards the next solicitation (AIST-18) that is due out in November of this year. Little closed with a discussion of considerations relevant to the theory of data systems that mostly had to do with heterogeneities among data sets and the need for uncertainty quantification.

## 2.2 Foundations

The Foundations session was intended to bring together experts working in key (existing) areas which need to be combined in order to develop a theory of data systems. The areas so identified were 1) the "computational-statistical trade-off" made well-known by Michael Jordan and others; 2) approximate likelihoods and sufficient statistics; 3) systems and software architecture principles that guide the development of data systems at NASA and NOAA, and other creators and distributors of scientific data; and 4) frameworks and tools for understanding the structure of complex systems.

The first speaker was Dr. Venkat Chandrasekaran (Michael Jordan's former post-doc, currently a professor at Caltech) who described the statistical-computational trade-off as a question of when it pays to use a simpler, less computationally intensive analysis algorithm that can process more data faster than a more accurate and complex, but slower and more costly algorithm. In the first case, the algorithm is less accurate, but using more data can drive down error. The second case is the reverse. Chandrasekaran demonstrated with an example in which a complex optimization problem is solved approximately by a computationally fast convex program.

The second speaker was Dr. Richard Smith (UNC and former Director of SAMSI). Smith reported on research he and former student Petrutza Caragea (Iowa State) did on blocking methods for spatial statistics. Blocking methods use approximations to the likelihood function for a spatial data set formed by factoring the likelihood into terms that are then assumed to be independent. He described traditional blocking and his and Caragea's idea: to use blocking structures based on both within-block and between-block likelihood approximations. He analyzed the computational savings and asymptotic efficiencies of various choices through a set of examples, and commented on the utility of the approach in a distributed context.

Dan Crichton (JPL) followed with an introduction to concepts of system and software architecture as they pertain to the theory of data systems problem. The key point of Crichton's talk was that data systems are networks of computers that can be viewed at different levels of abstraction depending on what characteristics are to be emphasized or of are interest. These may be called "views" of the system. There is a "hardware view" that emphasizes the physical properties of the computers and data transfer mechanisms. "Software views" emphasize the organization of software modules, and so on. Similarly, different users of these systems may want more or less abstraction in certain parts of the system. For instance, a corporate customer contracting for computer usage does not want to know about system-level details of data transfer; just that data moves as if along a network "edge" that has a certain speed or capacity. Crichton described what these systems look like to a NASA data system user, and illustrated that just because two arbitrary computers are connected in principle to one another via the internet, does not mean they are connected in a useful sense for the theory of data systems. Special connections made possible through an underlying software infrastructure will be necessary to achieve the kind of transparent interconnectivity required. He closed with a review of work done previously at JPL on achieving better understanding of the benefits of analyzing candidate system topologies for new data systems.

The session's fourth speaker was SAMSI post-doc (and future JPL post-doc) Maggie Johnson. Dr. Johnson presented a framework for analyzing the trade-off between uncertainty in an analysis result, and the cost of performing that analysis, developed by the SAMSI Theory of Data Systems Working Group (a subsidiary of the Remote Sensing Working Group in SAMSI's 2017-2018 Program on Mathematical and Statistical Methods for Climate and the Earth System, which is led by Braverman and Matthews). Maggie showed a very simple example using the case of a simple mean computed from data on five remote servers, with computational resources only available on a single user node. In the example, it was assumed that correlation existed among data both within the same server, and between servers. The example showed how, even in the simplest of cases, the structure of the data has an influence on optimal data system design.

The final speaker in this session was Manlio De Domenico from the Bruno Kessler Foundation. Dr. De Domenico is an expert on multilayer networks: a generalization of ordinary networks in which links may exist between elements both within and between the constituent "layers". A layer is an ordinary network that expresses relationships relative to a certain "view" or aspect of the problem at hand. De Domenico gave a tutorial talk on how multilayer networks have been used in other domains including ecology, communications, and transportation and gave a brief overview of the tensor-based mathematics that generalizes simpler, ordinary network concepts.

## 2.3 Allied Problems: Optimization, Emulation, and Retrievals

In this session, leaders of the Remote Sensing Working Group subgroups reviewed the goals and progress in their respective areas. Since the entire workshop was devoted to the ToDS problem, that subgroup did not report in this session. Optimization, emulation, and retrievals are all problems that, at some level, require bringing data together from distributed sources. Speakers were asked to comment on how this connection might be made in the future.

Dr. Jessica Matthews (NOAA/CICS) reported on the activities of the Optimization subgroup. She reviewed the role of optimization methods in remote sensing retrieval algorithms, and a number of specific topics the group approached through journal paper reviews: optimization in standard optimal estimation theory applied to microwave retrievals of temperature, water vapor, and surface emissivity and of sea ice concentration; optimization in hyperspectral unmixing; parallel and interacting stochastic approximation annealing algorithms for global optimization; and optimization for retrievals from the High Resolution Infrared Radiometer Sounder (HIRS) instrument aboard NOAA's polar orbiting satellite series using neural networks. Dr. Matthews closed with a list of areas for future research, including satellite intercalibration, and some ideas about the intersection between ToDS and optimization.

Dr. Jon Hobbs (JPL) co-led the spatial retrieval subgroup along with Dr. Matthias Katzfuss (Texas A&M). Hobbs reported on the efforts of the two sub-subgroups that comprise this working group: so-called "Spatial-X" and "Spatial-Y". The Spatial-X group is tackling the problem of performing optimal estimation retrievals when the a priori used in the retrieval is a distribution on the entire spatial field, not on a single footprint alone as is currently the case for most

missions. It is reasonable to believe that there is substantial spatial correlation in the true field of atmospheric state variables and in the radiances that remote sensing instruments observe. The group is researching ways of incorporating that information via the a priori distribution. The Spatial-Y group is concentrating on the spatial dependence structure of the radiances themselves; in particular, how this might be exploited in the spectral unmixing problem. Spatial-X is using OCO-2 CO2 concentration as its test case, and Spatial-Y is using SMOS soil moisture as its test case.

Dr. Emily Kang (University of Cincinnati) described the activities of the Emulation subgroup in her talk, "Statistical Emulation with Dimension Reduction for Complex Physical Forward Models". This group focussed on building a low-dimensional emulator for the OCO-2 forward model. Such an emulator would produce estimates of the radiance vectors obtained when a state vector is input, more quickly and efficiently but less accurately than a full-physics algorithm. The group is investigating the use of Gaussian Process emulators for this purpose, and the near-term focus is dimension reduction. For high-dimensional input state vectors, the group is exploring the use of active subspace. For the high-dimensional output radiances, the group is studying functional principal component analysis.

## 2.4 Distributed Analysis Methods and Technologies

The morning session on Tuesday was devoted to show-casing several state-of-the-art methods and technologies for performing analysis on distributed data. The first three talks were intended as examples of statistical methodologies that are representative of the kinds of innovations that are necessary to bring modern spatial statistical methods to massive, distributed data. The last two talks covered considerations related to high-performance computing environments and a description of perhaps the most successful federated data delivery and analysis system in wide use today: the Earth System Grid Federation.

Dr. Matthias Katzfuss (Texas A & M University) presented a talk on the use of multiresolution approximations of spatial covariance functions for "big" spatial data sets. He reviewed a number of recent approaches for decomposing spatial covariance functions using basis function approaches. He then introduced a new, data adaptive approach called the MRA (Multi-resolution Approximation) for both stationary and non-stationary models. The spatial field is successively split into higher-resolution sub-domains, and the behavior of the underlying geophysical process is represented by an weighted sum of basis function values over all resolutions. The method has some similarities to wavelet models, but implies a valid Gaussian process which is then suitable for further use in a probabilistic modeling framework, especially uncertainty quantification. The talk finished with a discussion of computational complexity considerations, and other advantages of the method.

Professor Rajarshi Guhaniyogi (UC Santa Cruz) discussed his recently published work on distributed kriging for distributed data. He motivated the problem by introducing the usual (centralized) approach to kriging and its underlying Gaussian process model. He reviewed the literature on

inference for "big" spatial data, and especially on low-rank approximations. Then, Dr. Guhaniyogi identified the needs for the next generation of spatial statistical models and methods: scalability, avoiding storage of all the data, divide and conquer, parallelism, and theoretical justification. Building on these ideas, he introduced DISK (DIStributed Kriging) as a method to obtain separate posterior distributions from different parts of the data. These are aggregated after-the-fact to form a single "best" posterior distribution. This was followed by examples using both simulated and real (sea-surface temperature) data.

Dr. Zhengyuan Zhu (Iowa State University) gave a talk on asynchronous optimization, concentrating on asynchronous stochastic gradient descent. After describing the algorithm, he looked at some of the assumptions that are necessary to ensure convergence and statistical properties of time-to-convergence. The ideas were illustrated with several examples, such as computation of maximum likelihood estimates of the parameters of a multivariate normal covariance matrix.

After a break, Dr. Dorit Hammerling of NCAR gave an overview of special considerations required for computations in high-performance computing (HPC) environments. She discussed why one might want to work in an HPC environment: not just because large amounts of data are to be processed, but because many *tasks* are to be carried out. Dr. Hammerling described NCAR's HPC capabilities, and the system architecture that makes it more than merely having lots of parallelism. She also reviewed special tools for the R programming language that are written for HPC. Finally, she provided results from a set of benchmarking experiments that quantify the strengths of weaknesses of various choices in the architecture.

The last talk in this session was presented by Dr. Luca Cinquini (JPL), who is the lead architect for the Earth System Grid Federation (ESGF). He began with an overview of the ESGF: its purpose and relationship to the Intergovernmental Panel on Climate Change (IPCC), and the Coupled Model Intercomparison Project (CMIP6). He then described how search and access to climate model simulations works. The main part of Dr. Cinquini's talk concerned the analytical capabilities of the ESGF: analysis of data sets on the servers where they reside. These analyses are triggered by users with calls over http. Simple analytical tools (e.g., graphs of maps and time series) are currently available. So-called "server-side computation" is planned to be available within the next year or so.

## 2.5 Case Study Problems

This session was comprised of four talks that focused on science and application problems where it is typical for information to be brought together from multiple sources. While distributed analysis methods were not specifically used in these studies, the talks were intended to highlight how distributed analysis capabilities could have made these efforts more productive and/or more efficient.

Dr. Veronica Berrocal (University of Michigan) discussed environmental epidemiological studies, which seek to establish an association between a health outcome and an environmental

exposure. Health data is found from a variety of sources: National Center for Health Statistics, local and state health departments, hospital databases, Department of Health and Human Services, individual cohort studies, etc. Environmental data also is found from a variety of sources (e.g. NASA, NOAA, EPA). Given the disparate data locations, these type of studies may be suitable for applying the theory of data system framework for solving problems. Berrocal described some challenges faced by analysts prior to the analysis itself: incongruent spatial and temporal resolutions of data, different approaches to measuring uncertainty, and heterogeneous data formats. Unresolved statistical issues include: developing spatially-resolved metrics of exposure with associated uncertainty quantification, handling multiple forms of environmental exposures simultaneously (e.g. temperature along with humidity), and surmounting the computational challenges posed in the big-data era.

Next, Dr. Hui Su (JPL) presented multiple examples of how climate model simulations may be evaluated or constrained with relevant satellite data. Given that climate studies require large amounts of data, she stressed the importance of developing innovative tools to facilitate data processing, especially in a distributed setup. A key message was the concept of "emergent constraints" which are physically-motivated empirical relationships between the current climate and long-term climate projections. Here, targeting improvement in the model representation of individual physical processes may yield conflicting results in other areas. Su suggests that optimal estimates of multiple criteria may improve the overall uncertainty in climate model projections.

Dr. Carmen Boening (JPL) spoke about the challenges faced in sea level science. The primary scientific goal is to improve the quality of sea level rise projections by incorporating new physics knowledge gleaned from modern remote sensing data, coupled with better statistical representations of uncertainty sources in both observations and model outputs. Echoing the other speakers concerns, Boening highlighted common analysis challenges such as the requirement to combine large heterogeneous data from different sources in different formats and with differing temporal characteristics. A key challenge is dealing with varying, or absent, approaches to uncertainty quantification.

The final talk of the workshop was given by Dr. Vineet Yadav (JPL) who provided an overview on inverse modeling of atmospheric trace gas fluxes. He described a tiered observation strategy for trace gases including in situ, aircraft, and satellite measurements. Yadav stressed the importance of knowing both your problem *and* your computational architecture. That is, knowing if the application is computationally bound by I/O, processing speed, memory availability, or something else. He reiterated that collaborations between scientists and IT specialists are necessary, since it is not possible to be an expert on the complexities of both domains.

# 3 Conclusions

The main intent of this workshop was to bring together relevant threads of research, and identify where new work is required to integrate them and fill gaps. This information is crucial to

setting an agenda for future work in the theory of data systems. The talks motivated dynamic discussion periods where it became apparent that we stand at the very beginning of this research. A cohesive framework tying together all perspectives – as represented by statisticians, applied mathematicians, computer scientists, data system architects, remote sensing technology experts, and Climate and Earth System scientists – was not apparent going in. However, as a result of the Workshop, we reached the following conclusions.

- Theory of Data Systems is a "meta-problem" in that such a theory must tie together principles from a number of different domains into a common, holistic framework. Statistics, applied math, and system design all involve optimization problems and these must be pulled together. That requires being able to express costs in a common unit: computational costs, infrastructure costs, communications costs are all important. At the same time, data-driven science requires a formal approach to inference and (at minimum) a coherent metric for measuring uncertainty as a cost. The most obvious starting point is to use the variance of an estimator.
- Currently, data system design is done in isolation from the question of uncertainty in the resulting science data products.
- Science analysis using distributed data products is still in the mode of downloading all relevant data to a single server before beginning analysis.
- The idea of an "analytic center" seems to suggest centralizing both computation and data storage.
- The computational-statistical trade-off and the use of likelihood approximations are two approaches to the same problem. In the former, one contemplates using more data and weaker algorithms to estimate an unknown quantity, while in the latter, one contemplates approximating complex likelihoods with more tractable expressions. The two concepts are linked by the relationship between computational complexity of the algorithm/likelihood optimization and the amount of data required to achieve a maximum allowable uncertainty/estimator variance.
- Multilayer networks provide a potential analytical and visualization framework in which to solve ToDS problems. Each "layer" can be identified with different "view" of the problem. For instance, hardware relationships seem to lend themselves to description by network graphs. So do relationships among random variables (see literature on graphical models). It's less clear how one would represent algorithms, software, and science use cases. That said, if it could be done, a suite of visualization and tensor-based methods exist for examining the properties of multilayer networks.
- None of the attendees were experts in measuring network communication costs.

In view of these conclusions, we make the following recommendations for the research agenda.

1. A holistic mathematical framework needs to be fleshed out. The framework should facilitate more quantitative understanding of how component problems relate to one another, and it should permit focussing in on substantive issues within those component problems without leaving the framework.

2. The framework should give rise to a companion visualization environment. It is critical to exploit visualization in understanding relationships among components since our intuition is immature in that area.

3. Undertake research to determine whether, or to what extent, existing work in multilayer networks is relevant. It seems likely that extensions of these methods will be required. For example, the current formalism does not explicitly represent costs or probabilistic relationships.

4. Investigate state-of-the-art optimization strategies for complex empirical problems without closed-form expressions for cost functions. This applies to both statistical and other optimization problems.

5. Investigate and develop the connections between the computational-statistical trade-off and approximations for likelihoods.

6. Examine the underlying principles of software and system architecture, and find the optimization problems at which these principles are directed. Can this relationship be made to fit into the holistic mathematical framework/multilayer network formalism?

7. Collaborate with IT network experts to define appropriate metrics and methods for measuring communication costs. Undoubtedly, these already exist.

8. Find some science problems, involving spatial statistical analysis of distributed data, to serve as case study examples. These case study problems should involve science questions for which conclusions from the analyses are desired at a variety of uncertainty levels. They should also involve non-trivial application of spatial statistical inference (e.g., are spatial patterns changing over time).



This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

# Agenda, Participants, and Abstracts

| | | |
|---|---|---|
| **Monday, February 12** | | |
| 8:00 – 8:30 | Registration and continental breakfast | |
| 8:30 – 8:40 | Opening remarks | Amy Braverman and Jessica Matthews |
| 8:40 – 8:50 | Welcome on behalf of SAMSI | David Banks and Richard Smith |
| 8:50 – 9:00 | Welcome on behalf of CD3 and CDST | George Djorgovski and Dan Crichton |
| 9:00 – 9:20 | Distributed access and analysis: NASA | Mike Little (NASA) |
| 9:20 – 9:40 | Distributed access and analysis: NOAA | Jay Morris (NOAA) |
| **Foundations** | (Chair: Jenny Brynjarsdottir) | |
| 9:40 – 10:10 | The statistical–computational trade-off | Venkat Chandrasekaran (Caltech) |
| 10:10 – 10:40 | Approximate likelihoods | Richard Smith (UNC/SAMSI) |
| 10:40 – 11:00 | Break | |
| 11:00 – 11:30 | Data system architectures | Dan Crichton (JPL) |
| 11:30 – 12:00 | The ToDS problem | Maggie Johnson (NCSU) |
| 12:00 – 12:30 | Multilayer networks | Manlio De Domenico (FBK) |
| 12:30 – 2:00 | Lunch | |
| **Allied problems: optimization, emulation, and retrievals** | (Chair: Sandy Burden) | |
| 2:00 – 2:30 | Optimization working group | Jessica Matthews (NOAA) |
| 2:30 – 3:00 | Emulators working group | Emily Kang (U of Cincinnati) |
| 3:00 – 3:30 | Spatial retrieval working group | Jon Hobbs (JPL) |
| 3:30 – 4:00 | Break | |
| 4:00 – 5:00 | Discussion | Discussants: Sanso (UCSC), Chatterjee (U of MN), and Banks (Duke) |
| 5:00 – 7:00 | Poster session and reception | |
| **Tuesday, February 13** | | |
| 8:00 – 8:30 | Continental breakfast | |
| **Distributed analysis methods and technologies** | (Chair: Jim Rosenberger) | |
| 8:30 – 9:00 | Distributed spatial statistics | Matthias Katzfuss (Texas A & M) |
| 9:00 – 9:30 | Bayesian large-scale kriging | Rajarshi Guhaniyogi (UCSC) |
| 9:30 – 10:00 | Asynchronous optimization | Zhengyuan Zhu (Iowa State) |
| 10:00 – 10:30 | Break | |
| 10:30 – 11:00 | HPC for distributed analysis | Dorit Hammerling (NCAR) |
| 11:00 – 11:30 | The ESGF | Luca Cinquini (JPL) |
| 11:30 – 12:00 | Discussion | |
| 12:00 – 1:30 | Lunch | |
| **Case study problems** | (Chair: Mike Turmon) | |
| 1:30 – 2:00 | Climate and health | Veronica Berrocal (U of Michigan) |
| 2:00 – 2:30 | Climate science | Hui Su (JPL) |
| 2:30 – 3:00 | Sea-ice modeling and analysis | Carmen Boening (JPL) |
| 3:00 – 3:30 | Carbon cycle science | Vineet Yadav (JPL) |
| 3:30 – 4:00 | Break | |
| 4:00 – 5:00 | Discussion: agenda for ToDS research | |
| 6:00 – 8:00 | Workshop Dinner | |
| **Wednesday, February 14** | | |
| **Discussion: priorities** | | |
| 9:00 – 11:00 | Discussion and wrap-up | |

## Participant List

| *Name* | *Affiliation* | *Domain* | *email* |
|---|---|---|---|
| David Banks* | Duke | Statistics | banks@stat.duke.edu |
| Veronica Berrocal* | U of Michigan | Statistics | berrocal@umich.edu |
| Carmen Boening | JPL | Science/Remote sensing | Carmen.Boeing@jpl.nasa.gov |
| Amy Braverman | JPL | Statistics/Remote sensing | Amy.Braverman@jpl.nasa.gov |
| Jenny Brynjarsdottir* | Case-Western | Statistics | jxb628@case.edu |
| Sandy Burden† | U of Wollongong | Statistics | sburden@uow.edu.au |
| Venkat Chandrasekaran | Caltech | Statistics/CS | venkatc@caltech.edu |
| Snigdhansu Chatterjee* | U of MN | Statistics | chatterjee@stat.umn.edu |
| Luca Cinquini | JPL | Data systems | Luca.Cinquini@jpl.nasa.gov |
| Dan Crichton | JPL | Data systems | Daniel.Crichton@jpl.nasa.gov |
| Manlio De Domenico† | FBK | Mulitlayer networks | mdedomenico@fbk.eu |
| George Djorgovski | Caltech | Astronomy/CD3 | george@astro.caltech.edu |
| Rich Doyle | JPL | Computational science | richard.j.doyle@jpl.nasa.gov |
| Isabelle Grenier* | UCSC | Statistics/Applied Math (GS) | igrenier@ucsc.edu |
| Matthew Graham | Caltech | Astroinformatics | mjg@cd3.caltech.edu |
| Rajarshi Guhaniyogi† | UCSC | Statistics/Applied Math | rguhaniy@ucsc.edu |
| Michael Gunson | JPL | Remote sensing/Climate | Michael.Gunson@jpl.nasa.gov |
| Dorit Hammerling† | NCAR | HPC/Statistics | dorith@ucar.edu |
| Jon Hobbs | JPL | Statistics/Remote sensing | Jonathan.M.Hobbs@jpl.nasa.gov |
| Maggie Johnson* | NCSU/SAMSI | Statistics (PD) | mcjohn22@ncsu.edu |
| Emily Kang* | U of Cinncinnati | Statistics | Kangel@ucmail.uc.edu |
| Matthias Katzfuss* | Texas A&M | Statistics | katzfuss@stat.tamu.edu |
| Alex Konomi | U of Cincinnati | Statistics | KONOMIBR@ucmail.uc.edu |
| Otto Laminpaa† | FMI | Statistics/Remote sensing (GS) | Otto.Lamminpaa@fmi.fi |
| Kyo Lee | JPL | Data science/Remote sensing | Huikyo.Lee@jpl.nasa.gov |
| Mike Little | NASA/ESTO | Remote sensing | m.m.little@nasa.gov |
| Pulong Ma* | U of Cincinnati | Statistics (GS) | mapn@mail.uc.edu |
| Ashish Mahabal | Caltech | Astroinformatics | aam@astro.caltech.edu |
| Jessica Matthews* | NOAA/CICS | Applied Math/Remote sensing | jessica.matthews@noaa.gov |
| Charlie McElroy | Caltech | Data science (PD) | cmcelroy@caltech.edu |
| Anirban Mondal* | Case-Western | Statistics | anirbanstat@gmail.com |
| Jay Morris† | NOAA | Data systems | jay.morris@noaa.gov |
| Hai Nguyen | JPL | Statistics/Remote sensing | Hai.Nguyen@jpl.nasa.gov |
| Houman Owhadi | Caltech | Mathematical statistics/UQ | owhadi@caltech.edu |
| Derek Posselt | JPL | Remote sensing/Meteorology | Derek.Posselt@jpl.nasa.gov |
| Gavino Puggioni* | U of RI | CS/Statistics | gpuggioni@uri.edu |
| Jim Rosenberger† | PSU/NISS | Statistics | JLR@psu.edu |
| Bruno Sanso* | UCSC | Statistics/Applied Math | bruno@soe.ucsc.edu |
| Florian Schafer | Caltech | Mathematical Statistics/UQ (GS) | Florian.Schaefer@caltech.edu |

* = funded by SAMSI; † = funded by JPL; GS = graduate student; PD = post-doc

## Participant List (cont'd)

| *Name* | *Affiliation* | *Domain* | *email* |
|---|---|---|---|
| Sarah Sernaker* | U of MN | Statistics (GS) | serna022@umn.edu |
| Richard Smith | SAMSI/UNC | Statistics | rls@email.unc.edu |
| Massimo Stella† | FBK | Multilayer networks (PD) | massimo.stella@inbox.com |
| Andrew Stuart | Caltech | Math/Statistics/UQ | astuart@caltech.edu |
| Hui Su | JPL | Climate science | Hui.Su@jpl.nasa.gov |
| Joao Teixeira | JPL | Climate science | Joao.Teixeira@jpl.nasa.gov |
| Joaquim Teixeira | JPL | Data science/Remote sensing | Joaquim.P.Teixeira@jpl.nasa.gov |
| Michael Turmon | JPL | Statistics/Remote sensing | Michael.Turmon@jpl.nasa.gov |
| Paul von Allmen | JPL | Computational science/Remote sensing | Paul.A.Vonallmen@jpl.nasa.gov |
| Vineet Yadav | JPL | Science/Statistics | Vineet.Yadav@jpl.nasa.gov |
| Zhengyuan Zhu* | Iowa State | Statistics | zhuz@iastate.edu |

* = sponsored by SAMSI; † = sponsored by JPL; GS = graduate student; PD = post-doc

# Oral Abstracts

## Environmental exposure in environmental epidemiological studies: modeling approaches and challenges

Veronica Berrocal (University of Michigan)

### Abstract

A typical problem in environmental epidemiological studies concerns environmental exposure assessment. In this talk, we will discuss challenges to environmental exposure assessment and we will showcase and discuss statistical methods that have been developed to obtain estimates of environmental exposure (e.g. air pollution, temperature). Further we will discuss whether and how uncertainty in the environmental exposure has been and can be incorporated in health analyses.

## Data and Model Analysis and Uncertainty Quantification for Sea Level Science

Carmen Boening (Jet Propulsion Laboratory, California Institute of Technology)

### Abstract

Sea level change is a complex scientific problem involving many Earth system components. Not only are processes in the ocean important to understand for evaluating past, present, and future of sea level change, but sea level is also driven by external sources such as melting ice sheets, land hydrology, large scale changes in precipitation and evaporation and many more. NASA satellites and Earth system models provide a vast source of understanding these physical processes. However, analysis and uncertainty quantification of data and models are often challenging because of the size of the data, a large variety of storage locations to pull from, different data formats, and disparate error sources. In this talk, particular challenges of sea level science with a focus on water mass transport data from GRACE, sea level prediction uncertainties from ice and ocean models, and enabling analyses through web-based tools ([http://sealevel.nasa.gov)]http://sealevel.nasa.gov) will be discussed.

## Computational and statistical trade-offs in data analysis

Venkat Chandrasekaran (Caltech)

### Abstract

The rapid growth in the size and scope of datasets in science and technology has created a need for novel foundational perspectives on data analysis that blend computer science and statistics. That classical perspectives from these fields are not adequate to address emerging challenges with massive datasets is apparent from their sharply divergent nature at an elementary level ? in computer science, the growth of the number of data points is a source of "complexity" that must be tamed via algorithms or hardware, whereas in statistics, the growth of the number of data points is a source of "simplicity" in that inferences are generally stronger and asymptotic results can be invoked. In classical statistics, one usually considers the increase in inferential accuracy as the number of data points grows (with little formal consideration of computational complexity), while in classical numerical computation, one typically analyzes the improvement in accuracy as more computational resources such as space or time are employed (with the size of a dataset not formally viewed as a resource). In this talk we describe some of our research efforts towards addressing the question of trading off the amount of data and the amount of computation required to achieve a desired inferential accuracy. This is joint work with Michael Jordan, Yong Sheng Soh, and Quentin Berthet.

## The Earth System Grid Federation as a testbed for global, distributed data analytics

Luca Cinquini (Jet Propulsion Laboratory, California Institute of Technology)

### Abstract

The Earth System Grid Federation (ESGF) is a large international collaboration that operates a global infrastructure for management and access of Earth System data. Some of the most valuable data collections served by ESGF include the output of global climate models used for the IPCC reports on climate change (CMIP3, CMIP5 and the upcoming CMIP6), regional climate model output (CORDEX), and observational data from several American and European agencies (Obs4MIPs). This talk will present a brief introduction to ESGF, describe the data access and analysis methods currently available or planned for the future, and conclude with some ideas on how this infrastructure could be used as a testbed for executing distributed analytics on a global scale.

## An introduction to systems and software architecture considerations for scaling data analysis

Dan Crichton (Jet Propulsion Laboratory, California Institute of Technology)

### Abstract

Architectural decisions in designing data and computation intensive systems can have a major impact on the ability of these systems to perform statistical and other complex calculations efficiently. The storage, processing, tools, and associated databases coupled with the networking and compute infrastructure make some kinds of computations easier, and other harder. This talk will provide an introduction to software and data systems components that are important for understanding how these choices impact data analysis uncertainties and costs, and thus for developing system and software designs best suited to statistical analyses.

## Multilayer modeling and analysis of complex (systems) data

Manlio De Domenico (Fondazione Bruno Kessler)

### Abstract

Recently, we have discovered that a new level of complexity characterizes a variety of natural and artificial systems, where units interact, simultaneously, in distinct ways. For instance, this is the case of multimodal transportation systems (e.g., metro, bus and train networks) or of biological molecules, whose interactions might be of different type (e.g. physical, chemical, genetic) or functionality (e.g., regulatory, inhibitory, etc.). The unprecedented newfound wealth of multivariate data allows to categorize system's interdependency by defining distinct "layers", each one encoding a different network representation of the system. The result is a multilayer network model. Analyzing data from different domains we will show that neglecting or disregarding multivariate information might lead to poor results. Conversely, multilayer models provide a suitable framework for complex data analytics, including the challenging theory of data systems.

## DISK: a divide and conquer Bayesian approach to large scale kriging

Rajarshi Guhaniyogi (University of California, Santa Cruz)

### Abstract

Flexible hierarchical Bayesian modeling of massive data is challenging due to poorly scaling computations in large sample size settings. This talk is motivated by spatial process models for analyzing geostatistical data, which typically entail computations that become prohibitive as the number of spatial locations becomes large. We propose a three-step divide-and-conquer strategy within the Bayesian paradigm to achieve massive scalability for any spatial process model. We partition the data into a large number of subsets, apply a readily available Bayesian spatial process model on every subset in parallel, and optimally combine the posterior distributions estimated across all the subsets into a pseudo-posterior distribution that conditions on the entire data. The combined pseudo posterior distribution is used for predicting the responses at arbitrary locations and for performing posterior inference on the model parameters and the residual spatial surface. We call this approach "Distributed Kriging" (DISK). It offers significant advantages in applications where the entire data are or can be stored on multiple machines. Under the standard theoretical setup, we show that if the number of subsets is not too large, then the Bayes risk of estimating the true residual spatial surface using the DISK posterior distribution decays to zero at a nearly optimal rate. While DISK is a general approach to distributed nonparametric regression, we focus on its applications in spatial statistics and demonstrate its empirical performance using a stationary full-rank and a nonstationary low-rank model based on Gaussian process (GP) prior. A variety of simulations and a geostatistical analysis of the Pacific Ocean sea surface temperature data validate our theoretical results.

## High performance computing and spatial statistics: an overview of recent work at NCAR

Dorit Hammerling (NCAR)

### Abstract

While much of the recent literature in spatial statistics has evolved around addressing the big data issue, practical implementations of these methods on high performance computing systems for truly large data are still rare. We discuss our explorations in this area at the National Center for Atmospheric Research for a range of applications, which can benefit from large scale computing infrastructure. These applications include extreme value analysis, approximate spatial methods, spatial localization methods and statistically-based data compression and are implemented in different programming languages. We will focus on timing results and practical considerations, such as speed vs. memory trade-offs, limits of scaling and ease of use. This is joint work with Joseph Guinness, Marcin Jurek, Matthias Katzfuss, Daniel Milroy, Douglas Nychka, Vinay Ramakrishnaiah, Yun Joon Soon and Brian Vanderwende.

## Incorporating Spatial Dependence in Atmospheric Carbon Dioxide Retrievals from High-Resolution Satellite Data

Jonathan Hobbs (Jet Propulsion Laboratory, California Institute of Technology)

### Abstract

Earth-orbiting satellites that monitor atmospheric greenhouse gases, such as NASA?s Orbiting Carbon Observatory-2 (OCO-2), collect measurements of reflected sunlight at fine spatial and temporal resolution. The atmospheric constituent of interest, such as carbon dioxide (CO2) concentration, is estimated from these observations using a retrieval algorithm. A particular class of retrievals can be represented as hierarchical statistical models, and inference for the atmospheric state is achieved through the posterior distribution given the observed satellite radiances. The spatial retrieval subgroup will present an investigation of multi-pixel retrievals that combine nearby satellite observations for joint inference on a spatial field of atmospheric states. We illustrate the impact of true and assumed spatial dependence for different atmospheric variables and discuss needs and capabilities for a distributed approach to this spatial retrieval.

## A notional framework for a theory of data systems

Maggie Johnson (North Carolina State University)

### Abstract

Modern, large scale data analysis typically involves the use of massive data stored on different computers that do not share the same file system. Computing complex statistical quantities, such as those that characterize spatial or temporal statistical dependence, requires information that crosses the boundaries imposed by this partitioning of the data. To leverage the information in these distributed data sets, analysts are faced with a trade-off between various costs (e.g., computational, transmission, and even the cost building an appropriate data system infrastructure) and inferential uncertainties (bias, variance, etc.) in the estimates produced by the analysis. In this talk we introduce a framework for quantifying this trade-off by optimizing over both statistical and data system design aspects of the problem. We illustrate with a simple example, and discuss how it may be extended to more complex settings. This is joint work with Amy Braverman (JPL) and Brian Reich (NCSU).

## Statistical Emulation with Dimension Reduction for Complex Physical Forward Models

Emily Kang (University of Cincinnati)

### Abstract

The retrieval algorithms in remote sensing generally involve complex physical forward models that are nonlinear and computationally expensive to evaluate. Statistical emulation provides an alternative with cheap computation and can be used to calibrate model parameters and to improve computational efficiency of the retrieval algorithms. We introduce a framework of combining dimension reduction of input and output spaces and Gaussian process emulation technique. The functional principal component analysis (FPCA) is chosen to reduce to the output space of thousands of dimensions by orders of magnitude. In addition, instead of making restrictive assumptions regarding the correlation structure of the high-dimensional input space, we identity and exploit the most important directions of this space and thus construct a Gaussian process emulator with feasible computation. We will present preliminary results obtained from applying our method to OCO-2 data, and discuss how our framework can be generalized in distributed systems. This is joint work with Jon Hobbs, Alex Konomi, Pulong Ma, and Anirban Mondal, and Joon Jin Song.

## Distributed access and analysis: NASA

MIke Little (NASA Earth Science Technology Office)

### Abstract

Data systems in NASA?s Earth Science Division are primarily focused on providing stewardship of the products of remote sensing and are manifested as Digital Active Archive Systems. Each Instrument Team has a related Science Team which defines the algorithms and monitors the processing of the output of the instruments to produce the related data products and in an format and standards compliance of them. These teams are influenced also by the research and applied sciences components of the programs, but the primary focus is on proving the ongoing validity of the products. Across the distributed system, every product is different. However, this is not conducive to analytics. NASA?s Advanced Information Systems Technology (AIST) program is developing an entirely new approach to creating Analytic Centers which focus on the scientific investigation and harmonize the data, computing resources and tools to enable and to accelerate scientific discovery. Stay tuned to find out how. A major element, in today?s science interests, is the comparison of multi-dimensional datasets; this warrants considerable experimentation in trying to understand how to do so meaningfully and quantitatively; asked another way, "What do you mean by similar?" Uncertainty quantification has evolved considerably in the arenas of data reduction and full physics models; however, the emerging demand for machine learning and other artificial intelligence techniques has failed to keep uncertainty quantification and error propagation in mind and there is considerable work to be done.

## Optimization methods in remote sensing

Jessica Matthews (NOAA CICS)

### Abstract

Statistical estimation and inference for large data sets require computationally efficient optimization methods. Remote sensing retrievals are, in fact, estimates of the underlying true state, and their optimization routines must necessarily make compromises in order to keep up with large data volumes. A sub-group of the Remote Sensing Working Group of the SAMSI Program on Mathematical and Statistical Methods for Climate and the Earth System is investigating how optimization in Bayesian-inspired retrievals and off-line statistical methods could be made more computationally efficient. We will report on discussions held to-date and describe how progress in the theory of data systems research can positively impact optimization methodologies.

## Satellites and Stovepipes

Jay Morris (NOAA)

### Abstract

NOAA does an excellent job of generating an disseminating data to meet the primary mission of Preservation of Life and Property. There is an unrealized opportunity to exploit the data for research and profit. Much of the data is hidden deep in archives with community specific portals for access. Modern technologies allow new methods to expose more data to wider audiences in order to stimulate innovation and discovery. NOAA is currently experimenting with cloud technologies through the big data partnership by making high value data sets such as GOES East available on the cloud through cloud provider partners. Specifically: 1. To understand and predict changes in climate, weather, oceans and coasts; 2. To share that knowledge and information with others; and 3. To conserve and manage coastal and marine ecosystems and resources. There is an unrealized opportunity to exploit NOAA?s vast data holdings for research and profit. Much of the data is hidden deep in archives with community specific portals for access. Modern technologies allow new methods to expose more data to wider audiences in order to stimulate innovation and discovery. NOAA is currently experimenting with cloud technologies through the big data partnership by making high value data sets such as GOES East available on the cloud through the partners.

## Blocking methods for spatial statistics and potential applications to distributed data

Richard Smith (University of North Carolina, Chapel Hill)

### Abstract

When spatial data are distributed across multiple servers, there is an obvious difficulty with computing the likelihood function without combining all the data onto one server. Therefore, it would be of interest to compute estimates of the spatial parameters based on decompositions of the spatial field into blocks, each block corresponding to one server. Two methods suggest themselves, a "between blocks" approach in which each block is reduced to a single observation (or a low-dimensional summary) to facilitate calculation of a likelihood across blocks, or a "within blocks" approach in which the likelihood is calculated for each block and then combined into an overall likelihood for the full process. In fact, I argue that a hybrid approach that combines both ideas is best. Theoretical calculations are provided for the statistical efficiency of each approach. In conclusion, I will present some thoughts for optimal sampling designs with distributed data. This is joint work with Petrutza Caragea of Iowa State University.

## Evaluating and Constraining Climate Model Simulations Using Satellite Data

Hui Su (Jet Propulsion Laboratory, California Institute of Technology)

### Abstract

Climate projections rely on general circulation models that parameterize many physical processes that cannot be resolved by finite-sized grids and contain large uncertainties. Therefore, evaluations of the performance of models in simulating present-day climate are necessary to ensure the accuracy of the projections of future climate. Reanalysis datasets and satellite observations are routinely used for model evaluations. Furthermore, a number of metrics have been proposed to serve as ?emergent constraints? on future climate projections based on the correlations of present-day model simulations and future projections. Large ensemble members of model simulations are needed to minimize the effects of internal variabilities and extract robust signals driven by forced climate change. These climate science studies involve large amounts of climate model simulations and observational datasets. Access to and analysis of the climate model simulations and observational data often encounter difficulties in data transfer and reorganization. The increasing resolutions of climate models make the data processing even more challenging. My presentation will review some of the recent studies in evaluating and constraining climate model simulations using satellite data and seek innovative ideas to facilitate such climate studies to be more efficient and accurate.

## An Overview of the Computational Process for Generating Covariance Matrices for Atmospheric Inverse Modeling of Trace Gas Fluxes

Vineet Yadav (Jet Propulsion Laboratory, California Institute of Technology)

### Abstract

Trace gas batch inverse problems are often formulated in a Bayesian framework that require minimization of an objective function that takes as an input atmospheric measurements of trace gas concentrations, prior estimates of fluxes, and a transport operator that describes the influence of the sources of fluxes on measurements. As part of minimization, batch inverse problems require computation of covariance matrices that describes the error in measurements and prior fluxes. Most of the computational/data bottlenecks in these inverse problems occur in estimating the transport operator that require processing of terabytes of output generated from a Weather model. Typically, this output is stored on tape storage system that needs to copied or moved into an intermediary storage system for computing the transport operator and finally the covariance matrices that are used in inverse problems. This operation of bringing data to the algorithm is an inefficient and time-delaying way to solve these problems and therefore necessitates development of methods that can work on partitioned observations and transport operator and compute covariance matrices and inverse estimates of fluxes at locations of data storage.

## Optimization for Distributed Data Systems: An Overview and Some Theoretical Results

Zhengyuan Zhu (Iowa State University)

### Abstract

The asynchronous parallel algorithms are developed to solve massive optimization problems in a distributed data system, which can be run in parallel on multiple nodes with little or no synchronization. Recently they have been successfully implemented to solve a range of difficult problems in practice. However, the existing theories are mostly based on fairly restrictive assumptions on the delays, and can not explain the convergence and speedup properties of such algorithms. In this talk we will give an overview on distributed optimization, and discuss some new theoretical results on the convergence of asynchronous parallel stochastic gradient algorithm with unbounded delays. Simulated and real data will be used to demonstrate the practical implication of these theoretical results.

# Poster Abstracts

## Functional ANOVA comparison of $CO_2$ Flux Predictions

Sandy Burden (University of Wollongong)

### Abstract

There are presently at least nine different flux-inversion (FI) models that produce spatially detailed CO2 flux field predictions based on XCO2 retrievals obtained from the OCO-2 Mission. This ensemble of predictions is a valuable resource for understanding FI models and for investigating and reducing prediction uncertainty. However, summarizing and evaluating the ensemble is not straightforward. FI models are frequently based on the same, or similar, sources of information, and hence their output may not be independent. For inference it is crucial to account for this dependence to avoid underestimating prediction uncertainty. This poster demonstrates the use of Functional ANOVA for comparing predicted spatial fields from multiple FI models, taking into consideration shared assumptions, parameters and/or data. Since the predictions are modeled as observed realizations from an underlying smooth random field with a common basis, the approach also reduces the amount of data required for analysis and facilitates comparison of multiple, potentially distributed, spatio-temporal fields.

## Optimal Estimation versus MCMC for CO2 retrievals

Jenny Brynjarsdottir (Case-Western Reserve University), Jonathan Hobbs, Amy Braverman, and Lukas Mandrake (Jet Propulsion Laboratory, California Institute of Technology)

### Abstract

The Orbiting Carbon Observatory-2 (OCO-2) collects infrared spectra from which atmospheric properties are retrieved. OCO-2 operational data processing uses Optimal Estimation (OE), a state-of-the-art approach to inference of atmospheric properties from satellite measurements. One of the main advantages of the OE approach is computational efficiency, but it only characterizes the first two moments of the posterior distribution of interest. Here we obtain samples from the posterior using a Markov Chain Monte Carlo (MCMC) algorithm, and compare this empirical estimate of the true posterior to the OE results. We focus on 600 simulated soundings that represent the variability of physical conditions encountered by OCO-2 between November 2014 and January 2016.

We treat the two retrieval methods as ensemble and density probabilistic forecasts, where the MCMC yields an ensemble from the posterior and the OE retrieval result provide the first two moments of normal distribution. To compare these methods we apply both univariate and multivariate diagnostic tools and proper scoring rules. The general impression from our study is that when compared to MCMC, the OE retrieval performs reasonably well for the main quantity of interest, the column averaged CO2 concentration XCO2, but not for the full state vector X which includes a profiile of CO2 concentrations over 20 pressure levels, as well as several other atmospheric properties.

## Modelling Precipitation Levels in California due to Atmospheric Rivers

Isabelle Grenier and Bruno Sanso (University of California, Santa Cruz)

### Abstract

Atmospheric Rivers are elongated regions in the atmosphere that transport water vapor out of the tropics. In California, these are responsible for the heavy rainfalls we observe during the winter. Due to climate change, we expect the number and the intensity of atmospheric rivers to increase. The goal of our research is to model the precipitation levels due to atmospheric rivers to assess the impact of climate change on the water supply in California. We first developed a low resolution model which aggregates precipitation over California at the monthly level. The covariates include the number of atmospheric rivers observed, a seasonality factor and the maximum integrated water vapor transport recorded during the month. The average prediction error for the winter months is between 10% and 30%. However, the accuracy is much lower for months out of the peak rain season. Our future work will focus on increasing the time and spatial resolution to increase the predictive accuracy.

## Dimension reduction for remote sensing of atmospheric methane profiles

Otto Lamminpaa (Finnish Meteorological Institute)

### Abstract

Determining the density profiles of trace gases from measured absorption spectra is an ill-posed inverse problem, in which the measurement typically contains limited amount of information. We consider ground based Fourier transform infrared spectrometer (FTIR, part of TCCON network) solar absorption measurements from FMI Arctic Research Centre, to invert atmospheric methane (CH4) density profiles.

This problem is computationally costly, which motivates the development of a dimension reduction scheme. In this study, we use Bayesian framework adaptive MCMC to characterize the full posterior distribution of the solution and the related uncertainties. As a main result, we present a dimension reduction method based on splitting the problem into informative and non-informative subspaces.

## Multi-resolution investigation of climate models using high-end computing resources: a parallel version of the Regional Climate Model Evaluation System powered by HEALPix

Huikyo Lee, Krysztof Gorski, and Brian Wilson
(Jet Propulsion Laboratory, California Institute of Technology)

### Abstract

While systematic, multi-model experimentation and evaluation have been undertaken for years (e.g., the CMIP5), the development and application of infrastructure for systematic, observation-based evaluations of spatial patterns in key climate variables simulated with various spatial resolutions are less mature, owing in part to the needed advances and synergies in both climate and data sciences. One of the main challenges in using existing analysis tools is to carry out the multi-resolution investigation of climate models. Given this, the principal science objective of my work is to provide quantitative and robust evaluations of spatial patterns simulated by climate models across multiple scales: comparison of spatial features at coarse (e.g. 100 km) and fine scales (e.g. 1-10 km) separately between observations and models. Model evaluation also critically rests on data science and technology infrastructure, including access to datasets, storage, and computation. The vast amounts of model and observational data at high resolution required by the model evaluation process have to be brought together in a high-performance, service-based cyber-infrastructure to support large-scale Earth science analytics– a challenge that the current study will address.

I introduce the Jet Propulsion Laboratory's Regional Climate Model Evaluation System (RCMES) powered by the Hierarchical Equal Area isoLatitude Pixelization (HEALPix) as a web-based service for evaluating climate models at different spatial resolutions. The unique capabilities of HEALPix include open-source libraries to facilitate the handling and distribution of massive datasets at different resolutions using parallel computing, and fast and robust analysis of spatial patterns from observational and model datasets regridded into HEALPix pixels, which have been widely used by astronomers and planetary scientists. Both RCMES and HEALPix are open-source software toolkits with a broad user base. We will maximize the utility of RCMES optimized with its parallel processing capabilities as a service through high-end computing (HEC) resources. Our preliminary result indicates that RCMES enhanced with HEALPix could contribute to the ESGF computing application program interface (API) in order to enhance the visibility and utilization of NASA satellite observations in CMIP6.

## Dynamic Fused Gaussian Process for Massive Sea Surface Temperature Data from MODIS and AMSR-E Instruments

Pulong Ma and Emily Kang (University of Cincinnati)

### Abstract

Sea surface temperature (SST) is a key climate and weather measurement, which plays a crucial role in understanding climate systems. Massive amount of SST datasets can be collected from satellite instruments each day with the advance of new remote-sensing technologies. However, these data are often sparse, irregular, and noisy. In addition, different instruments will produce SST data with incompatible supports and distinct error characteristics. For instance, the Moderate Resolution Imaging Spectroradiometer (MODIS) is able to produce SST data at 9km spatial resolution each day, whose quality is subject to weather conditions such as cloud; the Advanced Microwave Scanning Radiometer-Earth Observing System (AMSR-E) is able to produce SST data at 25km spatial resolution each day, whose quality is subject to radio frequency interference. Statistical methods for combining different sources of remote-sensing data will give much more accurate uncertainty analysis. In this article, we propose a Dynamic Fused Gaussian Process (DFGP) model, which extends the Fused Gaussian Process (FGP) in Ma and Kang (2017) to a spatio-temporal model that enables fast statistical inference such as smoothing and filtering for massive datasets. The change-of-support problem is also explicitly addressed in DFGP when statistical inference is made based on different sources of data whose spatial resolutions are incompatible. We also develop a stochastic Expectation-Maximization (EM) algorithm to allow fast parameter estimation in a distributed computing environment. The proposed DFGP is applied to a total of 3.5 million SST datasets in a one-week period in tropical Pacific Ocean area from MODIS and AMSR-E instruments.

## Spatial Statistical Downscaling for Constructing High-Resolution Nature Runs in Global Observing System Simulation Experiments

Pulong Ma and Emily Kang (University of Cincinnati) and
Amy Braverman and Hai Nguyen
(Jet Propulsion Laboratory, California Institute of Technology)

### Abstract

Observing system simulation experiments (OSSEs) have been widely used as a rigorous and cost-effective way to guide development of new observing systems, and to evaluate the performance of new data assimilation algorithms. Nature runs (NRs), which are outputs from deterministic models, play an essential role in building OSSE systems for global atmospheric processes because they are used both to create synthetic observations at high spatial resolution, and to represent the "true" atmosphere against which the forecasts are verified.

However, most NRs are generated at resolutions coarser than actual observations. Here, we propose a principled statistical downscaling framework to construct high-resolution NRs via conditional simulation from coarse-resolution numerical model output. We use nonstationary spatial covariance function models that have basis function representations. This approach not only explicitly addresses the change-of-support problem, but also allows fast computation with large volumes of numerical model output. We also propose a data-driven algorithm to select the required basis functions adaptively, in order to increase the flexibility of our nonstationary covariance function models. In this article we demonstrate these techniques by downscaling a coarse-resolution physical NR at a native resolution of 1-degree latitude by 1.25-degree longitude of global surface CO2 concentrations to 655,362 equal-area hexagons.

## Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity

Florian Schäfer (Caltech)

### Abstract

Many popular methods in machine learning, statistics, and uncertainty quantification rely on priors given by smooth Gaussian processes, like those obtained from the Matérn covariance functions. Furthermore, many physical systems are described in terms of elliptic partial differential equations. Therefore, implicitly or explicitly, numerical simulation of these systems requires an efficient numerical representation of the corresponding Green's operator. The resulting kernel matrices are typically dense, leading to (often prohibitive) $O(N^2)$ or $O(N^3)$ computational complexity.

In this work, we prove rigorously that the *dense* $N \times N$ kernel matrices obtained from elliptic boundary value problems and measurement points distributed approximately uniformly in a $d$-dimensional domain can be Cholesky factorised to accuracy $\epsilon$ in computational complexity $O\big(N \log^2(N) \log^{2d}(N/\epsilon)\big)$ in time and $O\big(N \log(N) \log^{d}(N/\epsilon)\big)$ in space. For the closely related Matérn covariances we observe very good results in practise, even for parameters corresponding to non-integer order equations. As a byproduct, we obtain a sparse PCA with near-optimal low-rank approximation property and a fast solver for elliptic PDE. We emphasise that our algorithm requires no analytic expression for the covariance function. Our work connects the probabilistic interpretation of the Cholesky factorisation, the *screening effect* in spatial statistics, and numerical homogenisation. In particular, results from the game theoretic approach to numerical analysis ("Gamblets") allow us obtain rigorous error estimates.

## Multi-layer ecological data processing for modelling pathogen spread: the ecomultiplex model

Massimo Stella (Fondazione Bruno Kessler),
Sanja Selakovic (University of Utrecht),
Alberto Antonioni (Universidad Carlos III de Madrid),
and Cecilia S. Andreazzi (Fiocruz Foundation)

### Abstract

Multiple routes of transmission for many diseases are investigated separately despite their potential interplay. As a unifying framework for understanding parasite spread through interdependent transmission paths, we present the "ecomultiplex" model, where multi-layer ecological data about predator-prey and parasite-host interactions are processed, combined and represented as a spatially embedded multiplex network. We adopt this framework for designing and testing potential control strategies for parasite spread in two empirical host communities. We base our simulations on the distributed spread of the parasite between multiplex layers. Our results show that the ecomultiplex network model is an efficient and low data-demanding method for identifying which species promote parasite spread, offering mechanistic interpretation of preliminary empirical findings and opening new insights in designing efficient control strategies for parasite containment.

## Uncertainty Propagation for a Large Scale Hydrological Routing Model

Michael Turmon, Jonathan Hobbs, JT Reager, Cedric David,
and Jay Famiglietti (Jet Propulsion Laboratory, California Institute of Technology)

### Abstract

Hydrological routing models use river connectivity information to propagate the localized lateral inflows of surface and subsurface water runoff into downstream flows. The resulting modeled flows can be used for planning and risk analysis, which has motivated the determination of standard errors for flows. We describe computational tradeoffs among several approaches for determination of streamflow uncertainties, which generally correspond to different assumptions about the spatial/temporal covariance of inflows from runoff. We introduce a "reach random effects" model to account for large-scale error correlation, as may be caused by spatially-correlated errors in precipitation forcing. We describe implementation of uncertainty propagation using RAPID (David et al. 2011) applied over the 650,000 reaches of the Western Contiguous United States covered by the NHDPlus network. Finally, we observe that new space missions should provide novel remote-sensing observations of flows at sparsely-sampled points in the river network. We use the accessibility of the full space-time flow covariance to understand the constraints on network flows offered by these new observations.

## The OCO-2 Retrieval Algorithm: Sensitivity to Choice of Prior Covariances

Joaquim Teixeira, Jon Hobbs, and Michael Gunson
(Jet Propulsion Laboratory, California Institute of Technology)

### Abstract

We present the results of an investigation into the sensitivity of the OCO-2 retrieval algorithm, to choices made in setting the retrieval's prior covariance matrix, using a Monte Carlo framework. The OCO-2 retrieval algorithm is an implementation of Bayes' Rule, and the prior covariance weights the cost function towards the prior mean. We wish to understand the effect of different prior covariance matrices on the retrieved CO2 profile. After constructing a set of alternative prior covariances (by manipulating lag correlation over the elements the CO2 profile vector, the variance of the total column average, and the level-by-level variances) we run Monte Carlo simulations with these alternatives across a set of marginal distributions chosen to represent different geophysical conditions. We observe that the choice of the prior covariance matrix can have a substantial effect on the retrieved CO2 profile, while leaving the total column average unchanged. These effects are invariant across different choices of marginal distributions used to generate synthetic "true" state vectors. We see that the choice of lag correlation in the CO2 profile, and standard deviation are the main drivers of these effects.

# Data Sciences for Climate and Environment

### Alan Turing Institute, March 26, 2018

### Jointly sponsored by the Lloyds Register Foundation and SAMSI

# About the event

Organisers: Michel Tsamados (University College London); Chris Oates (Newcastle University), Richard Smith, (University of North Carolina), and Ruth Petrie, (Rutherford Appleton Laboratory)

**Agenda**

The **Lloyd's Register Foundation** (LRF) programme on **data-centric engineering** at the Alan Turing Institute is delighted to partner with the **Statistical and Applied Mathematical Sciences Institute** (SAMSI) to present this topical workshop on Data Sciences for Climate and Environment.

The LRF programme, which brings together world-leading researchers from around the UK, aims to address the data-centric engineering needs of society and industry - an important component of which is to better understand the risks posed to infrastructure and society by the natural environment.

Collectively, we are modelling and monitoring our planet better than we have ever done in our history, as a result of sustained efforts from the climate modelling community and space agencies and the private sector worldwide.

Climate and weather models can now be run at finer spatial resolutions (10km or better), therefore enabling more realistic simulations of smaller and smaller scale processes (i.e. tropical cyclones in the atmosphere or eddies in the ocean) that can have severe impacts on our planet. At the same time there is a rapid growth in the number of satellites orbiting the Earth (221 launched in 2015, around 5000 in total) with a significant fraction of these satellites dedicated to Earth Observation using a large variety of sensors working at different electromagnetic frequencies (optical, radar, infrared, etc.). Our ability to store, process and share efficiently the vast amounts of data that are produced (~Pb yearly) by the modelling and remote sensing communities is a pre-requisite for the good functioning of these often publicly funded large programmes.

In this one-day workshop our speakers will present on how the new tools developed in data sciences can be applied to questions relating to climate and the environment to

help us address the great challenges that our society is facing in a rapidly changing planet.

Our event will be structured around five keynote speakers highlighting five separate topics described below and followed by a panel dialogue between our experts and the audience on the topic of Data Sciences for the Climate and the Environment.

**Bryan Lawrence**

Director of Models and Data at the UK National Centre for Atmospheric Science, Professor of Weather and Climate Computing at the University of Reading, and the Director of the STFC Centre for Environmental Data Analysis (CEDA).

**Jeremy Walton**

Jeremy joined the NERC-Met Office UKESM core group in 2014 as Scientific Systems Manager and is the Lead Computational Scientist for the UK Earth System Model (UKESM)

**Doug Nychka**

Director of the Institute for Mathematics Applied to Geosciences (IMAGe) at the US National Center for Atmospheric Research, and also a Senior Scientist in the Statistics and Data Sciences Section.

**Emily Shuckburgh**

Climate scientist and deputy head of the Polar Oceans Team at the British Antarctic Survey. Emily also holds a number of positions at the University of Cambridge (fellow of Darwin College, fellow of the Cambridge Institute for Sustainability Leadership, associate fellow of the Centre for Science and Policy and member of the Faculty of Mathematics)

**Prabhat**

Leads the Data and Analytics Services team at the US National Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Lab

**Amy Braverman**

Principal Statistician at the NASA Jet Propulsion Laboratory, California Institute of Technology, working in the Multi-angle Imaging SpectroRadiometer (MISR) science team and the Atmospheric Infrared Sounder (AIRS) science development team.

Agenda:

10:00-10:15: Introduction
10:15-11:00: Douglas Nychka - National Center for Atmospheric Research, CO, USA
11:30-12:15: Amy Braverman - Jet Propulsion Lab, California Institute of Technology, USA
13:15-14:00: Bryan Lawrence - UK National Centre for Atmospheric Science
14:00-14:45: Prabhat - Lawrence Berkeley National Lab, CA, USA
15:15-16:00: Jeremy Walton - Met Office, UK
16:00-16:30: Panel discussion - chaired by Emily Shuckburgh, British Antarctic Survey, UK.
16:30-18:00: Drinks reception

The Alan Turing Institute

# Data sciences for climate and the environment

## How new tools in data sciences can address the growing global challenges

26 March 2018
10:00 - 18:00
The Alan Turing Institute

**Agenda and speakers**

10:00-10:15: Introduction
10:15-11:00: Doug Nychka - National Center for Atmospheric Research, CO, USA
11:30-12:15: Amy Braverman - Jet Propulsion Lab, California Institute of Technology, USA
13:15-14:00: Bryan Lawrence - UK National Centre for Atmospheric Science
14:00-14:45: Prabhat - Lawrence Berkeley National Lab, CA, USA
15:15-16:00: Jeremy Walton - Met Office, UK
16:00-16:30: Panel discussion
16:30-18:00: Drinks reception

To register visit
**turing.ac.uk/events**

Follow us
@turinginst
#datacentricengineering

samsi
NSF•Duke•NCSU•UNC

Lloyd's Register Foundation

The Alan Turing Institute, British Library, 96 Euston Road
London NW1 2DB
**turing.ac.uk**

samsi
NSF·Duke·NCSU·UNC

# Climate Program Transition Workshop
## May 14-16, 2018
### SCHEDULE

**Monday, May 14th**
*SAMSI*

| | |
|---|---|
| | Remote Sensing |
| 8:30-8:40 | *"Introduction to Remote Sensing Working Group"*<br>**Amy Braverman**, JPL/Caltech |
| 8:40-9:05 | *"Optimization Methods in Remote Sensing"*<br>**Jessica Matthews**, NCSU/NOAA |
| 9:05-9:30 | *"Incorporating Spatial Dependence in Remote Sensing Inverse Problems"*<br>**Jonathan Hobbs**, JPL |
| 9:30-9:55 | *"Statistical Emulation with Dimension Reduction for Complex Physical Forward Models"*<br>**Emily Kang**, University of Cincinnati |
| 9:55-10:20 | *"A Notional Framework for a Theory of Data Systems"*<br>**Maggie Johnson**, SAMSI |
| 10:20-10:45 | "*Statistical Approaches for Un-Mixing Problem and Application to Satellite Remote Sensing Data*"<br>**Zhengyuan Zhu**, Iowa State |
| 10:45-11:05 | Break |
| | Data Assimilation |
| 11:05-11:35 | *"Sea Ice, Unstable Subspace, Model Error: three topics in data assimilation"*<br>**Amit Apte**, ICTS |
| 11:35-11:55 | *"Projected Data Assimilation"*<br>**Erik Van Vleck**, University of Kansas |
| 11:55-12:15 | *"Understanding Sea Ice Data for Data Assimilation"*<br>**Christian Sampson**, SAMSI |
| 12:15-1:15 | Lunch |
| | Detection and Attribution |
| 1:15-1:40 | *"Overview and New Modeling Directions"*<br>**Dorit Hammerling**, NCAR |

| | |
|---|---|
| 1:40-2:00 | *"Tunable Testbed for Detection and Attribution Methods"*<br>**Nathan Lenssen**, Columbia University |
| 2:00-2:25 | *"Inference on the Future Climate States from Multiple Ensembles using Bayesian Hierarchical Models"*<br>**Huang Huang**, SAMSI |
| | Ice Dynamics |
| 2:25-3:00 | *"Overview of Ice Dynamics Working Group"*<br>**Murali Haran**, Penn State |
| 3:00-3:20 | *"Overview of Sea-Ice Modeling and Statistical Challenges"*<br>**Deborah Sulsky**, New Mexico |
| 3:20-3:40 | *"Ice Model Calibration using Zero-Inflated Continuous Spatial Data"*<br>**Won Chang**, University of Cincinnati |
| 3:40-4:00 | *"Metrics for Evaluating Sea Ice Models"*<br>**Yawen Guan**, SAMSI |
| 4:00-4:20 | Break |
| | Food Systems |
| 4:20-4:50 | *"Overview of Food Systems Working Group"*<br>**Hans Kaper**, Georgetown University |
| 4:50-5:10 | *"Resilience of Food Networks"*<br>**Kaitlyn Hill**, University of Minnesota |
| 5:10-5:30 | *"Agent-based Modeling for Food Systems"*<br>**Adway Mitra**, ICTS |
| 5:45 | Return to Hotel |

**Tuesday May 15th**
*SAMSI*

| | |
|---|---|
| | Environmental Health |
| 8:30-9:00 | *"Accomplishments of the Environmental Health Working Group"*<br>**Brian Reich**, NCSU |
| 9:00-9:20 | *"Projecting the Impact of Temperature Changes on Human Mortality"*<br>**Veronica Berrocal**, Michigan |

| | |
|---|---|
| 9:20-9:40 | *"The Relationship between Extreme Events and Human Health in a Changing Climate".*<br>**Jesse Bell**, NOAA |
| 9:40-10:00 | *"On the Probability Distributions of the Intensity and Duration of Heatwaves"*<br>**Sohini Raha**, NCSU |
| 10:00-10:20 | *"Multiple Health Risk Factors"*<br>**Kimberly Kaufeld**, LANL |
| 10:20-10:40 | *"Bayesian Spatial Propensity Score Analysis: unmeasured and spatial confounding"*<br>**Bo Li**, UIUC |
| 10:40-11:00 | Break |
| | Parameter Estimation |
| 11:00-11:30 | *"Activities and Progress in the "Parameter Optimisation in Climate Modeling" Working Group"*<br>**Ben Timmermans**, LBNL |
| 11:30-12:00 | *"A Fast Particle-Based Approach for Computer Model Calibration"*<br>**Ben Lee**, Penn State |
| 12:00-1:00 | Lunch |
| | Stochastic Parameterization and Machine Learning |
| 1:00-1:30 | *"Stochastic Parameterization of Subgrid-Scale Air-Sea Fluxes"*<br>**Adam Monahan**, University of Victoria |
| 1:30-1:50 | *"Bayesian Approach to Climate Reconstructions using Stochastic Energy Balance Models"*<br>**Fei Lu**, Johns Hopkins |
| 1:50-2:20 | *"A Generative Adversarial Network Stochastic Parameterization of the Lorenz '96 Model"*<br>**David John Gagne**, NCAR |
| | Risk and Coastal Hazards |
| 2:20-2:35 | *"Introduction and Overview of the SAMSI CLIM Risk and Coastal Hazards Working Group"*<br>**Brian Blanton**, UNC/RENCI |
| 2:35-3:00 | *"The Nexus of Climate Data, Insurance, and Adaptive Capacity Located at The Collider (https://thecollider.org/) in Asheville, NC, November 8-9, 2018"*<br>**Rob Erhardt**, Wake Forest |

| | |
|---|---|
| 3:00-3:25 | *"Modeling Weather-induced House Insurance Risks"*<br>**Asim Dey**, University of Texas, Dallas |
| 3:25-3:50 | *"Estimating Extreme Strom Surge Levels: A Statistical Perspective"*<br>**Whitney Huang**, SAMSI |
| 3:50-4:10 | Break |
| | Statistics in Oceanography |
| 4:10-4:40 | *"Statistical Analysis of Oceanographic Data from Argo Profiling Floats: Progress and Challenges".*<br>**Mikael Kuusela**, SAMSI |
| 4:40-5:00 | *"Estimating Oxygen in the Southern Ocean using Argo Temperature and Salinity"*<br>**Donata Giglio**, Scripps Institution of Oceanography |
| 5:00-5:20 | Discussant: **Michael Stein**, University of Chicago |
| 5:30 | Buses return to hotel |

**Wednesday May 16th**
*SAMSI*

| | |
|---|---|
| | Extremes |
| 8:30-9:00 | *"Advances in Understanding of Climate Extremes"*<br>**Ken Kunkel**, NCSU/NOAA |
| 9:00-9:20 | *"Investigating Precipitation Extremes in the US Gulf Coast through the use of a Multivariate Spatial Hierarchical Model"*<br>**Brook Russell**, Clemson |
| 9:20-9:40 | *"Networks and Extremes: Review and Further Studies"*<br>**Ansu Chatterjee**, Minnesota |
| 9:40-10:00 | *"Semiparametric Models for Extremes"*<br>**Surya Tokdar**, Duke |
| 10:00-10:20 | Break |
| 10:20-11:30 | Plan for Final Report / Group Discussion |
| 11:30-12:30 | Special Lecture: *"Climate Extremes and Max-stable Processes"*<br>**Raphael Huser**, KAUST |
| 12:30 | Adjourn / Box Lunch |
| 1:00 | Buses to RDU Airport |

samsi
NSF·Duke·NCSU·UNC

# Climate Program Extremes Workshop
# May 16-17, 2018
## SCHEDULE

**Wednesday, May 16th**
*SAMSI*

| | |
|---|---|
| 11:30-12:30 | Special Lecture: *"Climate Extremes and Max-stable Processes"*<br>**Raphaël Huser**, KAUST |
| 12:30-1:00 | Boxed Lunch |
| | Climate and Weather Extremes |
| 1:30-2:00 | *"Historical Perspective on Hurricane Harvey Rainfall"*<br>**Ken Kunkel** (NC Climate Center, Asheville) |
| 2:00-2:30 | *"Extreme Values of Vertical Wind Speed in Doppler LIDAR ARM Measurements"*<br>**Charlotte Haley** (Argonne) |
| 2:30-3:00 | *"Employing a Multivariate Spatial Hierarchical Model to Characterize Extremes with Application to US Gulf Coast Precipitation"*<br>**Brook Russell** (Clemson) |
| 3:00-3:30 | *"The Dependence between Extreme Precipitation and Underlying Indicators of Climate Change"*<br>**Richard Smith** (UNC) |
| 3:30-4:00 | Break |
| | Spatial Extremes |
| 4:00-4:30 | *"Max-Infinitely Divisible Models for Spatial Extremes Using Random Effects"*<br>**Ben Shaby** (Penn State) |
| 4:30-5:00 | *"Assessing Models for Estimation and Methods for Uncertainty Quantification for Spatial Return Levels"*<br>**Bo Li** (Illinois) |
| 5:00-5:30 | **"***Decompositions of Dependence for High-Dimensional Extremes: Applied to Spatial Precipitation Extremes"*<br>**Yujing Jiang** (Colorado State) |
| 5:30-6:00 | *"Spatial Modeling for Improving Estimates of Extreme Precipitation Statistics at Weather Stations"*<br>**Mark Risser** (LBNL) |
| 6:15 | Return to Hotel |

**Thursday May 17th**
*SAMSI*

| | |
|---|---|
| | Networks and Extremes |
| 8:30-9:00 | *"Networks and Extremes: a review and further studies"*<br>**Ansu Chatterjee** (Minnesota) |
| 9:00-9:30 | "*Network Analysis of Gulf Coast Extreme Precipitation"*<br>**Whitney Huang** (SAMSI) |
| 9:30-10:00 | *"Extreme Rainfall Events of Indian Monsoon: A Network-based Analysis"*<br>**Adway Mitra** (ICTS; visiting SAMSI/UNC) |
| 10:00-10:30 | *"Life Cycle of Extreme Events Revealed by Network"*<br>**Chen Chen** (University of Chicago) |
| 10:30-11:00 | Break |
| | New Models for Extremes |
| 11:00-11:30 | *"Semiparametric Density Estimation for Heavy Tailed Data"*<br>**Surya Tokdar** (Duke) |
| 11:30-12:00 | *"A Semiparametric Bayesian Clustering Model for Spatial Extremes"*<br>**Brian Reich** (NCSU) |
| 12:00-12:30 | *"Semiparametric Models for Densities that Have Flexible Tail Behaviors"*<br>**Michael Stein** (University of Chicago) |
| 12:30-1:00 | *"Extreme Value Theory and the Reassessment in the Caribbean: lessons from hurricanes Irma and Maria"*<br>**David Torres**, University of Puerto Rico |
| 1:00 | Adjourn / Box Lunch |
| 1:30 | Shuttle to RDU Airport |

# Final Report
# Remote Sensing Working Group

**May 16, 2018**

1. Name of working group and program:

Remote Sensing Working Group in the Program on Mathematical and Statistical Methods for Climate and the Earth System.

2. Main working group participants and their affiliations:

Working Group Leaders: Amy Braverman (Jet Propulsion Laboratory) and Jessica Matthews (NCICS).

Webmaster: Maggie Johnson (SAMSI/NCSU Statistics).

Post-docs: Maggie Johnson (SAMSI/NCSU Statistics), Christian Sampson (SAMSI/UNC Mathematics).

Graduate students: Xinyue Chang (Iowa State Statistics), Colin Lewis-Beck (Iowa State Statistics), Pulong Ma (U of Cincinnati Math and Statistics), Si Cheng (U of Cincinnati Math and Statistics), Isabelle Grenier (UC Santa Cruz).

Other active members: Veronica Berrocal (U of Michigan), Jenny Brynjarsdottir (Case-Western Math and Statistics), Ansu Chatterjee (U of Minnesota), Jonathan Hobbs (Jet Propulsion Laboratory), Emily Kang (U of Cincinnati Math and Statistics), Georgios Karagiannis (Durham U), Matthias Katzfuss (TAMU Statistics), Alex Konomi (U of Cincinnati Math and Statistics), Anirban Mondal (Case-Western Math and Statistics), Jim Rosenberger (Penn State), Bruno Sanso (UC Santa Cruz), Joon Jin Song (Baylor Statistics), Zhengyuan Zhu (Iowa State Statistics)

3. Topics and Goals of the Working Group:

The Remote Sensing Working Group (RSWG) goal is to explore new problems in the analysis of large remote sensing data sets, and to develop appropriate statistical theory, methods, and algorithms for solving these problems. Following the Opening Workshop for the Program in August, and subsequent discussion in the main working group meeting, RSWG formed five sub-groups around the following partially overlapping topics:

1. **Spatial Retrievals:** Current operational remote sensing retrieval methodology is usually non-spatial. Inferences about the true geophysical state are made on the basis of observed radiance spectra at one location and time without considering nearby, most likely spatially (and temporally) correlated observations. The Spatial Retrieval, or Spatial X, subgroup has developed a spatial statistical framework for joint inference of the true state across nearby satellite pixels (footprints). This framework is particularly applicable for remote sensing retrievals that implement a Bayesian

hierarchical model for inference, such as the Orbiting Carbon Observatory-2 (OCO-2) retrieval of atmospheric carbon dioxide. The subgroup has surveyed similar multi-pixel retrieval approaches in the literature and has performed simulation experiments on a small spatial domain with a linear version of the OCO-2 physical forward model. With anticipated support from the OCO-2 project, the work will be extended to the operational nonlinear model used for the mission.

2. **Unmixing problem:** The objective of this working group is to model remote sensing data from satellites as high dimensional multivariate spatial temporal processes and use the models to address problems such as un-mixing inhomogeneous pixels and gap-filling missing data. Remote sensing data with high temporal resolution typically has lower spatial resolution, with one pixel often spanning over several square kilometers. The signal recorded by such a satellite pixel is typically a mixture of reflectance from different types of land covers within the pixel, resulting in a mixed pixel. Disaggregating the signal into its distinct components is referred to as the un-mixing problem. Using data from the Soil Moisture and Ocean Salinity (SMOS) satellite, we developed a Bayesian parametric model to un-mix pixels observed over an intensively cultivated agricultural region in the Midwest. A Bayesian non-parametric model is under development to deal with hyper-spectral un-mixing problems such as those in the OCO-2 (Orbiting Carbon Observatory 2) mission. Motivated by the need to impute missing data in the OCO-2 mission, we developed a spatial functional model to gap-fill and smooth the hyper-spectral data to reduce measurement error. These methods are applied to data from the OCO-2 mission.

3. **Intersatellite calibration:** A fundamental challenge for climate-quality remote sensing retrievals is ensuring the stability and accuracy of initial sensor radiance observations. Among the multiple instruments on board NOAA's Polar-orbiting Operational Environmental Satellite (POES) is the High-resolution Infrared Radiation Sounder (HIRS). To date, the POES series has launched 12 satellites where HIRS data is captured. Given the extreme conditions of space, the instrument itself degrades in sensitivity over time which impacts the radiances measured. In order to characterize any climatological trends from retrievals derived from these radiances across satellites in the series, sensor degradation must be accurately characterized. Nearest Neighbor Gaussian Process methods embedded in Bayesian hierarchical models are being explored as a dimension-reducing technique to quantify the discrepancy between observations made on separate satellites. Pilot work has examined one day of data for one channel on two overlapping satellites. Future work for this subgroup, led by Jessica Matthews and with about 5 more active participants, will focus on accelerating the algorithm to expand the analysis to examine all channels, longer time periods, and more satellites in the series.

4. **Emulators:** This subgroup, led by Emily Kang and with another 5 active participants, have been developing efficient statistical emulation for complex physical forward models in retrieval algorithms in remote sensing. Compared to heuristic or machine learning methods, a statistical emulator provides emulated values with associated uncertainty, and thus can be applied in a wider range of studies including uncertainty quantification and calibration. This subgroup develop a framework of combining dimension reduction of both input and output spaces and Gaussian process emulation and apply the method to emulate the forward model in OCO-2 retrieval. This work is closely related to SAMSI's 2018-2019 program Model Uncertainty: Mathematical and Statistics (MUMS). Pulong Ma will join SAMSI as a Postodoctoral Fellow this summer, and several subgroup members plan to continue the collaboration and further investigate the application and improvement of the current emulation method in the forthcoming program.

5. **Theory of Data Systems:** The ToDS sub-group (with about eight members and led by Amy Braverman) aimed to develop a framework for connecting statistical methodologies for inference from distributed data sources with ideas from systems and software architectures. The group spent much of the fall semester reviewing literature from the wide variety of discipline areas that are relevant to this endeavor: approximate likelihoods and sufficiency, the statistical-computational trade-off, software and systems architecture, theory of complex systems and networks, and spatial statistical methods for distributed data. Using a very simple case study (estimation of a population mean using correlated data from multiple servers), the group fleshed-out a simple framework for analyzing the trade-off between costs and estimation variance. In the second semester, the group moved on to consider a more realistic problem; that of estimating the parameters of a spatial covariance function with distributed data. That work is continuing, and driving towards a manuscript to be submitted for publication. Maggie Johnson, the group's post-doc, will move to JPL for her second year where she will continue to concentrate on this work.

4. Publications in development:

**Statistical Emulation with Dimension Reduction for Physical Forward Models in Remote Sensing**
Authors (alphabetically ordered): Jon Hobbs (JPL), Emily Kang (U of Cincinnati), Alex Konomi (U of Cincinnati), Pulong Ma (U of Cincinnati), Anirban Mondal (Case Western Reserve U), Joon Jin Song (Baylor U).

**A Nearest Neighbor Gaussian Process Approach to Quantify Intersatellite Differences in the HIRS time series**
Authors (alphabetically ordered): Si Cheng (U of Cincinnati), Jon Hobbs (JPL), Georgios Karagiannis (Durham U), Alex Konomi (U of Cincinnati), Jessica Matthews (NCICS), Christian Sampson (SAMSI/UNC Mathematics).

**A theory of data analysis systems for remote sensing.**
Authors (alphabetically ordered): Amy Braverman (JPL), Ansu Chatterjee (U of Minnesota), Isabelle Grenier (UC Santa Cruz), Maggie Johnson (NC State/SAMSI), Brian Reich (NC State), Jim Rosenberger (Penn State), Bruno Sanso (UC Santa Cruz), Zhengyuan Zhu (Iowa State).

*Paper on spatial retrievals (method introduction): Review of the previous work on spatial (multipixel) retrievals. Work published in remote sensing literature is re-cast in our mathematical framework where possible. Simulation setup using a simplified linear model for OCO-2 will be presented with canonical results contrasting spatial and non-spatial retrievals. Plan to submit by December 2018.*
Authors (alphabetically ordered): Veronica Berrocal (U of Michigan), Jenny Brynjarsdottir (Case-Western Math and Statistics), Jon Hobbs (JPL), Matthias Katzfuss (TAMU Statistics), Anirban Mondal (Case-Western Math and Statistics).

*Spatial retrieval paper using the full nonlinear forward model as a realistic OCO-2 application. Plan to submit by August 2019.*
Authors (alphabetically ordered): Veronica Berrocal (U of Michigan), Jenny Brynjarsdottir (Case-Western Math and Statistics), Jon Hobbs (JPL), Matthias Katzfuss (TAMU Statistics), Anirban Mondal (Case-Western Math and Statistics).

*Paper on using a Bayesian parametric model to un-mix SMOS pixels observed over an intensively cultivated agricultural region in the Midwest US, combining ground level prior information with satellite data. Plan to submit by December 2018.*
Authors (alphabetically ordered): Jenny Brynjarsdottir (Case-Western Math and Statistics), Xinyue Chang (Iowa State Statistics), Jon Hobbs (JPL), Maggie Johnson (NC State/SAMSI), Colin Lewis-Beck (Iowa State Statistics), Anirban Mondal (Case-Western Math and Statistics), Joon Jin Song (Baylor Statistics), Zhengyuan Zhu (Iowa State Statistics).

*Paper on nonparametric approach to the un-mixing problem where observed and aggregated data are represented as a linear combination of distinct components that represented as functional data using B-spline basis functions. Plan to submit by June 2019.*
Authors (alphabetically ordered): Jenny Brynjarsdottir (Case-Western Math and Statistics), Xinyue Chang (Iowa State Statistics), Jon Hobbs (JPL), Maggie Johnson (NC State/SAMSI), Colin Lewis-Beck (Iowa State Statistics), Anirban Mondal (Case-Western Math and Statistics), Joon Jin Song (Baylor Statistics), Zhengyuan Zhu (Iowa State Statistics).

*Paper on gap-filling of hyper-spectral satellite imagery using functional spatial models. Plan to submit by June 2019.*
Authors (alphabetically ordered): Jenny Brynjarsdottir (Case-Western Math and Statistics), Xinyue Chang (Iowa State Statistics), Jon Hobbs (JPL), Maggie Johnson (NC State/SAMSI), Colin Lewis-Beck (Iowa State Statistics), Anirban Mondal (Case-Western Math and Statistics), Joon Jin Song (Baylor Statistics), Zhengyuan Zhu (Iowa State Statistics).

5. Conference and workshop presentations:

See report on "Remote Sensing, Uncertainty Quantification, and a Theory of Data Systems" workshop held at California Institute of Technology, February 12-14, 2018.

See report on "CLIM Transition Workshop" held at SAMSI, May 14-16, 2018.

Emily Kang at "ASA ENVR Workshop on Statistics for the Environment: Research, Practice and Policy" to be held in Asheville, NC, October 2018.

Amy Braverman at "Data Science for Climate and the Environment" held at the Alan Turing Institute in London, UK, March 26, 2018.

6. Publicly available software developed:

Software will be provided as supplementary materials for the manuscripts in preparation.

7. Grant proposals:

Proposals to NSF-Division of Mathematical Sciences are planned.

Solicited proposal submitted to NASA's OCO-2 mission to support spatial retrieval research activities from June 2018-August 2019. Awards to Matthias Katzfuss and Jenny Brynjarsdottir. Total: $170K.

Solicited proposal submitted to NASA's Advanced Information Systems Technology Program (AIST) to support continued research in theory of data systems from May 2018-April 2019. Awarded to Amy Braverman. Total: $100K.

8. Organization of follow-up workshops:

This research of this working group will be presented at:

*2018 JSM (Vancouver, British Columbia, Canada, July 28 - Aug 2):* Two sessions have been proposed and are scheduled: "Statistical Methods for Remote Sensing Data" and "Advances in Inference for Massive Spatio-Temporal Environmental Data with Applications in Remote Sensing". There will also be a roundtable discussion, "Statistical Challenges in the Analysis of Distributed Remote Sensing Data".

*2018 SIAM Conference on Mathematics of Planet Earth (Philadelphia, PA, September 13-15):* Two sessions have been proposed and accepted to address "Challenges in Environmental Remote Sensing".

# SAMSI WORKING GROUP ON PARAMETER ESTIMATION

## Group Leader:

Ben Timmermans, Lawrence Berkeley National Laboratory

## Major group members:

Ben Seiyon Lee, Pennsylvania State University
Andrew Gettelman, NCAR
Murali Haran, Pennsylvania State University
Charles Jackson, University of Texas at Austin

## Motivation and Research Questions (from the Opening Workshop)

Most computer models for environmental systems, including climate models, contain physical parameters that may not be known a priori. The motivation for this group was to explore methods for estimating unknown parameters. This was divided into several subthemes:

(a) Current "state of the art:"

• What are key motivations & research priorities?
• Does best practice / "industry standard" exist?
• What are the key challenges and barriers to use?

(b) What progress in analysis / estimation given:

• High dimension input / (temporal) output?
• Stochastic or extremal output?
• Complex / uncertain observations.
• (Semi) automated methods for climate problems?
• Multiple sources of information (multi-resolution)

(c) Motivating questions:

• To what extent can parameter investigaton / estimation inform about the relationship between physical processes and output statistics (or model deficiency)?
• What is the relationship between parameter estimation, optimization and UQ? (cf data assimilation)

## Summary of Transition Workshop Presentations

Ben Timmermans gave the lead-off presentation. He first cited a paper of Rhoades et al. (2018) as motivation – a paper that calculated temperature and precipitation statistics under two microphysics models. He then reviewed the literature, beginning with classical references on statistical models for computer experiments (Sacks et al. 1989, Santner et al. 2003) before turning to modern Bayesian approaches beginning with the pathbreaking papers of Kennedy and O'Hagan (2001) and Oakley and O'Hagan (2004). Traditional MCMC approaches to Bayesian inference are often too slow in this setting so recent research has focused on alternatives, such as data assimilation through the Ensemble Kalman Filter approach (Carassi et al., 2017) or learning approaches (Schneider et al, 2017). He concluded with three examples, including recent work on hurricane modeling under global warming (Patricola and Wehner, 2018; Timmermans, Patricola and Wehner, 2018).

The second presentation was by Ben Seiyon Lee (Penn State) entitled "A Fast Particle-Based Approach for Computer Model Calibration." This was based on his PhD work joint with Murali Haran and several geoscientists at Penn State. The science motivation was projecting the contribution of Antarctic sea ice to global sea level rise (DeConto and Pollard, 2016). The methodological contribution was a new method of Bayesian inference intermediate between MCMC methods that are too slow to run with large models, and Gaussian emulation techniques which work well only for low-dimensional parameter spaces. The objective was a method that would work well for medium-fast models in high-dimensional parameter spaces.

## References


Carrassi A., Bocquet M., Hannart A., Ghil M. 2017. Estimating model evidence using data assimilation. Q. J. R. Meteorol. Soc., 143: 866–880.

DeConto, R. M. and Pollard, D. (2016). Contribution of Antarctica to past and future sea-level rise. Nature, 531(7596):591.

Kennedy & O'Hagan (2001) "Bayesian calibration of computer models" J.R. Statist. Soc. B

Oakley & O'Hagan (2004) "Probabilistic sensitivity analysis of complex models: a Bayesian approach" J. R. Statist. Soc. B

Patricola, C. and Wehner, M.F. (2018), Anthropogenic influences on major tropical cyclone events. Nature volume 563, pp. 339–346.

Rhoades, Ullrich, Zarzycki, Johansen, Marguli, Morrison, Xu and Collins (2018), Sensitivity of Mountain Hydroclimate Simulations in Variable-Resolution CESM to Microphysics and Horizontal Resolution. Journal of Advances in Modeling Earth Systems, Volume 10, Issue 6

Sacks et al. (1989) "Design and Analysis of Computer Experiments" Statistical Science

Santner et al. (2003) "The Design and Analysis of Computer Experiments" Springer, New York

Schneider et al. (2017) “Earth System Modeling 2.0: A Blueprint for Models That Learn From Observations and Targeted High-Resolution Simulations” GRL

Timmermans, Patricola and Wehner (2018), Simulation and Analysis of Hurricane-Driven Extreme Wave Climate Under Two Ocean Warming Scenarios. Oceanography. 31. 10.5670/oceanog.2018.218.

# Working Group on Data Assimilation

**Leaders:** Amit Apte, Mathematics, International Center for Theoretical Sciences (ICTS) – TIFR; Chris Jones, Mathematics, University of North Carolina at Chapel Hill (UNC); and Erik Van Vleck, Mathematics, University of Kansas (KU).

**Webmaster and postdoc:** Christian Sampson, SAMSI and UNC

**Graduate Students:** Colin Guider, UNC; Paul Cornwell, UNC;

**Other active members:** Elaine Spiller, Mathematics, Marquette University; Anugu Sumith Reddy, ICTS-TIFR; John Maclean, UNC; Sreekar Vadlamani, TIFR Center for Applicable Mathematics; Michal Branicki, Mathematics, University of Edinburgh; Laura Slivinski, NOAA/ESRL and U. of Colorado (CIRES); Aneesh Subramanian, UCSD; Cassidy Krause, KU; Yuxin Chen, Northwestern

## Topics and goals of the WG:

The three main themes on which the working group focused are the following:

1. Model error
2. Methodology for assimilation in presence of adaptive meshes
3. Approaches to problems in high dimensions

The specific goals of the group were to develop or understand the following three aspects, each of which would have an overlap with more than one of the themes above.

1. Advanced, fully nonlinear methods such as particle filters
2. Combining dynamical information such as instabilities within probabilistic framework
3. Lagrangian data assimilation as a framework to study the above issues

These goals are further tied into specific projects that are described below.

Two related issues are under discussion:

1. Flow over sills in deep ocean circulation
2. Noise induced tipping in hurricane models and the role of shear

## Projects/topics chosen by SAMSI postdocs and SAMSI grad fellows.

Christian Sampson (postdoc) and Colin Guider (student): neXtSIM is a Lagrangian sea ice model that makes use of a novel adaptive mesh in its numerical simulations. This mesh is particularly suited to problem of numerically simulating large scale sea ice phenomena, however, it makes data assimilation difficult due to the changing dimension of the state space. We aim to build upon initial work done by Colin Guider to develop a data assimilation

scheme specifically for use in the neXtSIM model.

Paul Cornwell (student): We aim to study the flow of water over a sill in the deep ocean. In the case where the flow is one-dimensional and there is no viscosity, a known relationship exists between the height of the water over the sill and its velocity. Our goal is to understand the analogous relationship when the flow is considered instead through a channel, possibly including small viscosity and/or the Coriolis effect. It is expected that this relationship will be manifested in a "critical condition," which corresponds to a loss of normal hyperbolicity for a slow manifold in a function space. This problem is of concern to data assimilation because the depth near the sill can be observed, and this in turn used to calculate the quantity of interest (the velocity).

## Projects/topics chosen by SAMSI postdocs and SAMSI grad fellows.

This was one of the most active working groups with at least seven papers at various stages of publication based on the group's activities. In addition, group leaders Apte, Jones and Van Vleck have given numerous conference proceedings about topics related to data assimilation. Details are in the "Program Products" section of the main report.

## Overview

The extremes working group divided into three subgroups, all of each operated semi-autonomously, but the groups did interact during the first several weeks of the program (through September) and also during a meeting after the first of the year.

One group focused on networks and extremes. The topic for this subgroup arose from a discussion at the 2017 SAMSI Program on Mathematical and Statistical Methods for Climate and the Earth System (CLIM) Opening workshop. Namely, a group of researchers working mainly on extremes talked to a group of researchers working on climate networks.
The goal of this sub-group was to study extreme events and their relation to networks. To this end, the group studied networks in general and climate networks in particular, and the handful of papers that define and work with network-structures that are based on extreme events. In addition to identifying and working on several projects the group also reviewed interesting parts of the related literature. Three papers on in preparation: an overview paper led by Imme Ebert-Uphoff (CSU), the second a network study of Gulf Coast precipitation led by Whitney Huang (SAMSI postdoc) and a third study of the Indian monsoon led by Adway Mitra (India).

A second group focused on semi-parametric approaches to modeling the bulk and tail(s) of a probability distribution. The main idea is to develop a model that can accommodate flexible bulk part of the distribution while allowing a smooth transition to asymptotic regime (in the limit) suggested by extreme value theory. The group started by reviewing recent related work on the Log-Histospline method of Huang (SAMSI postdoc) and Nychka (NCAR) for estimating precipitation extremes using the full range of (non-zero) precipitation observations. One of two complementary projects was led by Surya Tokdar (Duke) and student Erika Cunningham and the other was a collaboration between Brian Reich (NCSU), Ben Shaby (Penn State), and NCSU student A. Hazra.

A third group studied climate/weather applications of extremes methods focusing on extreme precipitation. The group reviewed the most recent methodologies for modeling spatial extremes and then advanced three parallel projects. Brook Russell (Clemson) is leading a project spatially modeling extreme precipitation in the Gulf Coast which accounts for spatial dependence in the data. A second project led by Mark Risser (Lawrence Berkeley Labs) is also spatially modeling extreme precipitation while considering effects due to data dependence. A third project motivated by Hurricane Harvey and led by Ken Kunkel (NCEI) and Richard Smith (SAMSI/UNC) is studying the largest Gulf Coast precipitation events.

## Publications in preparation

Climate Networks for Extremes. Lead author Imme Ebert-Uphoff with other authors (order not yet finalized): Imme Ebert-Uphoff, Adway Mitra, Whitney Huang, Dan Cooley, Singdhansu Chatterjee, Chen Chen, Zhonglei Wang. Potential journals: *Chaos, Computer & Geosciences, Journal of Climate*.

Network Analysis of Gulf Coast Extreme Precipitation Lead: Whitney Huang (order not yet finalized, alphabetical order now) Singdhansu Chatterjee, Chen Chen, Dan Cooley, Imme Ebert-Uphoff, Whitney Huang, Adway Mitra, Zhonglei Wang Potential journals: Advances in Statistical Climatology, Meteorology and Oceanography (ASCMO), Environmetrics, Geophysical Research Letters, Journal of Climate

Semiparametric Analysis of Extremes , Erika Cunningham and Surya Tokdar.

A semiparametric Bayesian model for spatial extremes. Hazra A, Reich BJ, Staicu A, Shaby BA. To be submitted to the Annals of Applied Statistics.

Study of extreme rainfall events of Indian monsoon using networks. Lead: Adway Mitra Authors: (order not yet finalized, alphabetical order now) Singdhansu Chatterjee, Chen Chen, Dan Cooley, Imme Ebert-Uphoff, Whitney Huang, Adway Mitra, Zhonglei Wang. Potential journals: Weather and Climate Extremes, Computers and Geosciences, Journal of Climate.

A probabilistic gridded product for daily precipitation extremes over the United States Author list: Mark D. Risser, Christopher J. Paciorek, Michael F. Wehner, Travis A. O'Brien, William D. Collins. Target journal: TBD

Investigating the link between Gulf of Mexico sea surface temperatures and US Gulf Coast precipitation extremes with focus on Hurricane Harvey. Brook Russell, Mark Risser, Ken Kunkel, Richard Smith. Target journal: TBD

Ken's manuscript

**Software**

R package 'sbde' currently available through git (https://github.com/tokdar/sbde) and is being prepared for a CRAN submission soon

**Connections**

Networks: From the outset, this group formed to explore connections between researchers familiar with networks (Ebert-Uphoff, Mitra, Chatterjee), those working in extremes (Huang, Cooley), and climate scientists (Chen).

Precipitation: This group brought together statisticians working with extremes (Russell, Cooley, Smith), spatial statisticians and an interest in extremes (Risser), and atmospheric scientists (Kunkel) interested in extreme precipitation.

Semiparametrics: While this group was primarily composed of statisticians, a number of collaborations were initiated that would not have been without the SAMSI program. In particular, Erika Cunningham and Surya Tokdar are working with Michael Wehner on wind data analysis; Brook Russels and Surya Tokdar have initiated a project on analyzing gulf coast precipitation data; Michael Stein and Surya Tokdar's group are exploring development of the semi-parametric density models. Brian Reich and Bo Li have begun a collaboration on calibration of misspecified Bayesian models for spatial extreme value analysis

**Conference Sessions**

While no conference sessions are yet organized from this working group, two of the networks projects will be presented at the Climate Informatics Workshop in September 2018.

# Final Report for the Stochastic Parameterization and Climate Informatics Working Groups

Adam Monahan

## 1 Stochastic subgrid-scale parameterizations of weather and climate models

All weather and climate models have finite spatial resolution. As a result, some physical processes are resolved while others are not. However, the unresolved processes cannot be neglected: as a result of dynamical nonlinearities, the unresolved scales have an upscale influence on the resolved scales. This influence must be parameterized in terms of resolved scale quantities.

Traditional parameterizations are deterministic, so that a unique configuration of the resolved scales implies a unique upscale influence from the unresolved. This representation is appropriate if the separation between resolved and unresolved scales is large, but is not necessarily so if the scale separation is small - which is generally the case in weather and climate models. In this case, it is more appropriate to represent the influence of unresolved scales on the resolved ones as random fields conditioned on the resolved scales. Such representations are referred to as stochastic parameterizations.

The stochastic parameterization working group consisted of the following regular participants:

- Adam Monahan (Lead)
  (School of Earth and Ocean Sciences, University of Victoria)
- Julie Bessac
  (Mathematics and Computer Science Division, Argonne National Laboratory)
- Hannah Christensen
  (Climate and Global Dynamics Division, NCAR/Department of Physics, University of Oxford)
- David John Gagne II
  (Advanced Study Program, NCAR)
- Fei Lu
  (Mathematics, Johns Hopkins University)

- Aneesh Subramanian
  (Center for Western Weather and Water Extremes, Scripps Institute of Oceanography)
- Nils Weitzel
  (Meteorological Institute, Universitaet Bonn)

The working group was composed of the following three subgroups, which met on an approximately weekly basis.

## 1.1 Bayesian Climate Reconstructions using Stochastic Energy Balance Models

**Participants: Weitzel, Lu, Monahan**

Deriving spatio-temporally resolved climate reconstructions of the last deglaciation (from ∼21ky BP to ∼6ky BP) from proxy data is important to understand the structure of gradual and abrupt climate changes and test hypotheses on the mechanisms behind those changes. As the available proxy data come from indirect measurements, and are therefore sparse and noisy, a physically constrained stochastic process for interpolation and a statistically rigorous quantification of uncertainty are important building blocks of reconstructions. Therefore, we use a Bayesian framework and a non-linear stochastic energy balance model (SEBM) for interpolation.
We started by defining the SEBM, which is motivated by the Fanning and Weaver (1996) model implemented in the UVic EMIC (Earth System Model of Intermediate Complexity); and the observation operator for proxy data containing spatio-temporally integrated information on the temperature evolution. Subsequently, we decided to use particle MCMC (Andrieu et al. 2010, Lindsten et al. 2014) for joint inference of the high dimensional climate state (dimension $10^3$ to $10^8$) and a low dimensional set of model parameters (order $10^0$ to $10^1$). The algorithm efficiently combines sequential Monte Carlo methods with MCMC techniques and exploits the forward structure of the SEBM to create MCMC proposals for the high dimensional, non-Gaussian posterior. Currently, we are testing the algorithm in idealized settings to understand stability and well-posedness of the inverse problem. Preliminary results were presented at the SAMSI CLIM Transition Workshop.
After further testing the algorithm, we plan to write a methodological paper describing the state and parameter estimation for the chosen non-linear stochastic dynamic model. Subsequently, we plan to perform pseudo-proxy experiments with the CCSM3 TraCE21K simulation of the last deglaciation and apply the method to a state-of-the-art proxy data synthesis. The results of these applications are envisaged to be published in a (paleo)climatology journal.
The SAMSI CLIM program brought together the three participants of this subproject, who have not worked together previously and combine backgrounds in maths, physics, and (paleo)climatology. This combination of know-how facilitated the definition of the paleoclimate problem to be studied, the specification

of the physically motivated stochastic model for interpolation, and the design of an efficient inference algorithm. This research project would likely not have occurred without the SAMSI program.

## 1.2 Stochastic Parameterization of Air-Sea Fluxes

**Participants: Monahan, Bessac, Christensen, Weitzel, Subramanian**

Air-sea fluxes of mass, momentum, and energy are influenced by the surface wind speed, $s$. Numerical weather and climate models have finite spatial resolution, and require surface fluxes averaged over model gridboxes. Flux parameterizations are generally a convex nonlinear function of wind speed. It follows that the flux averaged over a region of space (such as a gridbox) does not equal the flux that would be computed from the averaged wind speed. Furthermore, the gridbox-averaged wind speed itself is not available from the weather or climate model. Rather, the models directly simulate the gridbox-averages of the horizontal wind components $(u, v)$. Denoting model gridbox-averaging by angle brackets, we have the following set of inequalities:

$$\overbrace{\langle \text{flux}(s) \rangle \geq \underbrace{\text{flux}\left(\langle s \rangle\right) \geq \text{flux}\left(\sqrt{\langle u \rangle^2 + \langle v \rangle^2}\right)}_{\text{subgrid-scale velocity variations}}}^{\text{convexity of flux parameterization}} \tag{1}$$

The left-hand quantity in inequality (1) is what is desired, while the quantity on the right is what is directly available from the resolved flow in models. Furthermore, because of a lack of scale separation, the true flux $F^{(T)}$ is not necessarily a deterministic function of the resolved flux $F^{(R)}$.

With an eye toward ultimately developing a stochastic parameterization, air-sea fluxes working group modelled the logarithmically transformed difference $\varepsilon = \log_{10}\left(F^{(T)} - F^{(R)}\right)$ between the true and resolved fluxes. This analysis was conducted using high-resolution (4 km $\times$ 4 km, hourly output of 9 days duration) fields over the Indian Ocean - West Pacific domain produced by the UK Natural Environment Research Council 'Cascade' project (Holloway et al., 2012). By coarse-graining the high-resolution model output, we were able to compute both "true" and "resolved" surface fluxes. This analysis was carried out for a range of averaging scales from $0.125^\circ \times 0.125^\circ$ to $4^\circ \times 4^\circ$.

We found that the conditional dependence of the log-10 error process $\varepsilon$ on the resolved flux $F^{(R)}$ could be represented by a simple regression model:

$$\varepsilon = \sum_{j=0}^{N} A_j \left[\log_{10}\left(F^{(R)}\right)\right]^j + \zeta \tag{2}$$

such that to a first approximation the residual process $\zeta$ is a locally-correlated Gaussian spacetime field independent of $F^{(R)}$. This result was found to hold

across the entire domain, as well as within model subdomains. Previous studies have focused on the deterministic parameterization of $\varepsilon$. The novelty of our work was the focus on the residual process $\zeta$ as the basis of a stochastic parameterization. As coarse-graning scales increase, the variance of $\zeta$ is found to decrease (more averaging reduces the need for explicit stochasticity). We further found that some of the variability of $\zeta$ can be reduced by modelling its dependence on precipitation, a quantity which is associated with mesoscale near-surface velocity variations.

The results that have so far been obtained are currently being assembled in a manuscript to be submitted for publication in an atmospheric science/climate journal. These results will also be presented by J. Bessac at the Joint Statistical Meeting in Vancouver in July, 2018. Future work will involve developing a more detailed model of the spacetime structure of the field $\zeta$; extending this analysis to a larger spatial domain and longer time period; and implementing this parameterization in a comprehensive weather/climate model. J. Bessac will take the lead on the first of these projects. A. Monahan will seek funding to support the second of these future projects through an upcoming NSERC Discovery Grant application. H. Christensen is currently recruiting an MSc student to work on the third of these projects. We expect that most of the participants in the subgroup will continue to work on this project. Each of these projects is expected to lead to a separate publication.

Most of the members of this subgroup met through the SAMSI program, without which this research would not have occurred.

## 1.3 Machine Learning

**Participants: Gagne, Christensen, Subramanian, Monahan**

Machine learning and deep learning models have the potential to serve as sub-grid parameterizations for numerical weather and climate models. By fitting to data about the relationship between the coarse grid and a subgrid process, machine learning models could learn a more optimal fit than physics-based methods. Because there is often not a single sub-grid forcing associated with particular coarse grid conditions, incorporation of stochasticity into the machine learning modeling process is necessary to capture the full range of possible conditions. An initial feasibility study for this approach has been conducted with the Lorenz 96 chaotic dynamical system model. Generative adversarial networks (GANs), a technique for using a discriminative neural network to train a generative neural network to produce conditional samples of arbitrary distributions, are used to model the sub-grid forcing in the Lorenz 96 model. The GANs are compared with polynomial regression on both weather and climate time-scales. Stochasticity was added to the GANs through dropout and Gaussian noise layers in addition to random inputs correlated in time. The GANs are able to reproduce the true sub-grid forcing distribution in offline evaluations. When integrating the model with the GAN and polynomial parameterizations, the GAN and polynomial produce similar distributions of coarse grid values to

the true distribution, but the polynomial consistently outperforms all GAN settings tested (although some settings are competitive). Including more stochastic noise leads to better coverage of the range of forcing values, and using little to no noise leads to the model trajectory becoming fixed in one regime. In the weather evaluations, the GAN and polynomial parameterizations have competitive RMSE, but the GAN has slightly higher ensemble spread. Further work is continuing on including more temporally correlated noise in the machine learning model, evaluating the effects of different numerical integration methods for the forecast model, and submitting an article for publication in the Journal of Advances in Modeling Earth Systems.

While most of the participants of this subgroup knew each other before the start of the SAMSI program, this program encouraged the pursuit of this research project and provided necessary support (such as teleconferencing facilities).

# Final Report for the Environmental Health Working Group

1. Name of working group and program: Environmental Health is a working group in the CLIM program

2. Main WG participants and their affiliations

WG leader: Brian Reich, NCSU, Statistics
Webmaster: Yawen Guan, NCSU, Statistics/SAMSI
Post-docs: Yawen Guan, NCSU, Statistics /SAMSI
Graduate Students: Amanda Muyskens, NCSU, Statistics
Other active members:

David Banks, Duke University, Statistical Science
Jesse Bell, ICS-NC
Veronica Berrocal, University of Michigan, Biostatistics
Howard Chang, Emory University, Biostatistics & Bioinformatics
Sujit Ghosh, NCSU, Statistics
Staci Hepler, Wake Forest University, Mathematics and Statistics
Margaret Johnson, SAMSI post-doc
Emily Kang, University of Cincinnati, Mathematical Sciences
Matthias Katzfuss, Texas A&M University, Statistics
Kimberly Kaufield, Los Alamos National Lab
Bo Li, UIUC, Statistics
Donghai Liang, Emory Univeristy
Pulong Ma, University of Cincinnati, Mathematical Sciences
Elizabeth Mannshardt, US Environmental Protection Agency
Kyle Messier, University of Texas - Austin
Richard Smith, SAMSI, Statistics
Joon Jin Song, Baylor University, Statistical Science
Michael Wehner, Lawrence Livermore National Lab

3. Topics and goals of the WG:

In the beginning of the program we identified several topics (subgroups) with goals given below

(1) **Multiple risk factors**: Estimate the joint effect of multiple exposure variables and/or define an interpretable index that captures their joint effect

(2) **Heat waves**: Estimate the effect of heat waves on mortality and morbidity

(3) **Mobile/wearable devices**: Develop new statistical methods applicable to watches that measure air pollution, and mobile devices that provide fine-resolution spatial information

(4) **Data fusion**: Develop new methods to blend multiple data sources to estimate the spatiotemporal distribution of air pollution

(5) **Causal inference**: Extend methods of causal inference to the spatial setting while accounting for interference and spillover effects

(6) **Infectious disease**: Determine the influence of climate on the spread of infectious diseases and use this information for short-term forecasting

4. Summary of each working group

We have made substantial progress on most of these topics. The current status of each working group and plans for carrying the research forward are given below.

(1) **Multiple risk factors**: Understanding the multivariate effects from climate models and/or pollutant models requires a lower dimension specification due to the correlations between climate models or pollutants. It is difficult to decipher the effects of pollutants in terms of the effect in modeling health outcomes. Additionally, speciated pollutant data often has several missing values which is usually resolved by averaging the values. Multiple methods have been used to assess the correlation between climate models and pollutants, PCA, Factor Analysis, Weighted Quantile Regression. This discussion has led to two projects (1) we have put together a review paper to compare methods for pollutants as related to multiple health outcomes. This includes assessing the impacts of supervised vs. unsupervised learning methods. (2) We have started to build a multivariate approach to simulate pollutants, assess imputation approaches by randomly taking out observed values and the impact on multiple health outcomes.

Plans for future work: Each of these three projects will result in a paper. Tentative titles are

Hepler, S, Kaufeld, KA, Simmons, S. Comparison of multivariate spatial temporal methods for modeling air pollutants.

Hepler, S, Kaufeld, KA, Ghosh, S, Berrocal, V. Imputation approach for multivariate pollutants with multivariate health outcome.

(2) **Heat waves**: Many studies on climate change have indicated that one of the consequences of global warming could be an increase in the frequency and duration of heat waves. As epidemiological studies have shown a relationship between temperature and mortality, with an increase in the risk of mortality associated with heatwaves, this subgroup set to focus on heatwaves and study how heatwaves and high temperature affect the risk of mortality, as well as how to define heatwaves. The subgroup discussed 6 projects but worked actively on the following 3 projects:

(i) *Heatwave definition:* Currently, there is not a formally universally accept definition of heat wave. This project aims to come up with a data-driven definition of heat wave, in terms of threshold as well as persistence (number of days a certain condition has to be satisfied) using threshold based stochastic process theory (similar to what is known in the literature as up crossings and down crossings times of martingales). Basically, given a stochastic process (say for apparent temperatures) at a location, this project will study (a) what functional of the stochastic process captures the two major aspects of a heatwave, viz., that it's over a given threshold (e.g., 110F) and that it sustained over that threshold for a given period of time (e.g., 5 days etc.); and (b) what are the statistical properties of the functional as we vary the threshold (say over locations).

(ii) *Improving inference on the effect of heatwave on health using data fusion of yearly and daily data:* Improve on the estimation of the county-level effect of heat on health by combining two health datasets: a daily dataset that reports mortality counts over a limited set of areal units (counties/cities) and a yearly dataset that reports mortality counts over a large set of areal units (counties). The idea is to leverage the finer temporal resolution of the first one and the greater spatial extent of the second one.

(iii) *Studying the impact of heat on human mortality using a detection and attribution framework:* Heat-mortality studies use observational data on mortality and temperature data (from weather stations) to estimate the relationship between heat exposure and health. Extrapolating the relationship to the future to project the mortality burden of heat involves using the output of climate models that come in the form of gridded output. This project aims to perform a detection and attribution type of analysis of the impact of temperature on mortality by using climate model output runs, both historical and future ones to study the relationship between temperature and mortality.

*Plans for future work*: Each of these three projects will result in a paper. Tentative titles are:

S. Raha, S. Ghosh, J. Bell and H. Chang. On the intensity and duration of heatwaves.

M. Winkel, B. Reich, J. Bell and V.J. Berrocal. Fusing multiple health datasets to improve inference on the effect of heatwaves on mortality.

V. J. Berrocal, R. Smith and M. Weiner. A detection and attribution analysis of the effect of temperature on mortality.

(3) **Mobile/wearable devices**: People are increasingly concerned with understanding their personal environment, including exposure to harmful air pollutants, and its effect on their health. They are interested in real-time information on a localized scale order to make informed decisions on their day-to-day activities. In addition, The US Environmental Protection Agency reports that over 45 million people live in close proximity to a major roadway. It is important to understand and communicate pollutant patterns on a finer scale, as microenvironment effects can be extremely varied due to factors such as meteorology and local traffic patterns. Fine-scale air quality measurements and methodology to provide real-time air pollution maps as well as short-term air quality forecasts on a fine-resolution spatial scale represent a paradigm shift in measurements and are vital to increasing public awareness. This is becoming even more relevant as portable sensors are increasingly more affordable leading to use by local governments and individual citizens.

The Mobile/Wearables Sensors Subgroup developed two projects working on methodology for very fine scale measurements. The data were provided from the original project leads. The Google Car project provides a unique source of highly detailed data with spatial and temporal complexities. It can provide information about commuter exposure, hot spots within high-trafficked city streets, as well as complex patterns due to meteorological effects and microenvironments. This fine-scale spatial and temporal information could also lead to the methodology and trends information needed to start to understand acute exposure. The DRIVE study (Dorm Room Inhalation to Vehicle Emissions) is a Georgia Tech and Emory University

coordinated study (Jeremy Sarnat co-lead). Conducted at GA Tech, GPS and PM2.5 data was collected every 10 seconds from 40 students who completed 2 sessions of 48-hr personal sampling, including activity logs and co-location check-ins with stationary Federal Equivalency Monitors. The Mobile/Wearable Sensors subgroup also included weekly biomonitoring. These very preliminary metabolomics could provide information such inflammatory responses and stress-related endpoints. A first step will be to evaluate these metabolomics as possible indicators for air pollution stress response.

*Plans for future work*: The google car project is wrapping up and a paper should be submitted in the summer of 2018. The tentative title and target journal are

> Johnson M, Guan Y, Messier K, Reich B, Katzfuss M, Song JJ, Mannshardt E. Using mobile monitors for fine-scale spatiotemporal air pollution analysis. To be submitted to *The Journal of the American Statistical Association*.

The Emory/GA Tech/DRIVE student has compiled a data set and finalize scientific objectives and analysis plan. The team is hiring a summer RA to continue moving the work forward.

(4) **Data fusion**: Understanding complex environmental processes often requires incorporating multiple data sources and sophisticated statistical methods to fuse the data while accounting for biases and uncertainties in each data source. For example, to map ambient air pollution one might combine data from monitoring stations, numerical model simulations, meteorological networks and satellite retrievals. The subgroup discussed several problems but focused on the three described below. (i) The first project is a review and comparison of current approaches to spatial modeling of ambient air pollution in the US. We compared prediction accuracy for several modeling approaches including linear regression, machine learning and geostatistics using different combinations data streams. We found that the best predictions came from a geostatistical model that uses numerical model output and station data but ignores other data sources. (ii) In a second project, we used geostatistical models for simultaneous prediction of multiple pollutants. A multi-pollutant approach is promising because the pollutants are strongly correlated and some are measured only sparsely. We proposed a novel spectral method that models correlation between pollutants and numerical models differently for different spatial resolutions and thus borrows information across data sources only when it is appropriate. We applied the proposed method to produce maps of daily speciated fine particulate matter across the US. (iii) Real-time numerical model forecasts of air quality in the vicinity of a forest fire are crucial for minimizing the impacts of these increasingly-common events. A major source of error in these forecasts is spatial misalignment of the forecasted smoke plume. To calibrate real-time forecasts we developed a Bayesian spatiotemporal model that includes a dynamic image-warping process to model and propagate spatial misalignment errors in the forecast.

*Plans for future work*: Each of these three projects will result in a paper. Tentative titles and target journals are

> Berrocal V, Guan Y, Muyskens A, Wang H, Chang HH, Reich BJ. Comparison of spatial data fusion methods for mapping fine particulate matter across the US. To be submitted to *Environmental Health Perspectives*.

Guan Y, Reich BJ, Chang HH. Multivariate spectral downscaling for speciated particulate matter. To be submitted to *Biometrics*.

Majumder S, Gaun Y, Rappold A, Reich BJ. Short-term forecasting of forest-fire-attributed particulate matter using forecast warping and smoothing. To be submitted to the *Journal of Agricultural, Biological and Environmental Statistics*.

In addition, Suman Majumder will base his PhD thesis on the forest-fire forecasting problem under the supervision of Brian Reich and Yawen Guan.

(5) **Causal inference**: Observational data are increasingly being used for causal inference in social science and public health studies. Outcome regression and various versions of propensity score analyses are the most commonly used parametric methods for causal inference. Propensity score is the probability that an individual would have been treated based on that individual's observed pretreatment variables. It is a one-dimensional variable that summarizes the multidimensional pretreatment covariates. Spatial confounding is that given the fixed covariates the spatial unmeasured random effects still affect both the propensity score and outcome. We developed rigorous Bayesian causal inference methods to account for spatial unmeasured confounding by including spatial random effects in the propensity score estimation. Our methods have shown great advantages over alternatives, however, we think our results are still biased possibly due to the "feedback" between the outcome stage and the propensity score stage in the full Bayesian estimation, so the group is currently working to remove the feedback effects and improve the Bayesian estimates of treatment effects.

*Plans for future work*: We are in the process of writing the manuscript, and the tentative title and target journal are

Song JJ, Berrocal V, Guan Y, Li B, Yang S. Spatial propensity score modeling for environmental health studies. To be submitted to *Biometrics*.

(6) **Infectious disease**: This group quickly dissolved and is not described further.

# SAMSI Working Group on Food Systems Final Report

June 1, 2018

## 1 Background

The agricultural establishment (farmers, agribusiness, and the agricultural research community) has made significant progress in its efforts to improve agricultural productivity and efficiency. Yet, with about one billion people hungry [12], two billion people with insufficient nutrients [25], and over two billion already overweight or obese [19], undernutrition and malnutrition are affecting more than half the world's population. Clearly, when enough food is produced but sizable fractions of the population suffer from malnutrition or are overweight, we need to get a better understanding of the global food system. These observations motivate the study of food systems in the SIAM Activity Group on Mathematics of Planet Earth (SIAG/MPE) and the formation of a Working Group (WG) on Food Systems as part of the SAMSI Climate Program in 2017–'18.

This report summarizes the activities and findings of the WG during the period August 2017–May 2018. Section 2 lists the participants, Section 3 the objectives, and Section 4 the activities of the WG; Section 5 summarizes the findings for each of the objectives listed in Section 3. Based on these findings, the WG has identified a number of projects for future research, which are listed in Section 6.

Note that the subject of "food systems" has many aspects and that the literature on the subject is vast. The participants of the WG are well aware of the fact that they have only scratched the surface and that many aspects have been left out of the discussion. This report reflects only the discussions that have occupied the WG during the reporting period. The authors make no claims of comprehensive coverage; the best we can say is that we have made significant progress learning about some aspects of "food systems."

## 2 Participants

The WG had seven active participants and several outside participants. The active participants joined the WG at the SAMSI Climate Opening Workshop in August 2017 and collaborated during the entire reporting period. The outside participants joined the WG occasionally to report on scientific activities and consult on technical aspects of food systems. The WG was supported by a postmaster, Huang Huang at SAMSI.

- Active WG members
  - Hans Engler* (Georgetown U)
  - Kaitlin Hill (U Minnesota)
  - Maggie Johnson (SAMSI)
  - Hans Kaper* (Georgetown U)
  - Adway Mitra (IIT Bhubaneswar, India)
  - Jim Rosenberger (Penn State U)
  - Mary Lou Zeeman (Bowdoin College)

  * Coordinator
- Consultants (incl. MPE 2013+ workshop participants)
  - Patrick Canning (USDA/ERS)
  - Gidon Eshel (Bard College)
  - Jesse Milzman (U Maryland College Park)
  - Richard Sowers (U Illinois Urbana-Champaign)
  - Frederi Viens (Michigan State U)
  - Han Wang (Michigan State U)

# 3 Objectives

The WG agreed on the following objectives for 2017–'18:

- Become familiar with the literature
- Explore various data sources
- Identify themes and problems of interest
- Explore applications of mathematics and statistics

Findings for each of these objectives are presented in Section 5.

# 4 Activities

The WG met regularly during the fall and spring semesters, usually on Thursday mornings, 10:00-11:00 a.m. (ET), using the SAMSI WebEx account. The WG maintained a record of its activities and findings on Google Docs.

The WG organized a two-day workshop at Georgetown U (April 21-22) to discuss progress and identify challenging issues. The workshop was attended by several of the outside participants (listed under Consultants in Section 2). The workshop was supported by a grant from the MPE2013+ program at DIMACS (Rutgers U).

# 5 Findings

A schematic representation of a food system is shown in Figure 1. It shows the three levels

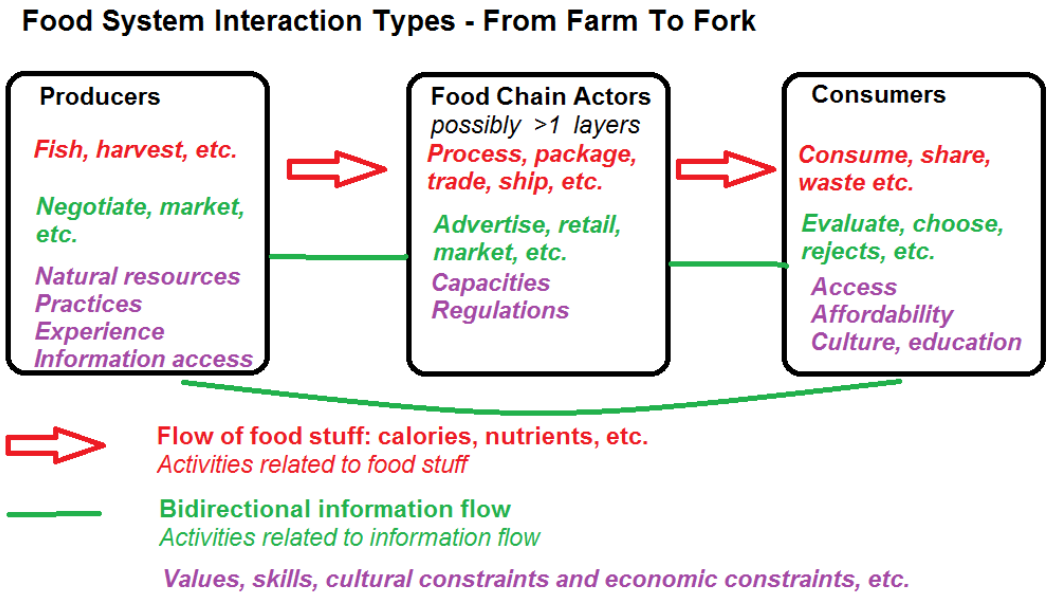



Figure 1: Schematic representation of a food system (generated by the WG).

of participants: producers, food chain actors, and consumers. Food stuff generally flows from producers to food system actors to consumers, as indicated by the red arrows, while information flows along the green lines in both directions between the three levels. Food stuff can be characterized in various ways; for example, by its energy content (measured in calories), its nutrient content (grams of protein), its mass, volume, monetary value, etc. Information is much more difficult to characterize; it embodies the laws of supply and demand and involves decision making at every level—decisions which often reflect cultural constraints and individual preferences.

The following sections summarize the findings for each of the initial objectives listed in Section 3. Based on these findings, the WG has developed some ideas for future research projects, which are listed in Section 6.

## 5.1 Literature Review

The WG considered many articles that were judged to be relevant to food systems and food system modeling. References to specialized topics are given in Section 5.3. Material of general interest can be found, for example, in Refs. [1, 2, 12, 14, 19, 25, 27].

In the course of its investigations, the WG has gathered a number of web sites with definitions and descriptions of commonly used concepts:

- Food energy, https://en.wikipedia.org/wiki/Food_energy

- Food groups, https://en.wikipedia.org/wiki/Food_group
- Food and food groups, https://www.mynetdiary.com/food-catalog/
- Food nutrients (tables), https://en.wikipedia.org/wiki/Table_of_food_nutrients
- Food security, https://en.wikipedia.org/wiki/Food_energy
- Nutrient, https://en.wikipedia.org/wiki/Nutrient
- Nutrient density, https://en.wikipedia.org/wiki/Nutrient_density
- Nutrition, https://en.wikipedia.org/wiki/Nutrition#Macronutrients
- Pasture-based cattle farming, http://www.grass-fed-solutions.com/
- Interstate livestock movement, https://www.ers.usda.gov/webdocs/publications/37685/15376_ldpm10801_1_.pdf?v=41022

## 5.2 Data Sources

Food system modeling is a data-driven activity. There are numerous data sources on all possible topics. They are heterogeneous, in various formats, and spread all over the web. Here are some of the data sources that the WG found useful:

- The US Department of Agriculture/Economic Research Services (USDA/ERS, https://www.ers.usda.gov) tracks interstate trade of agricultural products, including food and live animal trade.
- The United Nations Food and Agriculture Organization (FAO, http://www.fao.org/home/en/) tracks international trade of agricultural products, http://www.fao.org/faostat/en/#home
- Many state and local agencies maintain useful data bases; for example, the Texas Department of Agriculture publishes reports and statistical data on its web site https://www.texasagriculture.gov/About/TexasAgStats.aspx, and the city of Austin, Texas, publishes its own reports at http://www.austintexas.gov/page/austins-food-system-research-reports. Montgomery County MD has a visualization tool for showing dependencies between demographic and economic diversity and food security at https://countystat.maps.arcgis.com/apps/webappviewer/index.html?id=099052a140cd4bb38e99cbeb870ebce0.
- Non-government organizations (NGOs) such as producer associations (example: Texas Farm Bureau, http://texasfarmbureau.org/), associations of food chain actors, and not-for-profit organizations such as food banks (example: Feeding America, http://www.feedingamerica.org/research/) are useful sources of data and background information.

- Most land grant universities have academic units whose mission includes the study of food systems. Johns Hopkins University has a Center for a Livable Future, which lists a number of Food Policy Networks at http://www.foodpolicynetworks.org/.

- Scraping the web is a useful technique to obtain information that is not directly available in published form. An example is found in [5], where data on interstate live cattle trade were obtained by scraping online records of cattle auction houses.

- Data and appealing visualizations on international development issues of all kinds are available at the Gapminder website, https://www.gapminder.org/.

## 5.3 Themes and Problems of Interest

The WG identified several recurring themes and issues:

- **Economic modeling**
  - Input/output models (I/O). These are static (fixed time) or quasistatic (discrete time) models for the activity of sectors of the economy. They typically take the form of matrix equations with moderately large sparse coefficient matrices that are obtained by observing economic activity. Their granularity can be adjusted, and they are capable of predicting price structures and responses to small changes. The analysis is sometimes done with spreadsheets, resulting in a lack of transparency in published results.
  - Computable General Equilibrium models (CGEs). These are nonlinear versions of I/O models that take the behavior of economic agents into account and can incorporate the effects of external factors such as changes of market rules or environmental change. They typically use advanced techniques of linear algebra and optimization and often employ specialized software, resulting in "black box" characteristics that make them difficult to analyze [7, 13, 15, 16].

- **Mathematical modeling**
  - Food systems can be modeled at many levels of granularity. For example, one may consider the flow of energy (measured in calories) or of nutrients (measured in grams of protein) in a given food system, or one may consider food systems for individual commodities or food groups, as in life cycle analyses. Particularly difficult to model is the flow of information, which is an integral part of every food system—for example, information about current market conditions or shifting consumer preferences.
  - Agent-based models (ABMs). Such models employ computational models for "actors" with limited view of their environment and some limited rationality. They are particularly suitable when the system involves decisions based on individual preferences and cultural norms [17, 18, 26]. They are used in computational simulations and the results may then be analyzed statistically or in visualizations.

- Equation-based models (EBMs). These models follow the well-known modeling paradigm for physical systems that are completely determined by a set of state variables whose evolution is governed by physical laws, usually resulting in a deterministic model for a set of processes that is described by differential or difference equations.
- Statistical models (SMs). Such models describe aggregate properties of populations, similar to EBMs. The emphasis is on quantifying uncertainties due to errors in observational data, missing data, and effects of sub-scale phenomena that are not captured on the scale of the model. Typically, a process model is assumed to be given, and the emphasis includes stochastic modeling of the data, prediction of bulk behavior, and the assessment of uncertainties.

- **Themes**

  - Beyond standard economic theory. Food systems have been studied primarily from the perspective of economics, with emphasis on maximizing profit at some level or Gross Domestic Product (GDP) at the national level. The WG studied the new concept of "doughnut economics," first proposed by Raworth [21] to frame the discussion of concepts and goals that transcend national boundaries and integrate global constraints into economic modeling. The schematic in Figure 2 shows how the space for human action (green "doughnut") is constrained by an outer ring of global environmental processes and limitations and an inner ring of social objectives. Doughnut economics thus provides a framework for sustainable development within the planetary boundaries while aiming for social justice for all.

    Doughnut economics also embodies the Social Development Goals (SDGs) adopted by the United Nations, https://www.un.org/sustainabledevelopment/sustainable-development-goals/. This set of goals appears to be shaping the scientific debate in development economics and mathematical scientists need to be familiar with it.
  - Complex networks. Food systems are complex systems. They involve multiple layers (producers, food chain actors, consumers) and multiple scales in time (days, months, years) and space (individual, local, regional, national, international), with many feedback mechanisms (market mechanism, individual preferences, cultural norms). Globalization leads to increasing interconnectedness.
  - Network structure and dynamics. Existing methods refer mostly to static networks; food systems are inherently dynamic. What are the characteristics of a food system network, and how do they compare to, say, the characteristics of a power grid or a small-world network?
  - Environmental impact of diet choices. As the level of development of a society increases, diet preferences tend to shift toward an animal-based diet (beef, pork, poultry), resulting in increasing demands on the environment [8, 9, 10, 11, 23, 24]. What does a nutritionally equivalent plant-based diet look like? Are there

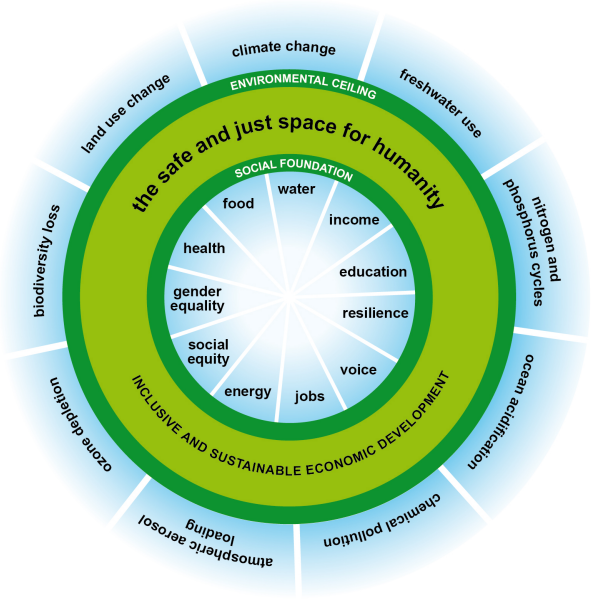



Figure 2: Doughnut economics.

pathways to shift from an animal-based diet to a plant-based diet? If so, how does one encourage such a shift?

- Food loss and food waste. Americans waste approximately 40% of all food that is fit for consumption. Not-for-profit food banks have been established to distribute surplus food to people in need, thus reducing food insecurity. Issues: how to collect food in sufficient quantities, how to distribute food equitably, given available transportation resources. Optimization techniques borrowed from logistics modeling have been applied successfully [20]. Can we adapt methods from financial mathematics?

- **Case studies**

  - Beef production systems. The WG studied several articles analyzing the characteristics and environmental impact of three systems: the conventional system (CON, beef cattle fed in feedlots with growth-enhancing technology), the natural system (NAT, beef cattle fed in feedlots without growth-enhancing technology), and the grass-fed system (GFD, free-ranging beef cattle) [3, 4]. No clear conclusions emerged, but the WG developed an interesting hybrid population model for a herd consisting of both breeding cattle and cattle raised for slaughter. This study may lead to a publication.

  - Livestock transportation networks. There are data on cattle movement from farm to auction houses to slaughter houses at various levels of granularity: per region, state, or county in the US. These data have been used, for example, to study propagation of disease among cattle [5].

- Global wheat transportation network. The FAO maintains data on global food transportation. These data have been used in the case of wheat to characterize transportation and trade networks and their dynamics and assess their response to shocks and attacks [11].
- Replacing an animal-based diet by a nutritionally equivalent plant-based diet. Our study followed mostly the articles by Eshel and collaborators who considered the environmental impact of such hypothetical changes as well as the range of possible pathways to achieve such a goal [8, 9, 10, 22].

# 6 Future Plans

The WG has identified a number of possible projects for further exploration. We suggest to use existing mathematical tools for networks (items 1 and 2), address issues of resilience (2 and 3), employ combinations of agent based and equation based models for promising tasks (3 and 4), and continue existing promising work by other researchers (5 and 6).

1. Investigate whether global food systems for different commodities can be classified by their network structure. In which way do the networks differ from, say, a network representing the power grid? In which way are they similar? Do the same for food systems at the national, regional, and local level. Which network features persist across different scales? Are there any apparent scaling laws? Is it possible to rank the network structures of different food systems in some hierarchical manner? Are there features that require new mathematics?

2. Use the network structure of a food system to characterize its resilience to perturbations, broadly defined. For example, characterize the resilience to climate change (a gradual perturbation) by evaluating the impact of past events like changing patterns of climate extremes on network connectivity, or the resilience to shocks (a sudden perturbation) by evaluating the impact of disruptions in a supply chain.

3. Investigate the resilience of various food systems to disease-based disturbances by modeling disease transmission across the network, using an equation-based model (dynamical systems approach) and an agent-based model. How does disease spread, and how does the network recover? What roles do decision makers play in the spread of disease across the food system? Are the results using EBMs and ABMs comparable? Are there advantages of using one model over the other?

4. Use agent-based models to explore various policy options to encourage farmers to use sustainable farming practices, or to encourage consumers to shift from an animal-based diet to a plant-based diet.

5. Evaluate optimal distribution strategies for deliveries by food banks to local distribution centers, based on variable supply and demand. This continues work in [20].

6. Investigate the environmental impact of different strategies (CON, NAT, GFD) to raise beef cattle in the presence of dynamic climate change. This continues work in [8].

# References


[1] Allen, T. and P. Prosperi, *Modeling Sustainable Food Systems*, Environmental Management (2016) 57:956–975 DOI:10.1007/s00267-016-0664-8.

[2] Beddington J., M. Asaduzzaman, A. Fernandez, M. Clark, M. Guillou, M. Jahn, L. Erda, T. Mamo, B.N. Van, C.A. Nobre, R. Scholes, R. Sharma, and J. Wakhungu, *Achieving food security in the face of climate change: Summary for policy makers from the Commission on Sustainable Agriculture and Climate Change*, CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS). Copenhagen, Denmark. http://www.ccafs.cgiar.org/commission.

[3] Capper, J.L. (2011). *The environmental impact of beef production in the United States: 1977 compared with 2007.* Journal of Animal Science, **89** (12), 4249-4261.

[4] Capper, J.L., *Is the Grass Always Greener? Comparing the Environmental Impact of Conventional, Natural and Grass-Fed Beef Production Systems*, Animals **2** (2012) 127–143, DOI:10.3390/ani2020127.

[5] Carroll, I.T. and S. Bansal, *Livestock market data for modeling disease spread among US cattle*, (2015) http://dx.doi.org/10.1101/021980.

[6] Carter, N., S. Levin, A. Barlow, V. Grimm, *Modeling tiger population and territory dynamics using an agent-based approach*, Ecological Modeling **312** (2015) 347–362.

[7] Dixon, P.B. and B.R. Parmenter, *Computable general equilibrium modelling for policy analysis and forecasting*, Handbook of computational economics, **1** (1996)3–85.

[8] Eshel, G., A. Shepon, T. Makovc, and R. Milo, *Land, irrigation water, greenhouse gas, and reactive nitrogen burdens of meat, eggs, and dairy production in the United States*, PNAS **111** (2014) 11996–12001, DOI: 10.1073/pnas.1523119113.

[9] Eshel, G., A. Shepon, T. Makovc, and R. Milo, *Environmentally optimal, nutritionally aware beef replacement plant-based diets*, Environmental Science and Technology, **50** (2016) 8164–8168.

[10] Eshel, G. and Y. Carmel, *Expanded view of ecosystem stability: A grazed grassland case study*, PLoS ONE **12** (6): e0178235 https://doi.org/10.1371/journal.pone.0178235.

[11] Fair, K.R., C.T. Bauch, and M. Anand, *Dynamics of the Global Wheat Trade Network and Resilience to Shocks*. Scientific reports, **7** (2017), 7177.

[12] FAO, IFAD, UNICEF, WFP and WHO, *The State of Food Security and Nutrition in the World 2017. Building resilience for peace and food security*, Rome, FAO.

[13] *General Algebraic Modeling System (GAMS)*, (2017), GAMS Software GmbH. P.O. Box 40 59. 50216 Frechen, Germany, https://www.gams.com/docs/intro.htm.

[14] Garnier, J. and M.A. Lewis, *Expansion Under Climate Change: The Genetic Consequences*, Bull Math Biol (2016) 78:2165–2185 DOI 10.1007/s11538-016-0213-x.

[15] Gillig, D. and B.A. McCarl, *Introduction to Computable General Equilibrium Model (CGE)*, Course Notes (no date), Dept. of Agricultural Economics, Texas A&M University.

[16] Harrison, W.J. and K.R. Pearson, *Computing solutions for large general equilibrium model using GEMPACK*, Computational Economics, **9** (1996) 83–127.

[17] Macal, C.M. and M.J. North, *Tutorial on agent-based modelling and simulation*, Journal of Simulation **4** (2010) 151–162.

[18] Miller, B.W., I. Breckheimer, A.L. McCleary, L. Guzmán-Ramirez, S.C. Caplow, J.C. Jones-Smith, and S.J. Walsh, *Using stylized agent-based models for population-environment research: a case study from the Gal'apagos Islands*, Population and Environment, **31** (2010), 401–426.

[19] Ng, M., T. Fleming, M. Robinson, B. Thomson, N. Graetz, C. Margono, E.C. Mullany, S. Biryukov, C. Abbafati, S.F. Abera, et al., *Global, regional, and national prevalence of overweight and obesity in children and adults during 1980–2013: A systematic analysis for the Global Burden of Disease Study 2013*, The Lancet, **384** (9945) (2014), 766–781.

[20] Orgut, I.S., Ivy, J. Uzsoy, R., Hale, C., *Robust optimization approaches for the equitable and effective distribution of donated food*, European J. of Operational Research, **269** (2018), 516–531.

[21] Raworth, K., *A safe and just space for humanity. Can we live within the doughnut?* Oxfam Discussion Paper, February 2012.

[22] Shepon, A., G. Eshel, E. Noor, and R. Milo (2018). *The opportunity cost of animal based diets exceeds all food losses*. Proceedings of the National Academy of Sciences, 201713820.

[23] Springmann, M., H.C.J. Godfray, M. Rayner, and P. Scarborough, *Analysis and valuation of the health and climate change cobenefits of dietary change*, PNAS **113** (2016) 4146–4151, DOI: 10.1073/pnas.1523119113.

[24] Tilman, D. and M. Clark, *Global Diets Link Environmental Sustainability and Human Health*, Nature (2014) 515:5180522, DOI:10.1038/nature13959.

[25] Tulchinsky, T.H., *Micronutrient Deficiency Conditions: Global Health Issues*, Public Health Reviews (2010) 32:BF03391600, https://doi.org/10.1007/BF03391600.

[26] Van Dyke Parunak, H., R. Savit, and R.L. Riolo, *Agent-Based Modeling vs. Equation-Based Modeling: A Case Study and Users' Guide*, Proceedings of Multi-agent systems and Agent-based Simulation (1998), 10–25, Springer, LNAI 1534.

[27] Zhang, D.D. H.F. Lee, C. Wang, B. Li, Q. Pei, J. Zhang, and Y. An, *The causality analysis of climate change and large-scale human crisis* PNAS **108** (2011) 17296–17301, http://www.pnas.org/cgi/doi/10.1073/pnas.1104268108.

# SAMSI Ice Dynamics Working Group: Annual Report
# May 2018

**1. Name of working group and program.**

Working group: Ice Dynamics
Program: Mathematical and Statistical Methods for Climate and Earth Systems

2. **Main WG participants and their affiliations (including academic departments) WG leader(s):**

WG leader: Murali Haran, Penn State University Statistics

Webmaster(s): Christian Sampson, SAMSI/UNC Chapel Hill, Applied Math

Post-doctoral fellows:
Yawen Guan, SAMSI/NC State, Statistics
Christian Sampson, SAMSI/UNC Chapel Hill, Applied Math

Graduate Student Members:
Colin Guider, UNC Chapel Hill, Mathematics

Active Members:

- Amit Apte, apte@icts.res.in, Tata Institute of Fundamental Research (TIFR), International Centre for Theoretical Sciences
- Jenny Brynjarsdottir, jenny.brynjarsdottir@case.edu Case Western Reserve University, Mathematics, Applied Mathematics and Statistics
- Won Chang, wonchang3@gmail.com, University of Cincinnati, Mathematical Sciences
- Yawen Guan, yawenguan@gmail.com, SAMSI/NC State Statistics
- Alex Bledar Konomi, alex.konomi@uc.edu University of Cincinnati, Mathematical Sciences
- Anirban Mondal, anirbanstat@gmail.com, Case Western Reserve University, Mathematics, Applied Mathematics and Statistics
- Joel Upston, jupston@math.unm.edu, University of New Mexixo, Mathematics.
- Christian Sampson, christian.sampson@gmail.comSAMSI/UNC Chapel Hill Applied Math
- Deborah Sulsky, sulsky@math.unm.edu University of New Mexico Mathematics and Statistics

**3. Topics and goals of the WG**

*Project 1: Developing metrics to evaluate sea ice models*

The main goal of this research project is to study the behavior of a sea ice model (developed by group member Deborah Sulsky) as a function of its key parameters, and compare the model to

observations. There are many challenges in this work. The model output and the observations are images with a complex structure; it is difficult to determine appropriate features to use to evaluate how well a model output is reproducing observations. For instance the leads (large fractures) in the ice sheet present features that are non-smooth. Aligning the images to the observations is also difficult. We have started studying good metrics for comparing sea ice model output to observations of concentration and ice thickness data. The methodology we are development is based on some recent work in warping functions and field-deformation approach. The warping functions will be useful for developing a distance between the observations of sea ice and model output at various parameter settings. This provides an approach to inferring model input/parameter settings. This work will be led by SAMSI postdoc Yawen Guan with SAMSI postdoc Christian Sampson.

*Project 2: Calibrating a model with zero-inflated positive spatial data*
This project will focus on developing a formal sea ice model calibration framework. The model takes two days to run for each parameter setting, so some form of emulation (approximation) of the computer model is important. Standard Gaussian process approaches for emulation may not work here due to the nature of the model output and observations. In particular, sea ice thickness results in lots of zeroes (no ice) and the rest of the thickness values are positive and continuous, spatially dependent data. No methodology exists for this problem, especially for high-dimensional data. We are working on a dimension-reduction approach for this problem, a two-stage model for the binary ice-no ice pattern and for ice thickness where there is ice. This project is led by group member Won Chang.

*Project 3: Adaptive design for studying a sea ice model*
This project will involve efficient adaptive design to obtain well selected points in parameter space. This work will be led by group member Alex Konomi.

**4. Projects/topics chosen by postdocs**.
SAMSI postdoc Yawen Guan is leading Project 1 jointly with postdoc Christian Sampson. She is also an active participant in Project 2. Postdoc Christian Sampson is working on a related problem in a separate project to melt ponds; the melt ponds project is currently listed under the data assimilation working group.

**5. Sources of data**
Sea ice model runs from group member Deborah Sulsky. Observations from various sources, mostly from satellites, will be used for verification and calibration. The data are available for download from NSIDC, but in some cases require additional processing in order to compare with modeled data. Specifically, we are using ice concentration, ice thickness and ice motion data sets.

**6. List of past as well as possible future discussion or presentation topics in WG meetings.**

We gave 2 talks at the CLIM Opening Workshop, August 21-25, 2017
Murali Haran: ``Some Statistical Challenges in Studying the West Arctic Ice Sheet”

Deborah Sulsky: ``Modeling Arctic Sea Ice”

We gave 4 talks in a session in the SAMSI Climate Transition workshop on May 14-16, 2018
Murali Haran: “Overview of Ice Dynamics Working Group”
Deborah Sulsky: ``Overview of Sea-Ice Modeling and Statistical Challenges''
Won Chang: ``Ice Model Calibration using Zero-Inflated Continuous Spatial Data''
Yawen Guan: ``Arctic Sea Ice Plays an Important Role in the Global Climate''

Deborah Sulsky will make a presentation at WCCM2018, July 22-27, 2018, New York, NY in the Minisymposium on ``Geomechanics of Land and Sea Ice”

Deborah Sulsky is organizing organizing a minisymposium on Geoscience and Natural Disasters for the USACM conference on Meshfree and Particle Methods: Applications and Theory (http://mfpm2018.usacm.org). The conference is September 10-12, 2018 in Santa Fe, NM.

**7. Any other past or proposed activities**

The working group has been meeting weekly throughout the program. We intend to continue meeting with the same frequency for the next few months and beyond, focusing on turning the above projects into manuscripts. In particular, we intend to submit a manuscript by August 2018 on ice metrics with calibration of ice strength parameters to the *Journal of Agricultural, Biological, and Environmental Statistics (JABES).*

A second paper on the methods developed to handle zero-inflated, positive spatial data is also planned. This paper will focus on concentration and thickness fields and their use in calibrating albedo parameters.

An additional project came out of discussions between Won Chang and Deborah Sulsky at the Transition Workshop. This paper would extend Chang, et al. (2016), which studies precipitation patterns, to identifying leads in Arctic observations and forecasts, and studying their physical characteristics (size, orientation, persistence) and their statistics.

We note that these are all new collaborations (between statisticians and applied mathematicians) that are a direct result of this SAMSI program. A white paper is being planned for submission in response to a call from the ONR MURI program (topic: Advanced Analytical and Computational Modeling of Arctic Sea Ice) due in June.

,

# Final Report of the Detection and Attribution Working Group

**Dorit Hammerling, Matthias Katzfuss and Richard Smith**

1. **Name of working group and program.** Detection and Attribution; Climate program.

2. **Main WG participants and their affiliations:**

WG leader(s): Dorit Hammering (NCAR, now at Colorado School of Mines), Matthias Katzfuss (Statistics, Texas A&M), Richard Smith (UNC and SAMSI)

Webmaster: Maggie Johnson (SAMSI)

Other active members included: Aurelien Ribes (Meteo France), Philippe Naveau (INRIA, France), Alexis Hannart (Ouranos, Montreal), Charles Jackson (University of Texas), Nils Wietzel (University of Bonn), Nathan Lenssen (NCAR), Veronica Berrocal (University of Michigan), Adway Mitra (ICTS-TIFR), Sandy Burden (University of Woolongong)

3**. Topics and goals of the WG**

Detection and attribution refers to the class of statistical techniques used in statements of the form "at least X% of the observed global warming since 1950 (or some other year) is due to human-caused greenhouse gas warming" (usually with some associated confidence or uncertainty statement). The method as currently used in the climate literature was defined in a series of papers in the climate science literature within roughly the period 1996-2006, but they raise many statistical questions associate with problems such as regression in very high-dimensional spaces, "errors in variables" and related techniques to take account of the uncertainty of projections derived from a single climate model, and multi-model ensembles to take account of the variability among different climate models. There is also a related, but largely separate, literature on the "single event attribution" problem which is often applied to the case of individual extreme events, e.g. to what extent can the very heavy rainfalls associated with Hurricane Harvey, in August 2017, be considered to have been exacerbated by greenhouse-gas warming? Recent statistical developments include links to the causal inference framework of Judea Pearl and others, and an attempt to reformulate detection and attribution in the language of Bayesian hierarchical models. The aims of the present group can be loosely summarized as trying to extend and strengthen the statistical foundations of detection and attribution by exploring these and other methodological developments in more detail, and to apply the methods to new datasets to better understand how these methodological issues play out in applications.

4. **If applicable: Projects/topics chosen by postdocs and SAMSI grad fellows. Please provide a short description of the individual research projects**

Maggie Johnson is exploring an alternative and more detailed Bayesian approach which uses ideas derived from some papers of Bhattacharya and Dunson to model high-dimensional datasets in a Bayesian factor analysis framework.

Another idea suggested by Veronica Berrocal is to use a principal components framework similar to what is already in the detection and attributions literature, but with Lasso-like penalties.

A third idea suggested by Matthias Katzfuss is to exploit recent developments in the literature on spatial statistics for the spatial analysis of high-dimensional datasets.

The group has not so far had a detailed discussion of the single-event attribution problem, but here also there are possibilities for using Bayesian hierarchical approaches for combining predictions of different models, or for more detailed connections with extreme value theory and the work of the Extremes group in this program.

## 5. **Sources of data**

The CMIP5 dataset is readily available as a public source of climate model output. Michael Wehner (Berkeley Lab) has proposed a specific problem related to US temperature series and has promised to help us compile the relevant data to look at this.

## 6. **Outcomes**

Three projects are ongoing and were the subject of presentations at the Transition Workshop.

First, Dorit Hammerling reviewed the theory of detection and attribution methods and the traditional approach via linear regression. She then outlined a Bayesian approach that had been developed in a paper by Katzfuss, Hammering and Smith (Geophysical Research Letters, 2017). In the final part of her presentation she outlined a new approach, being developed in collaboration with Maggie Johnson and Richard Smith, that extended the Bayesian approach by incorporating modern ideas of Bayesian factor analysis (Bhattacharya and Dunson, Biometrika 2011).

The second presentation was from Nathan Lennsen and described the development of a testbed of climate models and observations that could be used to evaluate statistical methods for detection and attribution. This is joint work performed at NCAR with input from Dorit Hammerling and Alexis Hannart. It is possible that this project will be combined with the previous one for eventual publication.

The third presentation was by SAMSI postdoc Huang Huang in collaboration with Dorit Hammerling, Bo Li and Richard Smith. The objective is to come up with a new method for multi-model ensembles when there is dependence among the ensemble members. This is potentially valuable in extending detection and attribution approaches to cases that involve multiple models with possibly different physical assumptions. Huang has continued to work on this project during the 2018-19 year at NCAR.

# Final Report for the Risk and Coastal Hazards Working Group

## 1. Name of working group and program

Program on Mathematical and Statistical Methods for Climate and the Earth System (CLIM), working group on Climate Risks and Coastal Hazards.

## 2. Main WG participants and their affiliations (including academic departments)

**WG leaders:**

Brian Blanton (Environmental Initiatives, RENCI)
Vyacheslav Lyubchich (Chesapeake Biological Laboratory, UMCES)
Richard Smith (SAMSI)

**Post-docs:**

Whitney Huang (Statistics, University of Chicago)

**Graduate Students:**

Asim Dey (Math Sciences, University of Texas at Dallas)

**Other active members:**

Jesse Bell (Cooperative Institute for Climate and Satellites-NC, NC State University)
Lelys Bravo de Guenni (Statistics, Northern Illinois University)
Robert Erhardt (Mathematics and Statistics, Wake Forest University)
Yulia Gel (Math Sciences, University of Texas at Dallas)
Chris Lenhardt (Marine Science, UNC Chapel Hill)
Wei Mei (Marine Sciences, UNC Chapel Hill)

## 3. Topics and goals of the WG

The WG has a focus on the end users of various climate, weather, and social models. Our goal is to bridge the gap between academic research and decision-making by elaborating approaches that can potentially provide meaningful information to the managers dealing with the impacts of climate change. Some of the specific topics are:

1. What do actuaries need from climate/weather/hazard models?
    a. What kind of data do they expect climate science to provide
    b. Uncertainty in climate model predictions? Reliability of predictions.
    c. How do climate models relate to changing probabilities of the drivers of losses?
2. How could insurance markets drive behavior on climate adaptation?

. What increases in risk could this group show that would flow through insurance rates and incentivize/disincentivize certain community-level behaviors (building codes, coastal development, etc.)?

a. They need to know how flood, hurricane probabilities are changing. How does this knowledge impact underwriting?

3. Nexus between future climates, risk management, and societal behavior.
4. Sensitivity of decisions to variance in climate models.

**4. If applicable, projects/topics chosen by postdocs and SAMSI grad fellows.**

Whitney Huang: *Modeling storm surge using a combination of physical and statistical approaches.* The project targets statistical issues with the so called joint probability method, the recommended approach by Federal Emergency Management Agency (FEMA) for tropical storm surge analysis. Collaborators: Richard Smith, Rick Luettich (UNC IMS), Brian Blanton, and Wei Mei.

Asim Dey: *Two-stage modeling of weather-induced home insurance risks with support vector machine regression*. The project explores the approach of two-stage analysis of distribution of claims and losses and their relationships with weather factors. Downscaled climate model projections are used to show possible risks due to climate change associated with different climate scenarios. Collaborators: Yulia Gel and Vyacheslav Lyubchich.

**5. Sources of data**

Observed data from: International Best Track Archive for Climate Stewardship (IBTrACS), NOAA's Severe Weather Data Inventory (SWDI) integrated database, National Climate Data Center data through NOAA's Climate Data Online (CDO) archive, and Environment Canada.

Model output from: ADCIRC runs in the coastline of North Carolina, North American Coordinated Regional Climate Downscaling Experiment (NA-CORDEX).

Actuarial data from: Verisk.

**6. Outputs of the working group**

Working group member Robert Erhardt was chief organizer of a workshop entitled “The Nexus of Climate Data, Insurance, and Adaptive Capacity” that took place in Asheville, NC, on November 9. Group members Jesse Bell, Brian Blanton and Richard Smith were also on the organizing committee, that developed out of the discussions within this working group but which also included representatives from the insurance industry and government. The workshop was funded by a separate grant from the National Science Foundation and attracted nearly 60 participants.

Lelys Bravo de Guenni introduced the group to the SHELDUS dataset (https://cemhs.asu.edu/sheldus) and completed a pilot project based on that dataset for South Carolina (Bravo de Guenni, paper in preparation). She has proposed am ore substantial project joint with other members of the group.

Visiting Fellow Slava Lyubchich and visiting graduate student Asim Dey, together with Dey's advisor Yulia Gel, worked on a project modeling home insurance claims from heavy precipitation events (Dey, Gel and Lyubchich, forthcoming).

Some of the group's work is continuing through a SAMSI working group on Storm Surge Hazard and Risk. This is part of the Model Uncertainty Mathematics and Statistics (MUMS) program, ongoing for 2018-19. Whitney Huang (SAMSI/CANSSI postdoc, now at University of Victoria) and Taylor Asher (UNC-Chapel Hill) are co-leaders of this group.

# Final Report of the SAMSI Working Group on Statistical Oceanography
# Michael Stein and Mikael Kuusela

1. Name of working group and program:

Statistical Oceanography, Program on Mathematical and Statistical Methods for Climate and Earth Systems

2. Main WG participants and their affiliations (including academic departments):

WG leader(s):
Michael Stein (Department of Statistics, University of Chicago)

Webmaster(s):
Mikael Kuusela (SAMSI / Department of Statistics and Operations Research, UNC Chapel Hill)

3. Active participants

Post-docs:
Mikael Kuusela (SAMSI / Department of Statistics and Operations Research, UNC Chapel Hill), Huang Huang (SAMSI / Department of Statistical Science, Duke University), Donata Giglio (Scripps Institution of Oceanography, University of California San Diego)

Graduate students:
Pulong Ma (Department of Mathematical Sciences, University of Cincinnati), Nils Weitzel (Meteorological Institute, University of Bonn)

Other active members:
Amit Apte (International Centre for Theoretical Sciences, Tata Institute of Fundamental Research), Fred Bingham (Department of Physics and Physical Oceanography, UNC Wilmington), Sarah Gille (Scripps Institution of Oceanography, University of California San Diego), Alison Gray (School of Oceanography, University of Washington), Jong-June Jeon (Department of Statistics, University of Seoul), Bo Li (Department of Statistics, University of Illinois at Urbana-Champaign), Matthew Mazloff (Scripps Institution of Oceanography, University of California San Diego), Anirban Mondal (Department of Mathematics, Case Western Reserve University), Dean Roemmich (Scripps Institution of Oceanography, University of California San Diego)

4. Topics addressed by the WG:

The overarching theme of this working group was to develop new statistical methods for analyzing data from Argo profiling floats. Argo is an array of almost 4000 floats that measure ocean temperature and salinity in the upper 2000 meters on a global scale. Argo has transformed physical oceanography by bringing the field to a new data-rich era. At the same time, these data are a rich source of fascinating

statistical problems. This working group has brough together statisticians and oceanographers to tackle some of these mutually interesting challenges.

A fundamental quantity in understanding climate change is the heat content of the ocean and how it changes over time. Although Argo floats do not go to the bottom of the oceans, they do provide a rich source of information for estimating this quantity as it varies over seasons and years and for providing defensible uncertainties for these temporal patterns. This work was the main project for Mikael Kuusela and more details on this work are given below.

Motivated in large part by the ocean heat content problem, another area that the group planned to pursue was the study of variation in ocean characteristics in the vertical dimension. However, as described below, we realized early in the year that the structure of the Argo data made it unnecessary to model this vertical structure in order to get sensible estimates of trends in ocean heat content. Thus, this problem was largely put on hold over the year, but it is a fundamental problem in both oceanography and spatial statistics and is a topic that group members will hopefully address, either individually or in groups, in the coming years.

Some newer Argo floats have sensors that measure biogeochemical variables, opening up a host of new scientific questions. Because these variables covary with temperature and salinity, there should be scope developing multivariate spatio-temporal models that could combine information from these new floats with the broader network of floats measuring temperature and salinity. Donata Giglio has been developing machine learning methods to estimate oxygen levels using this data and Huang Huang (see below) has begun exploring whether methods that more explicitly use the spatial information can lead to better estimates.

An important topic the group hoped to address was to develop statistical and dynamical approaches to mapping the flow field of the oceans from Argo data. Argo data provide two very different sources of information on these flows: the direct information based on the movement of the floats and the indirect information from temperature and salinity measurements, which determines the density, which in turn determines the flows at all depths once the flows at a single depth are known. Data assimilation methods in principle provide a natural way to take account of all of the available information, but there are great conceptual and computational challenges in developing such an approach. We had initially hoped to work with people in the data assimilation group to develop methods along these lines, but the two groups separately had more than enough to work on to make such a collaboration feasible. It could be worthwhile to consider including this topic in some appropriate future SAMSI program, for example, one dedicated to methods for combining multiple sources of data in a single analysis. Participants from Scripps and the University of Washington have continued to work on simpler statistical approaches to flow mapping and discussed this work during several group meetings. The statisticians contributed by providing ideas for better handling of the statistical mapping problem. Alison Gray is currently experimenting with the implementation of some of these ideas for mapping the flow in the Southern Ocean.

While visiting SAMSI, Jong-June Jeon became interested in developing machine learning methods for Argo data and Mikael suggested exploring the possibility of improving the quantification of mapping

uncertainty using neural network models. Jong-June has started to work with a student Sang Jun Moon to produce an implementation of a neural network-based quantile regression algorithm that is suitable for Argo-type space-time prediction problems.

The statisticians in the working group benefitted immensely from the active participation of the oceanographers who brought in expertise on Argo data and provided guidance on problems that are of scientific interest to oceanographers. Likewise, the oceanographers in the group have benefitted from this interaction with statisticians by learning more about statistical modeling and methodologies. Both sides see immense benefit in continuing this interaction beyond the year-long SAMSI program. Indeed, Donata Giglio aptly summerized after the Transition Workshop that "collaborations between statisticians and oceanographers have enormous potential to advance our understanding of how to use large data sets more effectively to learn about sea level/ocean circulation and variability, and it offers interesting challenges to the statisticians".

5. Projects/topics undertaken by postdocs and SAMSI grad fellows:

Mikael Kuusela: In addition to continuing his work on statistical interpolation of ocean temperatures using space-time models, Mikael has mainly worked on estimating the temporal trend in ocean heat content, a quantity of intense interest in understanding transient climate change. This problem is in principle a three-dimensional integration, but because essentially all Argo floats capture a detailed vertical profile of temperature and salinity to the same depth of 2000 m, it is possible to first integrate the heat content vertically along these profiles and then integrate the results horizontally to measure the ocean heat content from the surface to 2000 m, ignoring locations where Argo cannot get data due to shallow depths or surface ice. This key insight, to first integrate the profiles vertically and then integrate those values horizontally, turned what appeared to be a potentially intractable problem both in terms of modeling and computations into a merely challenging but realistic problem to attack. Most of the groudwork for producing new estimates of ocean heat content trends has been done and we hope to finish a paper on this topic by early fall. Along with Doug Nychka and Pulong Ma, Mikael has also been working on a method for carrying out conditional simulations of the temperature field that handles the nonstationarity of the process and is computationally feasible. This approach may also be used to produce uncertainty estimates on the ocean heat content values.

Huang Huang: The biogeochemical mapping project aims to use machine learning methods for biogeochemical value prediction. Temperature and salinity measurements are available from all Argo instruments. A small fraction of them are also equipped with biogeochemical sensors. In ongoing work, Huang is exploring the use of neural networks to combine the information from both types of floats to produce better interpolants of oxygen levels in space-time than methods that only use data from the biogeochemical sensors. This project improves upon previous work by Donata Giglio by incorporating information from nearby floats in the machine learning predictor.

Pulong Ma: Pulong Ma has been working together with Doug Nychka and Mikael Kuusela on developing moving-window conditional simulation methods for nonstationary random fields. This builds upon previous work by Doug on local unconditional simulations and by Mikael on local point predictions. The basic idea is to simulate a white noise realization over the spatial domain and then compute a

conditional simulation realization by multiplying a local subset of the white noise with a symmetric matrix square root of the local predictive covariance, keeping the white noise realization fixed when looping over the grid points. Doug is leading a write-up of this work, Pulong is working on simulation studies and Mikael is providing an application to Argo temperature mapping. The aim is to submit this work by early fall.

6. Sources of data:

* Argo profiling float data, available at ftp://usgodae.org/pub/outgoing/argo, see also http://www.argo.ucsd.edu/
* Preprocessed Argo data, available at https://github.com/mkuusela/PreprocessedArgoData, based on the May 8, 2017 snapshop of the Argo GDAC (http://doi.org/10.17882/42182#50059)
* Roemmich-Gilson Argo climatology, available at http://sio-argo.ucsd.edu/RG_Climatology.html
* CERES top-of-atmosphere flux data, available at https://ceres.larc.nasa.gov/order_data.php

7. List of discussion or presentation topics in WG meetings:

The weekly group meetings had an active core of regular participants and a larger group of people who participated sporadically. As the list of topics show, there were three active subgroups, one on ocean heat content, one on mapping biogeochemical quantities and one on flow mapping. As the list shows, group meetings have continued past the formal end of the year and the plan is to keep these meetings going for as long as they are productive.

Sep 7: Dean Roemmich (overview of Argo)
Sep 14: Mikael Kuusela (WG plans, data access, locally stationary interpolation of Argo data)
Sep 21: Matthew Mazloff (biogeochemical mapping)
Sep 28: Discussion of the Ninove (2016) paper: https://www.ocean-sci.net/12/1/2016/os-12-1-2016.html
Oct 5: Donata Giglio (Estimating Oxygen in the Southern Ocean using Argo Temperature and Salinity)
Oct 12: Sarah Gille (flow mapping)
Oct 19: Amit Apte (Lagrangian data assimilation)
Oct 26: Division into subgroups
Nov 2: Subgroup 1: Ocean heat content
Nov 9: Bo Li (nonstationarity in spatial statistics)
Nov 16: Subgroup 2: Flow mapping
Nov 30: Subgroup 3: Biogeochemical variables
Dec 7: Subgroup 1: Ocean heat content
Dec 14: Subgroup 2: Flow mapping
Jan 18: General meeting
Jan 25: Biogeochemical subgroup
Feb 1: Ocean heat content subgroup
Feb 22: Flow mapping subgroup (line-of-sight matrix)
Mar 1: Ocean heat content subgroup
Mar 8: Updates from all subgroups

Mar 15: Ocean heat content subgroup
Mar 22: Discussion on covariance functions
Mar 26: Ocean heat content subgroup
Apr 9: Flow mapping subgroup
Apr 23: Ocean heat content subgroup (Vertical integration studies by Anirban)
Apr 30: Biogeochemical subgroup (Spatial neural nets by Huang)
May 7: Preparation for the Transition Workshop

----- Summer meetings -----

Jun 5: Ocean heat content
Jun 11: Ocean heat content
Jun 25: Flow mapping (TBC)
Jul 2: Biogeochemical (TBC)

8. Other activities

* Mikael Kuusela visited the Scripps Institution of Oceanography on October 18-20, 2017
* Mikael Kuusela visited Profs. Tailen Hsing and Stilian Stoev at the University of Michigan on December 11-12, 2017 to discuss their interest in statistical analysis of Argo data
* Mikael Kuusela gave many talks (University of Cincinnati, Colorado State University, University of North Carolina at Chapel Hill, University of Washington, Simon Fraser University, University of Edinburgh, King's College London, Carnegie Mellon University, Rutgers University) based in part on work done at SAMSI as part of the process of interviewing for a faculty position. He has accepted a tenure-track position at Carnegie Mellon University.
* Mikael Kuusela gave a talk at the Ocean Sciences Meeting on February 16, 2018 based in part on the work done within this working group
* Mikael Kuusela presented a poster at the SIAM Conference on Uncertainty Quantification on April 16, 2018 based in part on the work done within this working group
* Fred Bingham (UNC Wilmington) visited SAMSI on April 4, 2018 and presented a seminar entitled "Cooking the GOOS: Our increasingly sophisticated network of ocean observations and its use in the detection of climate change"
* Mikael Kuusela gave SAMSI Postdoc Seminars on work done within this working group on October 25, 2017 and on May 2, 2018
* Mikael Kuusela, Donata Giglio and Michael Stein presented work done within this working group at the CLIM Transition Workshop at SAMSI on May 15, 2018
* A group of undergraduate students worked on detecting and characterizing El Nino using Argo data at the SAMSI Undergraduate Modelling Workshop on May 21-25, 2018 under the guidance of Mikael Kuusela
* Mikael Kuusela will give an overview talk on the work done within this working group at the TIES 2018 conference on July 16-21, 2018

9. Publications

* M. Kuusela and M. L. Stein. Locally stationary spatio-temporal interpolation of Argo profiling float data. arXiv:1711.00460 [stat.AP], under revision for the Proceedings of the Royal Society A, 2018.
* M. Kuusela, D. Giglio, A. Mondal and M. L. Stein. Statistical Methods for Ocean Heat Content Estimation with Argo Profiling Floats. In preparation, 2018.
* D. Nychka, P. Ma and M. Kuusela. Local Conditional Simulations for Inference of Spatial Fields. In preparation, 2018.

# Report on STATMOS/SAMSI Summer School


## Michael Stein and Richard Smith


The Summer School on Climate Datasets took place at the National Center for Atmospheric Research, Boulder, Colorado, from July 17-21, 2017. Co-organizers were Michael Stein (University of Chicago, co-leader of STATMOS), Richard Smith (SAMSI), Doug Nychka (NCAR) and Michael Wehner (Lawrence Berkeley Lab). For SAMSI, this was the first event of the 2017-18 Climate program.

This event, loosely referred to as the "Climate Palooza," took the form of a series of nine working groups, each charged with analyzing a specific dataset. Apart from the four organizers, there were thirty participants, ranging from junior researchers to senior professors, in disciplines covering both the mathematical and atmosphere/ocean sciences. An organizing principle was that each working group had to contain at least one mathematical scientist and at least one atmosphere or ocean scientist. Two incoming SAMSI postdocs (Yawen Guan and Mikael Kuusela) were among the participants.

The week's activities included an overview of climate research by Wehner and a panel discussion on interdisciplinary research to which all four organizers contributed. The rest of the time was spent in group work except for the final morning, when they reconvened to present their results.

Following are the outline schedule, topics and membership of each of the nine working groups, and reports from seven of them.

# Tentative Schedule as of 06/26/17

There are 9 groups in total. I encourage each group to do some preliminary work/discussion in advance of the meeting. Feel free to ask me or some of the other senior attendees any questions you might have.

I'm looking forward to seeing and working with all of you. I hope it will be fun and productive.

**In terms of a schedule, there isn't much of one and what there is could be changed on the fly if needed, but here is what we've got.**

--Michael

*We will have some kind of coffee break/mingling time each afternoon.*

## Monday

9-9:30, groups/individuals introduce themselves
9:30- approx 10:30, panel discussion on effective interdisciplinary collaboration

Rest of day. Group work.

---

## Tuesday

9-9:45 Plenary talk by Michael Wehner

Rest of day. Group work.

---

## Wednesday

3:15-4:45 Interim group presentations
5:30 Reception at Under the Sun. Doug Nychka will lead a walk from NCAR to Under the Sun. Pretty easy walk since all downhill.

If several groups are facing similar challenges, we may schedule one or more impromptu discussions on specific topics on Tuesday or Wednesday. We'll see.

---

## Thursday

Group work.

---

## Friday

9-12:30, group presentations. We will end promptly at 12:30, so each group will have no more than 22 minutes to present, leaving time for one short break.

---

# PROJECT PROPOSALS and WORKGROUPS

* No group may have more than four people
* Every group must have at least one climate and one stat person. If a group has four people, we encourage 2 climate and 2 stat.
* Please let Michael Stein and Mitzi know of all such agreements to join a group so we can track how things are going.

**A *tentative* Workshop Schedule follows after the list below, and may change at any time.**

---

**TITLE:** Statistics of top-of-atmosphere fluxes
**PROPOSER:** Jonah Bloch-Johnson, University of Chicago; [climate]; jsbj@uchicago.edu
**PARTICIPANTS:**
Nathan Lenssen, Columbia University; [statistics]; nathan.lenssen@nasa.gov
David Legates, University of Delaware; [climate]; legates@udel.edu
**COMMENTS:**

---

**TITLE:** Assessing skill in simulated precipitation distributions and extremes over regions
**PROPOSER:** Richard Grotjahn, Univ. California, Davis; [Climate]; grotjahn@ucdavis.edu
**PARTICIPANTS:**
Jose Garcia, Univ. Extremadura (Spain) [Physics]; agustin@unex.es
Matthias Katzfuss, Texas A&M University; [Statistics]; katzfuss@gmail.com.
Whitney Huang, Purdue University; [Statistics]; huang251@purdue.edu
**COMMENTS:**

---

**TITLE:** Analysis and interpretation of oceanographic decorrelation scales from Argo profiling float data
**PROPOSER:** Mikael Kuusela, postdoc, University of Chicago, Department of Statistics; kuusela@uchicago.edu
**PARTICIPANTS:**
Fred Bingham, UNC-Wilmington, binghamf@uncw.edu
Oksana Chkrebtii, Ohio State University Statistics, oksana@stat.osu.edu
Danielle Elie Touma, Stanford University, detouma@stanford.edu
**COMMENTS:**

---

**TITLE:** Covariability of temperature and precipitation extremes
**PROPOSER:** Karen McKinnon, ASP Postdoctoral Scholar, NCAR; mckinnon@ucar.edu
**PARTICIPANTS:**
Andy Poppick, Carleton College, apoppick@carleton.edu
Lynne Seymour, Department of Statistics, University of Georgia; seymour@uga.edu
Malte Stuecker, University of Washington, stuecker@atmos.washington.edu
**COMMENTS:**

---

**TITLE:** Representing extremes in high-resolution gridded climate products
**PROPOSER:** Jared Oyler, Postdoctoral Scholar, Penn State University; jared.oyler@psu.edu
**PARTICIPANTS:**
Liz Drenkard, Postdoctoral Scholar, Scripps Institution of Oceanography, liz.drenkard@gmail.com
Hossein Moradi Rekabdarkolaee, Postdoctoral Scholar, Department of Statistical Sciences and Operations Research,
Virginia Commonwealth University, moradirekabh@vcu.edu
**COMMENTS:**

---

**TITLE:** Inferring calving events of ice shelves from satellite data
**PROPOSER:** David Trossman, Research Associate, University of Texas-Austin, Institute for Computational Engineering and Sciences; david.s.trossman@gmail.com
**PARTICIPANTS:**
Yawen Guan, Penn State University, yig5031@psu.edu
**COMMENTS:**

---

**TITLE:** Understanding variability and extremes in ocean surface chlorophyll fluorescence
**PROPOSER:** Daniel Whitt, NCAR CGD, dwhitt@ucar.edu
**PARTICIPANTS:**
Joe Guinness, NC State University, jsguinne@ncsu.edu
Mikyoung Jun, Texas A&M University, mjun@stat.tamu.edu
**COMMENTS:**

---

**TITLE:** Quantifying near-term skill at predicting shifts in the likelihood of temperature and precipitation extremes
**PROPOSER:** Stephen Yeager, NCAR; yeager@ucar.edu
**PARTICIPANTS:**
Matz Haugen, University of Chicago, (?) matzhaugen@gmail.com
Pulong Ma, University of Cincinnati, mapn@mail.uc.edu
Vineel Yettella, University of Colorado, Boulder, vineel.yettella@colorado.edu
**COMMENTS:**

---

**TITLE:** Projection of global urban climate and their robustness and uncertainties
**PROPOSER:** Lei Zhao, Postdoctoral Research Associate, Princeton University; lei.zhao@princeton.edu
**PARTICIPANTS:**
Chen Chen, Postdoctoral Scholar, University of Chicago, chenchen1@uchicago.edu
Andrew Bray, Assistant Professor of Statistics, Reed College, abray@reed.edu
**COMMENTS:**

# Predicting near-term shifts in the likelihood of climate extremes

Steve Yeager, Matz Haugen, Pulong Ma, Vineel Yettella

Objectives:
Explore methods for quantifying the skill of an initialized decadal prediction simulation set performed with CESM (the CESM Decadal Prediction Large Ensemble, CESM-DP-LE) in hindcasting interannual-to-decadal variations in climate extremes. Evaluate skill enhancements (if any) associated with initialization by contrasting with a complementary historical simulation set forced with external radiative loadings (greenhouse gases, aerosols, etc) but having unsynchronized, non-historical internal climate variations across members (the CESM Large Ensemble, CESM-LE). Both simulation sets are large ensembles (40 members), permitting robust estimation of changes in the tails of climate distributions.

What we accomplished:
We focused on maximum daily surface temperature (TMAX) regionally-averaged over land in Western Europe (10°W-10°E, 40°N-54°N). We computed the climatological (1964-2014) 90th percentile ($Q_{0.9}$) of TMAX as a function of day of year from observations (HadGHCND) and from the CESM-LE. For CESM-DP-LE, we computed daily $Q_{0.9}$ as a function of the lead time of the 122-month (3712-day) hindcasts. This was done to correctly account for the hindcast drift in CESM-DP-LE associated with full field initialization. We then computed TMAX exceedance counts above $Q_{0.9}$ as a function of calendar time for observations, CESM-LE, and CESM-DP-LE. For the latter, a lead time choice is needed to define the calendar time of the hindcasts and the corresponding lead-time-dependent $Q_{0.9}$ is used as the threshold. The result (for annual DP at lead year 1) is shown in Figure 1. We also explored the sensitivity to different choices of season and time-averaging windows. While CESM-DP-LE shows evident skill in reproducing observed variability, it is not clear whether the skill is significantly different from CESM-LE in this region for this variable. We also characterized the CESM-DP-LE probability density function of TMAX (as a function of lead time) by fitting a GEV distribution to the average over ensemble members and start times. The results are shown in Figs. 2-4.

What we plan to do:
We will further develop the processing tools required to repeat the daily $Q_{0.9}$ computation at each point on the globe--a non-trivial computation that, for CESM-DP-LE, involves manipulation of a 1.5 TB data array for each field of interest. We can then assess how skill varies in space as well as (lead) time. We will investigate other fields that have shown promising skill when examined as seasonal means over land (e.g. precipitation). We will also explore different techniques for defining extremal indices

based on clusters of threshold exceedances, that might be relevant to phenomena such as heat waves, cold spells, flooding events, and droughts.

Figures:

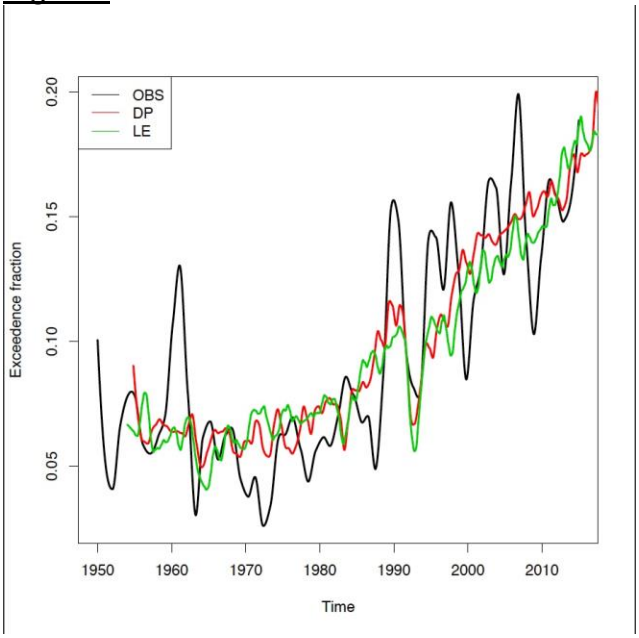



Fig 1: Exceedence fraction above the 90th percentile of daily maximum surface temperature over Western Europe (10°W-10°E, 40°N-54°N) from HadGHCND observations (OBS, black), the 40-member CESM Large Ensemble (LE, green), and the 40-member CESM decadal prediction large ensemble (DP, red). All time series have been smoothed. DP is plotted for lead year 1.

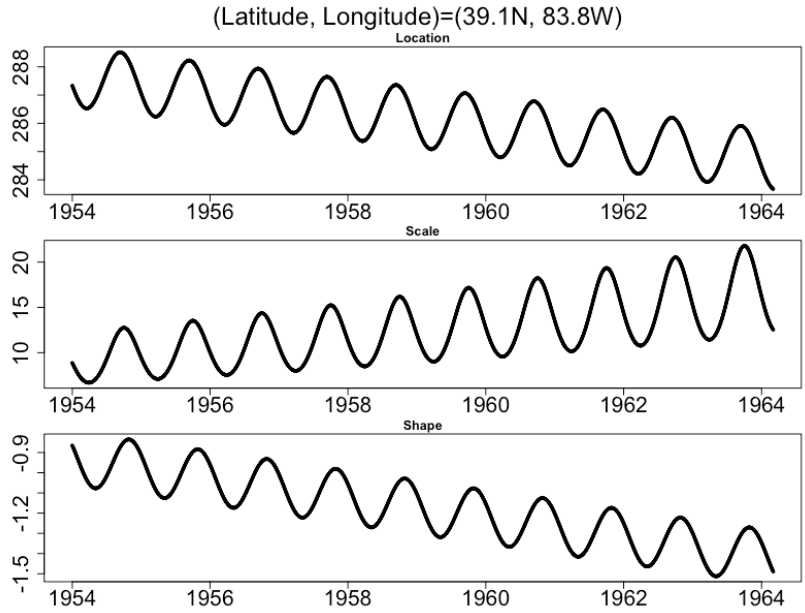



Fig 2: Parameter estimates in the proposed hierarchical model with GEV distribution for decadal hindcasts of daily maximum temperature over 10 years in Cincinnati. The x-axis represents hindcast lead time, not calendar time.

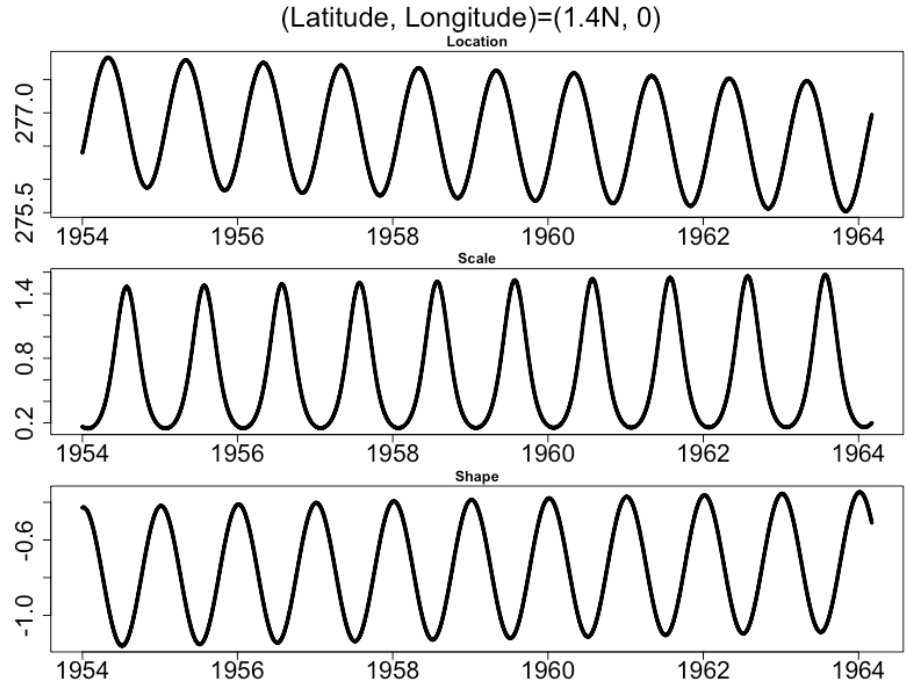



Fig 3: Parameter estimates in the proposed hierarchical model with GEV distribution for decadal hindcasts of daily maximum temperature over 10 years in equator. The x-axis represents hindcast lead time, not calendar time.

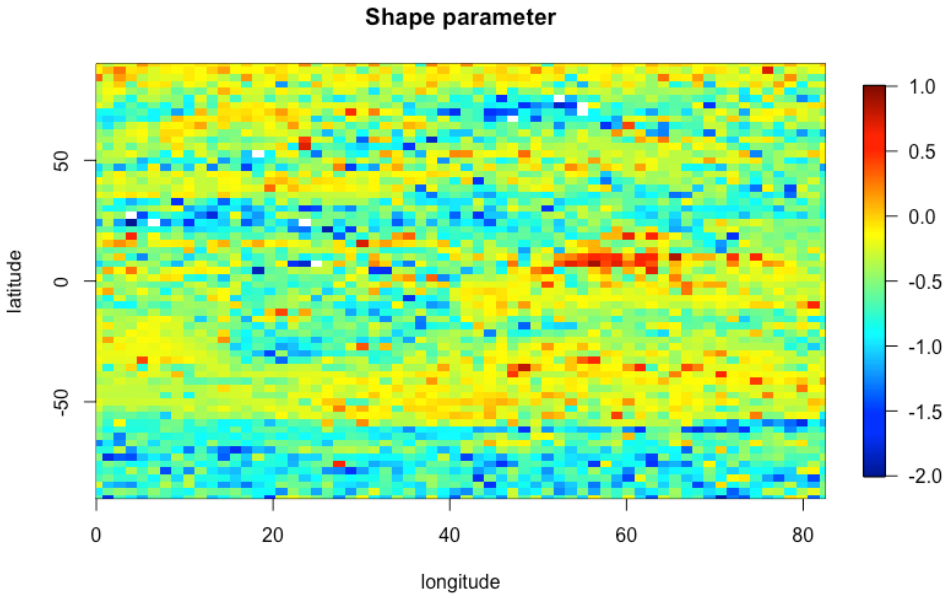



Fig 4: Parameter estimates for shape parameter in the proposed hierarchical model with GEV distribution for decadal hindcasts of daily maximum temperature over a large subregion at lead time 1.

# Global urban temperature projections in CMIP5

Lei Zhao[1], Andrew P. Bray[2], Chen Chen[3]
1 Program in Science, Technology and Environmental Policy (STEP), Princeton University, Princeton, NJ, 08544
2 Department of Mathematics, Reed College, Portland, OR, 97202
3 Department of Geophysical Sciences and Department of Statistics, University of Chicago, Chicago, IL, 60637

Heat stress associated with climate change is one of the most serious climate impacts. Its consequences are amplified for urban populations because of the urban heat island (UHI) effect. Because more than 50% of the world's population currently lives in cities, and that percentage is projected to increase to 70% by year 2050, there is a pressing need to understand how urban temperatures change under climate change and to assess the robustness and uncertainties associated with the projections. However, most of the CMIP5 Earth system models (ESMs) do not incorporate urban representation and parameterization in their modeling realms, with the NCAR's CESM as an only exception which has a process-based multi-layer urban land model. Our group seeks to provide multi-model urban temperature projections in CMIP5 and to assess their robustness and uncertainties.

To achieve this goal, instead of implementing an urban parameterization in each of the CMIP5 ESMs, we are building a statistical emulator based on CESM modeled outputs and applying the emulator to other ESMs to generate multi-model projections. Our statistical emulator takes the atmospheric forcings to CLM (the land component of CESM) as input, and output urban screen-height temperatures. Our analyses show that the atmospheric forcing variables are strong predictors for urban near-surface temperature. The statistical models are spatially heterogeneous because the urban morphological parameters are different for different urban grid cells. The temporal heterogeneity is significant as well (Figure 1), and should be explicitly represented in the statistical model.

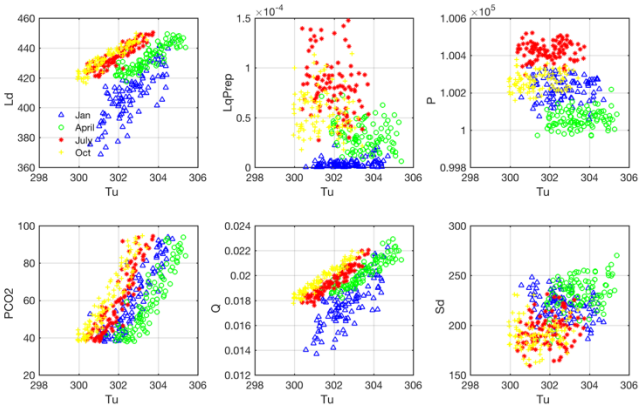



Figure 1 Relationships between urban temperature and long-wave radiation ($L_d$), rainfall (LiqPrecip), air pressure (P), $CO_2$ (PCO2), specific humidity (Q) and shortwave radiation ($S_d$).

Cross-validated by CESM's all 6 realizations in CMIP5, our emulator showed good agreement with the process-based modeled urban temperature, with RMSE less than 1 K (Figure 2).

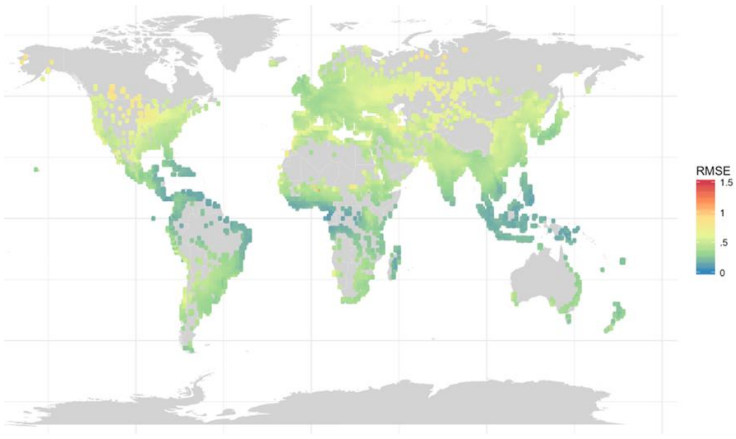



Figure 2 The root-mean-squared-error (RMSE) of predicted urban temperature using statistical emulator compared to the CLM-modeled urban temperature

# STATMOS/SAMSI Project Report by David Trossman and Yawen Guan

Objective:

We want to investigate how the processes that are of leading order importance to calving depends upon location and time and ultimately want to predict when a calving events occur.

What we did during the workshop:

We extracted observations of sea ice cover from two different datasets: the Level 4 Operational sea Surface Temperature and sea Ice Analysis (OSTIA; a gridded 1/4x1/4 degree, objectively interpolated product) and the Level 3 MODerate resolution Imaging Spectroradiometer (MODIS; a gridded 4x4 km product from using both the Aqua and Terra satellites). Four locations were chosen for case studies based on their locations and geometries: Jakobshavn and Helheim Glaciers, Greenland, and Thwaites and Larsen C Ice Shelves, Antarctica. The sea surface temperatures and sea ice fractions from the OSTIA dataset were found to be strongly anti-correlated over the seasons. Because this could imply an important role for basal melting of sea ice at its base, the decay rate of the sea ice fraction at the end of the Spring or beginning of the Summer seasons is an ambiguous proxy for whether calving occurs. The MODIS data were then examined by calculating the areal extent of ice-covered ocean and the number of chunks (closed contours) of ice-covered ocean. It was hypothesized that the number of chunks of sea ice sharply increases (sea ice areal extent remaining approximately constant) when calving events occur, and the number of chunks of sea ice remains approximately constant (sea ice areal extent steadily decreasing) when basal melting occurs. The time series of the sea ice areal extent and number of chunks of sea ice over April of 2011 through December of 2015 were extremely variable from one day to the next (see the attached figure). The reason for this was found to be the obscuring cloud coverage that the MODIS instruments cannot see through, particularly after controlling for nighttime impedance.

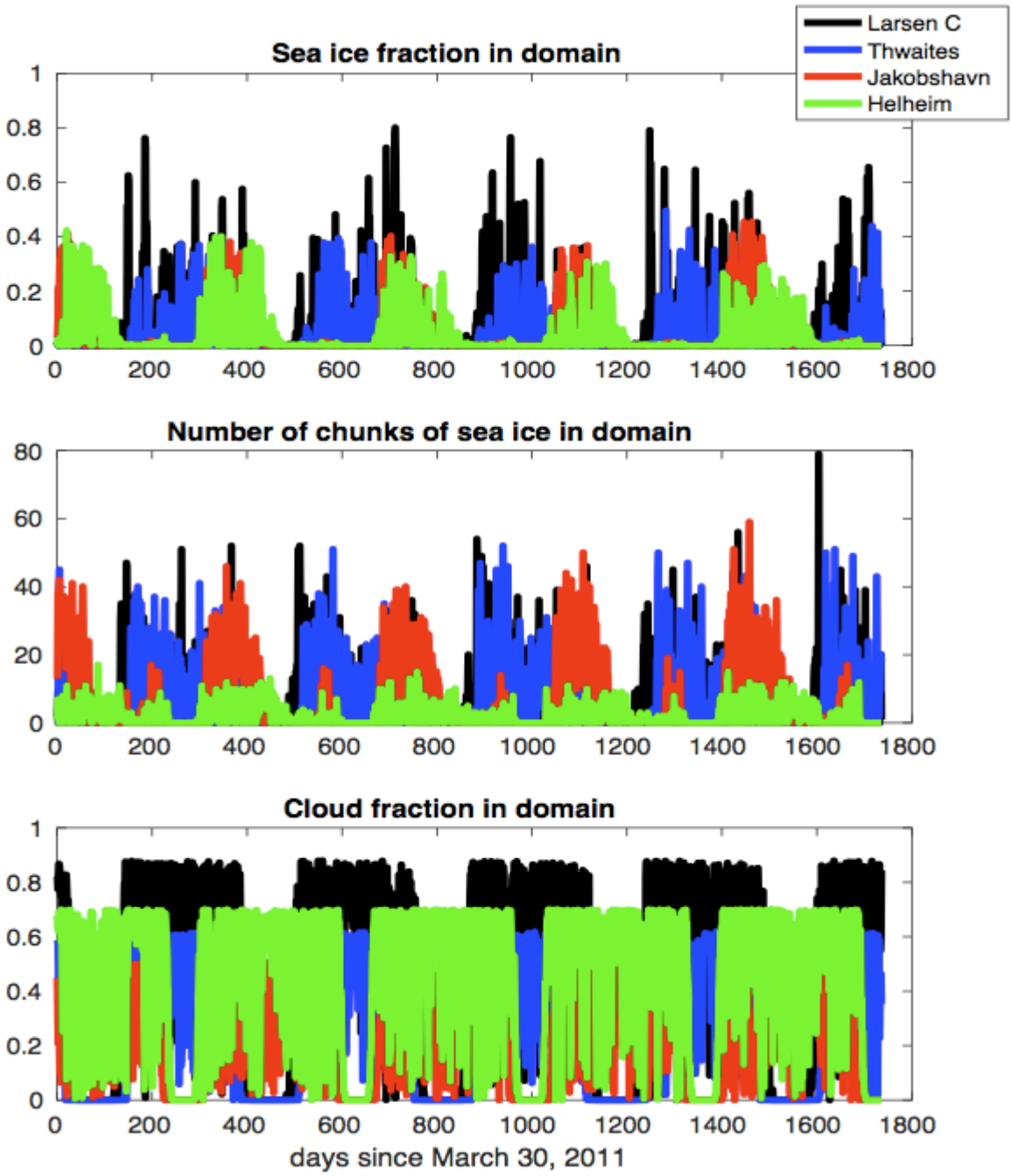
What we plan on doing after the workshop:

Existing theories suggest that iceberg calving events are related to a number of processes in a complicated manner. These include a floatation criterion, ocean waves/seismic events, longitudinal strain, damage mechanics, velocity of the mélange, basal crevassing, ocean-driven melting at the terminus, and hydrofracturing. We propose to investigate the mechanisms of ice calving by constructing binary event statistical models with the potential to account for as many of these processes as possible (given observational data constraints).

In the immediate future, we will focus on environmental variables that models can resolve, such as ice velocity, ocean temperatures where the ice makes contact with the ocean, and a floatation criterion that depends upon the ocean bathymetry and ice heights. We will explore the phase space of these variables to find thresholds that can be used in a binary event statistical model. In order to explore this phase space, we first need to detect whether a calving event has occurred, which we plan on doing by merging datasets whose measurements of sea ice extent are not primarily determined by cloud coverage. These datasets include Sentinel-1A and -2A synthetic

aperture radar visible and infrared band data at 5-20 meters resolution (but global coverage only every 12 days), Level 4 IceBridge passive microwave data from ICESat and SSM/I at 70 meter resolution (but only during intermittent time periods), and Landsat-7 and -8 data in the visible band data at 30 meter resolution (but global coverage only every 16 days).

After detecting ice calving events, we will investigate their relationships with other environmental variables using statistical models. For example, spatial generalized linear mixed models (SGLMMs) can be particularly useful for modeling ice calving. This regression framework allows us to investigate the relationships between the multiple processes and ice calving response; this framework also has the flexibility to account for the spatial dependence in the ice calving events. We plan to perform the statistical analysis for both Greenland and Antarctica to understand the mechanisms driving ice calving in these two distinct regions.


Sea ice fraction in domain
Larsen C
Thwaites
Jakobshavn
Helheim
1
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000
1200
1400
1600
1800
Number of chunks of sea ice in domain
80
60
40
20
0
0
200
400
600
800
1000
1200
1400
1600
1800
Cloud fraction in domain
1
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000
1200
1400
1600
1800
days since March 30, 2011

**Precipitation precursors to heat waves in the United States**

Karen McKinnon, Andrew Poppick, Lynne Seymour, and Malte Stuecker

Heat waves are primarily a meteorological phenomena resulting from processes such as increased radiation or advection of anomalously warm air. In this context, the predictability of heat waves will be limited to the 5-10 day timescales associated with initial-condition atmospheric predictability. It is of interest, however, to ask whether heat waves are also related to more slowly-varying boundary conditions that may allow us to extend our horizon of predictability beyond weather timescales.

An important boundary condition for heat waves is the land surface. Simplistically, a moister land surface will allow for a greater partitioning of incoming heat into latent, rather than sensible, heat, which will keep temperatures lower. This process, however, is generally thought to be relevant only in certain 'transitional' regions that are neither too wet nor too dry (Seneviratne et al., 2010).

Here, we examine the relationship between precipitation and heat waves at weather stations across the United States. We focus on precipitation rather than soil moisture itself because both real-time and historical soil moisture measurements are limited, and integrated metrics of precipitation have been shown to be reasonable proxies for soil moisture (Guttman, 1998).

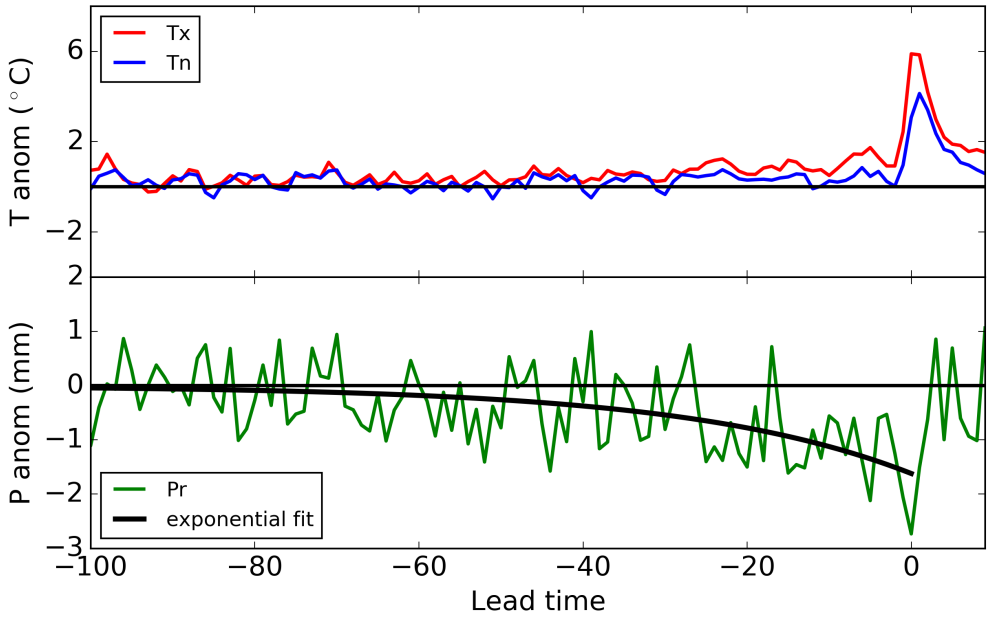



Figure 1: The composite behavior of daily maximum temperature (Tx, red), daily minimum temperature (Tn, blue) and precipitation (green) before a heat wave in Iowa City, Iowa. The black line shows the best fit of the composite precipitation before a heat wave to a decaying exponential function.

Figure 1 shows an example relationship between heat waves and precipitation over Iowa City, Iowa. Precipitation decreases before heat waves at lead times up to 80 days. A

straightforward interpretation of this behavior would be that the land surface dries over a period of 80 days, at which point it becomes more likely to have hot temperatures (due to a decrease of the patitioning of heat into latent heating).

Motivated by this finding, we quantify the timescale of precipitation decreases before heat waves over hundreds of stations across the United States. We identify a region in the central-eastern United States where a precipitation signal is evident at lead times of 40-60 days, suggesting the possibility of predictability on these timescales. To test this idea, we fit a statistical model to the data, and find that including precipitation information improves heat wave predictions at longer-than-weather lead times.

In future work, we plan to:

- improve the statistical model to better quantify the magnitude of improved predictability via integrated precipitation
- develop a basic physical model for the manner in which precipitation decays before heat waves
- develop and test a hypothesis that explains the spatial extent of the region that shows stronger links between precipitation deficits and heat waves
- perform our analyses for different types of heat waves, with a focus on humidity and nighttime temperatures

# References

Guttman, N. B., 1998: Comparing the Palmer drought index and the standardized precipitation index. *JAWRA Journal of the American Water Resources Association*, **34 (1)**, 113–121.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, 2010: Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Science Reviews*, **99 (3)**, 125–161.

# Phytoplankton variability in Earth system models and satellite observations

J. Guinness, M. Jun, and D. Whitt

July 31, 2017

Despite containing small biomass relative to land plants, marine phytoplankton carry out about half of all photosynthesis on the planet, form the base of essentially all marine ecosystems, and are an important component in global geochemical cycles of carbon, nitrogen and other elements. Given the importance of phytoplankton to the health of our planet as a whole, our oceans, and the industries that depend on them, it is important to understand how a changing climate will influence phytoplankton. The Community Earth System Model (CESM) incorporates dynamics of three phytoplankton species. We are studying the spatial and temporal distributions of these species in models of the current and future climate under the various emissions scenarios included in the CESM Large Ensemble (CESM-LE). For example, Figure 1 shows how anthropogenic climate change might alter the seasonal cycle of phytoplankton in the subpolar North Atlantic ocean by both reducing the magnitude of the maximum chlorophyll concentration and shortening the duration of the spring bloom.

Phytoplankton chlorophyll can also be observed from satellite sensors such as the Moderate Resolution Infrared Spectrometers (MODIS) on board NASA Aqua. However, satellite observations are often impeded by cloud cover, especially in the North Atlantic, where phytoplankton blooms are a prominent feature of the seasonal cycle. In fact, cloud cover is the norm—not the exception—in the North Atlantic. Due to the short duration of phytoplankton blooms, it is important to incorporate information about data missingness into estimates of weekly and monthly averages, rather than use simple averages of the available data, which is the current practice for constructing 8-day and monthly composites. In addition, due to the dependence of phytoplankton blooms on available sunlight, there is a potential that the missing data are not missing at random, which might cause climatological composites (i.e. inter-annual sample means) to be biased estimates of true climatologies. We are using climate model output to study how large these biases can potentially be. This approach, combined with Gaussian process spatial and temporal interpolation, will be used to develop bias-corrected 8-day and monthly composites as well as climatologies, complete with uncertainty measures.

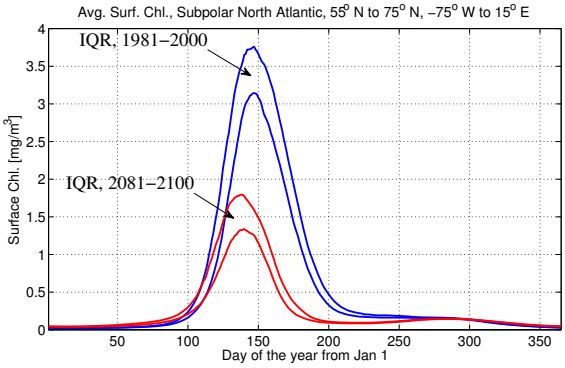



Figure 1: The interquartile ranges of surface chlorophyll in the subpolar North Atlantic on each day of the annual cycle. The range is derived from 26 members of the CESM-LE and from 1981-2000 (blue curves) or from 2081-2100 under the RCP8.5 emissions scenario (red curves).

**Representing Extremes in High-Resolution Gridded Climate Products**

**Project group members:** Liz Drenkard, *Scripps Institution of Oceanography*; Hossein Rekabdarkolaee, *Virginia Commonwealth University*; Jared Oyler, *Pennsylvania State University*

**Purpose of project:** High-resolution gridded climate products are an important source of locally relevant climate data for a wide range of climate impact analyses. Using station-based point observations, the products provide spatiotemporally continuous best estimates of climate variables. Because product interpolation methods are designed to minimize the error variance of their estimates, they smooth extremes and spatiotemporal variability. This smoothing is expected and acceptable for most analyses. However, for analyses that depend on accurate representations of the spatiotemporal dependence structure and magnitude of extreme events, the use of traditional gridded products is problematic. The objective of our project was to develop an interpolation approach that better captures spatiotemporal patterns of extremes while still maintaining adequate performance with respect to unconditional error variance.

**Study domains and datasets:** Our project consisted of two case studies: (1) interpolation of daily station observations of precipitation within a region surrounding the Chesapeake Bay watershed; and (2) interpolation of missing values in the Coral Reef Temperature Anomaly Database (CoRTAD) sea surface temperature (SST) data product for the Coral Triangle. CoRTAD is derived from the Pathfinder satellite SST project, and is used to predict regional coral stress, bleaching thresholds and likelihood of coral mortality. CoRTAD currently implements two infilling schemes for missing data grid cells: taking the median value from adjacent grid cells and temporal-spline fitting for remaining gaps.

**Workshop progress:** We developed and tested two interpolation approaches for the precipitation observations. In the first approach, we combined a machine learning random forest method for estimating conditional quantiles with geostatistical conditional simulation to produce unsmoothed interpolations of daily precipitation. We found that the approach was effective at capturing all quantiles of precipitation intensity, including extremes. However, two limitations remained: (1) high computational requirements due to the need to fit a unique model for every day; and (2) a lack of local predictors that decreased local accuracy. To address these limitations, we started development of a second neural network approach based on a single neural network model that incorporated nearest neighbor station spatial location and precipitation information. Using a binary interpolation of precipitation occurrence as a test case, we found the accuracy of the neural network method was near that of our first approach and significantly reduced computational requirements.

For CoRTAD, we took weekly SST data with missing values and applied a spatial kriging method to infill missing values prior to recalculating maximum degree heating weeks (DHWs; metric for accumulated thermal stress) for 1998. We found that, although large-scale patterns in estimated coral stress exposure were consistent with CoRTAD's original infilling methods, regional patterns varied considerably (Figure 1). This suggests that satellite SST products with substantial levels of missing data due to cloud contamination and along coastlines should not be the only guide in identifying potential regional coral reef conservation priorities.

**Next steps:** Based on validation in the U.S. Mid-Atlantic, our initial daily precipitation methods appear useful for creating accurate spatiotemporal representations of precipitation occurrence and extreme events. We will next expand the neural-network approach to precipitation amount and compare its performance against our quantile random forest approach. Our objective will be to increase local accuracy while at same time decrease computational requirements. To expedite processing of CoRTAD weekly SST, we worked on a smaller portion of the Coral Triangle domain and used a random subsample of 25,000 points for interpolation. Additionally, we only processed the year 1998 (i.e., global bleaching event) but still used the CoRTAD climatology (with the original infilling scheme) to calculate DHWs. We next plan to process all years in CoRTAD for the entire Coral Triangle domain without random subsampling. We will then recalculate the DHW metric and compare to the original CoRTAD product. Additional efforts entail using spatiotemporal models and neural network methods for infilling, as well as applying these methods to more recent Coral Reef Watch SST products, with the ultimate goal of illustrating the influence of gridded infilling methods on the prediction of thermal stress and bleaching events for coral reefs at local scales.

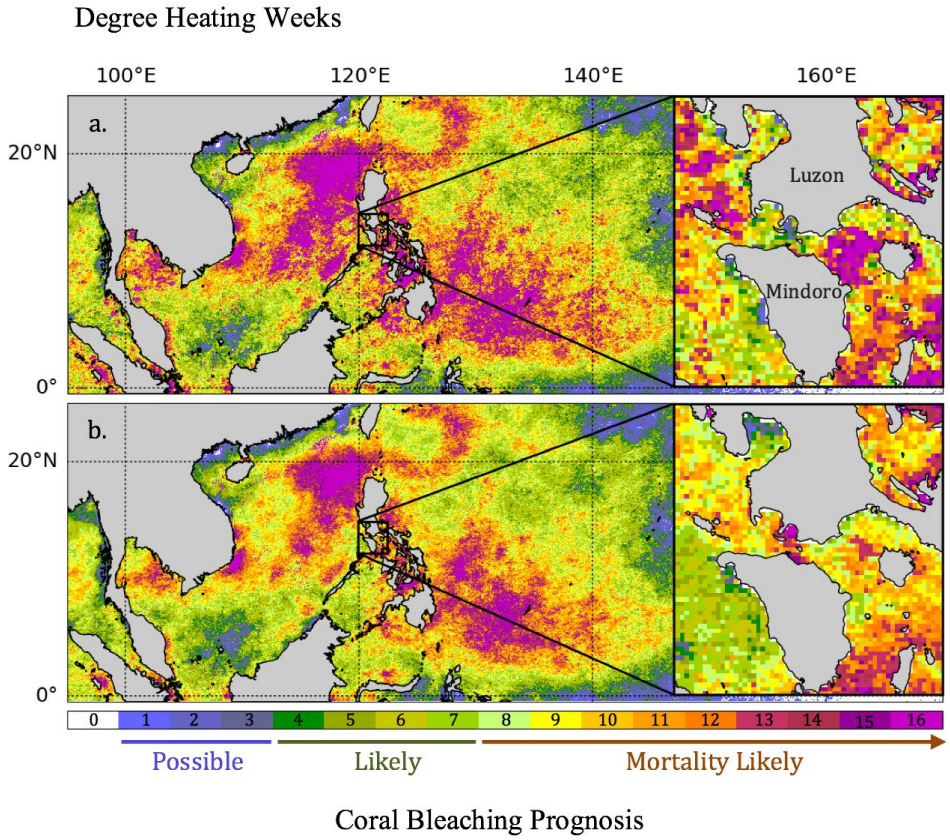



**Figure 1:** Comparison of the Degree Heating Weeks (DHWs) calculation using: **(a)** missing value infilled sea surface temperature (SST) from the original Coral Reef Temperature Anomaly Database (CoRTAD) product; and **(b)** missing value filled CoRTAD SST using spatial kriging. While the large-scale patterns are similar, regional DHWs (e.g., Verde Island Passage, inset) differ.

# Causes of skewness in Ocean salinity

Report of Salinity Skewness group for STATMOS/SAMSI Climateplaooza

24-July-2017

Frederick Bingham, Danielle Touma, Mikael Kuusela and Oksana Chkrebtii

## Introduction

Sea surface salinity (SSS) is an important variable for understanding the global freshwater cycle and surface fluxes. In the absence of other ocean processes at the surface, increasing SSS may indicate evaporation and decreasing salinity precipitation, or in other words freshwater fluxes.

Salinity has been measured by Argo floats since about 2005. The floats make measurements from ~2000 m depth up to the surface. The nominal Argo spatial separation is 3˚ x 3˚ with a sample interval of 10 days (http://argo.ucsd.edu). Additionally, between August 2011 and June 2015, salinity has been measured by the Aquarius satellite.

When rain falls on the surface of the ocean, it creates fresh pools with increased stratification and low salinity. These may be sampled by surfacing Argo floats, or by satellite. These fresh pools are mixed back into the bulk mixed layer by wind or other turbulent processes [e.g. Asher et al., 2014a]. An inverted, but usually more short-lived, effect may be observed resulting from evaporation [Yu 2011; Asher et al., 2014b]. Thus, low SSS anomalies tend to persist for longer than high ones, and are more likely to be observed either by floats or by satellites. For this reason, one might expect that the probability distribution of salinity would be negatively skewed. The purpose of this project was to investigate the potential negative skewness of SSS and to see if it could be statistically related to covariates such as rainfall, evaporation or wind speed.

## Data

We used two salinity datasets to compute skewness.

The Aquarius (v 4.0) data is a binned, level 3 product sampled every 7 days over the globe. The satellite was functioning from August 2011 until June 2015, providing about 201 weeks of data. To get the skewness, we calculated a monthly mean and subtracted it from individual measurements. To produce a smoother field, we used a moving window and calculated skewness for 3˚ x 3˚ boxes, producing estimates every 1˚.

Argo data were also used to compute skewness at 10 m and 300 m, again using a moving-window approach. The time period used was January 2007 – December 2016. Quality-controlled Argo salinity profiles were first linearly interpolated to the desired depth level. Skewness was then computed for the residuals obtained by subtracting the seasonally varying Roemmich-Gilson Argo climatology [Roemmich and Gilson, 2009]. The estimates were computed every 1˚ using a 10˚ x 10˚ moving window.

Evaporation data were from the OAflux dataset [Yu et al., 2008] which produces daily estimates on a 1˚ x 1˚ grid. Mean values were computed over the time periods September 2011 – June 2015.

Daily precipitation data were from the Tropical Rainfall Measuring Mission (TRMM) [Liu et al., 2012]. We calculated the mean over September 1, 2011 – May 31, 2015. We regridded the TRMM data to 1˚ x 1˚ from 1/4˚ x 1/4˚ using bilinear interpolation to match the other datasets.

Wind data were monthly values from the NCEP/NCAR Reanalysis I [Kalnay et al., 1996]. We averaged wind speeds over the same September 2011 – May 2015 period. We regridded the reanalysis data to 1˚ x 1˚ from 2.5˚ x 2.5˚ using bilinear interpolation to match the other datasets. Since bilinear interpolation is not recommended when increasing grid resolution, we will use a different reanalysis dataset in the future.

## Skewness

Skewness is a measure of the asymmetry of a probability distribution. Negative skewness indicates that the distribution has a longer, fatter tail on the left side. It also quantifies an aspect of the non-Gaussian nature of a probability distribution. There are a number of different estimators of skewness [Dodge, 2008]. We used one that quantifies differing widths of a probability distribution on either side of the median.

$$\kappa = \frac{[Q(.95) - Q(.5)] - [Q(.5) - Q(.05)]}{Q(.95) - Q(.05)}$$

where Q is the empirical quantile function, i.e., Q(.95) is the value of the 0.95 empirical quantile of a distribution. The primary advantage of this quantile-based measure of skewness over moment-based estimates is its robustness to outliers in the data.

## Regression model

In order to explore the data, we first considered a simple linear regression model relating the response (skewness) to functions of the covariates (precipitation, evaporation and wind speed).

Figure 3 shows fits to SSS from Argo and Aquarius data (left to right) for the models (top to bottom),

$$M1: \kappa(x_i, y_i) = \beta_0 + \beta_1 log\big(p(x_i, y_i)\big) + \epsilon(x_i, y_i), \epsilon(x_i, y_i) \sim iidN(0, \sigma^2), i = 1, \dots, n$$

$$M2: \kappa(x_i, y_i) = \beta_0 + \beta_1 log\big(p(x_i, y_i)\big) * w(x_i, y_i) + \epsilon(x_i, y_i), \epsilon(x_i, y_i) \sim iidN(0, \sigma^2), i = 1, \dots, n$$

where is p the precipitation and w the wind speed.

The assumption of spatially independent and identically distributed errors may not be realistic in a spatial setting where nearby observations are likely to be correlated. We therefore also considered a spatial linear mixed effects regression model to relate the response to functions of the covariates,

$$M3: \kappa(x_i, y_i) = \beta_0(x_i, y_i) + \beta_1 log\big(p(x_i, y_i)\big) + \beta_2 e(x_i, y_i) + \beta_3 w(x_i, y_i) + \beta_4 log\big(p(x_i, y_i)\big) * w(x_i, y_i) + \epsilon(x_i, y_i)$$

where the first term is now a function of both latitude and longitude, e denotes evaporation and the error $\epsilon$ is assumed to have zero mean and a stationary covariance with an exponential covariance function and a nugget term,

$$Cov\Big(\epsilon(x_i, y_i), \epsilon\big(x_j, y_j\big)\Big) = exp\left\{\frac{-d\Big((x_i, y_i), \big(x_j, y_j\big)\Big)}{v}\right\} + \sigma^2 1\big\{(x_i, y_i) = \big(x_j, y_j\big)\big\}$$

where d is the Euclidean distance between the spatial inputs. Estimated model coefficients for Argo and Aquarius SSS data are provided in Figure 4.

## Results

The skewness computed from the Aquarius and Argo near-surface data shows that it is overwhelmingly negative, as expected (Fig. 1). At the 300 m level, the skewness appears normally distributed with a mean near zero.

For Argo near the surface (Fig. 2, top), there is a large area of negative values in the tropical Pacific and Atlantic. In the Atlantic, the fresh plumes associated with the Amazon and Congo rivers, both of which discharge at the equator, can be seen. At the 300 m level (Fig. 2, middle), no distinct patterns are visible in the equatorial region. This supports the hypothesis that the negative patterns are caused by freshwater fluxes near the surface.

For Aquarius, negative values are seen north of the equator in the Pacific and Atlantic and off the equator in the South Pacific (Fig. 2, bottom). Positive values can be seen at the eastern boundaries of each ocean basin and in some areas of the western equatorial Pacific and tropical

Indian Ocean. Skewness computed from Argo data at the surface looks different from Aquarius, potentially because of the larger window size.

The model results (Fig. 3) show negative fits between the natural logarithm of precipitation and skewness for Aquarius and between the interaction of wind and the natural logarithm of precipitation for both Aquarius and Argo at 10 m. Although these statistics do not take into account spatial correlation or control for other variables, they suggest a possible underlying trend to be further investigated.

The spatially varying mixed effects model is fitted on data in the equatorial region of the Pacific (20˚S-20˚N, 160˚E-90˚W). Due to computational constraints, we fit this model to a subset of size 1000 of the total observations available in that region. Figure 4 shows that most of the spatial variability in SSS can be described by the spatially varying intercept function and spatial error variance. The covariates do not appear to contribute substantial predictive power under this model. A possible explanation for this is the flexibility of the mean process relative to the predictive ability of the covariates. We plan to investigate alternative spatial regression models in the future.

## Future work

In the future, we would like to continue to refine the comparison between the Aquarius and Argo products in a more "apples-to-apples" sense, looking at similar temporal and spatial ranges and smoothing scales. The regressions we have done have used mean covariates. We would like to do the same using seasonal values, as the primary mode of variability for SSS is seasonal. Beyond the seasonal scale, we would like to look at time variability of skewness using time-dependent regression techniques (e.g., functional regression). Finally, we would like to look at joint skewness between SSS and sea surface temperature (SST). Initial looks at SST indicate that it is positively skewed. This makes some sense as high values of SST are indicative of events at the surface that may decrease the density and make the water column more stably stratified. A joint measure of skewness may help to clarify if surface forcings change both SSS and SST in the same way.

# References


Asher, W. E., A. T. Jessup, R. Branch, and D. Clark (2014a), Observations of rain-induced near-surface salinity anomalies, *Journal of Geophysical Research: Oceans*, *119*(8), 5483-5500, doi:10.1002/2014JC009954.

Asher, W. E., A. T. Jessup, and D. Clark (2014b), Stable near-surface ocean salinity stratifications due to evaporation observed during STRASSE, *Journal of Geophysical Research: Oceans*, *119*(5), 3219-3233, doi:10.1002/2014JC009808.

Dodge, Y. (2008), *Concise Encyclopedia of Statistics, The*, 616 pp., Springer, New York.

Kalnay et al., (1996) The NCEP/NCAR 40-year reanalysis project, Bull. Amer. Meteor. Soc., 77, 437-470.

Liu, Z., D. Ostrenga, W. Teng, and S. Kempler, 2012: Tropical Rainfall Measuring Mission (TRMM) Precipitation Data and Services for Research and Applications. *Bull. Amer. Meteor. Soc.*, **93**, 1317–1325, https://doi.org/10.1175/BAMS-D-11-00152.1.

Roemmich, D. and J. Gilson (2009). The 2004–2008 mean and annual cycle of temperature, salinity, and steric height in the global ocean from the Argo Program. *Progress in Oceanography*, 82:81–100.

Yu, L. (2010), On Sea Surface Salinity Skin Effect Induced by Evaporation and Implications for Remote Sensing of Ocean Salinity, *Journal of Physical Oceanography*, *40*, 85-102, doi:110.1175/2009JPO4168.1171.

Yu, L., X. Jin, and R. Weller (2008), Multidecade Global Flux Datasets from the Objectively Analyzed Air-sea Fluxes (OAFlux) Project: Latent and Sensible Heat Fluxes, Ocean Evaporation, and Related Surface Meteorological Variables. *Woods Hole Oceanographic Institution Tech. Rep.*, 64pp.

# Figures

Figure 1

Distributions of skewness computed from (left) Argo data at 10 m depth, (center) Argo data at 300 m depth and (right) Aquarius data at the surface. Ordinate is the number of 1˚ squares with a given value of skewness.

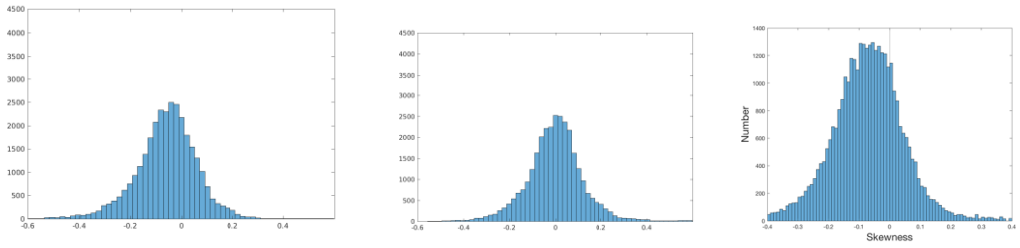


Figure 2

Global maps of skewness computed from (top) Argo data at 10 m depth, (middle) Argo data from 300 m depth and (bottom) Aquarius data. Unitless color scale is on the right. Note different color scales for the different products.

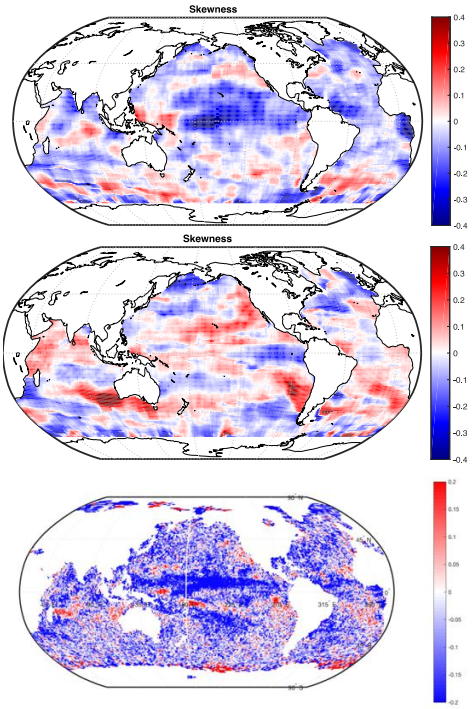


Figure 3

Univariate regression results showing fits to (left column) Argo at 10 m and (right column) Aquarius in the tropical Pacific. Top row shows skewness vs. log(p) and bottom row shows skewness vs. w * log(p).

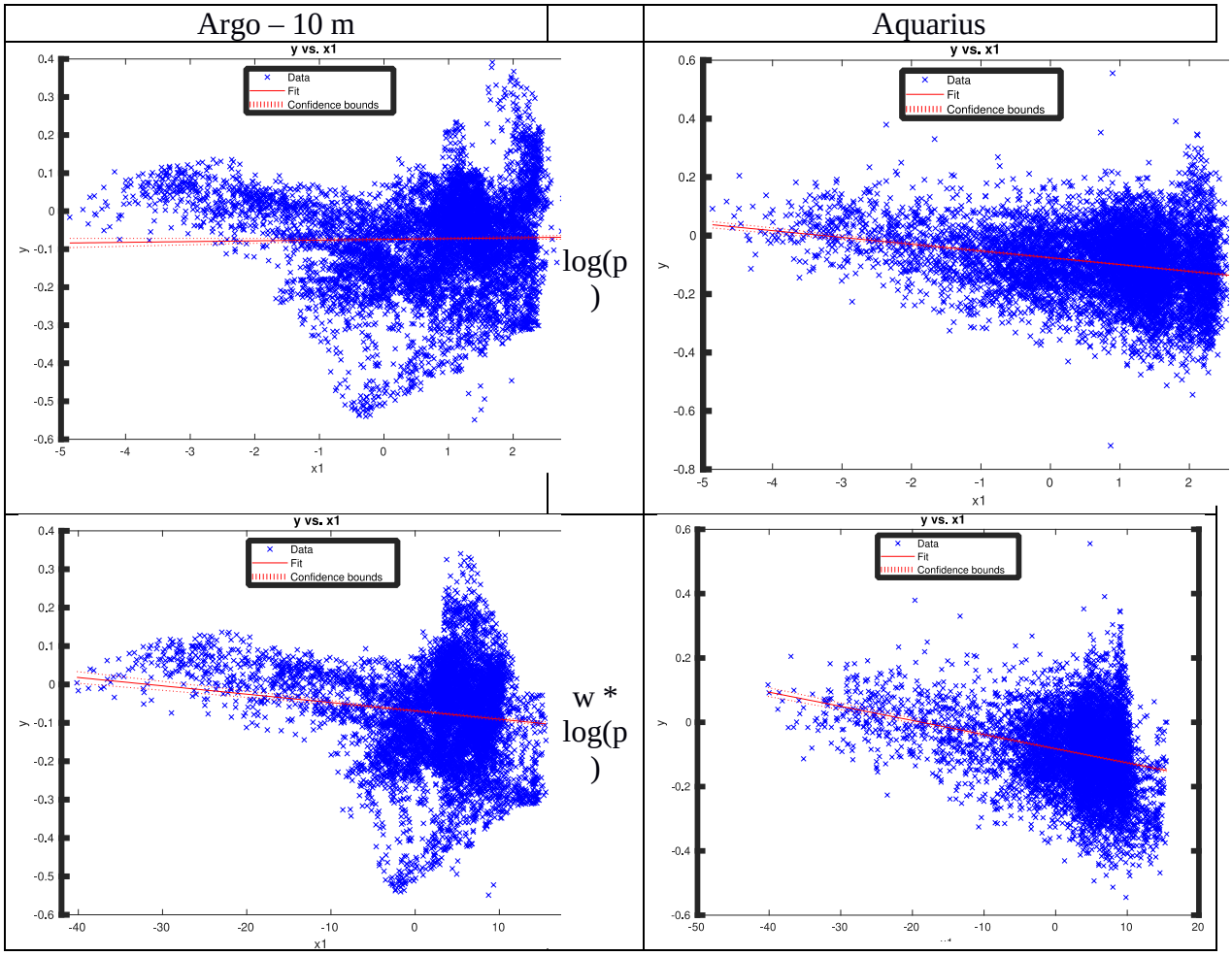
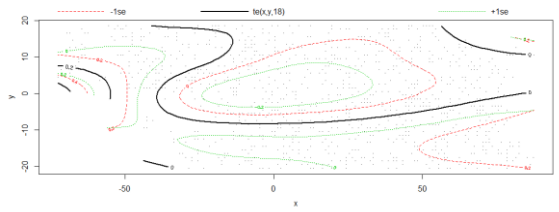
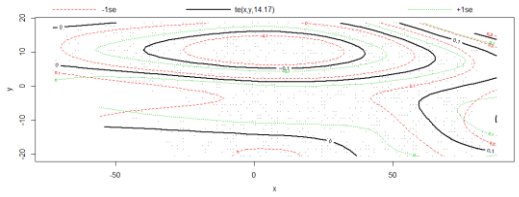


Figure 4

Spatial regression results for the tropical Pacific.

| Argo – 10 m | Aquarius |
|---|---|
| Formula:<br>kappa.ao ~ te(x, y) + evap.aq + wind.aq * lograin.aq<br><br>Parametric coefficients:<br>Estimate Std. Error t value Pr(>\|t\|)<br>(Intercept) -7.228e-02 5.902e-02 -1.225 0.2210<br>evap.aq 2.789e-05 1.484e-04 0.188 0.8509<br>wind.aq -6.973e-03 3.514e-03 -1.984 0.0475 *<br>lograin.aq 1.440e-02 7.907e-03 1.821 0.0689 .<br>wind.aq:lograin.aq -1.153e-03 1.262e-03 -0.913 0.3613<br>---<br>Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1<br><br>Approximate significance of smooth terms:<br>edf Ref.df F p-value<br>te(x,y) 18 18 7.721 <2e-16 *** | Formula:<br>kappa.aq ~ te(x, y) + evap.aq + wind.aq * lograin.aq<br><br>Parametric coefficients:<br>Estimate Std. Error t value Pr(>\|t\|)<br>(Intercept) -0.2160672 0.0632038 -3.419 0.000656 ***<br>evap.aq 0.0005941 0.0004448 1.336 0.181921<br>wind.aq 0.0050045 0.0088182 0.568 0.570495<br>lograin.aq 0.0049667 0.0230183 0.216 0.829214<br>wind.aq:lograin.aq -0.0006528 0.0034781 -0.188 0.851173<br>---<br>Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1<br><br>Approximate significance of smooth terms:<br>edf Ref.df F p-value<br>te(x,y) 14.17 14.17 4.772 1.24e-08 *** |
|  |  |

samsi
NSF·Duke·NCSU·UNC

# Undergraduate Workshop
## October 23-24, 2017
### SCHEDULE

**Monday, October 23, 2017**
**SAMSI**, RTP

| | |
|---|---|
| 8:30-8:45 | Shuttle from Hotel to SAMSI and Registration |
| 8:45--9:00 | *Opening Remarks*<br>**Elvan Ceyhan**, SAMSI/NCSU |
| 9:00-9:45 | *Tutorial on R Software*<br>**Maggie Johnson**; **Huang Huang**; **Whitney Huang**, SAMSI |
| 9:45-10:00 | Break |
| 10:00-11:00 | *Applications in Climate Context*<br>**Michael Wehner**, Lawrence Berkeley National Laboratory (LBNL) |
| 11:00-11:15 | Break |
| 11:15-12:15 | *Distributions and Statistics in Climate Data*<br>**Doug Nychka**, National Center for Atmospheric Research (NCAR) |
| 12:15-1:30 | Lunch |
| | Lectures and Hands-on Demos*:* Overview of Statistical Models/Methods for Climate Research |
| 1:30-2:15 | *"Introduction to Spatial Data Analysis with R"*<br>**Maggie Johnson**, SAMSI |
| 2:15-2:25 | Break |
| 2:25-3:10 | *"Extreme Value Analysis for Climate Research"*<br>**Whitney Huang**, SAMSI |
| 3:10-3:20 | Break |
| 3:20-4:00 | *Hands on Demos for Climate Extremes R Software*<br>**Maggie Johnson**; **Huang Huang**; **Whitney Huang**, SAMSI |
| 4:00-4:30 | Concluding Remarks and Wrap Up<br>**Maggie Johnson**; **Huang Huang**; **Whitney Huang**, SAMSI |

| | |
|---|---|
| 4:30-5:15 | *Panel on Career Opportunities*<br>**Jared Rennie**, Cooperative Institute of Climate and Satellites – North Carolina (CICS-NC) and National Centers for Environmental Information (NCEI)<br>**SAMSI: Postdocs** and **Graduate Fellows** |
| 5:15--6:30 | Dinner |
| 6:30-6:45 | Shuttle to Hotel |

**Tuesday, October 24, 2017**
**SAMSI**, RTP

| | |
|---|---|
| 9:15-10:15 | *Statistical Development and Challenges for Paleo-climate Reconstruction*<br>**Bo Li**, University of Illinois |
| 10:15-10:30 | Break |
| 10:30-12:00 | *How was this Made?: making dirty data into something usable at NCEI*<br>**Scott Stevens** and **Jared Rennie**, Cooperative Institute of Climate and Satellites – North Carolina (CICS-NC) and National Centers for Environmental Information (NCEI) |
| 12:00-12:15 | Adjourn and Box Lunch |
| 12:15→ | Shuttles Depart to RDU Airport |

samsi
NSF·Duke·NCSU·UNC

# Undergraduate Modeling Workshop
## May 20-25, 2018
### SCHEDULE

**Sunday, May 20**
**CENTENNIAL CAMPUS**, NC State University

| | |
|---|---|
| 10:00-4:30 | Participant check-in to NCSU **Centennial Campus- Valley Hall Wolf Ridge Apartments** NCSU vans will pick up from RDU. (GPS for drivers: 35.767847, -78.673144) |
| 5:00-7:00 | Welcome reception at "On the Oval Culinary Creations" on NCSU Centennial CampusAddress of the Building: **NCSU Centennial Campus Food Court** (GPS for drivers 35.770080,-78.674372) Coordinator: **Thomas Gehrmann** (NCSU) - Introductions, presentation expectations. **All SAMSI postdocs and graduate fellows** |

**Monday, May 21**
**Institute of Advanced Analytics (IAA)**, (Roanoke Room) - NC State Centennial Campus

| | |
|---|---|
| 7:30-9:00 | Individual Breakfast "On the Oval Culinary Creations." |
| 9:15-9:30 | Introduction and overview of SAMSI: **Elvan Ceyhan**, Deputy Director, SAMSI Logistics of week: **Thomas Gehrmann, NCSU** |
| 9:30-10:30 | **Setting the Scene:** Why we believe the climate is changing and why the research focus will be on its impacts. **Doug Nychka (NCAR)** Why applied mathematicians and statisticians need to be at the climate table. **Chris Jones (UNC-CH)** |
| 10:30-11:00 | Beverage Break in Vance Reception Lounge. Beverage available from 9:00am on (this day only). 11:00-12:00 Overview of program and getting to know each other |
| 12:00-1:30 | Group Lunch in Vance Reception Lounge. During lunch: Discussion of opportunities in analytics **Michael Rappa**, Institute of Advanced Analytics (NCSU) |
| 1:30-4:30 | **Project group meetings:** *Goal: Formulate problem and consider roles of individuals in groups* |
| 4:30-5:00 | **Reports from groups**: Problem formulation and group structure |
| 5:00-6:30 | **Individual Dinner** at "On the Oval Culinary Creations", or where each group may decide. |
| 7:00-8:30 | Intro to R Lab in **IAA Roanoke Classroom (no propping doors open after 5pm)**. |

**Tuesday, May 22**
**Institute of Advanced Analytics (IAA)**, (Roanoke Room & Breakout Rooms) - NC State Centennial Campus

| | |
|---|---|
| 8:45-9:00 | Recap and announcements: **Doug, Chris, Elvan and Thomas**<br>*Goal for day: Find data and prepare/formulate model as needed* |
| 9:00-10:15 | Group meetings |
| 10:15-10:45 | Break in Vance Reception Lounge. |
| 10:45-12:00 | Group meetings |
| 12:00-1:30 | Lunch break |
| 1:30-2:30 | **Guest Lecture:** Hierarchical models for sparsely sampled high-dimensional LiDAR and forest variables: An interior Alaska FIA case study**. Andrew Finley (Michigan State)** |
| 2:30-3:00 | Break in Vance Reception Lounge |
| 3:00-4:00 | Group meetings |
| 4:00-5:30 | **Progress Reports and Feedback:** Each group has 10 minutes |
| 5:30-6:46 | Individual Dinner |
| 6:45- | Open Group Work time at each group's discretion. |

**Wednesday, May 23**
**Institute of Advanced Analytics (IAA)**, (Roanoke Room & Breakout Rooms) - NC State Centennial Campus

| | |
|---|---|
| 8:45-9:00 | Recap and announcements: **Doug, Chris, Elvan and Thomas**<br>*Goal for day: reconcile models and data, develop code as needed* |
| 9:00-10:15 | Group meetings |
| 10:15-10:45 | Lounge |
| 10:45-12:00 | Group meetings |
| 12:00-1:30 | Lunch Break |
| 1:30-3:00 | **Progress Reports and Feedback:** Each group has 10 minutes |
| 3:00-3:30 | Break in Vance Reception Lounge. |
| 3:30-4:30 | Group meetings |
| 4:30-6:00 | SAMSI Graduate Fellow Poster Session at “On the Oval” |
| 6:00-7:30 | **Group Dinner and Dessert at “On the Oval”** |

**Thursday, May 24**
**Institute of Advanced Analytics (IAA)**, (Roanoke Room & Breakout Rooms) - NC State Centennial Campus

| | |
|---|---|
| 8:45-9:00 | Recap and announcements: **Doug, Chris, Elvan and Thomas**<br>*Goal for day: focus results, revisit original problem, distill lessons, message and any future* ideas for research and/or instruction |
| 9:00-10:15 | Group meetings |
| 10:15-10:45 | Break in Vance Reception Lounge |
| 10:45-12:00 | Group meetings |
| 2:30-3:00 | Break in Vance Reception Lounge |
| 3:00-5:30 | Group meetings |
| 5:30-6:45 | Individual Dinner |
| 6:45- | Open Group Work time at each group's discretion |

**Friday, May 25**
**Institute of Advanced Analytics (IAA)**, (Roanoke Room) - NC State Centennial Campus

| | |
|---|---|
| 8:45-9:00 | Announcements and wrap-up: **Elvan Ceyhan**, Deputy Director (SAMSI)<br>*Goal for day: Articulation of outcomes* |
| 9:00-10:15 | Group presentations and discussions<br>[25min each; bring slides loaded on a USB Thumbdrive] |
| 10:15-10:30 | Break in Vance Reception Lounge |
| 10:30-11:45 | Group presentations [25min each] |
| 11:45-12:00 | Vote of thanks and Project group photo with mentors |
| 12:00 | Adjourn and take box lunch from Vance Reception Lounge for on the road or shuttle to airport. Shuttle van to RDU is provided |

**SAMSI MODELING WORKSHOP PROJECTS**

**TOPIC 1: ARCTIC SEA-ICE**

Leader: Christian Sampson

Title: Low order models of Arctic sea ice and the effect of process parameterizations for predictions.

Outline: Covering 7-10% of the Earth's surface, Sea ice is a critical component of the climate system and is sensitive to changes in global temperature. In the paper "Nonlinear threshold behavior during the loss of Arctic sea ice" (Eisenmann & Wetlaufer 2009) a low order model for sea ice thickness is presented which exhibits hysteresis in sea ice loss as the climate warms. This model considers the most impactful processes affecting sea ice volume and parameterizes them, good representations of these processes is thusly paramount. One such important process is the ice albedo feedback. In the Arctic summer months, snow melt turns into dark ponds which sit atop the sea ice. Dubbed melt ponds, these ponds lower the over albedo (reflectance) of the ice causing the absorption of more incoming solar radiation, promoting more melting and further lowering the ice albedo and continuing. In this project we will investigate how different parameterizations of this process affect model out put and hysteresis. We can also examine how changing seasonal temperature variations, such as early warming and melting, will affect the fate of the Arctic sea ice pack.

**Group members: Madeleine Braye, Yasmin Eady , Nicole Jacobs, Jesse Liu, Ryan Norlinger, Jose San Martin,**

**TOPIC 2: AIR QUALITY**

Leader: Yawen Guan

Title: data analysis on air pollutant exposures. ( we will use R for data analysis)

Outline: Fine particulate matter (PM2.5) is a mixture of air pollutants that, at a high concentration level, has adverse effects on human health. An interesting statistics problem is to estimate these pollutant exposures for the entire US, such estimates can be used to inform policy and decision making. During the workshop, we will work on two major sources of air quality data that are used by the EPA to estimate pollutant exposures, including monitoring data and the Community Multiscale Air Quality (CMAQ) model. The monitoring stations provide fairly accurate measurements of the pollutants, however they are sparse in space and take measurements at a coarse time resolution, typically 1-in-3 or 1-in-6 days. On the other hand, the CMAQ model provides daily concentration levels of each component with complete spatial coverage on a grid; these model outputs, however, need to be evaluated and calibrated to the monitoring data. We will explore these air quality data for the summer of 2011 and brainstorm on statistical models to estimate air pollutant exposures.

**Group members: Meixi Chen, Vincent Gonzales, Alan Ji, Chandni Malhotra, Hongyu Mao, Sharon Sung**

**TOPIC 3: OCEAN TEMPERATURE**

Leader: Mikael Kuusela

Title: Spatial statistics for reconstructing ocean temperature fields using Argo float data

Outline: Description: Argo floats (http://www.argo.ucsd.edu/) measure ocean temperature and salinity in the upper 2000 meters on a global scale. Spatial statistics provides the tools for reconstructing the full temperature and salinity fields based on sparse point observations from the floats. The large size and complex spatio-temporal dependence structure of the Argo data set mean that state-of-the-art statistical techniques are needed for efficient reconstructions. The goal of this project is to learn to explore, visualize, model and spatially interpolate subsets of Argo data using R. The preliminary plan is to use the "fields" package and to see how far we can push the tools there when reconstructing regional ocean temperature fields.

**Group members: Ahmet Hatip, Alex Hayes, Alexa Maxwell, Joseph Struzeski, Ingrid Tchkaoua, Wengbo Wang**

**TOPIC 4: SOUTHEASTERN US RAINFALL**

Leader: Whitney Huang

Title: Gulf coast rainfall data analysis

Outline: In this project the group members will play with daily rainfall data collected in Gulf coast (535 stations in total) from 1949 to 2017. The purpose of this exercise are to:

1) to give students an idea of a typical example of a climate data set (spatio-temporal data) and some associated scientific questions (e.g. how rainfall extremes vary in space and time and how that might affected by other things like green house greenhouse gases or temperatures).

2) to get students familiar with data analysis using R including data manipulation, data visualization, and data summary.

3) to introduce some statistical methods (e.g. time series analysis, spatial statistics, extreme value analysis) to analyze this kind of data to "answer" (perform statistical inference) the questions of interest.

**Group members: Lin Ge, Jianan Jang, Jessica Robinson, Erin Song, Seth Temple, Adam Wu**

**TOPIC 5: VEGETATION**

Leader: Maggie Johnson

Title: Analysis of vegetation using remote sensing data

Outline: Imaging spectrometers housed on satellites are used to obtain data on vegetated surfaces by measuring reflectance from the Earth's surface. These data are very useful as they provide information on changes in vegetation over time on global scales, which is important to assess the impacts of changes in weather and climate, and the effects of agricultural practices. However, the information provided by these data can be limited due to the resolution of the sensors and inhibiting factors such as cloud cover. In this project we will use two remote sensing sources of the Enhanced Vegetation Index (EVI) to analyze vegetation over Nebraska. The first, Landsat EVI, is available at fine spatial resolution, but is sparse in time. The second, MODIS EVI, is obtained regularly in time, but is available at a much coarser spatial resolution. We will use these data to explore the relationships between vegetation and changes in temperature and landcover (e.g. corn fields versus grasslands), as well as to classify the landcover in unknown regions.

**Group members: Samuel Hood, Zhihan Lu, Rita Pradhudesai, Thomas Rechtman, Meghana Tatneni, Ganlin Ye**

**TOPIC 6: FOREST COVER**

Leader: Huang Huang

Title: inference on forest variables from complete-coverage LiDAR data and sparse observations

Outline: We have two sources for forest variables, from direct measurements, which are always expensive and would be sparse in space, and correlated LiDAR data that has complete coverage. The Bonanza Creek Experimental Forest (BCEF) is a Long-Term Ecological Research (LTER) site consisting of vegetation and landforms typical of interior Alaska. People are interested in three forest variables: above-ground biomass (AGB); tree density (TD); basal area (BA). The brightness, greenness, and wetness tasseled cap indices can be used as covariates to explain the forest variables. In the undergraduate workshop project, students can brainstorm from the easiest regression models to more sophisticated spatial models and compare the differences of inferences from different ideas.

**Group members: Richard Groenwald, Mehmut Hatip, Katrina Lewis, Jennifer Soter, Astride Tchkaoua, Sylvester Wiebeck**

SAMSI Public Lecture

# *The Storm Next Time: Hurricanes and Climate Change*

samsi
NSF•Duke•NCSU•UNC

*Monday, October 9, 2017 @ 7:30pm*

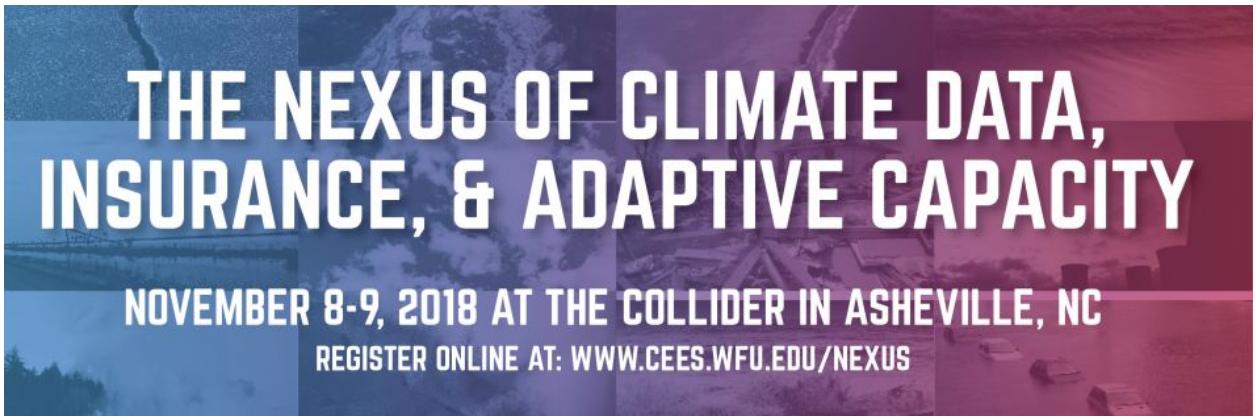
**Genome Sciences Building, G100 Auditorium**
**University of North Carolina - Chapel Hill**

*Lecture introduced by Carol L. Folt,*
*Chancellor, University of North Carolina at Chapel Hill*

PRESENTED BY:
**Dr. Kerry Emanuel**
Professor of Atmospheric Science
MIT

The recent tragedies of Hurricanes Harvey and Irma, together with earlier extreme events such as Hurricanes Katrina and Sandy, has raised the question whether the apparent increasing severity of such events can be attributed to the human influence on greenhouse gas warming. Dr. Emanuel will review the growing consensus that the incidence of the strongest storms will increase over time, even though there may be a decline of the far more numerous weaker events.


***This lecture is free and open to the public!***




"Can insurers extend their self-chosen historical role in addressing root causes (as founders of the first fire departments, building codes and auto safety testing protocols) to one of preventing losses at a much larger scale, namely, the global climate?"

-Evan Mills, *Science* 2005

## DESCRIPTION OF WORKSHOP

This workshop will facilitate a national, interdisciplinary scientific research discussion on modeling and managing climate change risks between three different but related research communities: the climate modeling and data community, statisticians, and researchers within the insurance and reinsurance industries. Each of these communities maintains an active research agenda addressing the measurement and management of climate-related risks, and each seeks to inform decision-makers and stakeholders in government, business, and scientific communities. However, the methods, emphases and collaborative connections differ in these three research communities. This workshop will combine all three streams of research, focus on the gaps between each, explore ways to shrink these gaps through novel interdisciplinary approaches that require collaboration across research communities. We will collectively seek to answer the question posed by Evan Mills (2005) in *Science*: "Can insurers extend their self-chosen historical role in addressing root causes (as founders of the first fire departments, building codes, and auto safety testing protocols) to one of preventing losses at a much larger scale, namely, the global climate?"

**THE VENUE**

The Collider
1 Haywood St, Asheville, NC 28801

**Tentative Schedule (*subject to change*)**

**Day 1: Schedule of Events**

| Time | Event | Location |
|---|---|---|
| 7:45 – 8:30 am | Registration and light refreshments | Lounge/Reception |
| 8:30 – 8:50 am | Opening Remarks: **Josh Dorfman**, CEO, The Collider. **Robert Erhardt**, Workshop Chairperson. | |
| | ***Session 1: The Current State (moderator: Megan Robinson, Chief Operating Officer, The Collider)*** | |
| 8:50 – 9:20 am | **Deke Arndt**, Chief, Climate Monitoring Branch, NOAA National Centers for Environmental Education | Technology Theater |
| 9:20 – 9:50 am | **Jennifer Jurado**, Chief Resilience Officer and Director, Broward County Government | Technology Theater |

| | | |
|---|---|---|
| 9:50 – 10:20 am | **Roy Wright**, President & CEO, Institute for Business & Home Safety; Former Deputy, FEMA | Technology Theater |
| 10:20 – 10:40 am | Discussion | |
| 10:40 – 11:00 am | Break | |
| | ***Session 2: Future States (moderator, Rob Erhardt, Associate Professor of Statistics, Wake Forest University)*** | |
| 11:00 – 11:30 am | **Doug Nychka**, Professor of Applied Mathematics & Statistics at Colorado School of Mines | Technology Theater |
| 11:30 am – 12:00 pm | **Mitch Roznik**, University of Manitoba | Technology Theater |
| 12:00 – 12:30 pm | **Jeremy Hess,** Associate Professor of Emergency Medicine, University of Washington | Technology Theater |
| 12:30 – 1:45 pm | Lunch | |
| | ***Session 3: Changing Extremes and Their Impacts (moderator: Richard Smith, Professor of Statistics, UNC- Chapel Hill)*** | |
| 2:00 – 2:30 pm | **Dan Cooley,** Associate Professor of Statistics, Colorado State University | Technology Theater |
| 2:30 – 3:00 pm | **Raghuveer Vinukollu,** Natural Catastrophe Solutions Manager at Munich Reinsurance America, Inc. | Technology Theater |
| 3:00 – 3:15 pm | Discussion | |

| | | |
|---|---|---|
| 3:15 – 3:45 pm | Breakout Group Formation | |
| 3:45 – 4:00 pm | Break | |
| 4:00 – 5:30 pm | Breakout Group Discussions: Current climate data products & their use in insurance; Climate projections and their use in insurance; Extreme flood risks and FEMA; Climate and Health | Overlook Lounge |
| 5:30 – 7:30 pm | Poster Session and Reception (co-sponsored by the Society of Actuaries and Casualty Actuarial Society) | Overlook Lounge |

---

**Day 2: Schedule of Events**

| Time | Event | Location |
|---|---|---|
| 7:45 – 8:30 am | Registration and Refreshments | |
| 8:30 – 8:40 am | Opening Remarks: Robert Erhardt | Technology Theater |
| | ***Session 4: Synthesis of Climate Data and Insurance (moderator: Jesse Bell, Claire M. Hubbard Associate Professor, University of Nebraska Medical Center)*** | |
| 8:40 – 9:10 am | **Adam Smith**, NOAA | Technology Theater |
| 9:10 – 9:40 am | **Steve Kolk**, Kolkulations | Technology Theater |
| 9:40 – 10:10 am | **Mathieu Boudreault,** Université du Québec à Montréal | Technology Theater |

| | | |
|---|---|---|
| 10:10 – 10:25 am | Discussion | |
| 10:25- 10:40 am | Break | |
| 10:40 – 11:00 am | Brief Reports from Day #1 Breakout Group Discussion Leads | Technology Theater |
| 11:00 am – 12:30 pm | Reconfigure Breakout Group Discussions: Current Climate Data Projections and Their Use, Climate Projections and Their Use, Extremes, Flood Risks and FEMA and Climate and Health | |
| 12:30 – 1:30 pm | Lunch | |
| 1:30 – 2:30 pm | Open Group Discussion and Wrap-Up | Technology Theater |
| 2:30 pm | Adjourn | |
| 2:30 – 4:00 pm | Organizing Committee Wrap-Up | |

## ORGANIZING COMMITTEE

**ROB ERHARDT** Associate Professor, Wake Forest | **JESSE BELL** Associate Professor, University of Nebraska Medical Center
**BRIAN BLANTON** Director of Earth Data Sciences, RENCI | **FRANK NUTTER** President, Reinsurance Association of America
**MEGAN ROBINSON** Chief Operations Officer, The Collider | **RICHARD SMITH** Professor UNC-CH, Former Director of SAMSI