samsi
NSF·Duke·NCSU·UNC

# Climate Program Remote Sensing Workshop
# February 12-14, 2018

## SPEAKER TITLES/ABSTRACTS

**Veronica Berrocal**
University of Michigan

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

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.

**Carmen Boening**
JPL

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

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, di_erent 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.

**Venkat Chandrasekaran**
Caltech

"Computational and Statistical Trade-offs in Data Analysis"

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.

**Luca Cinquini**
JPL

“The Earth System Grid Federation as a Testbed for Global, Distributed Data Analytics”

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.

**Dan Crichton**
JPL

“An Introduction to Systems and Software Architecture Considerations for Scaling Data Analysis”

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.

**Manlio De Domenico**
Fondazione Bruno Kessler

“Multilayer Modeling and Analysis of Complex (Systems) Data”

Complex systems are characterized by constituents -- from neurons in the brain to individuals in a social network -- which exhibit special structural organization and nonlinear dynamics. As a

consequence, a complex system cannot be understood by studying its units separately because their interactions lead to unexpected emerging phenomena, from collective behavior to phase transitions. 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 -- including molecular biology, neuroscience, urban transport, telecommunications -- 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, allowing to quantify the resilience of a system to perturbations (e.g., localized failures or targeted attacks), improving forecasting of spreading processes and accuracy in classification problems.

References

MDD et al, Phys. Rev. X 3, 041022 (2013)
MDD, A. Sole-Ribalta, S. Gomez & A. Arenas, PNAS 11, 8351 (2014)
MDD, Sole, Omodei, Gomez & Arenas, Nature Comms. 6, 6868 (2015)
MDD, Nicosia, Arenas & Latora, Nature Comms. 6, 6864 (2015)
Baggio, Burnsilver, Arenas, Magdanz, Kofinas & MDD, PNAS, 113, 13708 (2016)
MDD, Granell, Porter & Arenas, Nature Phys., 12, 901 (2016)

**Rajarshi Guhaniyogi**
University of California, Santa Cruz

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

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.

**Dorit Hammerling**
NCAR

"High Performance Computing and Spatial Statistics: an overview of recent work at NCAR"

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 bene_t 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

**Jonathan Hobbs**
JPL

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

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.

**Maggie Johnson**
SAMSI

"A Notional Framework for a Theory of Data Systems"

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)

**Emily Kang**
University of Cincinnati

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

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.

**Matthias Katzfuss**
Texas A&M University

"Multi-resolution Approaches for Big Spatial Data"

Remote-sensing instruments have enabled the collection of big spatial data over large spatial domains such as entire continents or the globe. Basis-function representations are well suited to big spatial data, as they can enable fast computations for large datasets and they provide flexibility to deal with the complicated dependence structures often encountered over large domains. We propose two related multi-resolution approximations (MRAs) that use basis functions at multiple resolutions to (approximately) represent any covariance structure. The first MRA results in a multi-resolution taper that can deal with large spatial datasets. The second MRA is based on a multi-resolution partitioning of the spatial domain and can deal with truly massive datasets, as it is highly scalable and amenable to parallel computations on distributed computing systems.

**Mike Little**
NASA Earth Science Technology Office

"Distributed Access and Analysis: NASA"

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 inuenced 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 _nd 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.

**Jessica Matthews**
NOAA CICS

"Optimization Methods in Remote Sensing"

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 o_-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.

**Jay Morris**
NOAA

"Satellites and Stovepipes"

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.

**Richard Smith**
SAMSI/UNC

"Blocking Methods for Spatial Statistics and Potential Applications to Distributed Data"

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 _eld 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 lowdimensional 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.

**Hui Su**
JPL

“Evaluating and Constraining Climate Model Simulations Using Satellite Data”

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.

**Vineet Yadav**
JPL

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

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.

**Zhengyuan Zhu**
Iowa State University

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

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 cannot 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.