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 ϵ 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.