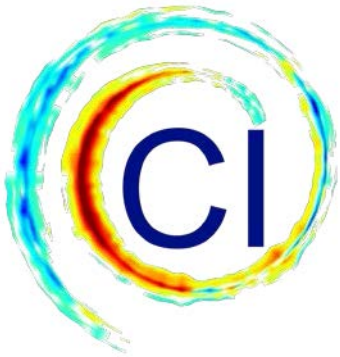


Proceedings of the 6th International Workshop on Climate Informatics: CI 2016

Volume editors
Arindam Banerjee
Wei Ding
Jennifer Dy
Vyacheslav Lyubchich
Andrew Rhines

Series editors
Imme Ebert-Uphoff
Claire Monteleoni
Doug Nychka

NCAR Technical Notes
NCAR/TN-529+PROC

National Center for Atmospheric Research
P. O. Box 3000
Boulder, Colorado
80307-3000
www.ucar.edu

NCAR IS SPONSORED BY THE NSF

NCAR TECHNICAL NOTES

http://library.ucar.edu/research/publish-technote

The Technical Notes series provides an outlet for a variety of NCAR Manuscripts that contribute in specialized ways to the body of scientific knowledge but that are not yet at a point of a formal journal, monograph or book publication. Reports in this series are issued by the NCAR scientific divisions, serviced by OpenSky and operated through the NCAR Library. Designation symbols for the series include:

EDD – Engineering, Design, or Development Reports
Equipment descriptions, test results, instrumentation, and operating and maintenance manuals.

IA – Instructional Aids
Instruction manuals, bibliographies, film supplements, and other research or instructional aids.

PPR – Program Progress Reports
Field program reports, interim and working reports, survey reports, and plans for experiments.

PROC – Proceedings
Documentation or symposia, colloquia, conferences, workshops, and lectures. (Distribution maybe limited to attendees).

STR – Scientific and Technical Reports
Data compilations, theoretical and numerical investigations, and experimental results.

The National Center for Atmospheric Research (NCAR) is operated by the nonprofit University Corporation for Atmospheric Research (UCAR) under the sponsorship of the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

National Center for Atmospheric Research
P. O. Box 3000
Boulder, Colorado 80307-3000

NCAR/TN-529+PROC
NCAR Technical Note

2016-09

Proceedings of the Sixth International Workshop on Climate Informatics: CI 2016

Editors
Arindam Banerjee
University of Minnesota, Minneapolis, MN
Wei Ding
University of Massachusetts, Boston, MA
Jennifer Dy
Northeastern University, Boston, MA
Vyacheslav Lyubchich
University of Maryland, College Park, MD
Andrew Rhines
Harvard University, Cambridge, MA

Series Editors
Imme Ebert-Uphoff
Colorado State University, Fort Collins, CO
Claire Monteleoni
George Washington University, Washington, DC
Douglas Nychka
National Center for Atmospheric Research, Boulder, CO

Computational and Information Systems Laboratory (CISL)
Institute for Mathematics Applied to the Geosciences (IMAGe)

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH
P. O. Box 3000
BOULDER, COLORADO 80307-3000
ISSN Print Edition 2153-2397
ISSN Electronic Edition 2153-2400
ISBN 9780997354812

How to Cite this Document

A. Banerjee, W. Ding, J. Dy, V. Lyubchich, A. Rhines (Eds.), I. Ebert-Uphoff, C. Monteleoni, D. Nychka (Series Eds.), Proceedings of the 6th International Workshop on Climate Informatics: CI 2016. NCAR Technical Note NCAR/TN-529+PROC, Sept 2016, doi: 10.5065/D6K072N6

The ISBN number (optional use) for this document is 978-0-9973548-1-2.

The CI logo on the cover page is courtesy of Michael Tippett. Colors show deviations of sea-surface temperatures from their climatological values in the equatorial Pacific from January 1997 to April 2000 with time going counter-clockwise.

6th International Workshop on Climate Informatics CI 2016

Table of Contents

Papers

Foreword by the CI 2016 Workshop Chairs

Climate informatics is an emerging interdisciplinary field that combines climate science with statistics, machine learning, and data mining. The Climate Informatics workshop series is a yearly workshop which started in 2011. The goal of the workshop series is to bring together researchers in these various areas, to stimulate discussion of new ideas, foster new collaborations, and thereby accelerate discovery across disciplinary boundaries.

This year's workshop was held on September 21-23, 2016, at the National Center for Atmospheric Research (NCAR), Boulder, CO. The workshop consisted of a one-day Hackathon event on the topic of Northern Hemisphere Sea Ice Prediction organized by Balazs Kegl, Camille Marini, and Andrew Rhines (Wednesday, September 21), followed by two days of the main workshop event (September 22-23). The main workshop had six invited talks from experts in both climate science and data science: Doug Nychka (National Center for Atmospheric Research), Pradeep Ravikumar (Carnegie Mellon University), Yulia Gel (University of Texas at Dallas), Jason Smerdon (Columbia University), Sudipto Banerjee (University of California at Los Angeles), and Allen Pope (National Snow & Ice Data Center). Furthermore, 34 peer-reviewed short papers were selected for presentation at the workshop. The short papers are included in these proceedings and were presented at the workshop as rapid fire spotlight presentations (2 minutes each) or spotlight presentations (10 mins each, papers by young scientists with strong positive reviews) and all papers were part of a 2-hour poster session with reception on Thursday, September 22 evening. The workshop also had a panel discussion focused on the Future of Climate Informatics on Friday, September 23, and a hiking event for community-building immediately following it.

There were 75 registered participants for the event. Thanks to the support from a NSF grant, we were able to provide travel fellowships to a total of 25 young career scientists to help cover their travel expenses to attend the workshop.

This workshop was made possible through the tireless work of the entire organizing committee. We would like to thank the program committee chairs, Slava Lyubchich and Andy Rhines, for doing an amazing job in handling the review process, selecting the high quality papers for the workshop, and helping with organizational details. They often went outside their call of duty to take care of several aspects of the workshop planning and organization. We thank Wei Ding, the publicity and publications chair, who has been extremely helpful in getting the word out about the workshop, handling general CI communications and for putting together these proceedings. With Wei's help, word about the CI workshop reached several communities, and we had numerous first time attendees. We thank Eniko Szekely, the budget chair, who has handled the budget and travel scholarships. Eniko graciously handled the complexities of handling the budget, and made sure several of the workshop attendees get their travel scholarships. All the committee members have worked hard over the past year to create an exciting program. Finally, we thank the steering committee, Imme Ebert-Uphoff, Claire Monteleoni, and Doug Nychka, for continually guiding us with various organizational aspects and for helping publish the CI2016 Proceedings. They have been extremely generous with their time and guidance, without which the workshop would not have been possible.

The support team at NCAR, Kathy Peczkowicz and Cecilia Banner, under the guidance of Doug Nychka, made sure participants did not starve (coffee breaks and reception), and stayed awake (coffee!), made

sure they had a roof over their head (hotel) and transportation to and from the workshop (bus). They put together the web page, the workshop hand-outs, and helped us with anything we needed. We also thank Jennifer Phillips of NCAR for setting up the proceedings in the NCAR OpenSky repository.

We thank all of our sponsors, The National Science Foundation, the National Center for Atmospheric Research, NVIDIA and STATMOS for their financial support.

Last, but not least, we'd like to thank the most important group of people - the participants - for contributing to, participating, and continuing to advance the state-of-the-art in climate informatics.

Arindam Banerjee
Jennifer Dy
CI2016 Workshop Co-Chairs

Foreword by the CI 2016 Steering Committee

Each passing day makes it more apparent that the threat of climate change is one of the greatest scientific and societal challenges of the 21st century. We are fortunate, however, that the amount of observational and model-simulated climate data has grown at an accelerating rate. This wealth of data creates a unique opportunity for data scientists (broadly defined, as researchers in machine learning, data mining, and statistics) to partner with climate scientists in the development of new methods for interdisciplinary knowledge discovery.

Climate Informatics is an emerging field at the interface of data science and climate science. In 2011, Claire Monteleoni and Gavin Schmidt launched the International Workshop on Climate Informatics, so that researchers from the fields of data science and climate science could learn from each other and start new collaborations. The first workshop was held at the New York Academy of Sciences and based on the initial success it rapidly evolved into an important annual event for this field. In 2012, the workshop venue moved to the Mesa Laboratory at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado. Doug Nychka, director of the IMAGe (Mathematics Applied to Geosciences) group, and his staff, have hosted the event at NCAR ever since. Overlooking the city of Boulder, and bordered by stunning cliffs, forests, and park land, this location has provided a wonderful setting for this workshop, reminding participants of the importance of protecting this planet. Within the first 5 years, the workshop attracted participants from over 19 countries and 30 U.S. states.

NCAR Mesa Laboratory in Boulder, CO

We feel very privileged to support the emerging climate informatics community and contribute to its growth. New collaborations have been formed and new ideas have taken flight, leading to new insights on climate change. We are looking forward to future workshops and the impact of our interdisciplinary research community.

The Climate Informatics Workshop Steering committee:

- Claire Monteleoni, George Washington University (co-founder)
- Doug Nychka, NCAR (2012-present)
- Imme Ebert-Uphoff, Colorado State University (2013-present)

Information about future workshops and other CI news can always be found on our website, www.climateinformatics.org.

Organizing Committee

Workshop Co-Chairs:
Arindam Banerjee, University of Minnesota
Jennifer Dy, Northeastern University

Program Committee Co-Chairs:
Slava Lyubchich, University of Maryland Center for Environmental Science (UMCES)
Andrew Rhines, Harvard University

Publicity and Publications Chair:
Wei Ding, University of Massachusetts Boston

Travel and Budget Chair:
Eniko Szekely, New York University

Steering Committee:
Imme Ebert-Uphoff, Colorado State University
Claire Monteleoni, George Washington University
Doug Nychka, National Center for Atmospheric Research

Local Administrative Support:
Kathy Peczkowicz, NCAR
Cecilia Banner, NCAR

Program Committee Members:
Arindam Banjeree, University of Minnesota
Rich Caruana, Microsoft Research
Timothy Delsole, George Mason University
Wei Ding, University of Massachusetts Boston
Julien Emile-Geay, University of Southern California
Carlos Felipe Gaitan Ospina, University of Oklahoma
Yulia Gel, University of Texas at Dallas, USA and University of Waterloo, Canada
Vipin Kumar, University of Minnesota, Twin Cities
Dong Liang, University of Maryland Center for Environmental Science
Stefan Liess, University of Minnesota
Vyacheslav Lyubchich, University of Maryland Center for Environmental Science
Nikunj Oza, NASA
Andrew Rhines, University of Washington
Eniko Szekely, New York University
Martin Tingley, Insurance Australia Group; Pennsylvania State University
Suz Tolwinski-Ward, AIR Worldwide

List of Registered Participants

First Name	Last Name	Affiliation
Arindam	Banerjee	University of Minnesota
Sudipto	Banerjee	University of California, Los Angeles
Jon	Barker	NVIDIA
Greg	Behm	Independent
Gabe	Bromley	Montana State University- LRES
Darin	Comeau	Los Alamos National Laboratory
Chintan	Dalal	Rutgers University
Wei	Ding	University of Massachusetts, Boston
Constantine	Dovrolis	Georgia Tech
Jennifer	Dy	Northeastern University
Imme	Ebert-Uphoff	Colorado State University
Scott	Farley	University of Wisconsin, Madison
David John	Gagne	National Center for Atmospheric Research
Yulia	Gel	University of Texas, Dallas
Cristina	Gonzalez-Haro	Telecom Bretagne
Sulagna	Gope	Indian Institute of Technology Kharagpur
Mohammad	Gorji	University of Arizona
Yanira	Guanche	FSU Jena
David M.	Hall	University of Colorado Boulder
Dorit	Hammerling	NCAR IMAGe
Alexander	Heye	Cray, Inc
Lasse	Holmstrom	University of Oulu
Whitney	Huang	Purdue University
Nikola	Jajcay	Institute of Computer Science, CAS
Mark	Jenne	Indiana University
Xiaowei	Jia	University of Minnesota
Arne	Johanson	Kiel University
Anuj	Karpatne	University of Minnesota

Balazs	Kegl	CNRS-Paris France
Hossein	Keshavarz	University of Michigan
Vipin	Kumar	University of Minnesota
Sai	Kumar Popuri	University of Maryland, Baltimore County
Guangyu	Li	University of Southern California
Yi	Li	Northeastern University
Carlos	Lima	University of Brasilia
Yunjie	Liu	Lawrence Berkeley Lab
Slava	Lyubchich	University of Maryland
Steve	Mauget	USDA-ARS
Seth	McGinnis	NCAR
Andrew	Meijers	British Antarctic Survey
Adway	Mitra	International Center for Theoretical Sciences
Claire	Monteleoni	George Washington University
Balu	Nadiga	Los Alamos National Lab
Guru	Nayak	University of Minnesota
Doug	Nychka	NCAR
Nikunj	Oza	NASA
Arthur	PAJOT	LIP6, UPMC
Emma	Pierson	Stanford University
Allen	Pope	National Snow & Ice Data Center
Stan	Posey	NVIDIA
Vinay	Ramakrishnaiah	UWYO/NCAR
Pradeep	Ravikumar	University of Texas, Austin
Andrew	Rhines	Harvard University
Melissa	Rishel	University of Northern Colorado
Steve	Sain	The Climate Corporation
Savini	Samarasinghe	Colorado State University
Todd	Sanford	Climate Central
Ethan	Schaeffer	Penn State University

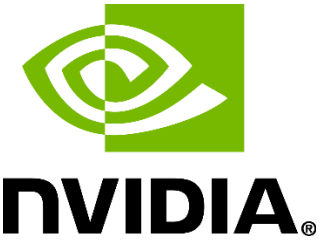
Erich	Seamon	University of Idaho
Bohar	Singh	George Mason University
Joanna	Slawinska	Rutgers University
Jason	Smerdon	Columbia University
Erica	Staehling	WeatherTiger, LLC
Eniko	Szekely	New York University
Craig	Tierney	NVIDIA
Suz	Tolwinski-Ward	AIR Worldwide
Ryan	Truchelut	WeatherTiger, LLC
Matthias	Tuma	World Climate Research Programme (WCRP)
Nathan	Urban	Los Alamos National Laboratory
Jorge R.	Urrego-Blanco	Los Alamos National Laboratory
Qi	Yu	University of Southern California
Tadesse	Zemicheal	Oregon State University
Haozhe	Zhang	Iowa State University
Zhengyuan	Zhu	Iowa State University
Yong	Zhuang	University of Massachusetts, Boston

Acknowledgements of Sponsors

We gratefully acknowledge the generous contributions of the following sponsors who have helped make CI 2016 possible:

The Research Network for Statistical Methods for Atmospheric and Oceanic Sciences

The Nvidia Corporation



The National Science Foundation

The National Center for Atmospheric Research

Hackathon and Workshop Agenda

Hackathon, Wednesday, September 21, 2016

9:30	Bus pickup at Marriott Courtyard
9:45	Bus pickup at Millennium Harvest House (overflow lodging)
10:05 – 10:45	Arrival at NCAR Mesa Lab: Registration, Coffee and Introduction
10:45 – 12:30	Session 1
12:30 – 1:30	Lunch and discussion of initial results
1:30 – 3:00	Session 2
3:00 – 3:15	Coffee Break
3:15 – 5:00	Session 3
5:00 – 6:00	Debriefing and closing
6:00	Bus departs NCAR for Marriott Courtyard

CI Workshop, Thursday, September 22, 2016

7:45	Bus pickup at Marriott Courtyard to NCAR
8:15 – 8:45	Registration and continental breakfast
8:45 – 9:00	Opening Remarks
9:00 – 10:00	**Invited Talk - Doug Nychka**: *Extremes in Regional Climate: What to do with 8000 Histograms?*

10:00 – 10:30 Coffee Break

10:30 – 11:30 **Invited Talk - Pradeep Ravikumar:** *Poisson Graphical Models With Rich Dependence Structures*

11:30 – 12:00 Spotlight Presentations:

Chintan Dalal: *Covariance Structure Analysis of Climate Model Output*

Haozhe Zhang: *Identifying Precipitation Regimes in China Using Model-Based Clustering of Spatial Functional Data*

Emma Pierson: *Detecting and Predicting Beautiful Sunsets Using Social Media Data*

12:00 – 12:10 Group Photo

12:10 – 1:30 Lunch (Cafeteria serves food from 11:30 am - 1:30 pm, cash only)

1:30 – 2:30 **Invited Talk - Yulia Gel:** *Where Statistics and Data Science Meet Climate Risk Insurance*

2:30 – 3:15 Poster Highlights, Part 1

3:15 – 3:45 Coffee Break

3:45 – 4:30 Poster Highlights, Part 2

4:30 – 6:30 Reception and Posters

6:30 Bus departs NCAR, return to Boulder Courtyard Marriott

CI Workshop, Friday, September 23, 2016

7:45 Bus departs Boulder Courtyard Marriott to NCAR

8:15 – 8:30 Continental Breakfast

8:30 – 9:30 **Invited Talk - Jason Smerdon**: *Gigabytes, Megadroughts and Two Kiloyears of Climate History*

9:30 – 10:00 Coffee Break

10:00 – 11:00 **Invited Talk - Sudipto Banerjee**: *Bayesian Modeling and Inference for High-Dimensional Spatial-Temporal Data*

11:00 – 12:00 Spotlight Presentations:

Joanna Slawinska: *Spatiotemporal Pattern Extraction with Data-Driven Koopman Operators for Convectively Coupled Equatorial Waves*

Yi Li: *A Nonparametric Copula Based Bias Correction Method for Statistical Downscaling*

Guruprasad Nayak: *Multiple Instance Learning for Burned Area Mapping Using Multi-Temporal Reflectance Data*

12:00 – 1:00 Lunch and Posters (Cafeteria serves food from 11:30 to 1:30pm, cash only)

1:00 – 2:00 **Invited Talk - Allen Pope**: *Snow and Ice from Space – how computing helps us harness Landsat 8 to study our cryosphere*

2:00 – 3:00 Panel Discussion

Panelists: *Sudipto Banerjee, Yulia Gel, Doug Nychka, Allen Pope, Jason Smerdon*

3:00 – 3:15 Concluding Remarks

3:15 – 3:45 Coffee Break

3:45 – 5:15 Community-Building via Hiking

5:30 Bus departs NCAR, return to Boulder Courtyard Marriott

Climate Informatics 2016 Hackathon

Balázs Kégl and Andrew Rhines

The Rapid Analysis and Model Prototyping Platform

The Rapid Analysis and Model Prototyping (RAMP) is a versatile management and software tool for connecting data science to domain sciences, which is the main mission of the Paris-Saclay Center for Data Science (CDS). It grew organically out of our experience with data challenges, and evolved through the dozen iterations that we carried out in our research and training activities. The RAMP is developed as an in-house tool at the CDS, in collaboration with the Center for Scientific Management (CGS) at Ecole des Mines.

It was originally designed as a collaborative prototyping tool that makes efficient use of the time of scientists in solving the data analytics segment of high-impact domain science problems. We then realized that it is equally valuable for training novice data scientists, for networking, for communication, and as a social science observatory. It has been rapidly becoming a standard educational tool, used in three UPSaclay data science masters, but also in other programs in Paris and Lille. It has been used six times at Saclay, and in five hackatons outside Saclay (Paris School of Economics; French National Museum of Natural History; twice at NCAR, Colorado; Epidemium, Paris).

The RAMP is used in the following operational context. Similarly to a data challenge, the data provider arrives with a prediction problem and a corresponding data set. An experienced data scientist then cleans and curates the data and formalizes the problem. This process can take two weeks to six months, and results in a starting kit, typically an ipython notebook that introduces the domain science problem, describes the data, and shows a first untuned solution (benchmark). The problem is then set up using the RAMP software, and a RAMP event is organized with 30-50 data scientists and domain scientists. The RAMP event usually takes a single day to attract data scientists who do not wish to engage for a longer period of time learning the domain problem. We have been experimenting with other formats: data challenges usually take several months, and course projects can take several weeks. When the data science problem requires the mastering of a specific tool, the RAMP event can be preceded by a Training Sprint. Part of the Training Sprint can also be devoted to introducing the domain science problem, otherwise this introduction takes place at the beginning of the RAMP.

In an ongoing project we will stabilize the software tool and gradually open it so RAMP events can be organized also independently of the CDS core team. We will continue integrating the tool and the format into training programs. The goal is to make the RAMP a standard tool in data science education.

Hackathon Topic: Predicting Arctic Sea Ice Cover

Arctic sea ice cover is one of the most variable features of Earth's climate. Its annual cycle peaks at around 15 million square kilometers in early spring, melting back to a minimum of about 6 million square kilometers in September. These seasonal swings are important for Earth's energy balance, as ice reflects the majority of sunlight while open water absorbs it. Changes in ice cover are also important for marine life and navigation for shipping.

In recent years, Arctic sea ice cover has declined rapidly, particularly during the September minimum. These changes have outpaced the predictions of climate models, and forecasting extent remains a formidable challenge. Typically, skillful predictions are limited to 2–5 months in advance, while idealized experiments suggest that predictions up to two years in advance should be possible [2].

Better tools to predict ice cover are critical for seasonal and regional climate prediction, and would thus address grand challenges in the study of climate change [1].

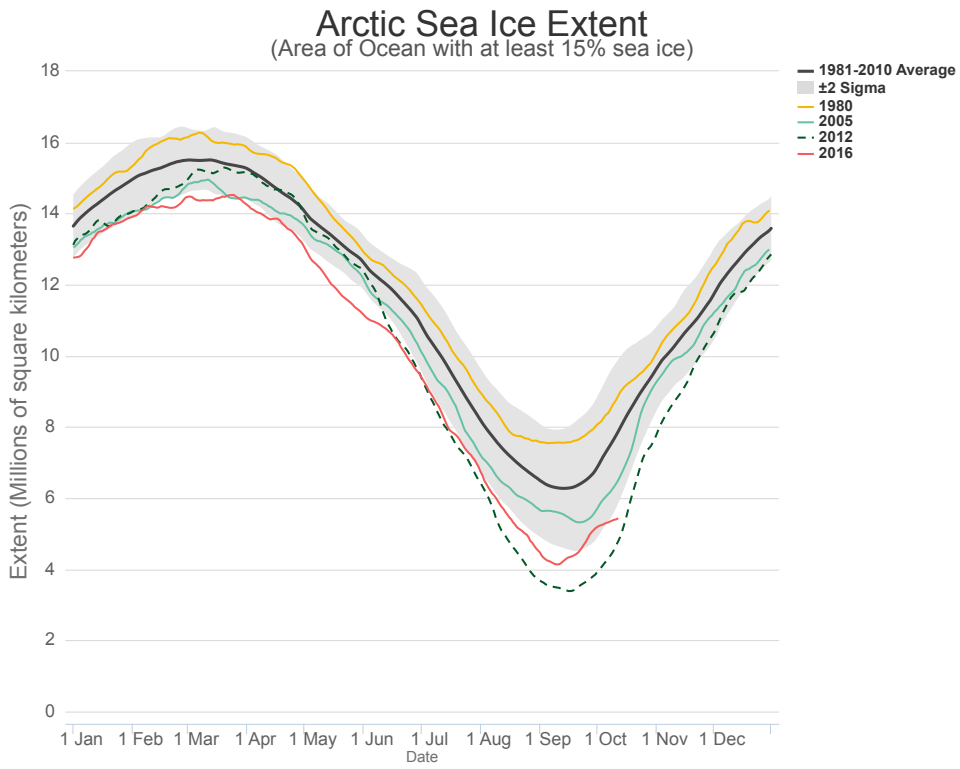


Figure 1: Recent declines in Arctic sea ice cover, courtesy of the National Snow and Ice Data Center.

The CCSM4 simulator

As a surrogate for observational data, the Climate Informatics 2016 RAMP used output from a 1,300 year simulation of the NCAR CCSM 4.0 climate model. The model was run in fully-coupled mode with interactive ocean, atmosphere, and sea ice. The simulation was also performed in an idealized pre-industrial mode, where greenhouse gas concentrations and other external forcings are held fixed to 1850 CE levels. This allows us to access a stationary climate over a 1,000+ year period, which makes the evaluation of the predictor more robust than if we used real measurements that are both non-stationary and limited to several decades.

The data is a time series of maps, z_t, consisting of different physical variables on a regular grid on the Earth, indexed by lon(gitude) and lat(itude) coordinates. The variables made available are:

- `ice_area` — the Northern Hemisphere sea ice area, in millions of squared kilometers.
- `ts` — surface temperature, most important over the oceans which have a very high heat capacity.
- `taux` — zonal (x-direction) surface wind stress. This is the frictional effect of winds on the sea surface and sea ice.
- `tauy` — meridional (y-direction) surface wind stress.
- `ps` — surface pressure.
- `shflx` — Surface sensible heat flux, the amount of heat transferred from the surface to the atmosphere.
- `cldtot` — Total cloud cover (fractional), which has strong effects on radiative energy balance at the surface.

We stress that these fields are a small subset of the variables that define the model state, representing a conservative scenario where only certain features can be remotely observed in real time — primarily from satellites and numerical weather prediction models that can accurately estimate the large scale pressure and wind patterns. Notably, we excluded several important variables including sea ice thickness and the temperature and stratification of the ocean mixed layer.

The fields are monthly averages for 1,300 years, giving a total of 15,600 time points. To be conservative with respect to stationarity, the first 400 years were excluded as the deep ocean takes some time to fully equilibrate in the model.

The prediction task

The goal was to predict the Northern Hemisphere sea ice area 4 months in advance. Since the most important prediction is the minimum area in September, we also output the RMSE over predictions in May, predicting that years (minimum) ice area in September.

The pipeline consists of a time series feature extractor and a predictor. Since the task is regression, the predictor is a regressor, and the score to minimize will be the root mean square error. The feature extractor will have access to the whole dataset. It will construct a classical feature matrix where each row corresponds to a time point. You should collect all information into these features that you find relevant to the regressor. The feature extractor can take anything from the past, that is, it will implement a function $x_t = f(z_1, \ldots, z_t)$. Since you will have access to the full data, in theory one could cheat (even inadvertently) by using information from the future. We implemented a randomized test to find such bugs, but it is still important to avoid this since it would make the results irrelevant.

Domain-knowledge, results, and lessons for future hackathons

Participants were free to explore any regression technique to improve the prediction. Since the input dimension is relatively large (2000+ dimensions per time point even after subsampling to 5-degree spatial resolution) sparse regression techniques were effective in limiting the subset of data that were used. However, sparse methods also revealed that a surprisingly small subset of the variables were necessary to obtain the best predictions, often with fewer than 50 dimensions contributing (i.e., 50 unique variable-location combinations). Surface sensible heat flux in parts of the Arctic was often found to be a key element, suggesting that it was being used in conjunction with surface temperature to 'learn' sea ice thickness despite it not being provided as an input variable. Nevertheless, even predictions using more complex methods did not improve substantially over an initial baseline set by simple sparse regression methods and even a 1-D autoregressive model on sea ice extent alone. It appears likely that the provided fields placed too stringent a limit on 4-month predictability.

A key goal of climate science is in connecting what is understood from simulations to real-world observations, which are intrinsically limited to several decades or at most a few centuries. In this hackathon we addressed predictability within the context of a perfect, stationary model, and it will be useful to consider future hackathon problems in which generalizability between different climate models — or from climate models to observations — can be tested.

References

[1] Ghassem R Asrar, James W Hurrell, and Antonio J Busalacchi. A need for "actionable" climate science and information: summary of wcrp open science conference. *Bulletin of the American Meteorological Society*, 94(2):ES8–ES12, 2013.

[2] Virginie Guemas, Edward Blanchard-Wrigglesworth, Matthieu Chevallier, Jonathan J Day, Michel Déqué, Francisco J Doblas-Reyes, Neven S Fučkar, Agathe Germe, Ed Hawkins, Sarah Keeley, et al. A review on arctic sea-ice predictability and prediction on seasonal to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society*, 2014.

Invited Talks

Doug Nychka, National Center for Atmospheric Research
Extremes in Regional Climate: What to do with 8000 Histograms?
Link to presentation

Abstract: As attention shifts from broad global summaries of climate change to more specific regional impacts there is a need for the data sciences to quantify the uncertainty in regional predictions. A regional climate model (RCM) is a large and complex computer code based on physics that simulates the detailed flow of the atmosphere in a particular region from the large scale information of a global climate model. Part of the value of these simulations is to explore the potential extremes in weather that are due to natural variation and also to climate change. Here we present an application that combines logspline density estimates to discern tail behavior in a distribution with spatial methods for large data sets (LatticeKrig). This is applied to estimate return levels for daily precipitation from a subset of the North American Regional Climate Change and Assessment Program. Here the regional models comprise about 8000 grid locations over North America and so pose challenges for the statistical analysis of functional data. Besides efficient algorithms this application also explores using embarrassing parallel steps using the Rmpi package on the NCAR supercomputer (Yellowstone).

Pradeep Ravikumar, University of Texas at Austin
Poisson Graphical Models With Rich Dependence Structures
Link to Presentation

Abstract: Undirected graphical models, such as Gaussian, Ising, and discrete/multinomial graphical models, are widely used in a variety of applications for modeling distributions over a large number of variables. These standard instances, however, are ill-suited to modeling count data, which are increasingly ubiquitous in climate studies, and spatial incidence data, as well as other big-data settings such as genomic sequencing data, user-ratings data, and site visits. Existing proposals for distributions for multivariate count data have a crucial caveat: the dependence structures they model are largely restrictive, with solely negative or positive dependencies in some cases.

Can we devise multivariate distributions that can capture rich dependence structures between count-valued variables? We address this question via a series of multivariate extensions of the univariate Poisson distribution, providing a new class of Poisson graphical models. We also provide tractable schemes with guarantees for learning our class of Poisson graphical models from data, and demonstrate the performance of our methods by learning simulated networks as well as a network from microRNA-Sequencing data.

Joint work with Eunho Yang, Genevera Allen, Zhandong Liu, David Inouye, Inderjit Dhillon.

Yulia Gel, University of Texas at Dallas
Where Statistics and Data Science Meet Climate Risk Insurance
(Presentation not recorded)

Abstract: Last few years were particularly volatile for the insurance industry in North America and Europe, bringing a record number of claims due to severe weather. According to the 2013 World Bank study, annual average losses from natural disasters have increased from $50 billion in the 1980s to about $200 billion nowadays. Adaptation to such changes requires early recognition of vulnerable areas and the extent of the future risk due to weather factors. Despite the well documented impact of climate change on the insurance sector, there exists a relatively limited number of studies addressing the effect of the so-called ``normal'' extreme weather (i.e., higher frequency, lower individual but high cumulative impact events) on the insurance dynamics. In this talk we discuss utility and limitations of statistical and machine learning procedures to address modelling and forecasting of such weather-related insurance losses and the potential impact of uncertainty quantification on the insurance sector and policy holders.

Jason Smerdon, Columbia University
Gigabytes, Megadroughts and Two Kiloyears of Climate History
Link to Presentation

Abstract: Paleoclimatology spans many different timescales and incorporates a vast array of natural archives that serve as proxies for past climate variability and change. Among the time periods of study, the Common Era (CE; the last two thousand years) contains the most abundant collection of high-resolution (seasonal to annual) proxy records spread globally across land and sea. The CE is also becoming an increasingly common target for transient simulations using fully-coupled climate and earth system models. The abundance of proxy records and the growing number of climate simulations, in conjunction with the fact that the CE is a period when natural forcing occurred under background conditions that were not too different from today, make it a compelling and critical period of focus.

A consequence of the large amounts of proxy archives and the expanding number of model simulations is the growing need for techniques that facilitate data inquiry, comparison and ensemble characterization across the large collection CE datasets. This talk will highlight two principal examples in the study of CE climate: 1) the reconstruction of hemispheric and global climate fields using a now large body of regression and missing data methods; and 2) data-model comparisons between spatiotemporal hydroclimate reconstructions and ensembles of coupled model simulations. These two examples will be explored in the context of how they have been approached to date and the opportunities that they offer for new approaches within climate informatics.

Sudipto Banerjee, University of California Los Angeles
Bayesian Modeling and Inference for High-Dimensional Spatial-Temporal Data
Link to Presentation

Abstract: With the growing capabilities of Geographic Information Systems (GIS) and user-friendly software, statisticians today routinely encounter geographically referenced data containing observations from a large number of spatial locations and time points. Over the last decade, hierarchical spatial-temporal process models have become widely deployed statistical tools for researchers to better understanding the complex nature of spatial and temporal variability. However, fitting hierarchical spatial-temporal models often involves expensive matrix computations with complexity increasing in cubic order for the number of spatial locations and temporal points. This renders such models unfeasible for large data sets. In this talk, I will present some approaches for constructing well-defined spatial-temporal stochastic processes that accrue substantial computational savings. These processes can be used as "priors" for spatial-temporal random fields. Specifically, we will discuss and distinguish between two paradigms: low-rank and sparsity and argue in favor of the latter for achieving massively scalable inference. We construct a well-defined Nearest-Neighbor Gaussian Process (NNGP) that can be exploited as a dimension-reducing prior embedded within a rich and flexible hierarchical modeling framework to deliver exact Bayesian inference. Both these approaches lead to algorithms with floating point operations (flops) that are linear in the number of spatial locations (per iteration). We compare these methods and demonstrate their use in a number of applications and, in particular, in inferring on the spatial-temporal distribution of air pollution in continental Europe using spatial-temporal regression models in conjunction with chemistry transport models.

Allen Pope, National Snow & Ice Data Center
Snow and Ice from Space – how computing helps us harness Landsat 8 to study our cryosphere
Link to Presentation

Abstract: The polar regions and the wider cryosphere are quickly evolving harbingers of worldwide change, driving shifts in water resources and global sea level rise. Therefore, it is important that we measure and monitor snow, glaciers, and ice sheets in a consistent and repeatable manner over time. The spatial, temporal, and radiometric resolution of Landsat 8 support a quantitative measure of polar change over decameter spatial scales and weekly to seasonal timescales. To date, Landsat 8 has collected over three years of imagery of exceptional data. In this presentation, I will provide a tour of various applications of Landsat 8 to study the cryosphere, each of which has been facilitated by cyberinfrastructure – in particular I will discuss dust deposition on snow in Colorado, supraglacial lake monitoring in Greenland, measurement of Antarctic-wide ice velocities, and progress on the building of a Landsat 8 mosaic of Antarctica. Each of these has been facilitated by new practices in research computing, opening a new world of cryospheric research.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

BAYESIAN MODELS FOR CLIMATE RECONSTRUCTION FROM POLLEN RECORDS

Lasse Holmström[1], Liisa Ilvonen[1], Heikki Seppä[2], Siim Veski[3]

***Abstract*—We report progress in using Bayesian multinomial regression models to reconstruct the past climate from fossil pollen records. Three model variants and associated example reconstructions are described: a single-core reconstruction, a multi-core reconstruction with spatial dependence, and a single-core reconstruction with time uncertainty.**

I. BACKGROUND

Instrumental records of past climate variation often cover only the last 100-200 years. In contrast, climate proxies can provide a continuous record of climate change extending thousands of years into the past [1]. The data needed for paleoclimate reconstruction consist of a "training set" of the modern values of the climate variable of interest together with the associated modern proxy data, as well as a past record of the same proxy.

We consider temperature reconstruction from lake sediment pollen compositional data. The basic idea is time-for-space substitution where the unknown temporal variation of past temperature is mimicked by the varying modern temperatures in a large geographical area that covers a wide range of environmental conditions. Data from modern training lakes are used to capture the relationship between the pollen abundances of various plants and the temperature. Then this relationship and sediment core pollen abundances are used to reconstruct the unknown past temperatures. For useful reviews on climate reconstruction methodologies, see [2], [3] and [4]. Bayesian approaches to pollen-based reconstructions are discussed in [5] and [6] proposes a unifying Bayesian framework for the paleoclimate reconstruction problem.

Corresponding author: L Holmström, lasse.holmstrom@oulu.fi
[1]Department of Mathematical Sciences, University of Oulu, Finland
[2]Department of Geosciences and Geography, University of Helsinki, Finland, [3]Institute of Geology, Tallinn University of Technology

II. RECONSTRUCTION MODELS

A. Reconstruction from a single core

Detailed Bayesian modelling for past environmental reconstruction was first proposed in [7], [8], and [9]. The Bum model of [8] can be viewed as a Bayesian version of some earlier reconstruction techniques, such as the weighted-averaging and Gaussian logit models [10], [11]. A modification of Bum, the Bayesian hierarchical multinomial regression model referred to as the Bummer was described in [7]. The Bummer model was further discussed and modified in [12], [13], [14] and [15].

To describe the original Bummer model, let us denote temperature by x and pollen abundance by y. Superscripts m and f are used to indicate modern (training) values of a variable and f is used for their past (fossil) values. Thus, x_i^m for example denotes current temperature at training lake i and y_{ij}^f denotes the abundance of pollen taxon j in sample i of the sediment core. Let the vector $\mathbf{y}_i^m$ include the observed pollen taxon abundances in training lake i. The multinomial model assumes that

$$\mathbf{y}_i^m = (y_{i1}^m \ldots, y_{il}^m | y_{i+}^m, \mathbf{p}_i^m) \sim \text{Mult}(y_{i+}^m, \mathbf{p}_i^m),$$

where $\mathbf{p}_i^m = (p_{i1}^m, \ldots, p_{il}^m)$ and p_{ik}^m is the probability that a random pollen grain from lake i represents taxon k. The taxon occurrence probabilities are drawn from a Dirichlet distribution,

$$(p_{i1}^m, \ldots, p_{il}^m | x_i^m, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \sim \text{Dirichlet}(\lambda_{i1}^m, \ldots, \lambda_{il}^m),$$

where $\lambda_{i1}^m, \ldots, \lambda_{il}^m$ represent unimodal Gaussian responses of pollen to temperature,

$$\lambda_{ik}^m = \alpha_k \exp\left\{-[(\beta_k - x_i^m)/\gamma_k]^2\right\}.$$

Analogous models are assumed for past quantities but the taxon-specific parameters α_k (scaling factor), β_k (optimum temperature), and γ_k (tolerance) are taken to be the same both for training and fossil data. The idea is that the response for taxon k is high when the temperature (x_i^m or x_i^f) in the environment is close

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

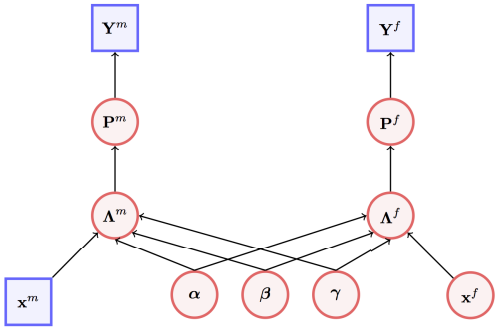


Fig. 1. The basic multinomial regression model (Bummer). Squares denote data and circles are model parameters. Capital letters indicate aggregate quantities, $\mathbf{Y}^m = [\mathbf{y}_1^m, \ldots, \mathbf{y}_n^m]$ with n the number of training lakes etc.

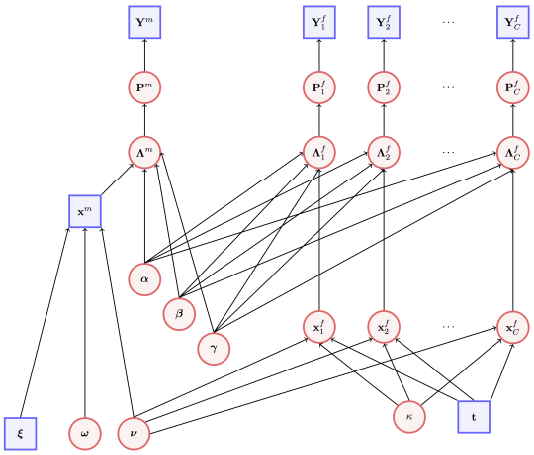


Fig. 2. The spatio-temporal model for reconstructions from several cores simultaneously. A spatial linear trend is modeled by covariates $\boldsymbol{\xi}$ and the associated parameters $\boldsymbol{\omega}$. The parameter $\boldsymbol{\nu}$ is associated with the spacial covariance matrix. Union chronology is denoted by $\mathbf{t}$.

to the taxon's optimal temperature β_k, and low when it is not. The randomness of the multinomial probabilities models the fact that other factors besides the temperature can affect the pollen abundances. Figure 1 is a graphical representation of the basic multinomial regression model.

The original Bummer assumed a Gaussian iid prior for the past temperatures x_i^f but we use a smoothing prior elicited from a numerical climate simulation of the annual mean temperature for the area where the core lakes are located. Prior temporal smoothness is imposed by assuming that

$$x_{i+1}^f = x_i^f + \frac{1}{\sqrt{\kappa}}(t_{i+1} - t_i)\varepsilon_i, \; i = 1, \ldots, N, \qquad (1)$$

where $x_1^f \sim \mathrm{N}(\mu, 1)$, μ is the current temperature, the ε_i's are iid standard Gaussian variables, and t_i's are the ages of the core sediment samples (the chronology). In principle, $x_1^f = \mu$, and the prior variance 1 can be thought to describe e.g. measurement error. The parameter $\kappa > 0$ controls the smoothness of sample paths and a vague prior for it is elicited from a 1150 year long NCAR numerical climate model simulation of mean annual temperature for northern Europe [16], restricted to the core chronology ages.

B. Multiple cores

Next, let us consider reconstruction from several cores simultaneously [14]. More reliable reconstructions should result by taking into account not only temporal correlations along each core but also the spatial correlation between cores. The sediment layer ages of the cores are first combined into a single chronology by forming their union and the reconstruction for each core is based on this union chronology. Denoting by N the number of distinct dates in the union chronology, let $\mathbf{x}_c^f = [x_{c1}^f, \ldots, x_{cN}^f]^T$ denote the unknown past temperatures for core $c = 1, \ldots, C$, and let $\mathbf{X}^f = [(\mathbf{x}_1^f)^T, \ldots, (\mathbf{x}_C^f)^T]^T$ be the concatenated vector of past temperatures for all cores. Then the spatio-temporal prior for the past temperatures is

$$\mathbf{X}^f | \mathbf{\Sigma} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{\Sigma}),$$

where $\boldsymbol{\mu} = [\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_C]^T$, $\boldsymbol{\mu}_c = [\mu_c, \ldots, \mu_c]^T \in \mathbb{R}^N$ and μ_c is the modern temperature at the location of core c. The covariance matrix $\mathbf{\Sigma}$ is assumed to be separable,

$$\mathbf{\Sigma} = \mathbf{C}_S \otimes \mathbf{C}_T \in \mathbb{R}^{CN \times CN},$$

where $\mathbf{C}_S$ is spatial covariance and $\mathbf{C}_T$ is temporal covariance defined by (1). The spatial dependency is modeled with a linear trend and additive isotropic noise characterized by a parametric exponential covariance function. Priors for the spatial dependency model parameters are elicited by considering the known mean annual temperatures in the part of northern Europe where the training lakes are located and training data contributes to the estimation of these parameters. A graphical representation of the spatio-temporal model is depicted in Figure 2.

C. Including time uncertainty

In reality, the precise ages t_i of the core sediment layers are not known. In fact, we only know their depths in the core and only a small number of the depths have

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

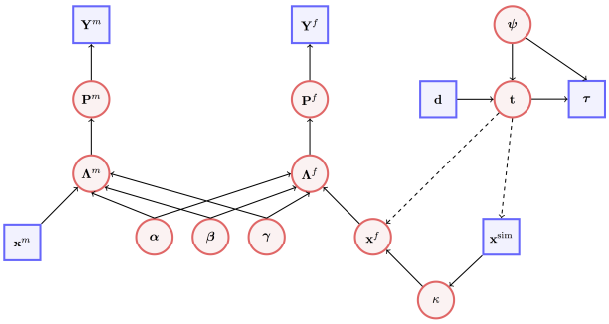


Fig. 3. The model with time uncertainty. In the Bchron module, $\mathbf{d}$ stands for sample depths, τ for radiocarbon dates and ψ for age-depth model parameters. $\mathbf{x}^{\text{sim}}$ is the NCAR climate model simulation.

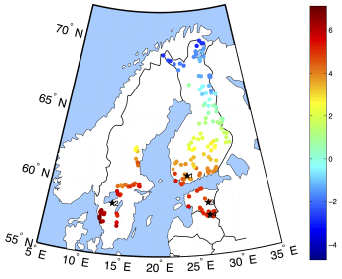


Fig. 4. The locations of training (dots) and core lakes (stars): 1 = Arapisto, 2 = Flarken, 3 = Raigastvere, 4 = Rõuge. Color indicates current annual mean temperature.

been dated with radiocarbon analysis. Estimates for the ages of the remaining sediment samples are obtained by applying some interpolation method. In order to get a more realistic idea of the overall credibility of the results, the uncertainty associated with the core chronology should be factored in to the estimation process. We do this in [17] in a way similar to [18] by adding a separate chronology module based on the Bchron age-depth model [19]. Realizations of chronologies are generated by Bchron and for each random chronology a reconstruction is made using the multinomial regression model. The model with time uncertainty is depicted in Figure 3.

III. RESULTS

The data included 173 training and four core lakes in Finland, Sweden and Estonia (Figure 4). 104 different pollen taxa were counted from the sediment samples.

Two independently made temperature reconstructions are shown in Figure 5 and Figure 6 shows reconstructions from the same cores but with spatial dependence between all four cores included in the model. Both types of reconstructions exhibit the well-known features

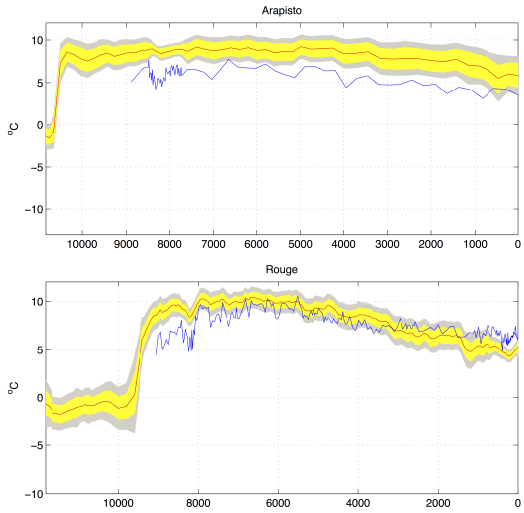


Fig. 5. Temperature reconstructions from individual cores for two lakes. Time on the horizontal axis is in years before present (BP). Red: posterior mean. Blue: WA-PLS. Yellow/gray: pointwise/simultaneous 95% credible intervals.

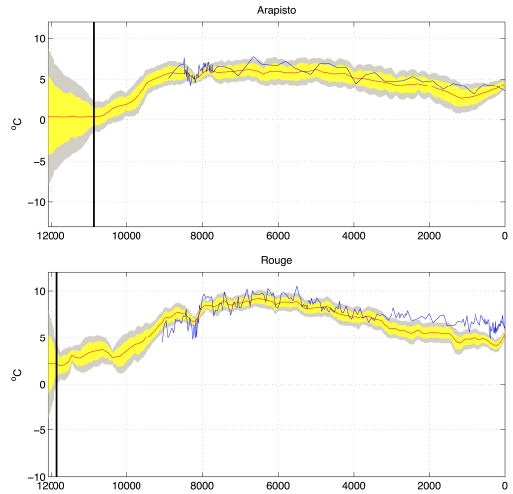


Fig. 6. Spatio-temporal temperature reconstructions. The black line marks the oldest date in the core chronology.

of Holocene temperature history but spatial correlation (and shared environmental parameters) makes the latter reconstructions more similar which can be considered an improvement given that the lakes are located in the same general area. The rate of early warming leading to the Holocene is also more plausible in the spatio-temporal reconstructions and they also match the standard WA-PLS reconstructions better.

Figure 7 shows part of the reconstruction for lake Rõuge when time uncertainty is included in the model. The posterior mean temperature is very smooth and for example misses the well-known cold episode around

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

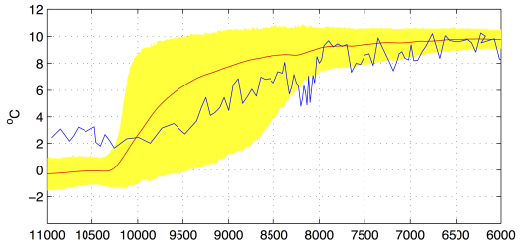


Fig. 7. Rõuge temperature reconstruction with time uncertainty.

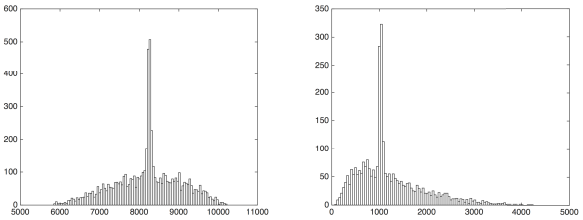


Fig. 8. Lake Rõuge posterior distributions (histograms based on 6000 sample points) of timing of the 8200 BP cold event (left) and length of warming leading to the Holocene (right).

8200 years before present. This episode does show in the individual posterior sample temperature histories but their timing varies and therefore the episode disappears from the posterior mean. On the other hand, including time uncertainty allows one to analyze for example the uncertainty in the timing of this episode and the length of initial Holocene warming (Figure 8).

REFERENCES

[1] V. Masson-Delmotte *et al.*, "Information from paleoclimate archives," in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (T. Stocker, G.-K. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, and P. Midgley, eds.), pp. 383–464, Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, 2013.

[2] P. D. Jones *et al.*, "High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects," *The Holocene*, vol. 19, no. 1, pp. 3–49, 2009.

[3] H. J. B. Birks, O. Heiri, H. Seppä, and A. E. Bjune, "Strengths and weaknesses of quantitative climate reconstructions based on late-quaternary biological proxies," *The Open Ecology Jounal*, vol. 3, pp. 68–110, 2010.

[4] S. Juggins and H. J. B. Birks, "Quantitative environmental reconstructions from biological data," in *Tracking Environmental Change Using Lake Sediments, vol. 5, Data Handling and Numerical Techniques* (H. J. B. Birks, A. F. Lotter, S. Juggins, and S. J. P., eds.), pp. 431–494, Dordrecht: Springer, 2012.

[5] C. Ohlwein and E. R. Wahl, "Review of probabilistic pollen-climate transfer methods," *Quaternary Science Reviews*, vol. 31, pp. 17 – 29, 2012.

[6] M. P. Tingley, P. F. Craigmile, M. Haran, B. Li, E. Mannshardt-Shamseldin, and B. Rajaratnam, "Piecing together the past: statistical insights into paleoclimatic reconstructions," *Quaternary Science Reviews*, vol. 35, p. 122, 2012.

[7] K. Vasko, H. T. T. Toivonen, and A. Korhola, "A Bayesian multinomial Gaussian response model for organism-based environmental reconstruction," *Journal of Paleolimnology*, vol. 24, pp. 243–250, 2000.

[8] H. T. T. Toivonen, H. Mannila, A. Korhola, and H. Olander, "Applying Bayesian statistics to organism-based environmental reconstruction," *Ecological Applications*, vol. 11, no. 2, pp. 618–630, 2001.

[9] A. Korhola, K. Vasko, H. T. T. Toivonen, and H. Olander, "Holocene temperature changes in northern Fennoscandia reconstructed from chironomids using Bayesian modelling," *Qaternary Science Reviews*, vol. 21, no. 16–17, pp. 1841–1860, 2002.

[10] C. J. ter Braak and H. van Dame, "Inferring pH from diatoms: a comparison of old and new calibration methods," *Hydrobiologia*, vol. 178, no. 3, pp. 209–223, 1989.

[11] H. Birks, J. Line, S. Juggins, A. Stevenson, and C. Ter Braak, "Diatoms and pH reconstruction," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 327, no. 1240, pp. 263–278, 1990.

[12] P. Erästö and L. Holmström, "Selection of prior distributions and multiscale analysis in Bayesian temperature reconstructions based on fossil assemblages," *Journal of Paleolimnology*, vol. 36, pp. 69–80, 2006.

[13] J. S. Salonen, L. Ilvonen, H. Seppä, L. Holmström, R. J. Telford, A. Gaidamavičius, M. Stančikaitė, and D. Subetto, "Comparing different calibration methods (WA/WA-PLS regression and Bayesian modelling) and different-sized calibration sets in pollen-based quantitative cllimate reconstructions," *The Holocene*, vol. 22, no. 4, pp. 413–424, 2012.

[14] L. Holmström, L. Ilvonen, H. Seppä, and S. Veski, "A Bayesian spatiotemporal model for reconstructing climate from multiple pollen records," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1194–1225, 2015.

[15] J. Li, L. Ilvonen, Q. Xu, J. Ni, L. Jin, L. Holmström, X. Cao, Z. Zheng, H. Lu, Y. Luo, Y. Li, C. Li, X. Zhang, and H. Seppä, "East Asian summer monsoon precipitation variations in monsoonal China over the last 9500 years: a comparison of pollen-based reconstructions and model simulations," *The Holocene*, vol. 26, no. 4, pp. 592 – 602, 2016.

[16] C. M. Ammann, F. Joos, D. S. Schimel, B. L. Otto-Bliesner, and R. A. Tomas, "Solar influence on climate during the past millennium: Results from transient simulations with the NCAR climate system model," *Proceedings of the National Academy of Sciences USA*, vol. 104, no. 10, pp. 3713–3718, 2007.

[17] L. Ilvonen, L. Holmström, H. Seppä, and S. Veski, "A Bayesian multinomial regression model for paleoclimate reconstruction with time uncertainty." To appear in *Environmetrics*, 2016.

[18] A. C. Parnell, J. Sweeney, T. K. Doan, M. Salter-Townshend, J. R. Allen, B. Huntley, and J. Haslett, "Bayesian inference for palaeoclimate with time uncertainty and stochastic volatility," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 64, no. 1, pp. 115–138, 2015.

[19] J. Haslett and A. Parnell, "A simple monotone process with application to radiocarbon-dated depth chronologies," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 57, no. 4, pp. 399–418, 2008.

6th International Workshop on Climate Informatics
September 22-23, 2016
CI 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

ON INFORMATION CRITERIA FOR DYNAMIC SPATIO-TEMPORAL CLUSTERING

Ethan D. Schaeffer[1], Jeremy M. Testa[2], Yulia R. Gel[3], Vyacheslav Lyubchich[2]

***Abstract*—Modern climate data sets, including paleo-reconstructions, long-term weather monitoring records, and remote sensing data, contain a wealth of space-time information that leads to a variety of challenges related to data storage, management, and analysis. This has sparked an interest in dynamic space-time clustering algorithms that are particularly suitable for the analysis of large data streams. The trend-based clustering algorithm TRUST allows segmentation of space-time processes in real time, but requires the user to set multiple tuning parameters, and this step is usually performed in a subjective manner. Here we propose a data-driven automatic approach to simultaneously select the tuning parameters based on a penalized loss function. We focus on the two most important parameters of the TRUST algorithm, which define short-term closeness of observations across locations and long-term persistence of such closeness within an analyzed time window. We demonstrate the performance of the enhanced clustering procedure using simulated time series, and illustrate its applicability using long-term records of water temperature in Chesapeake Bay.**

I. MOTIVATION

Contemporary environmental data sets often exhibit resolution at disparate spatial and temporal scales. Moreover, the data structure evolves in space and time, and thus standard assumptions of covariance separability and conventional inference tools for spatio-temporal processes might be inappropriate for these data.

One of the potential approaches to address this problem is to cluster environmental data so that groups are relatively homogeneous but allowed to evolve in space and time. TRend based clUstering algorithm for Spatio-Temporal data stream (TRUST) [1] allows dynamic clustering of spatio-temporal data, based on the sliding window argument, thus, TRUST is suitable for analysis of data streams. The algorithm (R code is available from [2]) uses all available information without aggregation and allows shapes and number of clusters to change over time. Hence, the biggest clusters provide insight into the dominant behavior of the spatio-temporal data, whereas outlying observations that potentially cause bias in aggregated data can be classified separately. A disadvantage of TRUST is the number of tuning parameters that substantially influence the clustering performance.

In this paper, we propose a simple and computationally efficient data-driven approach to select optimal values of the TRUST tuning parameters using the notion of stability [3], [4]. We further extend the approach of [5] and [6] by allowing the automatic selection of multiple TRUST tuning parameters. While we primarily focus on TRUST, our approach is also applicable to optimal selection of tuning parameters in other static and dynamic clustering algorithms. We validate our approach with Monte Carlo simulations and illustrate its utility for analysis of the Chesapeake Bay water quality monitoring data.

II. METHOD

Consider a set of geo-referenced data recorded at N locations during a period T and stored in a matrix $X_{T\times N}$. Each row of this matrix is a *layer* or a snapshot, and p consecutive rows constitute a *slide* [1]. TRUST identifies trend clusters in the data based on homogeneity measures for short-term (slide-level) and long-term (window-level) clustering. Assume that data arrive in slides (e.g., monthly data come in yearly chunks of $p = 12$) and windows include w slides (e.g., the number of considered years).

The key feature of TRUST is that a number of clusters K and shapes of clusters are not fixed a-priori and can vary over time and space, which is achieved by adapting a sliding-window model to multiple spatially distributed data sources. The TRUST algorithm consists of the two main steps. The first step is to identify trend-clusters over the slide time (i.e., *slide-level* clustering), based on closeness (homogeneity) of sources within a layer, where a level of homogeneity is controlled by a threshold δ. Hence, the most important role in the TRUST slide-level clustering is played by the tuning parameter δ: that is, δ defines how close two time series should be in a slide in order to be clustered

Corresponding author: V. Lyubchich, lyubchic@umces.edu
[1]Pennsylvania State University, USA [2]University of Maryland Center for Environmental Science, USA [3]University of Texas at Dallas, USA

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

together [1]. The second step is to approximate trend-clusters by combining the slide-level trend cluster sets (i.e., *window-level* clustering). This is acheived by grouping time series that have been clustered together at least $\varepsilon \times w$ number of instances within a window of w slides. Thus, the parameters δ and ε are the dominant parameters to control clustering performance of the TRUST algorithm (see earlier study by [5] on sensitivity of TRUST in respect to δ). With a smaller δ, only a few time series can be clustered together, which leads to a larger number of small clusters at the slide level. A higher ε requires more time series to belong to the same slide-level cluster (for instance, under the extreme condition of $\varepsilon = 1$, time series are to be classified as one cluster for all w slides). Hence, a higher ε also leads to a larger number of small clusters but now at the window level.

Let $K(\delta, \varepsilon)$ be the number of obtained clusters, and $L(\delta, \varepsilon)$ be a loss function evaluating homogeneity of obtained clusters. (For the sake of notations we further on omit dependence of K on δ and ε.) A range of suitable loss functions includes, for instance, the Rand index, F-measure [7], the gap statistic [8], the consensus index [9] as well as various conventional and robust versions of ANOVA, e.g. Levene's statistics for homogeneity of group variances [10] and rank-based ANOVA [11]. In this paper we consider $L(\delta, \varepsilon)$ as the residual variance $\sigma^2(\delta, \varepsilon)$, where residuals are the differences between clustered values and the mean value of the corresponding cluster. Clearly, the result with zero variance (the lowest loss) is attained at $K = N$, which does not meet the goal of clustering. To avoid the extreme of $K = N$, we can use a penalized loss function of clustering performance. Here as a proof of concept, we employ a simple and easily tractable penalized loss function, which is one of the most widely adopted information criterions in statistics, namely, Bayesian information criterion (BIC):

$$L^*(\delta, \varepsilon) = (N - K) \ln \sigma^2(\delta, \varepsilon) + K \ln(N - K).$$

The idea is to search for arguments δ_{opt} and ε_{opt} that minimize $L^*(\delta, \varepsilon)$ over a suitable range, which allows to simultaneously select the optimal tuning parameters (see Algorithm 1).

We set a possible range for δ based on interquartile range (IQR), as suggested in [1]. However, IQR calculated on whole $X_{T\times N}$ can be very large if differences between clusters are substantial. Therefore, we calculate IQR for each time series and use a median $\widetilde{IQR}$ from N values. Selecting a possible range for ε is more straightforward, since this proportion changes depending on how many slides fit a window (see Algorithm 1).

Fig. 1 shows a simulated example of selecting the optimal parameters based on BIC: δ_{opt} and ε_{opt} correspond to the darkest cell (minimal BIC) and also determine the final number of clusters K_{opt}. Yellow cells correspond to the combinations of δ and ε yielding $K = 1$ and the highest BIC. White cells are with the combinations yielding $K = N$, thus, $\sigma^2(\delta, \varepsilon) = 0$ and $L^*(\delta, \varepsilon)$ cannot be computed.

Algorithm 1: Optimized TRUST algorithm.

input : Data matrix $X_{T\times N}$, time series in columns; number of slides in a window w; number of layers in a slide p.

output: An N-vector of cluster associations; optimal δ_{opt} and ε_{opt}.

Let $\widetilde{IQR}$ be a median interquartile range for the N time series;

$\delta = \widetilde{IQR}/N, 2\widetilde{IQR}/N, \ldots, \widetilde{IQR}$;

$\varepsilon = 1/w, 2/w, \ldots, 1$;

for $i = 1, \ldots, length(\delta)$ **do**

  for $j = 1, \ldots, length(\varepsilon)$ **do**

    run TRUST algorithm using δ_i and ε_j;

    calculate $L^*(\delta_i, \varepsilon_j)$ for the TRUST output;

  end

end

let optimal parameters be those corresponding to the minimal L^*:

$$[\delta_{opt}, \varepsilon_{opt}] = \arg\min_{\delta,\, \varepsilon} L^*(\delta, \varepsilon);$$

run TRUST algorithm using δ_{opt} and ε_{opt} to obtain the final cluster associations.

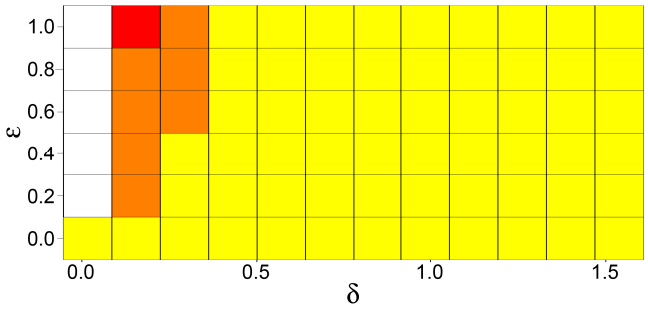


Fig. 1. Selection of δ_{opt} and ε_{opt} based on the lowest BIC (δ and ε corresponding to the red cell). Yellow indicates the highest BIC.

III. SIMULATION STUDY

We validate our approach with synthetic autoregressive (AR) time series. Particularly, we simulate N AR(1) series with an autoregressive coefficient $\phi_1 = 0.5$ and standard normal innovations. A 'burn-in' period

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

TABLE I
PURITY OF CLUSTERING WITH THE OPTIMIZED TRUST ALGORITHM, FOR THE VARYING MEAN SHIFT, NUMBER OF TIME SERIES N, AND THEIR LENGTH T

			Mean shift			
N	T	w	0	2	4	8
24	36	3	0.946	0.403	0.899	1.000
	60	5	0.970	0.544	0.961	1.000
50	36	3	1.000	0.411	1.000	1.000
	60	5	1.000	0.444	1.000	1.000

for simulation of each time series is 300 values. We form two clusters ($J = 2$) by adding a mean shift to half of the simulated time series. Thus, with varying the number of time series N in the simulations, we also vary the number of time series in each cluster ($N/2$). Similar to [5], we set the shift to be 0, 2, 4, or 8, where the zero shift means no difference between clusters and all time series are expected to be clustered together. The number of Monte Carlo simulations for each combination of the parameters is 1000.

At each Monte Carlo step, we apply the optimized TRUST Algorithm 1 with fixed $p = 12$ layers in a slide (as we are using monthly data) and a single window covering all T observations, so the number of slides in a window ($w = T/p$) varies with T. To evaluate the clustering accuracy, we calculate purity as the proportion of time series correctly assigned to the clusters [7]:

$$Purity(\Omega, C) = \frac{1}{N} \sum_{j=1}^{J} \max_{k=1,\ldots,K} |\omega_k \cap c_j|,$$

where $\Omega = \{\omega_1, \ldots, \omega_K\}$ is the set of identified clusters and $C = \{c_1, \ldots, c_J\}$ is the set of simulated clusters (in our simulations, $J = 2$). That is, within each simulated cluster $j = 1, \ldots, J$ we find the size of the most populous cluster from the K clusters we identified. Then, we sum together the J sizes we found and divide by N.

The results in Table I indicate that the proposed approach allows us to achieve accurate clustering. The purity increases with sample size T, and rapidly rises to 1 with the growing mean shift. Remarkably, the purity also rises by increasing the number of time series in each cluster (as we change N from 24 to 50, the number of time series in each cluster goes from 12 up to 25). The results with the mean shift 0 describe the limit case when all time series belong to a single cluster and, on average, more than 94.6% of the time series were correctly assigned to that one cluster.

IV. CHESAPEAKE BAY TEMPERATURE TRENDS

We apply our method to the data set collected by the Chesapeake Bay Program at 116 bay monitoring stations throughout Chesapeake Bay and its tributaries. Water temperature is a key indicator of climate change in the area and one of the major factors influencing biological and biogeochemical processes within the Bay [12]. We used monthly averages of surface water temperature over the period 01.1985–12.2014 (30 years) to study the primary spatio-temporal trends using Algorithm 1.

We pre-processed the data by filtering out the stations with more than 15% of missing data (97 stations remained) and filled-in remaining missing values with an interannual average value for corresponding combination of month and station. To focus on the temporal dynamics rather than on individual level differences between the stations, we scaled the data for each station to zero mean and unit variance.

The results of applying the clustering Algorithm 1 to surface water temperature are shown in (Fig. 2) and reveal several interesting patterns. For example, note that the northern and southern regions of the Potomac River estuary do not covary, where the southern section clusters with the temperature dynamics of the adjacent main-stem (Cluster 1). This difference could be due to the strong freshwater influence in the northern Potomac, which contrasts with the influence of main-stem Bay on the southern Potomac. Similar differences between the upper and lower reaches of the Patuxent, Rappahannock, and York Rivers emphasize the regional differences in the relative influence of freshwater versus the main-stem Bay. Another interesting result from the clustering output is the main-stem clusters separately in the far north (Cluster 3) from the south (Cluster 1). This could be due to increased ocean influence on the south region in contrast to freshwater influences from the Susquehanna River in the north. Finally, the tributaries of the upper Bay and middle-eastern Bay cluster together (Cluster 9), where these systems are characterized by shallow depths and low freshwater inputs.

In an attempt to gain more information about the individual clusters we ran a non-parametric WAVK test [13], [14] to identify possible parametric trends in the different clusters. The WAVK test on the deseasonalized scaled data found three nearly significant results (p-value is slightly above 0.05) in clusters 1, 4, and 10 which could indicate possible significant trends in the clusters. Further research into the direction (shape) of these trends may be of interest.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

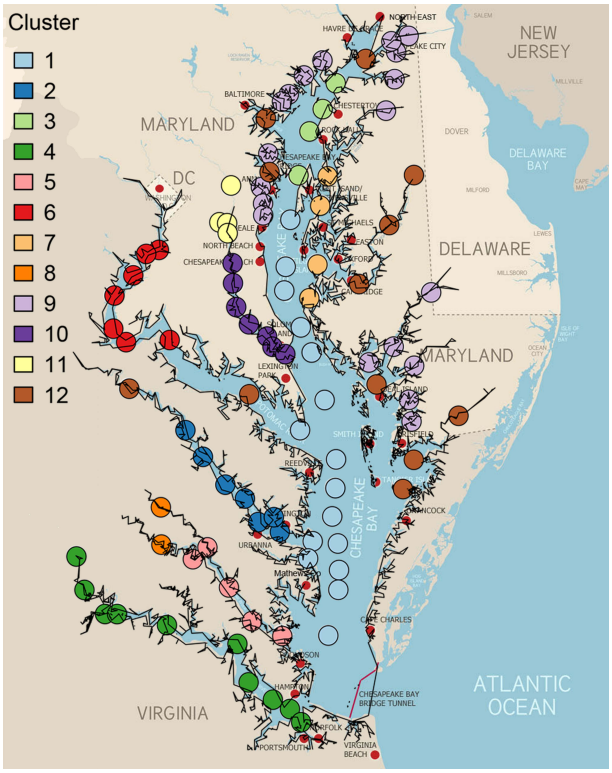


Fig. 2. Clusters of the Chesapeake Bay stations (97 stations total) based on the dynamics of monthly surface water temperature in 01.1985–12.2014. Cluster 12 contains stations that did not join any other cluster.

V. CONCLUSION

In conclusion, our method for automatic selection of tuning parameters has delivered high accuracy of clustering assignment, including the case of multiple clusters and the null case of a single cluster. Application of the new clustering approach to the dynamics of monthly surface water temperature in the Chesapeake Bay revealed a number of interesting patterns due to regional differences in freshwater streams. In the future we plan to investigate utility of other penalized loss functions in conjunction with the tuning parameter selection and extend the proposed clustering approach to other space-time algorithms.

ACKNOWLEDGMENTS

The authors thank D. Liang (UMCES, USA) for the help with spatial visualizations. The work of E. Schaeffer and V. Lyubchich was supported by the Maryland Sea Grant.

REFERENCES

[1] A. Ciampi, A. Appice, and D. Malerba, "Discovering trend-based clusters in spatially distributed data streams," in *International Workshop of Mining Ubiquitous and Social Environments*, pp. 107–122, 2010.

[2] V. Lyubchich, Y. R. Gel, X. Wang, and C. Chu, *funtimes: Functions for Time Series Analysis*, 2016. R package ver. 2.2.

[3] S. Ben-David and U. Von Luxburg, "Relating clustering stability to properties of cluster boundaries," in *Proceedings of the 21st Conference on Learning Theory*, pp. 379–390, 2008.

[4] J. Wang, "Consistent selection of the number of clusters via crossvalidation," *Biometrika*, vol. 97, no. 4, pp. 893–904, 2010.

[5] X. Huang, V. Lyubchich, A. Brenning, and Y. R. Gel, "Analysis of dynamic trend-based clustering on Central Germany precipitation," in *Proceedings of the Fifth International Workshop on Climate Informatics* (J. G. Dy, J. Emile-Geay, V. Lakshmanan, and Y. Liu, eds.), 2015.

[6] X. Huang, I. R. Iliev, A. Brenning, and Y. R. Gel, "Space-time clustering with stability probe while riding downhill," in *Proceedings of the SIGKDD'16 Workshop on Mining and Learning from Time Series (MiLeTS)*, 2016.

[7] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[8] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the number of clusters in a data set via the gap statistic," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[9] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, "Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data," *Machine Learning*, vol. 52, no. 1-2, pp. 91–118, 2003.

[10] J. L. Gastwirth, Y. R. Gel, and W. Miao, "The impact of Levene's test of equality of variances on statistical theory and practice," *Statistical Science*, pp. 343–360, 2009.

[11] D. D. Boos and L. A. Stefanski, *Essential Statistical Inference: Theory and Methods*. Springer, 2013.

[12] H. Ding and A. J. Elmore, "Spatio-temporal patterns in water surface temperature from Landsat time series data in the Chesapeake Bay, USA," *Remote Sensing of Environment*, vol. 168, pp. 335–348, 2015.

[13] L. Wang, M. G. Akritas, and I. Van Keilegom, "An ANOVA-type nonparametric diagnostic test for heteroscedastic regression models," *Journal of Nonparametric Statistics*, vol. 20, no. 5, pp. 365–382, 2008.

[14] V. Lyubchich, Y. R. Gel, and A. El-Shaarawi, "On detecting non-monotonic trends in environmental time series: a fusion of local regression and bootstrap," *Environmetrics*, vol. 24, no. 4, pp. 209–226, 2013.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

DETECTING MULTIVARIATE BIOSPHERE EXTREMES

Yanira Guanche García [1,3], Erik Rodner[1,3], Milan Flach[2], Sebastian Sippel[2], Miguel Mahecha[2,3], Joachim Denzler[1,3]

***Abstract*—The detection of anomalies in multivariate time series is crucial to identify changes in the ecosystems. We propose an intuitive methodology to assess the occurrence of tail events of multiple biosphere variables.**

I. MOTIVATION

Satellite remote sensing measurements provide valuable data for monitoring the earth system. In this direction, international research projects like BACI[1] and CAB-LAB[2] are making a great effort developing unified, high resolution, and open-access Earth Observations (EOs). The availability of multivariate EOs time series covering decadal periods allows the application of different techniques to detect changes or abnormal events in an unprecedented way. Anomalies in EOs may reflect changes in the dynamical system but need to be distinguished from spurious features such as sensor changes or processing artefacts, thus, its detection is an essential task in climate and ecosystem research, [1–3].

By combining different techniques, this study proposes an intuitive approach to assess the occurrences of multivariate tail events in terrestrial biosphere variables. This will allow the detection of sensitive regions or more severe phases during our historical records.

II. DATA DESCRIPTION

A preliminary version of the Earth System Data Cube developed within the CAB-LAB project[2] has been used in this study. The Earth System Data Cube is a practical way of storing spatio-temporal data. It encompasses 14 atmospheric and biosphere variables from different sources that expand 2001-2012 with 8-day resolution and 1° grid of global spatial resolution . We have used those variables related to biosphere processes [4]: Fraction of photosynthetic active radiation (*Fpar*); Leaf surface temperature (*LST*); Gross primary productivity (*GPP*); Terrestrial ecosystem respiration (*TER*); Heat flux (*H*); and Latent heat flux (*LE*); (Figure 1).

Corresponding author: Yanira Guanche García, yanira.guanche.garcia@uni-jena.de [1]Computer Vision Group, Friedrich Schiller University of Jena, Germany. [2]Max Planck Institute for Biogeochemistry, Jena, Germany. [3]Michael Stifel Center for Data-driven and Simulation Science, Jena, Germany.

[1]see: *baci-h2020.eu/*

[2]see: *earthsystemdatacube.net/*

III. METHODOLOGY

In 2009, [5] classified anomaly detection methods into 6 groups. One of them is an intuitive strategy based on statistical modeling, where anomalies are assumed to be points that are not well represented by a previously estimated statistical model [6], [7]. Following this idea, we have selected one location in Mid-Europe and we initiate our methodology by fitting a univariate Autoregressive Moving Average - ARMA(p, q) model for each variable. In order to avoid latter inconsistencies before fitting the model, the 6 variables have been locally deseasonalized and normalized (μ=0, σ=1) although we are aware that these techniques can mask certain anomalies. In this particular case, after checking the autocorrelation of the residuals [8], we have chosen an ARMA$(3, 1)$. By means of the residuals of the univariate regression models, we identify the extreme events within the multivariate time series with two procedures: *a)* estimating the coexceedances over a certain threshold and *b)* estimating the Mahalanobis distance of the residuals to the mean.

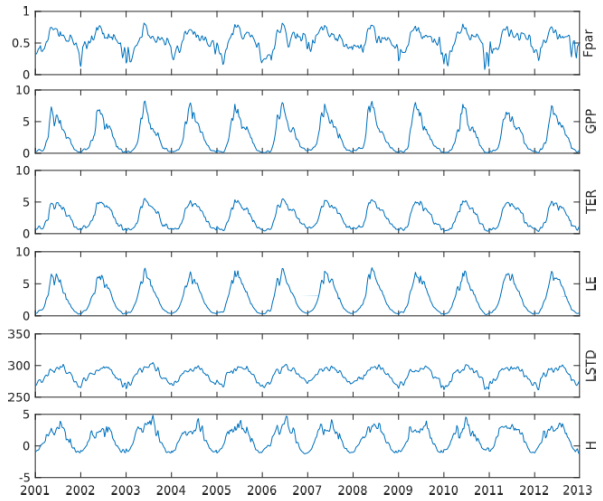


Fig. 1. Terrestrial biosphere variables at 50.5°N, 12.5°E.

a) Extreme residuals coexceedances: We define an extreme residual as one for which its absolute value lies above the 85% percentile of the residuals distribution. By considering the absolute value of the residuals for the selection of extremes we are focusing on all possible combinations of positive and negative extremes. Since our data set has 552 observations, the 15% highest residuals count 83 exceedances for each variable. Next, we count the number of variables simultaneously presenting an extreme residual for each timestep – referred to as coexceedances or as coincidences [9–13]. Usual thresholds for extreme events might be 90% for top-tail events and 10% for bottom-tail events. But we have considered at the same time positive and negative events, so a 10% cut-off to the absolute value of the residuals would give us few observations for a meaningful analysis, therefore we set the threshold at the 85% percentile. This method apart from being very intuitive and understandable, allows a direct interpretation of the extremes detected.

b) Mahalanobis distance: Coexceedances is a very direct and understandable approach, but it does not take into account the shape of the joint residuals distribution. An alternative way to define the extreme events is based on the Mahalanobis distance [14]. This metric, also known as Hotellings T^2 [15], considers the mean and the covariance matrix of the distribution based on a multivariate Gaussian distribution.For the sake of simplicity, we have used this technique although the residuals distribution do not follow a multivariate Gaussian distribution. More complex options would be the use of Support Vector Data Description (SVDD) techniques [16] or statistical depth functions [17] but it is out of the scope of this paper.

Direct comparison of both approaches is difficult as thresholds and thus number of extreme events have to be comparable, which is not the key interest of this study. Anyways, for benchmarking we will focus on the heatwave in Europe of 2003, which experienced the warmest summer record so far [18–21].

IV. GLOBAL APPLICATION

We have extended the methodology applied at one location to all the locations encompassed in the Earth System Data Cube. For each location, we have locally deseasonalized and normalized the 6 variables before fitting an univariate ARMA$(3, 1)$ model. We assume the same kind of model for all the locations for comparison reasons, although we are aware that this might be a strong assumption and will need further investigation.

With the residuals at each location we have: a) counted the coexceedances over a local threshold at the 85% percentile and b) estimated the Mahalanobis distance for all the timesteps considering the local covariance matrix and a critical distance above which extremes are significant. This critical distance has been estimated as the 99% quantile of the χ^2 with 6 degrees of freedom [22]. Figure 2 presents the percentage of observations with *strong* coexceedances (upper plot) and Mahalanobis distance scores above the critical value (lower plot). Strong coexceedances are those timesteps where at least 4 variables present values above its 85% percentile simultaneously. About 30-40% of the observations with extreme Mahalanobis scores present coexceedances in at least 4 variables.

This general overview allows us to detect regions where to focus for applying an attribution scheme trying to elucidate the role of meteorological drivers behind. However, it also presents some open issues: in the northern latitudes (*i.e.* Russia, Canada and Alaska), the amount of extreme residuals detected by both methods is higher to other areas in the globe. This can be related to the fact that in northern latitudes, biosphere variables present big changes in its variance along the year. This heteroscedastic behaviour overestimates the number of extremes. Additional error sources might be the globally fixed (p, q) parameters of the ARMA model, trends in the data or issues related to the projection error from the satellites. Another alternative might be the use of models that include the heteroscedasticity in the seasonal cycle like Generalized Autoregressive Conditional Heteroscedasticity - GARCH models. These issues need further investigations and we are currently working on them.

Coming back to the historical event of 2003 in Europe, Figure 3 represents the results obtained through both approaches over Europe for a certain timestep: 5^{th} of August 2003. At this time, Europe was under the extreme conditions of the hottest summer record since the XVI century [23]. The Mahalanobis distance represented in Figure 3 clearly shows the pattern of an extreme event in central Europe. On the other hand, the coexceedances are detecting 2-3 of 6 variables being extreme. However, with the coexceedances method it can be easily seen that LSTD and H are extreme along mid-Europe while the 2003 heat wave (Figure 4); showing that variables that are closer to the atmosphere (LSTD and H) present distinctly different behaviour than the biosphere one (GPP, TER, LE).

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

V. DISCUSSION AND CONCLUSIONS

We have presented a method to detect abnormal events in multivariate time series. The basic idea behind our approach is that an anomaly is a point in the time series which cannot be well represented by a statistical model that describes the entire time series. This assumption allowed us to investigate the occurrence of extreme observations in biospheric variables with two approaches: coexceedances and Mahalanobis distance of extremal residuals. Both methods present advantages and disadvantages that make them useful and complementary. The coexceedances approach is a method very easy to interpret and allows a direct explanation of the detected extremes. The Mahalanobis distance is able to provide more information of the multivariate joint distribution by losing the direct information about which are the variables exhibited an extreme event.

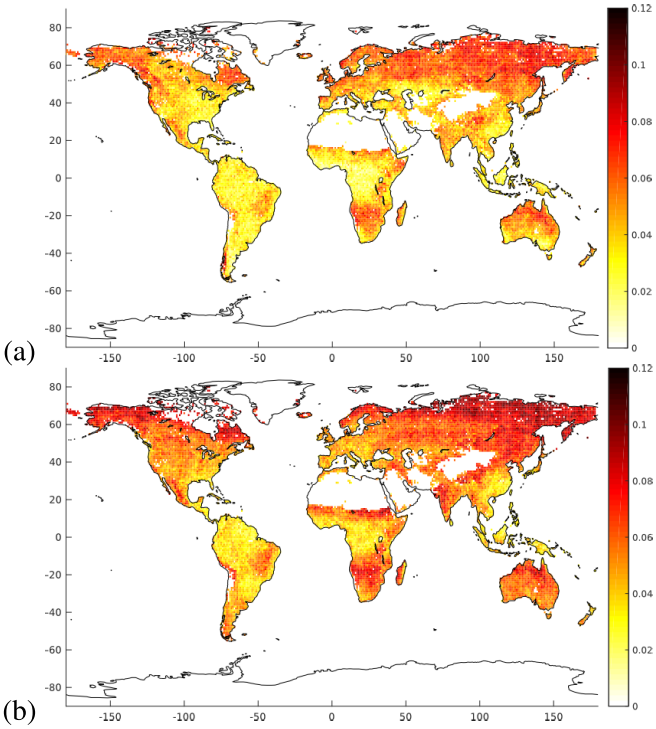


Fig. 2. (a) % of strong coexceedance events (4-6 residuals extremes). (b) % of Mahalanobis distance above a critical value.

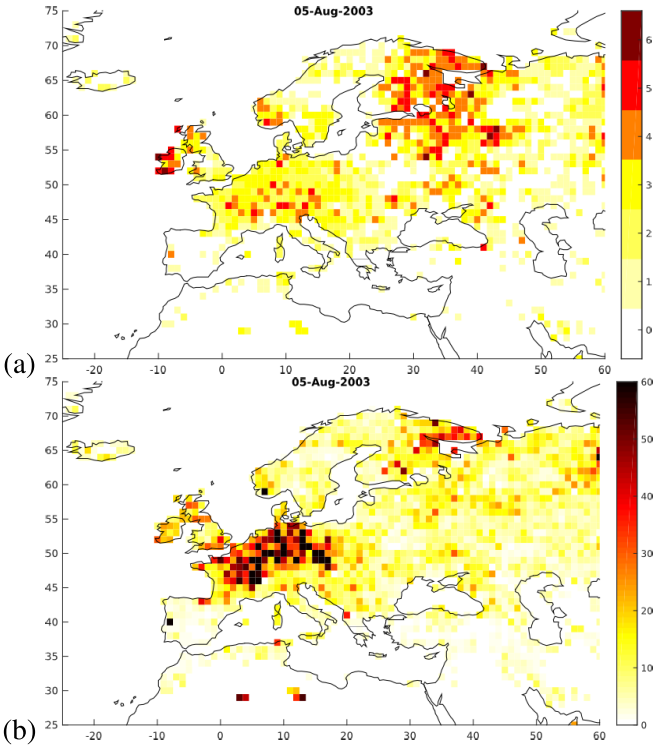


Fig. 3. (a) Coexceedances on 5^{th} of August 2003. (b) Mahalanobis distance on 5^{th} of August 2003.

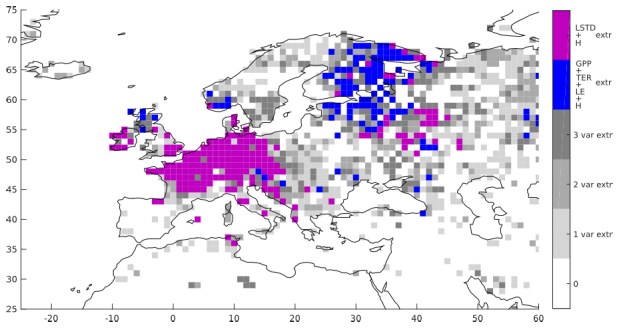


Fig. 4. Variables coexceeding on 5^{th} of August 2003. Grey scale-up to 3 unclassified variables being extreme, purple pixels-extreme LSTD and H and blue pixels- extreme GPP, LE, TER and H.

Multivariate methods allow the detection of insights unreachable from univariate methods; like the different behavior between variables closer to atmosphere and those closer to biosphere shown in the 2003 heat wave.

Future work will be focused on different aspects:

i) A better threshold selection needs to be done to make both methods comparable;

ii) Resolving the overestimation of extremes in the northern latitudes;

iii) Regionalization into areas with similar behavior. Consequently, adapting specific ARMA models for each similar region;

iv) Attribution scheme. This step is crucial to understand the processes causing abnormal events.

ACKNOWLEDGMENTS

Data used in this study were kindly provided by the ESA project CAB-LAB-Coupled atmosphere biosphere virtual laboratory. The support of the EU project BACI-Towards a Biosphere Atmosphere Change Index, contract 640176 is gratefully acknowledged.

REFERENCES

[1] J. Zscheischler, M. Reichstein, S. Harmeling, A. Rammig, E. Tomelleri, and M. D. Mahecha, "Extreme events in gross primary production: a characterization across continents," *Biogeosciences*, vol. 11, no. 11, pp. 2909–2924, 2014.

[2] A. W. R. Seddon, M. Macias-Fauria, P. R. Long, D. Benz, and K. J. Willis, "Sensitivity of global terrestrial ecosystems to climate variability," *Nature*, vol. 531, pp. 229–232, 2016.

[3] S. Sippel, J. Zscheischler, M. Heimann, F. E. L. Otto, J. Peters, and M. D. Mahecha, "Quantifying changes in climate variability and extremes: Pitfalls and their overcoming," *Geophysical Research Letters*, vol. 42, no. 22, pp. 9990–9998, 2015. 2015GL066307.

[4] G. Tramontana, M. Jung, C. R. Schwalm, K. Ichii, G. Camps-Valls, B. Ráduly, M. Reichstein, M. A. Arain, A. Cescatti, G. Kiely, L. Merbold, P. Serrano-Ortiz, S. Sickert, S. Wolf, and D. Papale, "Predicting carbon dioxide and energy fluxes across global fluxnet sites with regression algorithms," *Biogeosciences*, vol. 13, no. 14, pp. 4291–4313, 2016.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[6] F. J. Anscombe, "Rejection of outliers," *Technometrics*, vol. 2, no. 2, pp. 123–146, 1960.

[7] R. Mínguez, B. Reguero, A. Luceño, and F. Méndez, "Regression models for outlier identification (hurricanes and typhoons) in wave hindcast databases," *Journal of Atmospheric and Oceanic Technology*, vol. 29, no. 2, pp. 267–285, 2012.

[8] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 3rd ed., 1994.

[9] B. Algieri, M. Kalkuhl, and N. Koch, "A tale for two tails: Explaining extreme events in financialized agricultural markets," in *Conference Paper for the Australian Agricultural and Resource Economics Society*, 2015.

[10] A. Rammig, M. Wiedermann, J. F. Donges, F. Babst, W. von Bloh, D. Frank, K. Thonicke, and M. D. Mahecha, "Coincidences of climate extremes and anomalous vegetation responses: comparing tree ring patterns to simulated productivity," *Biogeosciences*, vol. 12, no. 2, pp. 373–385, 2015.

[11] J. F. Siegmund, T. G. Sanders, I. Heinrich, E. van der Maaten, S. Simard, G. Helle, and R. V. Donner, "Meteorological drivers of extremes in daily stem radius variations of beech, oak, and pine in northeastern germany: An event coincidence analysis," *Frontiers in Plant Science*, vol. 7, p. 733, 2016.

[12] J. Donges, C.-F. Schleussner, J. Siegmund, and R. Donner, "Event coincidence analysis for quantifying statistical interrelationships between event time series," *The European Physical Journal Special Topics*, vol. 225, no. 3, pp. 471–487, 2016.

[13] J. Zscheischler, R. Orth, and S. I. Seneviratne, "A submonthly database for detecting changes in vegetation-atmosphere coupling," *Geophysical Research Letters*, vol. 42, no. 22, pp. 9816–9824, 2015.

[14] P. Mahalanobis, "On the generalised distance in statistics (vol.2, pp.49–55)," *Proceedings National Institute of Science, India. Retrieved from http://ir. isical. ac. in/dspace/handle/1/1268*, 1936.

[15] H. Hotelling, "Multivariate quality control," *Techniques of statistical analysis*, 1947.

[16] D. M. Tax and R. P. Duin, "Support vector data description," *Machine learning*, vol. 54, no. 1, pp. 45–66, 2004.

[17] Y. Zuo and R. Serfling, "General notions of statistical depth function," *Annals of statistics*, pp. 461–482, 2000.

[18] Deutscher Wetterdienst, *Wetterrekorde in Deutschland und weltweit*, (5)000/08.15, ed., 2013.

[19] P. Ciais, M. Reichstein, N. Viovy, A. Granier, J. Ogée, V. Allard, M. Aubinet, N. Buchmann, C. Bernhofer, A. Carrara, *et al.*, "Europe-wide reduction in primary productivity caused by the heat and drought in 2003," *Nature*, vol. 437, no. 7058, pp. 529–533, 2005.

[20] M. Reichstein, P. Ciais, D. Papale, R. Valentini, S. Running, N. Viovy, W. Cramer, A. Granier, J. Ogee, V. Allard, *et al.*, "Reduction of ecosystem productivity and respiration during the european summer 2003 climate anomaly: a joint flux tower, remote sensing and modelling analysis," *Global Change Biology*, vol. 13, no. 3, pp. 634–651, 2007.

[21] C. Schär, P. L. Vidale, D. Lüthi, C. Frei, C. Häberli, M. A. Liniger, and C. Appenzeller, "The role of increasing temperature variability in european summer heatwaves," *Nature*, vol. 427, no. 6972, pp. 332–336, 2004.

[22] R. D. Cook and D. M. Hawkins, "Unmasking multivariate outliers and leverage points: comment," *Journal of the American Statistical Association*, vol. 85, no. 411, pp. 640–644, 1990.

[23] P. A. Stott, D. A. Stone, and M. R. Allen, "Human contribution to the european heatwave of 2003," *Nature*, vol. 432, no. 7017, pp. 610–614, 2004.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

SPATIO-TEMPORAL GENERATIVE MODELS FOR RAINFALL OVER INDIA

Adway Mitra[1]

***Abstract*—Stochastic Rainfall Generators have been explored by the water resources community to simulate daily, monthly or annual rainfall over compact geographical regions. In this work, we attempt to build a generative model for rainfall over India, a climatologically diverse region with substantial effects of orography. We simulate gridded rain-gauge observations over a $1^\circ - 1^\circ$ grid covering most of India, by using a few statistics estimated from the rain-gauge data as model parameters. We evaluate several generative models by drawing samples from them, computing summary statistics from the samples and comparing these with corresponding estimates from observed data. The challenge is to select a small set of statistics of the true observed data as input parameters, while maximizing the estimation accuracy of other statistics simulated by the generative model. We consider a sequence of models of increasing sophistication for this purpose. These sophistications are motivated by various relationships observed in the data. Results are compared with CMIP5 models. Applications of this generative modelling framework, for understanding the statistical characteristics of spatiotemporal variability, and for isolating reasons of models' successes and failures in reproducing these statistics, are discussed.**

I. MOTIVATION

Modelling and analysis of rainfall over India is a very important problem since rainfall affects the lives and livelihoods of a billion people, directly or indirectly. The formulation and evaluation of decisions in various public and private sector domains needs to take rainfall into account, and so it is important to have detailed realistic rainfall simulations. Such simulations can be provided by General Circulation Models (GCMs), like the ones in the Coupled Model Intercomparison Project (CMIP-5), but these models are unable to capture spatiotemporal variations in rainfall over this region, since they have not necessarily been tuned for regional climate. So a dedicated rainfall simulator for India is necessary, and this is what we have attempted in this work. Since magnitude of rainfall always involves high uncertainty, such simulators must be stochastic. Earlier papers related to Indian rainfall prediction and modeling ([1]) have recognized that it has a stochastic component. The proposed simulator is based on Bayesian generative processes. Stochastic rainfall simulators have been studied earlier ([2], [3], [4], [5]), but mostly for small homogeneous regions unlike a large diverse landscape like India. Here, we enlist the significant spatiotemporal statistics of observed data. The aim is then to use some of these statistics as input parameters for the generative model, while the remaining parameters are sought to be estimated from simulation output of the generative model. Also, reduction of input parameter size makes models more general. To this end, we investigate a series of models, and evaluate their simulation results. To progressively develop more sophisticated generative models, we include various spatiotemporal statistics that appear to be fundamental to constructing good models for our problem (as suggested in [6]). We find that the proposed models can outperform CMIP5 models on a wide range of relevant test statistics.

Corresponding author: Adway Mitra, adway.cse@gmail.com
[1]International Center for Theoretical Sciences, Bangalore, India

II. DATA AND SPATIO-TEMPORAL STATISTICS

Various statistics can be defined for any spatio-temporal process by domain experts, depending on the aspects of the process they are keen to understand. Such statistics can be computed from the observations or data-points of the process obtained using sensors. Any simulation model for the system need not reproduce every observation, but it should be able to reproduce statistical values. A simulation can be realistic only if it is initialized or parameterized realistically. Some of these statistics are used as input for Stochastic rainfall generators as parameters of probability distributions used in the model. The remaining statistics are computed from the simulation output, and compared to the observation statistics, which gives an evaluation of the model.

Consider rainfall data of the form Y_{stm}, where s is one grid location, t is a year and m is a month.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

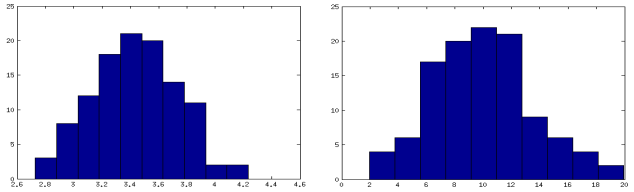


Fig. 1. Left: histogram of all-India Annual Mean rainfall over 1901-2011. Right: histogram of Grid-level Monthly Mean rainfall at a single random grid point, for month of June over 1901-2011. X-axis: rainfall in mm/day, Y-xis: frequency. Right plot does NOT represent all grid-locations and months.

Consider there are observations from S locations, T years and 12 months each year. The data we use is $1^\circ - 1^\circ$ daily precipitation for the period 1901-2011 ($T = 111$) at $S = 357$ grid locations over India, obtained from Indian Meteorological Department. This data can be converted to grid-wise annual form Y_{st}^M, by computing the average across all 12 months each year, and to all-India annual form, Y_t^{SM} by further averaging across all S locations for each year. Similarly, it can be converted to monthly form Y_{tm}^S by averaging across all locations. For each of the cases, we can compute the *mean and standard deviation* over the T years. The evaluation statistics are:

1) Mean and standard deviation (μ, σ), from All-India Annual mean denoted by $\{Y_t^{SM}\}$
2) Mean and standard deviation (μ_m, σ_m), from All-India Monthly Mean denoted by $\{Y_{tm}^S\}$
3) Mean and standard deviation (μ_s, σ_s), from Grid-wise Annual Mean denoted by $\{Y_{st}^M\}$
4) Mean and standard deviation (μ_{sm}, σ_{sm}), from Gridwise Monthly Mean denoted by $\{Y_{stm}\}$

In our models we use Gaussian distributions for monthly or annual rainfall (as in [4]). As the histogram of all-India Annual Mean shows (Fig. 1), Gaussian is clearly a good model. For gridwise and monthly means, the fit is less clear. So we use a Truncated Gaussian Distribution (truncated at 0) for these cases.

In the literature, spatial correlations in rainfall have been modeled by Gaussian Processes[3], but these include a Covariance function which is difficult to parametrize. Instead, we define *spatial patterns* θ and *monthly patterns* ϕ. Each spatial pattern is a S-dimensional probability distribution, indicating how total rainfall in a fixed period is distributed across all the locations. It may be specific to a year or a month (θ_m). A monthly pattern is a 12-dimensional probability distribution, indicating how the total rainfall at a location is distributed across the 12 months of a year. It may be specific to a single grid location (ϕ_s), or to the entire landmass (spanning S grids). We can compute *representative spatial and monthly-patterns* from the data Y_{stm} and they can be used as input parameters for some of the models.

These representines can be extracted by the process of *clustering*, where for each (t, m)-pair we compute a S-dimensional distribution from Y_{stm}, and carry out clustering of all the $12 \times T$ spatial pattern vectors, to get K representative spatial patterns (for any K). The clustering can be done using *spectral clustering*[7], which requires a similarity measure between every pair of data-points. Since in our case the data-points are probability distributions, we can use negative-exponential Kullback-Leibler Divergence- a measure of similarity between distributions. Similarly, we can compute represenetative monthly patterns ϕ_s specific to each location s.

III. Simple Models

In this section, we describe a few simple generative models, where a subset of the statistics described above are used as model parameters. The first model $M0$ is the basic Point Process model[2], where the rainfall at each grid-location and month are simulated independently, using the corresponding mean and standard deviation (Statistic 4 in the above list). The process is as follows:

- M0: $Y_{stm} \sim \mathcal{N}(\mu_{sm}, \sigma_{sm}) \forall s \in \{1, \ldots, S\}, m \in \{1, \ldots, 12\}, t \in \{1, \ldots, T\}$

The parameter complexity of the model is given by $2 * S * 12 = 24S$, which is quite high. We can alternatively simulate the annual rainfall for each location, and distribute it across all the months as follows:

- M1: $Y_{st}^M \sim \mathcal{N}(\mu_s, \sigma_s); Y_{stm} = \phi_{sm} * Y_{st}^M$

The parameter complexity for the local mean and standard deviation are now $2S$, while for the location-specific monthly patterns it is $12S$, i.e. $14S$ in total.

On the other side of the spectrum, we can simulate the total annual rainfall (M2) or total monthly rainfall (M3), over the entire Indian landmass, and distribute it over all the grids using the spatial patterns.

- M2: $Y_t^{SM} \sim \mathcal{N}(\mu, \sigma); Y_{tm}^S = \phi_m * Y_t^{SM}; Y_{stm} = \theta_{ms} * Y_{tm}^S$
- M3: $Y_{tm}^S \sim \mathcal{N}(\mu_m, \sigma_m); Y_{stm} = \theta_{ms} * Y_{tm}^S$

For M2, the parameter complexity is $12S$ for the month-specific spatial patterns, 12 for the monthly pattern and 2 for the total annual mean and standard deviation. For M3, it is $12 * S$ for the month-specific spatial pattern and 24 for the Gridwise Monthly parameters.

When the simulation outputs of these models are processed and the spatio-temporal statistics mentioned

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

above are computed, we find that the standard deviations are captured quite poorly. For example, model M0 gives almost the same value of Y_t^{MS} for each year. Hence μ^{M0} is quite close to μ^{DATA}, but σ^{M0} is very less compared to σ^{DATA}. This is because, there are *flood and drought years* where several locations simultaneously have high or low rain, but this is not captured if rainfall is simulated independently at the locations. On the contrary in M2 and M3, all locations in the country are simultaneously assigned high or low rainfall in the years when the annual Spatial Mean rainfall is high or low respectively. This is also not true, as from the true data we find that not all locations follow the spatial mean in extreme rainfall years. Also, for M2,M3 standard deviations like σ_s^{M2} and σ_{sm}^{M2} are scaled down according to θ_s, making them much smaller than true values. Thus, these simplistic models are insufficient and we need to explore more sophisticated models.

Model	M0	M2	M3	M4	M5	M6	M7	CMIP5
μ	0.02	input	input	0.3	0.02	0.02	0.06	0.98
σ	0.25	input	input	0.18	0.13	0.03	0.03	0.06
μ_m	0.01	0.06	0.01	0.03	0.04	0.01	0.08	0.87
σ_m	0.61	0.43	0.42	0.50	0.57	0.44	0.40	0.22
μ_s	0.07	0.05	0.02	0.3	0.08	0.05	0.08	1.38
σ_s	0.14	0.63	0.62	0.44	0.06	0.19	0.14	0.39
μ_{sm}	input	0.03	0.02	0.36	0.07	0.06	0.15	1.99
σ_{sm}	input	1.85	1.84	1.25	1.22	1.33	0.60	1.07
PC	24S	12S+14	12S+24	81S+3	27S+3	7S+9K +3	17S+12L +9K+3	-

TABLE I
ERRORS IN SPATIO-TEMPORAL STATISTICS AND PARAMETER COMPLEXITY (PC) FOR DIFFERENT MODELS

IV. ADVANCED MODELS

By analysis of the true data, we observe that there are some years when many locations receive much more rain than they usually receive, and in some years, many locations receive much less rain than usual. We approximate the annual rainfall in most locations using tri-modal Gaussian distributions- High (Mode 1), Low (Mode 2) and Normal (Mode 3). Also, in most years a large number of locations are in the same mode. Accordingly, we can assign to each year a type (1,2,3) based on the mode which most locations are in. Also, while most locations are in Mode p (1 or 2) in a year of type p, there are some locations which are in the opposite mode. Hence, for each location s we can have a probability distribution β_s for its mode Q_{st} in year t, conditioned on the year type P_t, i.e. $\beta_{sp}(q)$ is the probability of location s being in mode q in any year, conditioned on the year being of type p. At each location, we also need mean and standard deviation parameters for each mode (μ_{sq}, σ_{sq}). These can be either month-specific (M4) or annual (M5) along with a monthly pattern. In both cases, the complexity for β-distributions is $3 + 3 * 3 * S = 3 + 9S$. Complexity for the rainfall statistics are $2 * 12 * S * 3 = 72S$ for M4 and $12 * S + 2 * S * 3 = 18S$ for M5.

- $M4 : P_t \sim \beta; Q_{st} \sim \beta_{sp}; Y_{stm} \sim \mathcal{N}(\mu_{sqm}, \sigma_{sqm})$ where $p = P_t, q = Q_{st}$
- $M5 : P_t \sim \beta; Q_{st} \sim \beta_{sp}; Y_{st} \sim \mathcal{N}(\mu_{sq}, \sigma_{sq}); Y_{stm} = Y_{st}\phi_{sm}$ $(p = P_t, q = Q_{st})$

Our next insight from the true data is that there are groups of locations which are in the same mode in the same year. This is due to the presence of *homogeneous regions* having special rainfall characteristics such as the Western coast, the North-western desert region, the North-eastern regions, the Tamilnadu coast, the Central plains etc. We define *clusters of locations* based on co-occurring modes, and assign a cluster index C_s to each location s. Two locations are in the same cluster if they are in the same mode in most of the years. Such clusters can be found by Spectral Clustering[7], where the similarity between any two locations is defined as the number of years they are in the same mode (see Fig. 2). So we define cluster-specific conditional mode distributions, which also reduces parameter complexity.

- $M6 : P_t \sim \beta; Q_{kt} \sim \beta_{kp}; Y_{st} \sim \mathcal{N}(\mu_{sq}, \sigma_{sq}); Y_{stm} = Y_{st}\phi_{sm}$ $(p = P_t, q = Q_{kt}, k = C_s)$

The parameter complexity for the β-distributions now comes down to $3 + 3 * 3 * K = 3 + 9K$, where K is the total number of location-clusters. Additionally there is a complexity S for the cluster indices of each location, and $2 * 3 * S = 6S$ for rainfall statistics.

We observe that most models perform poorly in reproducing the local monthly standard deviation, but using this as an input parameter severely increases complexity. We observe in the true data that there are clusters of locations having similar values of monthly standard deviation, and these locations are also spatially coherent, like the previous set of clusters. So once again, we perform a clustering of the locations based on their monthly standard deviation values, using K-means clustering 12-dimensional vectors of monthly standard deviation for the locations. Each location is assigned to a cluster index D_s, and monthly standard deviations specific to these clusters are used (see Fig. 2). The parameter complexity now is $3 + 9K$ for the β-distributions, $3S + 12L$ for the rainfall magnitude statistics (L: number of variance-based clusters), $2S$ for C_s, D_s-indices, and $12S$ for ϕ-s.

- $M7 : P_t \sim \beta; Q_{kt} \sim \beta_{kp}; Y_{stm} \sim \mathcal{N}(\phi_{sm}\mu_{sq}, \sigma_{rmq})$ $(p = P_t, q = Q_{kt}, k = C_s, r = D_s)$

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

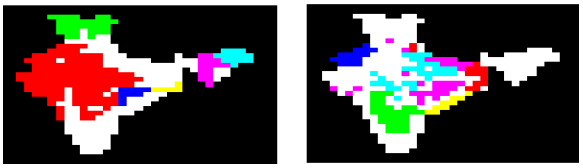


Fig. 2. A few selected large clusters of grid-points, according to co-occurring annual modes (left), and monthly standard deviation (right). Each colour represents one cluster.

V. Results

We run each of the models mentioned above for $T = 111$ years (size of the true data), and compute the spatio-temporal statistics listed in Section II. We report the absolute error in each of these statistics, with respect to the values estimated from the true data. We also compute these statistics from 15 CMIP-5 models which have been found to be most suitable for Indian rainfall simulation[8], and compute the average absolute errors across these models. The results in Table I show the relative performances. M7 achieves best results on several statistics while reducing parameter complexity. But it should be understood that none of these models can be called "best", since for each model we make a trade-off between estimation accuracy of different statistics and parameter complexity. The choice of model for simulation will depend on what aspects the simulation wants to focus on most strongly.

VI. Future Work

It is possible to use the models M4-M7 for inference, rather than simulation. Instead of providing the input parameters of years with low and high rainfall, and clusters of locations based on mode or monthly variance, it is possible to use these models on actual observed data (or data from some other simulation), and find out these statistics from the data by inference using Gibbs Sampling and E-M algorithm. This process may allow us to gain new insights about the data.

We have so far considered the variables for different months and years to be IID, i.e. the time-series of different variables are stationary. However at various locations the mean rainfall has been changing over the years, due to urbanization, deforestation etc. Moreover, global warming has already been affecting Indian rainfall, and may do so more drastically in future. The changes are likely to be spatially uneven, and a reliable simulator must faithfully reflect the spatio-temporal properties. It should be possible to incorporate such effects into the proposed rainfall simulator by making some variables conditionally dependent on variables in previous years (Markov process). Some variables in the model can be made dependent on other meteorological variables, and reanalysis products or GCMs could be used in combination with such a rainfall generator for making hindcasts/forecasts of rainfall over a large region. Furthermore the properties of generative models might have application as a diagnosis tool for testing and comparing simulations from GCMs.

Such stochastic rainfall simulators have been used for studying various policies in other places, for example, rainwater harvesting policies in Africa ([9]). Along these lines, we hope to integrate our stochastic simulators into other models in future, for hydrological planning related to water control and conveyance, land-use management etc.

Acknowledgments

The research was done when the author was at Divecha Center for Climate Change, Indian Institute of Science. The author is indebted to Dr. Ashwin K. Seshadri for his valuable suggestions.

References

[1] P. Dabral, A. Pandey, N. Baithuri, and B. Mal, "Stochastic modelling of rainfall in humid region of northeast india," *Water resources management*, vol. 22, no. 10, pp. 1395–1407, 2008.

[2] I. Rodriguez-Iturbe, D. Cox, and V. Isham, "A point process model for rainfall: further developments," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 417, pp. 283–298, The Royal Society, 1988.

[3] W. Kleiber, R. W. Katz, and B. Rajagopalan, "Daily spatiotemporal precipitation simulation using latent and transformed gaussian processes," *Water Resources Research*, vol. 48, no. 1, 2012.

[4] C. Glasbey and I. Nevison, "Rainfall modelling using a latent gaussian variable," in *Modelling Longitudinal and Spatially Correlated Data*, pp. 233–242, Springer, 1997.

[5] D. Wilks, "Multisite generalization of a daily stochastic precipitation generation model," *Journal of Hydrology*, vol. 210, no. 1, pp. 178–191, 1998.

[6] P. A. Mendoza, M. P. Clark, M. Barlage, B. Rajagopalan, L. Samaniego, G. Abramowitz, and H. Gupta, "Are we unnecessarily constraining the agility of complex process-based models?," *Water Resources Research*, vol. 51, no. 1, pp. 716–728, 2015.

[7] A. Y. Ng, M. I. Jordan, Y. Weiss, *et al.*, "On spectral clustering: Analysis and an algorithm," *Advances in neural information processing systems*, vol. 2, pp. 849–856, 2002.

[8] C. Jayasankar, S. Surendran, and K. Rajendran, "Robust signals of future projections of indian summer monsoon rainfall by ipcc ar5 climate models: Role of seasonal cycle and interannual variability," *Geophysical Research Letters*, vol. 42, no. 9, pp. 3513–3520, 2015.

[9] J. R. Cowden, D. W. Watkins, and J. R. Mihelcic, "Stochastic rainfall modeling in west africa: parsimonious approaches for domestic rainwater harvesting assessment," *Journal of Hydrology*, vol. 361, no. 1, pp. 64–77, 2008.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

A NONPARAMETRIC COPULA BASED BIAS CORRECTION METHOD FOR STATISTICAL DOWNSCALING

Yi Li[1], Adam Ding[1], Jennifer Dy[2]

***Abstract*—Global Climate Models (GCMs) currently provide coarse resolution outputs which preclude their application to accurately assess the effects of climate change on finer regional scale events that are important to inform stakeholders in making policy, management, or infrastructure planning decisions. Statistical downscaling are methods that use statistical models to infer the regional-scale or local-scale climate information from coarsely resolved climate models. One popular approach for statistical downscaling is the bias correction and spatial disaggregation (BCSD) method. BCSD utilizes quantile mapping to perform bias correction between the coarse resolution climate models to the fine resolution projection. In this paper, we analyze BCSD from a copula point of view and show that it is a restricted form of the copula function. Instead, we propose a nonparametric copula based BCSD method (NCBCSD) and empirically show that this more flexible method provides improved climate projection performance in terms of mean-squared-error compared to the traditional BCSD method.**

I. INTRODUCTION

Intense and frequent precipitation can have disastrous effects to society through the damage of agriculture, infrastructure, and economy. Over the past years there has been a significant increase in frequency and intensity of regional rainfall around the world [1]. Global Climate Models (GCMs) are used to project future climate scenarios many years into the future; however, GCMs only provide climate information at a coarse scale. Stakeholders need finer regional scale information to make time critical policy, management or planning decisions. Climate downscaling approaches are used to infer the regional-scale or local-scale climate information from coarsely aggregated climate models [2]. Statistical downscaling is a way that uses statistical methods for this purpose [3] [4] [5].

Predicting climate variables at a finer resolution from GCMs of local-scale and/or large-scale variables is not trivial. Bias correction and spatial disaggregation (BCSD) [6] is a popular approach for this climate statistical downscaling task. It is widely used in many applications [7]. It uses quantile mapping (QM) to correct the bias from the large scale GCMs to the finer regional scale.

In this work, we generalize the BCSD method from the copula point of view. The main contributions of this work are: (1) We show that the copula between the GCM and the local scale observations in the BCSD model achieves the Fréchet-Hoeffding copula upper bound, which assumes the strongest bivariate association; (2) We propose a nonparametric copula bias correction and spatial disaggregation (NCBCSD) method for statistical downscaling based on kernel density estimation (KDE). This method generalizes the BCSD in the way that it captures the joint copula between the GCM output and the local observation to be based on data, rather than assuming it to be the special copula upper bound as in BCSD. Consequently, NCBCSD has more flexility than BCSD in modeling the climate dependency structure which leads to higher accuracy in correcting the bias of the GCM output.

The rest of the paper is organized as follows. Section II provides a review on BCSD and introduces our proposed copula based approach. Section III provides empirical results on a precipitation data. Finally, we conclude in Section IV.

II. METHODOLOGY

In this section, we first briefly review the BCSD model, and then restate this problem from the copula point of view. Finally, we generalize the model with nonparametric density estimation with our proposed nonparametric copula bias correction and spatial disaggregation (NCBCSD) approach.

A. Review on BCSD

The Bias Correction and Spatial Disaggregation (BCSD) method is a widely used model in statistical downscaling. It relates the GCM output to the finer

Corresponding author: Yi Li, li.yi3@husky.neu.edu [1]Department of Mathematics, Northeastern University. [2] Department of ECE, Northeastern University.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

local scale resolution. The left subfigure in Fig 1 [8] displays an example GCM output of a climate variable (mean monthly temperature in January of 1950). Note that the resolution is at a coarse scale of $2.5^\circ \times 2.5^\circ$; while the right subfigure displays the same climate variable in a higher resolution at $\frac{1}{8}^\circ \times \frac{1}{8}^\circ$ on land area.

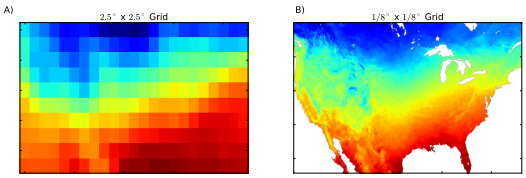


Fig. 1. Two separate datasets of mean monthly temperature in January 1950. A) Represents a coarse dataset at $2.5^\circ \times 2.5^\circ$. B) is a much higher resolution gridded dataset at $\frac{1}{8}^\circ \times \frac{1}{8}^\circ$ on land.

BCSD consists of two main steps. The first step involves the *bias correction* of the GCM output with the local scale information. Let X denote the output of the GCM, X^* be the bias corrected GCM output, Y be the local scale observation, and the cumulative distribution functions (CDFs) of X and Y are F_X and F_Y. In practice, the distributions for X and Y are observed to be different (*biased*). See Fig 2 for an illustration of this bias. BCSD *corrects* for this *bias* by applying quantile mapping (QM), a nonparametric way to match the distributions of X and Y (Fig 2) using the following equation

$$x^* = F_Y^{-1}(F_X(x)) \quad (1)$$

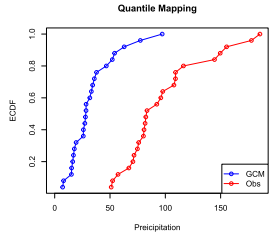


Fig. 2. Example of the empirical CDF (ECDF) used in QM of a climate variable in GCM and observational local scale.

The second step, *spatial disaggregation*, disaggregates (i.e., downscales) the bias-corrected GCM to local scale by using a local scaling factor learned from the historical local data. Local scaling factor is obtained first to reflect the interaction between the local station and the global climate status. Let X_t be the GCM output at some future time t, similar for $Y_t(s)$ as the future local scale observation at station s, downscaling of precipitation can be achieved with:

$$Y_t(s) = X_t^* \frac{\langle Y(s) \rangle}{\langle X^* \rangle}, \quad (2)$$

where $\langle \cdot \rangle$ is the historical average of the climate data.

B. Copula

From the form of (1), one can observe that it essentially characterizes a specific relationship between X and Y with their CDFs, i.e., the coarser resolution GCM output and the local observation. However, the joint dependence between two random variables can be better described with the concept of copula in statistics. For d random variables $X = (X_1, \cdots, X_d)$, their joint CDF can be decomposed by Sklar's Theorem [9]:

$$F(x_1, \cdots, x_d) = C[F_1(x_1), \cdots, F_d(x_d)], \quad (3)$$

where $F_j(x) = \mathbb{P}(X_j \leq x)$ is the marginal distribution of X_j and C is a copula – a joint distribution on the d-dimensional unit cube. $C(u_1, ..., u_d) = \mathbb{P}(U_1 \leq u_1, ..., U_d \leq u_d)$ is the CDF of copula-transformed, uniformly distributed, variables $U_j = F_j(X_j)$. This decomposition separates the dependence structure in the data from the marginals. All dependence information is contained in the copula only. Fig 3 shows the data from two distributions with different marginals but the same dependence structure.

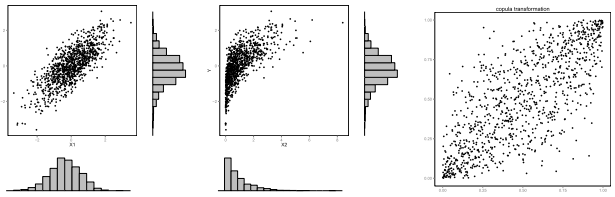


Fig. 3. (A) Bivariate Gaussian data with $\rho = 0.75$. (B) The data with exponential marginal for X. (C) The Gaussian copula. The first two distributions both have this copula.

Proposition 1 (Distribution of QM): The bias corrected GCM with quantile mapping in the BCSD model as defined in (1) follows the local distribution.

Proof: The distribution of X^* is $\mathbb{P}(X^* \leq y) = \mathbb{P}(F_Y^{-1}(F_X(X)) \leq y) = F_Y(y)$. ■

This implies that the bias corrected GCM with quantile mapping follows the distribution of Y, which is the distribution of the upscaled local observation. Proposition 1 and equation (1) imply that the copula-transformed variable $V = F_{X^*}(X^*) = F_Y(X^*) = F_X(X) = U$ so that the copula between X and X^* is $C(u, v) = \mathbb{P}(U \leq u, V \leq v) = \mathbb{P}(U \leq \min(u, v)) = \min(u, v)$. In the theory of copula [9], we can bound the copula by $\max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v)$. This implies that quantile mapping (using X^* to predict Y) assumes the strongest association between the local observation and the GCM on the copula scale.

Proposition 2 (Copula of QM): The copula function associated with the quantile mapping in the BCSD model (1) achieves the Fréchet-Hoeffding copula upper bound [9] $C(u, v) = \min(u, v)$.

C. Nonparametric Copula Based BCSD

Based on the result from Proposition 2, the dependence structure between the GCM output and the local

observations can be modeled more flexibly by doing the bias correction step with the joint copula density (without assuming it to be the Fréchet-Hoeffding upper bound), which captures the dependence between the two random variables.

In recent years, there has been growing interest in the research of modal regression in statistics, such as [10] [11], which solves the regression problem through the conditional mode of Y given X, i.e. $\texttt{mode}(Y|X=x)$, instead of the mean in conventional regression. The advantage of modal regression is that it can reveal structure that is missed by the conditional mean [10].

We incorporate modal regression and the copula density to perform statistical downscaling and propose the following modal regression model based on copula, which can be viewed as a generalization of quantile mapping in terms of modeling the joint behavior of the climate variable in GCM and the stations:

$$\begin{aligned}\texttt{mode}(Y|X=x) &= \texttt{argmax}_y f_{Y|X}(y|x) \\ &= \texttt{argmax}_y c(u,v) f_Y(y),\end{aligned} \tag{4}$$

where $U = F_X(X)$ and $V = F_Y(Y)$. Note that this is a nonparametric model and we do not assume the parametric or even linear structure ($\texttt{mode}(Y|X=x) = \beta^{\mathsf{T}}x$) of the mode as in [11].

Due to the low dimensionality of the above relationship, a nonparametric kernel density estimation method [12] is appropriate in practice to estimate the densities in (4), which enjoys nice theoretical consistency properties. Thus, the copula density and the marginal density can be estimated as: $\hat{c}(u,v) = \frac{1}{nh^2}\sum_{i=1}^{n} K(\frac{u-u_i}{h})K(\frac{v-v_i}{h})$, and $\hat{f}_Y(y) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{y-y_i}{h})$, where h is the bandwidth, which controls the smoothness, and $K(\cdot)$ is the kernel function. In practice, we apply the Gaussian kernel and select the optimal bandwidth based on the rule-of-thumb bandwidth estimator [13]: $h = (\frac{4\hat{\sigma}^5}{3n})^{\frac{1}{5}} \approx 1.06\hat{\sigma}n^{-\frac{1}{5}}$, where $\hat{\sigma}$ is the standard deviation of the samples. Thus, the full model is

$$\hat{Y} = \texttt{argmax}_y \hat{c}(u,v)\hat{f}_Y(y). \tag{5}$$

Intuitively, since this model does not assume the copula structure as in BCSD (see Proposition 2), and the model can learn the dependence structure between the GCM and the local observations based on the data. Therefore, we expect it to generally perform better than BCSD when correcting the bias for GCM output.

III. NUMERICAL EXAMPLES

In this section, we present a statistical downscaling task for monthly precipitation based on data from the south New England Area ($-73.15°$W to $-71.25°$W, and $41°$N to $43°$N), see Fig4. The GCM output is from the Geophysical Fluid Dynamics Laboratory's Coupled Physical Model (CM3). The local station observation data is from the University of Idaho Gridded Surface Meteorological Data [14].

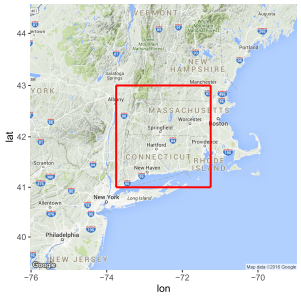


Fig. 4. The south New England Region (Massachusetts, Connecticut, Rhode Island and south New Hampshire) in one GCM grid with 2880 (60×48) local observations.

The left subfigure of Fig 5 presents the bias correction result for different methods. As we can see, the upscaled local observation (red) and raw GCM output (blue) is very different and the bias correction results with NCBCSD (black) is closer to the observation (red) than BCSD (blue) in the testing region.

The right right subfigure of Fig 5 plots RMSE: the square Root of MSE (mean squared prediction errors that is averaged over the 60×48 local stations). The NCBCSD method (black) yields the lowest RMSE compared with BCSD (blue) and the raw input of GCM without bias correction (red).

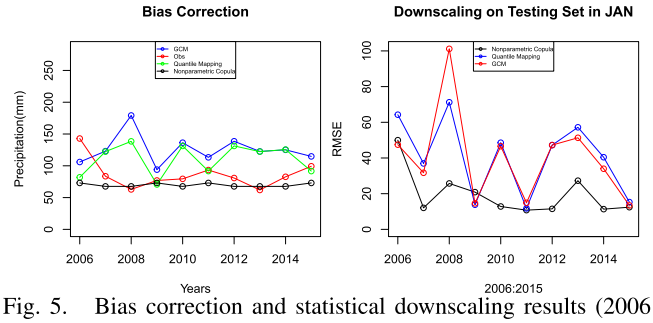


Fig. 5. Bias correction and statistical downscaling results (2006 - 2015) of the south New England Region in January.

IV. CONCLUSION

In this paper, we propose a nonparametric copula BCSD model for statistical downscaling, which relaxes the joint dependence structure and generalizes the idea of quantile mapping in BCSD. Results show that it has better performance in bias correction and consequently more accurate projection for statistical downscaling.

ACKNOWLEDGMENTS

We would like to acknowledge support for this project from the NSF grant CCF-1442728.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] H. Madsen, K. Arnbjerg-Nielsen, and P. S. Mikkelsen, "Update of regional intensity–duration–frequency curves in denmark: tendency towards increased storm intensities," *Atmospheric Research*, vol. 92, no. 3, pp. 343–349, 2009.

[2] J. T. Schoof, "Statistical downscaling in climatology," *Geography Compass*, vol. 7, no. 4, pp. 249–265, 2013.

[3] S. Ghosh, "SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output," *Journal of Geophysical Research: Atmospheres*, vol. 115, no. November 2009, pp. 1–18, 2010.

[4] "The statistical downscaling model: Insights from one decade of application," *International Journal of Climatology*, vol. 33, pp. 1707–1719, June 2013.

[5] J. T. Schoof, "High-resolution projections of 21st century daily precipitation for the contiguous u.s.," *Journal of Geophysical Research: Atmospheres*, vol. 120, no. 8, pp. 3029–3042, 2015. 2014JD022376.

[6] A. W. Wood, E. P. Maurer, A. Kumar, and D. P. Lettenmaier, "Long-range experimental hydrologic forecasting for the eastern United States," *Journal of Geophysical Research (Atmospheres)*, vol. 107, p. 4429, Oct. 2002.

[7] G. Brger, T. Q. Murdock, A. T. Werner, S. R. Sobie, and A. J. Cannon, "Downscaling extremesan intercomparison of multiple statistical methods for present climate," *Journal of Climate*, vol. 25, no. 12, pp. 4366–4388, 2012.

[8] T. Vandal, "Advancing statistical downscaling with data science." manuscript, 2015.

[9] R. B. Nelsen, *An introduction to copulas (Springer series in statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[10] Y.-C. Chen, C. R. Genovese, R. J. Tibshirani, and L. Wasserman, "Nonparametric modal regression," *Ann. Statist.*, vol. 44, pp. 489–514, 04 2016.

[11] W. Yao and L. Li, "A new regression model: Modal linear regression," *Scandinavian Journal of Statistics*, vol. 41, no. 3, pp. 656–671, 2014.

[12] D. Scott, *Multivariate density estimation: theory, practice, and visualization*. Wiley Series in Probability and Statistics, Wiley, 1992.

[13] B. W. Silverman, *Density estimation for statistics and data analysis*, vol. 26. CRC press, 1986.

[14] J. T. Abatzoglou, "Development of gridded surface meteorological data for ecological applications and modelling," *International Journal of Climatology*, vol. 33, no. 1, pp. 121–131, 2013.

6th International Workshop on Climate Informatics
September 22-23, 2016
CI 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

DETECTING AND PREDICTING BEAUTIFUL SUNSETS USING SOCIAL MEDIA DATA

Emma Pierson[1]

***Abstract*—Beautiful sunsets are one of the few pleasures freely available to everyone, but due to their subjective nature are difficult to quantitatively study on a large scale. Here, we use 1.2 million sunset posts on Instagram, a picture-sharing platform, to detect beautiful sunsets in 10 American cities over 7 months. We show that our metric of sunset quality correlates with human assessments, make sunset quality scores publicly available to allow more systematic study of sunsets, and use this dataset to answer a number of basic questions. Do some locations have more beautiful sunsets than others? Are there meteorological features which predict beautiful sunsets? Does a beautiful sunset today predict a beautiful sunset tomorrow? Is it possible to detect beautiful sunsets early enough to notify people so they can go outside to enjoy them? What visual features are people responding to when they call a sunset beautiful? We validate a widely used sunset prediction model developed by meteorologists, produce an algorithm which can visually discriminate between beautiful and mediocre sunsets, and provide a messaging service and web interface to notify users of beautiful sunsets.**

I. MOTIVATION

If beautiful sunsets could be reliably predicted, or detected in real time, millions of people could enjoy a free daily light show. In spite of this, beautiful sunsets have long been more the realm of artists and poets than of scientists; large-scale, quantitative data is not readily available, and previous research is sparse and based on small datasets or general application of physical laws [1], [2], [3] rather than analysis of large-scale datasets. Our contribution in this paper is two-fold: we use the image-sharing website Instagram to collect and make publicly available what is to our knowledge the first large-scale dataset on beautiful sunsets, comprising sunset quality scores in 10 large American cities over 7 months; second, we use this dataset to derive principles of beautiful sunset prediction and detection.

Corresponding author: Emma Pierson, epierson@cs.stanford.edu
[1]Department of Computer Science, Stanford University

II. METHOD

A. Dataset

Instagram is a social media platform with more than 300 million daily users on which people can post images with tags to describe image content (eg, “beautiful #sunset”). From October 2015 - May 2016 we used the Instagram Search API [4] to collect public Instagram posts tagged with latitude and longitude with tags relating to sunset: eg “#sunset” and “#instasunset”. To mitigate cultural differences in Instagram usage, we confined our analysis to posts within the United States. In total we collected 1.2 million posts, each with a location and sunset picture. Because reliably detecting spikes in Instagram activity requires a large number of posts, we focused our analysis on 10 large American cities: Los Angeles, New York, Boston, Chicago, Washington DC, Miami, San Diego, Seattle, Philadelphia, and San Francisco. For each city c, we computed the total number of Instagram posts n_{cd} within 0.5 degrees of the city on each day d.

Although it is intuitive that more people will take sunset photographs on days with beautiful sunsets, we performed three additional tests to validate our metric. First, we confirmed by hand inspection that images collected under these tags were sunset-related. Second, we confirmed that sunset posting activity does spike dramatically at sunset (Fig. 1A), implying that people are in fact reacting to local conditions as opposed to posting previously taken pictures. Third, we had three research assistants hand-code sunset quality in five cities in our dataset (Los Angeles, Miami, New York, Seattle, and Boston) as follows. For each city, we randomly selected 10 sunset pictures from the five days with the highest n_{cd}, and 10 pictures from days near the median n_{cd}; we refer to these below as “beautiful” and “average” sunsets. We presented the hand coders with one pair of sunset pictures at a time – one beautiful sunset, and one average sunset – in random order, and had them choose the one they felt had the more beautiful sunset. The three hand-coders agreed 86% of the time on average, indicating consensus about what constituted a beautiful sunset. Hand-coders overwhelmingly preferred sunsets

6th International Workshop on Climate Informatics
September 22-23, 2016
CI 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

from days with high n_{cd} to days with median n_{cd}: in cases where all three hand-coders agreed, they preferred the high n_{cd} sunset 80% of the time, and on average individual hand-coders preferred the high n_{cd} sunset 75% of the time. (We suspect that, in fact, the gap between the high n_{cd} and median n_{cd} sunsets is even more dramatic; low-quality pictures on high n_{cd} days appeared to be due to lack of photographer skill, not sunset quality. We provide the pictures used in the test in the Supplementary Information[1] (SI).) While assessing sunset quality is a gloriously subjective enterprise which we invite our readers to partake in, these quantitative checks confirm that n_{cd} do offer the first large-scale scores of sunset quality which correlate with human assessments.

We further increased the utility of our sunset scores by controlling time-of-year effects and differences in city size using the following procedure:

1) To account for differences in city size, we divided each city's daily count of sunset posts by the mean number of sunset posts in that city, so each city had mean $n_{cd} = 1$.
2) Social media usage varies by weekday and time of day: a sunset that occurs at 8 PM on Sunday may get more posts than one at 4 PM on Tuesday even if both are of the same quality. To account for this, we ran a linear regression where the dependent variable was n_{cd} and the covariates were a categorical indicator variable for each weekday and a second-order polynomial in sunset time. (We used a second-order polynomial because it captured the fact that the number of posts may increase or decrease non-linearly in sunset time, but does not provide too many degrees of freedom, which could produce overfitting). We set s_{cd}, the normalized sunset score, equal to the residual: ie, the variation in sunset quality which time effects did not account for. (We accounted for city size effects separately, rather than in our regression, because they are likely multiplicative rather than additive.)
3) Finally, we computed a binary "beautiful sunset" indicator variable b_{cd} which was 1 if a sunset had a score s_{cd} in the top 15% for city c and 0 otherwise. We focus our analysis on b_{cd} rather than s_{cd} because it is not that important to discriminate between sunsets at the 20th and 40th percentiles (neither is worth going outside for); the goal of interest is to find beautiful sunsets.

Features that correlate with b_{cd} predict that a city is unusually likely to have a sunset in the top 15% when controlling for time of sunset. While this metric is interpretable and relevant, future work should also investigate whether some cities have better sunsets than others. To facilitate future analysis, we make the dataset of our sunset quality scores publicly available (SI).

B. *Beautiful sunset prediction*

As a first illustration of the utility of our dataset, we evaluated the accuracy of a widely used sunset prediction algorithm. SunsetWx [5], which predicts sunset quality using the 4km NAM, is used by more than 20 TV stations across the country to forecast beautiful sunsets, but its developers have had to rely on anecdotal reports of sunset quality as validation [6]. They provided early access to their API, and we compared their predicted sunset quality to our sunset quality scores b_{cd} on 58 days across 10 cities (a total of 580 datapoints).

SunsetWx computes real-valued sunset quality predictions which it stratifies into "Poor", "Fair", "Good", and "Great"; very few sunsets are predicted to be "Great", so we exclude them from our analysis. We found that SunsetWx's stratified scores have predictive value. $b_{cd} = 1$ for 6% of sunsets predicted to be "Poor"; 15% of sunsets predicted to be "Fair", and 31% of sunsets predicted to be "Good".

We suspect, however, that predictions from SunsetWx could be improved if a systematic machine learning approach was used to develop a predictive algorithm, since even very simple predictors exhibit comparable accuracy. For example, using humidity as a univariate predictor of b_{cd} and running a logistic regression yields an out-of-sample AUC of 0.72, which is better than the SunsetWx AUC of 0.67. (AUC [7], or area under the curve, is a standard measure for assessing performance on a binary classification task – for example, discriminating between beautiful and average sunsets – and is defined as the probability that the model assigns a higher score to a positive example than to negative example. Higher AUCs denote better performance.) We thus believe there is room for improvement, and it is worth noting that SunsetWx's prediction algorithm has evolved since we collected the data for this paper. We believe that the most successful prediction algorithms will use a large number of sunset quality scores, such as those we have collected here, to train and assess a machine learning algorithm. Prediction algorithms which are tuned and developed using anecdotal reports of good sunsets are unlikely to yield optimal performance.

While developing a prediction algorithm that fully uses meterological data is a topic for future work, we evaluated the predictive utility of 20 weather features using meteorological data from ForecastIO [8] and pollution

[1] All supplementary information is available at http://cs.stanford.edu/~emmap1/sunsets/supplementary_information.zip

6th International Workshop on Climate Informatics
September 22-23, 2016
CI 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

data from the EPA [9]. Days with temperatures above 60 degrees are statistically significantly more likely to have $b_{cd} = 1$; so are days with lower cloud cover (with zero cloud-cover days having the highest probability), lower humidity, and calm winds (for detailed graphs, see SI). Interestingly, we also found that higher levels of pollution correlate positively with b_{cd}, with higher levels of carbon monoxide and nitrogen dioxide showing statistically significant correlations ($p < .01$); (we also assessed ozone and sulfur dioxide, which showed positive but not statistically significant correlations). This provides some evidence in the debate between those who claim that clear skies produce more brilliant sunsets and those who disagree [3], [10] although it is worth noting that our analysis is confined to cities and rural, unpolluted areas may have better sunsets in general, even if cities have better sunsets on days with more pollution.

We also find that a good predictor of today's sunset quality is *yesterday's* sunset quality. If yesterday's sunset was beautiful, the probability that today's sunset will be beautiful is 34%, as compared to 12% if yesterday's was not beautiful. We confirm that this is correlation is primarily due to short-term effects, not to longer-term time of year effects; the predictive value is much reduced if a week's separation rather than a day's separation is used.

In Table 1, we compare the predictive power of the features discussed above. (To reduce noise, we use our full dataset, rather than just the subset for which we have SunsetWx data, so the AUCs for these predictions are not directly comparable to the AUCs for SunsetWx predictions). For each feature set, we fit a logistic regression model on a train set comprising a random half of the dataset and assess model performance (measured by AUC) on a test set comprising the remaining half. (Using a test set avoids overfitting.) The strongest weather predictors are humidity and cloud cover, and the strongest pollution predictors are nitrogen dioxide and carbon monoxide. We find that weather features are more predictive than pollution features, and that a combined model using both weather and pollution features (and the previous day's sunset quality) slightly outperforms weather features alone. It is worth noting that we assess predictive power using features measured *at the time of sunset*, and that a true prediction algorithm would have to use the values of those features predicted ahead of time, rather than their measured values.

Features used in model	AUC
Humidity	0.64 ± 0.03
Cloud cover	0.64 ± 0.03
Temperature	0.57 ± 0.02
Visibility	0.57 ± 0.01
Wind speed	0.57 ± 0.03
Pressure	0.48 ± 0.02
Nitrogen dioxide	0.58 ± 0.02
Carbon monoxide	0.57 ± 0.05
Ozone	0.54 ± 0.03
Sulfur dioxide	0.53 ± 0.03
Previous day	0.58 ± 0.02
All features combined	0.71 ± 0.02
Weather features only	0.69 ± 0.02
Pollution features only	0.64 ± 0.02

Table I

How well does each model predict b_{cd}? AUCs reported are computed using a held-out test set, with the model fit on a separate train set. Errors are the standard deviation in AUC over multiple bootstrapped train sets.

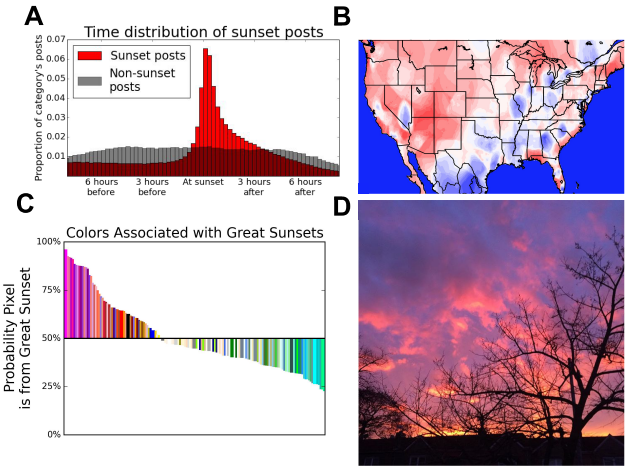


Fig. 1. A: Instagram sunset posts spike at sunset, while control posts do not. B: Density of sunset posts (red denotes higher densities) relative to control posts. C) colors associated with excellent sunsets. The vertical axis is the fraction of pixels of a particular color which are sampled from beautiful sunsets (as opposed to mediocre ones); D, the sunset algorithmically classified as most beautiful in a test set of 227 images.

C. Locations with beautiful sunsets

Merely plotting the density of sunset posts will not identify locations with beautiful sunsets because Instagram usage is heavily correlated with population density. We mitigate this problem by collecting a second dataset of Instagram posts with common tags (eg, "friends" and "smile") which we refer to as "control" posts; in Figure 1B, we plot the density of sunset posts relative to control posts across the United States. Sunset posts are particularly dense (red areas) near the coasts, in the mountains, and across much of the western United States. Results controlling for population density were qualitatively similar. (Because both methods control only imperfectly for location-based variation in social media usage, we used them only for this exploratory analysis rather than for the analysis in section B above.)

D. Visual features of beautiful sunsets

We next investigated whether we could quantitatively identify the visual features people associate with beautiful sunsets. We trained a classifier to classify sunsets as beautiful or average using a balanced dataset of beautiful and average sunsets selected as described in section A. We followed the following procedure: we randomly sampled 1,000 pixels from each image, computed the RGB representation of each pixel, and trained a K-nearest-neighbors classifier [11] to predict whether each pixel came from a beautiful or average sunset. (We used k-nearest-neighbors, which classifies a color depending on what fraction of the pixels nearest to it come from a beautiful sunset, because it allowed for flexible partitions of the color cube). In Figure 1C, we plot colors ranked by how strongly they are associated with beautiful sunsets (left, upward facing bars) or average sunsets (right, downward facing bars). The ranking matches intuition: pinks, purples, oranges, and reds are associated with beautiful sunsets, while greens and blues are associated with average sunsets. Strikingly, even this simple algorithm can classify a sunset as beautiful or average with accuracy comparable to our research assistants: on a test set of sunsets not used to train the algorithm, its AUC – ie, the probability that it assigns a higher score to a beautiful sunset than to an average sunset – is 0.70. (As discussed previously, the algorithm is unable to perfectly classify sunsets in part because photographer quality varies: some people take bad pictures on beautiful sunset days, or good pictures on mediocre sunset days.) In Figure 1D we show the picture in the test set to which the algorithm assigns the highest probability of being beautiful. It is easy to think of potential extensions and applications of this algorithm: it could be used to rapidly find particularly beautiful sunsets and improved by using a more sophisticated image-processing algorithm, like a convolutional neural network [12].

E. Real-time sunset notification

Beautiful sunset prediction is difficult because beautiful sunsets are rare: even if an algorithm can identify days when beautiful sunsets are twice as likely, they are still quite unlikely. An easier problem may be beautiful sunset *detection*: is it possible to detect beautiful sunsets as they are happening, using real-time social media data? We present a pilot of such a system, Sunset Nerd: we have created a website, http://sunsetfinder.herokuapp.com/, which monitors the real-time number of social media posts about the sunset in each city and assigns the sunset a quality score in real time by comparing to the historical number of sunset posts. We have built interfaces on Twitter and Facebook Messenger to allow users to receive beautiful sunset notifications so they can go outside. We will present results as they become available.

III. Discussion and Future Work

Though the sun sets on our present analysis of beautiful sunsets, many questions await a new dawn – and future work. Is it possible to use more sophisticated metereological features, like data on types of clouds, to improve sunset prediction? Will the real-time sunset detection system offer useful notifications? These open questions, and the analyses already performed, illustrate the utility of large-scale quantitative data in beautiful sunset prediction and detection.

Supplementary Information

All supplementary information is available at http://cs.stanford.edu/~emmap1/sunsets/supplementary_information.zip.

Acknowledgments

The author thanks the Hertz Foundation and NDSEGF for financial support; Justin Lowery and the other founders of SunsetWX for providing access to their API; Nick Mangus for EPA data; and Shengwu Li, Katherine Maslyn, and Andrew Suciu for research assistance and helpful conversations.

References

[1] C. Zerefos, P. Tetsis, A. Kazantzidis, V. Amiridis, S. Zerefos, J. Luterbacher, K. Eleftheratos, E. Gerasopoulos, S. Kazadzis, and A. Papayannis, "Further evidence of important environmental information content in red-to-green ratios as depicted in paintings by great masters," *Atmospheric Chemistry and Physics*, vol. 14, no. 6, pp. 2987–3015, 2014.
[2] R. E. Peterson, "Striking southern sunsets," *Weatherwise*, 1983.
[3] S. F. Corfidi, "The colors of sunset and twilight," 2014.
[4] "Instagram api." https://www.instagram.com/developer/, 2016.
[5] B. Reppert, J. DeFlitch, S. Hallett, and J. Lowery, "Sunsetwx: Innovation beyond the horizon." https://sunsetwx.com/, 2016.
[6] B. Reppert, J. DeFlitch, and S. Hallett, "Case study: Sunset model." https://www.sunsetwx.com/casestudy.pdf, 2016.
[7] J. Huang and C. X. Ling, "Using auc and accuracy in evaluating learning algorithms," *IEEE Transactions on knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.
[8] "Forecastio." https://developer.forecast.io/, 2016.
[9] "Epa air quality data." https://www3.epa.gov/airdata/ad_data.html, 2016.
[10] C. Ballantyne, "Fact or fiction: Smog creates beautiful sunsets," *Scientific American*, 2007.
[11] K. Fukunaga and P. M. Narendra, "A branch and bound algorithm for computing k-nearest neighbors," *IEEE transactions on computers*, vol. 100, no. 7, pp. 750–753, 1975.
[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

OCEANTEA: EXPLORING OCEAN-DERIVED CLIMATE DATA USING MICROSERVICES

Arne N. Johanson[1], Sascha Flögel[2], Wolf-Christian Dullo[2], Wilhelm Hasselbring[1]

***Abstract*—Ocean observation systems gather an increasing amount of climate-relevant time series data. To interactively explore and analyze such high-dimensional datasets, we developed the software OceanTEA. Our open-source tool leverages modern web technology to support interactive data visualization, spatial analysis of current patterns, and temporal pattern discovery via machine learning methods. The microservice architecture of OceanTEA ensures a maintainable implementation that seamlessly scales from desktop computers to cloud computing infrastructure.**

I. MOTIVATION

Ocean observation systems, such as the global array of more than 3000 free-drifting Argo floats belonging to the Global Ocean Observing System [1] or the modular ocean laboratory MoLab [2, 3], produce an increasing amount of time series data. Both statistical data mining techniques and manual exploration via visualization are necessary for oceanographers and climatologists to extract scientific knowledge from such vast datasets. For this purpose, we developed the software OceanTEA, which leverages modern web technology to support scientists in interactively exploring and analyzing high-dimensional datasets. By relying on a microservice architecture [4, 5], OceanTEA can not only be deployed on desktop computers but also on cloud computing infrastructure with built-in scalability. Making data available on the web can be useful for scientists collaborating on exploring a dataset (e.g., with limited access within an institute) as well as for providing interactive visualizations along with journal or conference publications. Since it has been shown that papers which feature published data receive higher citation counts [6], an interactive visualization of such data with OceanTEA could further improve the impact of a publication.

Corresponding author: A. Johanson, arj@informatik.uni-kiel.de [1]Software Engineering Group, Kiel University, Germany [2]GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany

The OceanTEA source code (along with a live demo of the tool) is available on GitHub[1] under the Apache 2.0 license [7].

A tool related to OceanTEA is Ocean Data View (ODV) [8], which is a proprietary (i.e., closed-source) desktop-only application used to produce a wide range of *static* figures from oceanographic datasets.

II. OCEANTEA

A screenshot of the web interface of OceanTEA (short for *Ocean*ographic *T*ime Series *E*xploration and *A*nalysis) is shown in Figure 1. The user interface is divided into four views (times series management, data exploration, spatial analysis, and temporal pattern discovery), which can be accessed via the tabs at the top of the page. The data exploration view (Figure 1) features options to filter the time series to be displayed according to:

1) study region
2) measurement device
3) measurement parameter (e.g., temperature)
4) depth range (multiple ranges are possible)

Furthermore, measurement stations can directly be selected via an interactive map displaying satellite images of the Earth's surface (provided by Google Maps [9]).

OceanTEA supports both univariate time series (e.g., temperature measured at a single site) and multivariate series (e.g., current direction and magnitude in several depth bins in the water column measured by an acoustic Doppler current profiler (ADCP)). Multivariate time series of currents (direction and magnitude) can be sliced along adjustable depth levels (see the right plot in Figure 1).

The interactive plots of OceanTEA are implemented using the CavasPlot [10] library built on top of D3.js [11]. The user can zoom into the plots and pan the axes (also by using touch gestures on devices that support them). At a high zoom level, the individual data points are displayed and tooltips are shown when

[1]https://github.com/a-johanson/oceantea

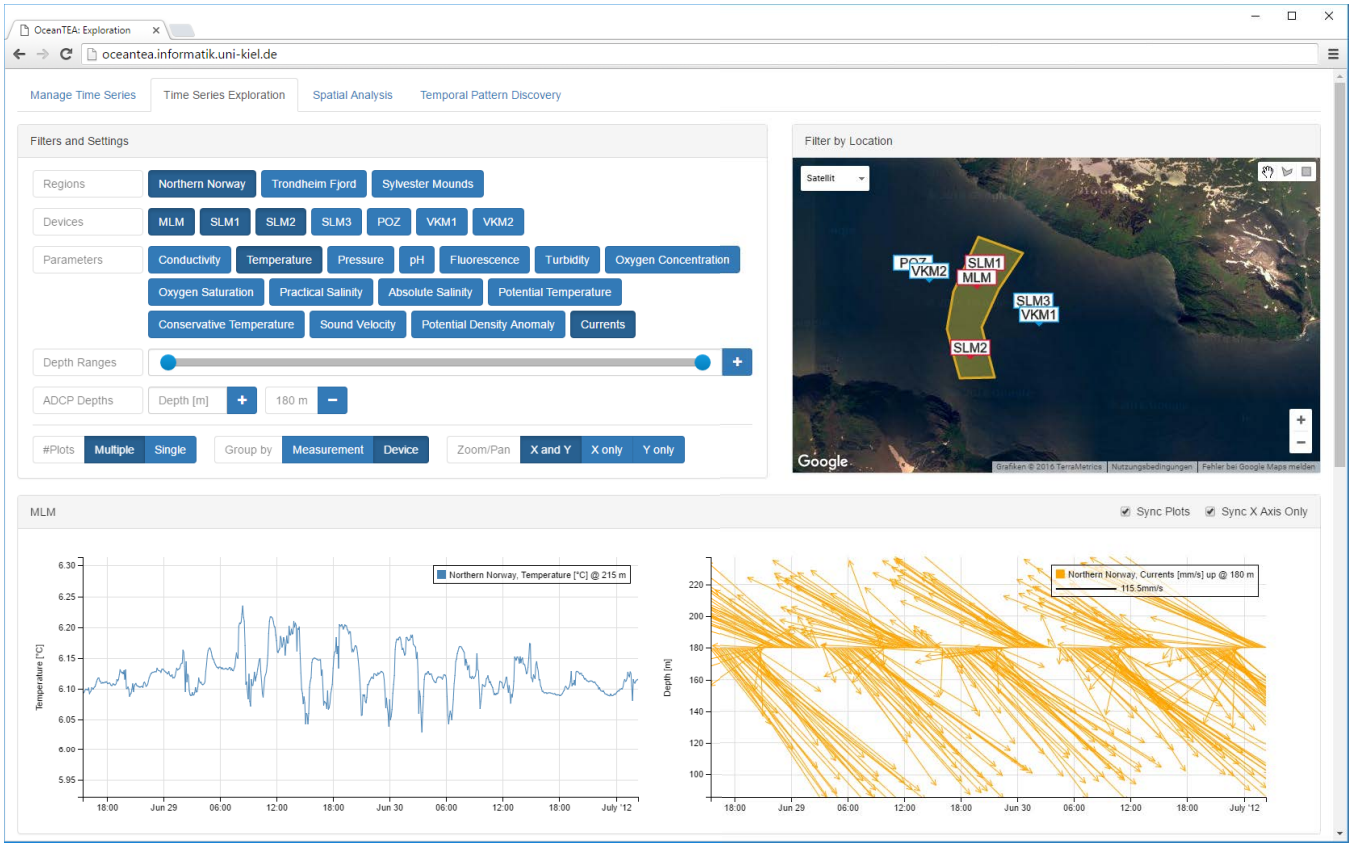


Figure 1. The data exploration view of OceanTEA.

the user hovers over the points (or touches them). The plots can be arranged by measurement parameter or by measurement device. It is possible to synchronize the axes of multiple plots and to join multiple graphs in a single plot.

In the management tab, time series can be added (e.g., from comma-separated values (CSV) files) and deleted. OceanTEA provides automatic unit conversion according to TEOS-10 [12] for several important oceanographic parameters (such as from *in situ* temperature to conservative temperature).

III. Microservice Architecture

The microservice architecture pattern [4, 5] partitions a software system into a set of so-called microservices. A microservice is a small, self-contained application that can be deployed independently and has a single responsibility [13]. In this context, small means that its complexity is low enough to be understood by a small team or even a single developer. That microservices are self-contained implies that they do not share code or database schemas with each other. In particular, each microservice can be implemented using the programming languages, middleware, and data storage facilities that suit the task of the service best (polyglot programming and persistence). As the whole software system is divided into microservices according to domain functionality (in the sense of bounded contexts in domain-driven design [14]), each service only has a single functional responsibility. Transaction-less communication—e.g., via RESTful protocols such as HTTP—is employed to coordinate tasks between the individual services.

While microservices incur the drawback of having to handle the additional complexity of distributed systems (e.g., ensuring fault tolerance), they provide the advantage of good maintainability and scalability [15]. As the complexity of a microservice is low, maintaining its code is easier than that of a large monolithic application (making it a feasible option to re-implement the whole service if necessary). Since microservices are self-contained and can be deployed independently, they can also be scaled independently as it is required by the current workload on the software system [15, 16, 17].

Figure 2 shows the microservice architecture of OceanTEA. The OceanTEA client, which contains the user interface and runs in the user's web browser, communicates with the server-side part of OceanTEA via

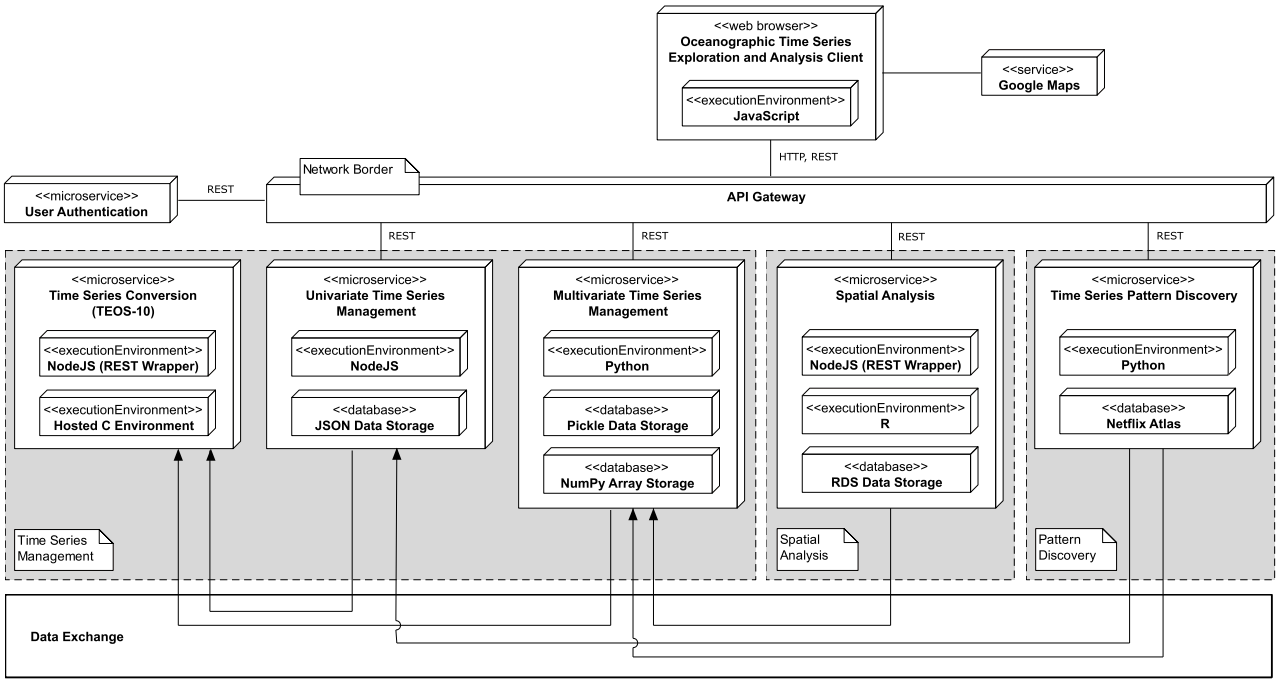


Figure 2. The microservice software architecture of OceanTEA.

an application programming interface (API) gateway. This gateway masks the complexity of communicating with different services and offers an integrated API to the client. The microservices comprising OceanTEA are divided into three so-called *verticals* that group services with related functionality. In the first vertical, we arranged microservices related to the management of time series. In the second vertical, we find the service for the spatial analysis of data, and the third vertical consists of the service for pattern discovery in time. Note, that we made use of the polyglot properties of microservices; for example, we reused an existing C implementation of TEOS-10 for the conversion microservice and implemented the multidimensional array handling of the multivariate time series management in Python, which allows to express the required slicing of arrays with a concise syntax.

We utilize Docker [18] to run each microservice in an isolated container. These containers can directly be deployed to private or public cloud infrastructure. Via Docker Machine [19], the same containers can also be executed on desktop computers running Mac OS X or Microsoft Windows. For these two platforms, we built installer applications to make the installation process user-friendly.

IV. FUTURE WORK

In the future, we plan to extend the spatial analysis view of OceanTEA with interactive 3D renderings of data fields in relation to structures on the ocean floor (work in progress). For the temporal analysis view, we are working on an implementation that leverages machine learning methods [20] to identify dependencies between different (lagged) time series.

OceanTEA will be used to interactively visualize, explore, and analyze oceanographic time series data for climate-relevant research concerning ocean physics, biology, and chemistry in a changing climate system. For example, OceanTEA is currently employed in studying the impact of ongoing climate change on cold-water coral reef ecosystems in order to assess whether factors such as ocean warming and acidification impact the physical and biogeochemical boundary conditions of these reefs [cf. 21]. We implemented an interactive illustration of modeling results with OceanTEA to accompany a publication which we prepared in this context [22]. In this way, our tool can be used to create dynamic visualizations of figures in papers to add value to publications reporting on data-driven research right from the beginning of the peer review process.

ACKNOWLEDGMENTS

This project was funded by the Cluster of Excellence 80 “The Future Ocean.” The “Future Ocean” is funded within the framework of the Excellence Initiative by the Deutsche Forschungsgemeinschaft (DFG) on behalf of the German federal and state governments.

REFERENCES

[1] D. Roemmich, G. C. Johnson, S. C. Riser, R. E. Davis, J. Gilson, W. B. Owens, S. L. Garzoli, C. Schmid, and M. Ignaszewski, "The argo program: Observing the global ocean with profiling floats," *Oceanography*, vol. 22, pp. 34–43, 2009.

[2] L. Rovelli, K. M. Attard, L. D. Bryant, S. Flögel, H. J. Stahl, M. Roberts, P. Linke, and R. N. Glud, "Benthic O_2 uptake of two cold-water coral communities estimated with the non-invasive eddy-correlation technique," *Marine Ecology Progress Series*, vol. 525, pp. 97–104, 2015.

[3] S. Flögel and W. Dullo, "High-resolution water mass measurements around cold-water corals: a comparative test study between repeated Conductivity-Temperature-Depth (CTD) casts and continuous data acquisition of bottom waters from the West Florida Slope, Gulf of Mexico," *Annalen des Naturhistorischen Museums in Wien. Serie A für Mineralogie und Petrographie, Geologie und Paläontologie, Anthropologie und Prähistorie*, pp. 209–224, 2011.

[4] S. Newman, *Building Microservices*. O'Reilly, 2015.

[5] E. Wolff, *Microservices: Flexible Software Architectures*. CreateSpace, 2016.

[6] H. A. Piwowar, T. J. Vision, and M. C. Whitlock, "Data archiving is a good investment," *Nature*, vol. 473, no. 7347, pp. 285–285, 2011.

[7] The Apache Software Foundation, "Apache license, version 2.0." https://www.apache.org/licenses/LICENSE-2.0, 2004.

[8] R. Schlitzer, "Ocean Data View." http://odv.awi.de, 2016.

[9] Google Inc., "Google Maps." https://maps.google.com, 2016.

[10] A. N. Johanson, "CanvasPlot: A JavaScript plotting library based on D3.js for visualizing large data sets." https://github.com/a-johanson/canvas-plot, 2016.

[11] M. Bostock, "D3.js: Data-Driven Documents." https://d3js.org, 2016.

[12] T. J. McDougall and P. M. Barker, "Getting started with TEOS-10 and the Gibbs Seawater (GSW) oceanographic toolbox," tech. rep., SCOR/IAPSO Working Group 127, 2011.

[13] J. Thönes, "Microservices," *IEEE Software*, vol. 32, no. 1, pp. 113–116, 2015.

[14] E. Evans, *Domain-Driven Design: Tackling Complexity in the Heart of Software*. Addison-Wesley, 2004.

[15] W. Hasselbring, "Microservices for scalability," in *Proceedings of the 7th ACM/SPEC on International Conference on Performance Engineering*, pp. 133–134, 2016.

[16] R. von Massow, A. van Hoorn, and W. Hasselbring, "Performance simulation of runtime reconfigurable component-based software architectures," in *Software Architecture (Proceedings ECSA 2011)*, vol. 6903 of *Lecture Notes in Computer Science*, pp. 43–58, Springer, 2011.

[17] A. van Hoorn, M. Rohr, I. A. Gul, and W. Hasselbring, "An adaptation framework enabling resource-efficient operation of software systems," in *Proc. of the Warm Up Workshop (WUP 2009) for ACM/IEEE ICSE 2010*, pp. 37–40, 2009.

[18] Docker Inc., "Docker." https://www.docker.com, 2016.

[19] Docker Inc., "Docker Machine." https://www.docker.com/products/docker-machine, 2016.

[20] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[21] S. Flögel, W. Dullo, O. Pfannkuche, K. Kiriakoulakis, and A. Rüggeberg, "Geochemical and physical constraints for the occurrence of living cold-water corals," *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 99, pp. 19–26, 2014.

[22] A. N. Johanson, S. Flögel, W. Dullo, P. Linke, and W. Hasselbring, "Modeling polyp activity of *Paragorgia arborea* using supervised learning." https://github.com/a-johanson/paragorgia-arborea-activity, 2016.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

DEPENDENCE OF INFERRED CLIMATE SENSITIVITY ON THE DISCREPANCY MODEL

Balu Nadiga[1], Nathan Urban[1]

***Abstract*—We consider the effect of different temporal error structures on the inference of equilibrium climate sensitivity[1](ECS), in the context of an energy balance model (EBM) that is commonly employed in analyzing earth system models (ESM) and observations. We consider error structures ranging from uncorrelated (IID normal) to AR(1) to Gaussian correlation (Gaussian Process GP) to analyze the abrupt 4xCO_2 CMIP5 experiment in twenty-one different ESMs. For seven of the ESMs, the posterior distribution of ECS is seen to depend rather weakly on the discrepancy model used suggesting that the discrepancies were largely uncorrelated. However, large differences for four, and moderate differences for the rest of the ESMs, leads us to suggest that AR(1) is an appropriate discrepancy correlation structure to use in situations such as the one considered in this article.**

Other significant findings include: (a) When estimates of ECS (mode) were differrent, estimates using IID were higher (b) For four of the ESMs, uncertainty in the inference of ECS was higher with the IID discrepancy structure than with the other correlated structures, and (c) Uncertainty in the estimation of GP parameters were much higher than with the estimation of IID or AR(1) parameters, possibly due to identifiability issues. They need to be investigated further.

I. INTRODUCTION

On the one hand, Earth System Models (ESMs) that comprise of atmosphere-ocean general circulation models (AOGCMs) coupled to other earth system components such as ice sheets, land surface, terrestrial biosphere, and glaciers are central to developing our understanding of of the workings of the climate system, and are proving to be the most comprehensive tool available to study climate change and develop climate projections [1]. Concomitantly, simple concepts such as climate sensitivities—metrics used to characterise the response of the global climate system to a given forcing—are central not only to climate modeling, but also to discussions of the ongoing global warming (e.g., see [2], [3]. Nevertheless, the immense computational infrastructure required and the cost incurred in running ESMs precludes the direct evaluation of such metrics.

Simple climate models (SCMs) on the other hand consider only integral balances of important quantities such as mass and/or energy, are computationally cheap and can be used in myriad differnt ways (unlike ESMs that are typically run only in the forward mode). It is for these reasons that the use of SCMs to estimate climate sensitivities, both in the context of ESMs and actual observations, is now well established (e.g., see [4], [5] and others).

II. METHODOLOGY AND RESULTS

A particular form of an SCM that has been popular in summarizing integral thermal properties of AOGCMs and/or ESMs is the anomaly-based upwelling-diffusion (UD) energy-balance model (EBM) [4]. To briefly describe such a model, consider a horizontally-integrated model of the climate system that is partitioned into two active layers in the vertical. An upper (surface) layer that that comprises the oceanic mixed layer, atmosphere and land surface and and a bottom layer that comprises the ocean beneath the mixed layer. Evolution of the upper surface heat content anomaly per unit area is given by

$$C_u \frac{dT_u}{dt} = \mathcal{F} - \lambda T_u - \gamma \left(T_u - T_d\right) \tag{1}$$

where λ represents. Exchange of heat between the surface layer and the ocean beneath is parameterized by the difference in temperature between the two layers. Similarly, evolution of the subsurface ocean heat content anomaly (again per unit area) is given by

$$C_d \frac{dT_d}{dt} = -\gamma \left(T_u - T_d\right). \tag{2}$$

Such two layer models have been used extensively to obtain point estimates of ECS of AOGCMs/ESMs (e.g., see [5], [6], and others). However, and to the best of our knowledge, the dependence of estimates of ECS on the assumed temporal structure of the discrepancy between ESM representation of the surface air temperature (SAT) and the above EBM's (Eqs. 1 & 2) representation of it has not been investigated:

$$T_u^{ESM}(t) = T_u^{EBM}(t) + \epsilon_t \tag{3}$$

Corresponding author: B.T. Nadiga, balu@lanl.gov
[1]Los Alamos National Lab, Los Alamos, NM 87544

[1]ECS is defined as the realized equilibrium surface warming—globally-averaged surface air temperature—for a doubling of CO_2

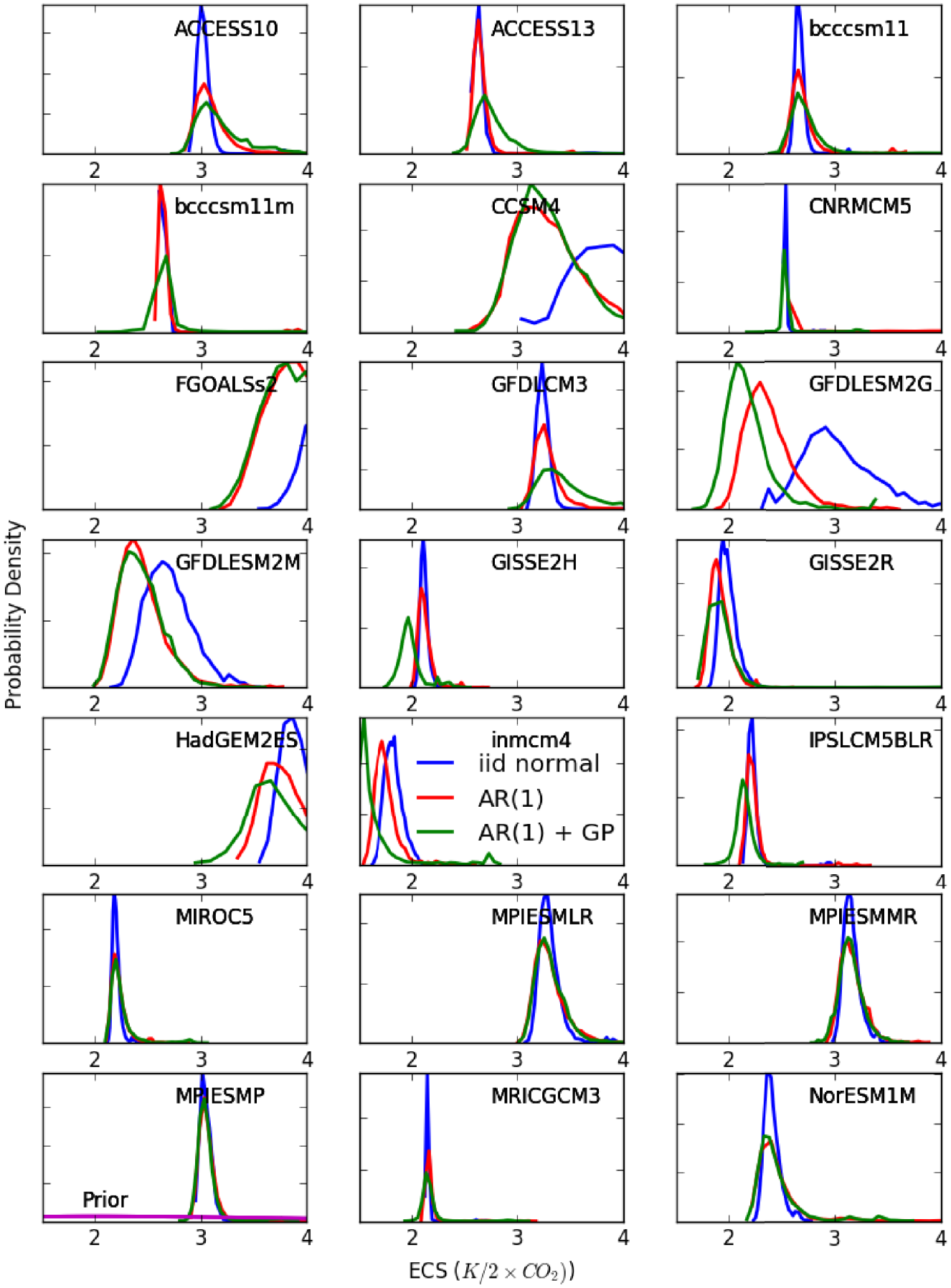


Fig. 1. Inferred ECS for the 21 models considered using CMIP5 experiment abrupt4xCO_2.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

The equation above arises from the fact that the anomaly-based EBM considered has no representation of climate variability, unlike the more comprehensive ESM that it is used to analyze. We consider three correlation structures for ϵ_t:

1) IID: $\Sigma(t-s) = \sigma^2\delta(t-s)$
2) AR(1): $\Sigma(t-s) = \sigma^2\rho^{|t-s|}$
3) AR(1)+GP: $\Sigma(t-s) = \sigma^2\exp(-\frac{(t-s)^2}{\lambda^2})$

where structure AR(1)+GP uses the sum of the covariances indicated in items 2 and 3 above

In the context of globally-averaged SAT (of which sea surface temperature or SST is a large component), we know, e.g., following the work of [7], that the mixed layer integrates (high-frequency) weather noise. Thus, SST (and therefore SAT) is expected to be auto-correlated in time, although the correlations themselves are highly spatiotemporally variable (e.g, winter SST anomalies are more persistent than summer SST anomalies, the tropical Pacific may display larger persistence than the tropical Atlantic, etc...). However, such correlations rarely exceed about six months and we are considering annual-averaged SAT. Physically, correlations on the interannual time scales are related to internal climate dynamics (phenomena such as the re-emergence of the winter mixed layer, delay-oscillations and others). While a casual inspection of some actual climate timeseries may suggest the unlikeliness of IID variability, it is not the case for the globally-averaged SAT timeseries in the abrupt4xCO_2 CMIP5 experiment that we analyze, and as we will see later. However, it should also be noted that if the discrepancies are actually correlated, then an inference of EBM parameters using the IID discrepency structure will result in estimates of uncertainty that are smaller than actual, and again as we will see later.

Figure 1 shows the posterior distribution of ECS with the three error structures (indicated in the legend) for the 21 ESMs. The prior is shown in the bottom-left panel. In this figure it is seen that

- For a substantial number of the ESMs, the three error structures lead to similar estimates of ECS (CNRMCM5, MIROC5, MPIESML/MR, MPIESMP, MRICGCM3, NorESM1M)
- When the estimates of ECS are differrent, estimates using IID tend to be higher (CCSM4, FGOALSs2, GFDLESM2G/M, HadGEM2ES, inmcm4)
- Differences in ECS estimates from that between AR(1) and AR(1)+GP tend to be smaller than that between either and IID
- In a majority of the ESMs considered, IID leads to smaller estimates of uncertainty in ECS suggesting that the discrepancies in those models are temporally correlated. The exceptions (CCSM4, GOALSs2, GFDLESM2G/M), are therfore surprising and need to be investigated further.

We also note that there was far more uncertainty in the estimation of GP parameters as opposed to estimateion of either IID and AR(1) parameters. This is likely due not only to the shortness of the ESM runs considered (150 years) from the point of low-frequency variability that the GP component was intended to capture, but may involve issues of identifiability and needs to be investigated further. However, when such problems occur, the parameters involved act more as nuisance parameters and do not prevent reasonable inference of ECS and other EBM parameters.

III. DISCUSSION

Simple climate models play a valuable role in helping interpret both observations and the responses of comprehensive ESMs. As such, we used a simple and popular EBM in a Bayesian framework to interpret the abrupt4xCO_2 CMIP5 experiment in 21 ESMs, in terms of their SAT response. We used three different statistical models to represent the discrepancy in the SAT response of the ESMs and SCMs. This discrepancy is largely due to natural variability—an aspect of climate that represented in the ESMs, but not in the SCMs. For seven of these models, the posterior distribution of ECS depended only very weakly on the discrepancy model used suggesting that the discrepancies were largely uncorrelated. For four of the models, the differences were large and for the rest of the models, the differences were moderate. Significant differences in a majority of the models, therefore, indicate the existence of temporal correlations in the discrepancies and the imprtance of accounting for them in a Bayesian inference framework.

Next, the differences in estimated ECSs were much smaller for inferences using AR(1) and AR(1)+GP as compared to differences between inferences using either of these models and IID. This coupled with the fact that the uncertainty in the estimation of GP parameters was much larger than that in the estimation of AR or IID parameters, leads us to conclude that AR(1) is a good choice[2] in situations such as the one considered in this article.

A number of other issues need to be investigated further: the higher estimates of ECS when using the IID structure for some of the ESMs, the higher uncertainty in the estimation of ECS when using the IID structure for some of the ESMs, and the increased uncertainty in the estimation of GP parameters as compared to that in the estimation of IID or AR(1) parameters.

[2]Additionally, the existence of an analytic inverse for the covariance of an AR(1) process makes it faster to compute with as compared to with a GP.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] G. Flato, J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. J. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, *et al.*, "Evaluation of climate models. in: Climate change 2013: The physical science basis. contribution of working group i to the fifth assessment report of the intergovernmental panel on climate change," *Climate Change 2013*, vol. 5, pp. 741–866, 2013.

[2] J. Mitchell, S. Manabe, V. Meleshko, and T. Tokioka, "Equilibrium climate change and its implications for the future," 1990.

[3] G. H. Roe and M. B. Baker, "Why is climate sensitivity so unpredictable?," *Science*, vol. 318, no. 5850, pp. 629–632, 2007.

[4] M. I. Hoffert, A. J. Callegari, and C.-T. Hsieh, "The role of deep sea heat storage in the secular response to climatic forcing," *Journal of Geophysical Research: Oceans*, vol. 85, no. C11, pp. 6667–6679, 1980.

[5] I. M. Held, M. Winton, K. Takahashi, T. Delworth, F. Zeng, and G. K. Vallis, "Probing the fast and slow components of global warming by returning abruptly to preindustrial forcing," *Journal of Climate*, vol. 23, no. 9, pp. 2418–2427, 2010.

[6] O. Geoffroy, D. Saint-Martin, D. J. Olivié, A. Voldoire, G. Bellon, and S. Tytéca, "Transient climate response in a two-layer energy-balance model. part i: Analytical solution and parameter calibration using cmip5 aogcm experiments," *Journal of Climate*, vol. 26, no. 6, pp. 1841–1857, 2013.

[7] K. Hasselmann, "Stochastic climate models part i. theory," *Tellus*, vol. 28, no. 6, pp. 473–485, 1976.

6th International Workshop on Climate Informatics
September 22-23, 2016
CI 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

COMBINING 15 YEARS OF MICROWAVE SST AND ALONG-TRACK SSH TO ESTIMATE OCEAN SURFACE CURRENTS

Pierre Tandeo[1], Aitor Atencia[2], Cristina Gonzalez-Haro[1]

***Abstract*—Ocean surface current is one of the main oceanographic variables. To estimate and track these currents, we use satellite measurements of Sea Surface Height (SSH), but these data are sparse in space and time, as they are collected along altimeter tracks. However, Sea Surface Temperature (SST) observations are much more complete in both space and time, and so the covariance of SST and SSH can be exploited to use SST datasets to help fill in the missing information about ocean currents where SSH data are lacking. Here, we test a new data-driven methodology combining SST and SSH information to estimate the ocean surface currents in the Agulhas current.**

I. MOTIVATION

The goal of this study is to estimate the geostrophic currents at the surface of the ocean using remote sensing data. In practice, to estimate those currents, we use satellite measurements of along-track altimetry to retrieve the Sea Surface Height (SSH) above geoid. We interpolate these along-track measurements using an optimal interpolation procedure, taking into account near past and future along-track data (typically 1 week before and after the current day). Finally, the geostrophic surface currents are estimated by calculating the spatial derivative of the SSH interpolated fields. Resulting currents are given in Fig. 1.

The problem is that along-track SSH measurements are very sparse in space and time. To improve the estimation of the surface currents, several works use the information of proxy variables such as Sea Surface Temperature (SST), salinity or ocean color also provided by satellite sensors. The advantage of satellite tracers like SST is that they have a better spatial and temporal resolution (typically, 1 to 25 km and hourly to daily). We distinguish different strategies to estimate the surface currents: (i) optical flow techniques between consecutive tracer images ([1]), (ii) physical properties to relate tracer information and displacements ([2]), (iii) statistical correlation between local information of tracers and displacements ([3]).

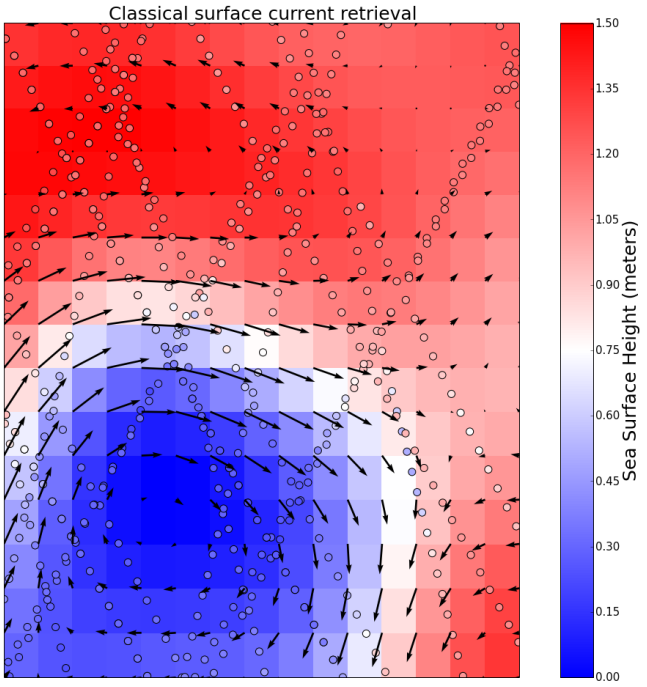


Fig. 1. Classical surface current retrieval from near past and future along-track SSH measurements, the 20^{th} of February 2004 in the Agulhas current (42°E to 46°E and 36°S to 40°S).

Here, we demonstrate the use of a new data-driven approach to exploit the synergy (i.e., the spatial consistency) between satellite measurements of microwave SST fields and along-track SSH measurements. The idea is to apply the analog method (also known as the nearest neighbor search, see [4]) on a large collocated SST/SSH database to artificially create pseudo-observations of along-track SSH from abundant SST. Details of this new methodology are given below.

Corresponding author: P Tandeo, pierre.tandeo@telecom-bretagne.eu [1]Telecom Bretagne, Brest, France [2]McGill, Montreal, Canada

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

II. REMOTE SENSING DATA

As SST fields, we use optimally interpolated microwave satellite data provided by Remote Sensing System (RSS), available online at http://www.ssmi.com//. It combines the signal of three microwave radiometers (Tropical rainfall measuring missions Microwave Imager, Advanced Microwave Scanning Radiometer Earth observing system, and WindSAT) which are robust to the presence of clouds. The spatial resolution is 1/4° × 1/4° and the temporal resolution is daily.

As along-track SSH, we use the Ssalto/Duacs Absolute Dynamic Topography provided by AVISO, available online at http://www.aviso.altimetry.fr/. It provides homogeneous observations from several altimetric missions (including for instance, Topex/Poseidon, Jason, Envisat). The spatial and temporal coverage of altimeters is very sparse compared to the microwave SST measurements.

In this study, we build a collocated database of SST and SSH between 1998 and 2014 in the Agulhas current. For each daily SST snapshot, we associate the corresponding along-track SSH measurements recorded in the region of interest the same day. We test our methodology on 2004 where the maximum of altimeters were available for the period 1998-2014 (see [5]).

III. METHOD AND PRELIMINARY RESULTS

The method is briefly schematized in Fig. 2. From a microwave SST image at a given time, we find analog situations in our SST database. They correspond to the best correlated situations in term of amplitude and spatial gradients. For instance, we identify $k = 8$ nearest situations from 1998 to 2014. For each of them, we extract the corresponding collocated along-track SSH measurements. Then, we aggregate all these selected along-track data to generate pseudo-observations of SSH. Finally, as in Fig. 1, we spatially interpolate these SSH data to retrieve the geostrophic surface currents.

Now, we propose to reconstruct the surface currents the 1^{st} of September 2004 in the Agulhas current. This period has been selected due to the high correlation between SST and SSH at the end of the austral winter (see [6] and [7]). We perform the experiment dividing the problem in $1° \times 1°$ small regions in order to capture the mesoscale dynamics of the current. For each snapshot of SST, we extract $k = 25$ analog situations with their corresponding collocated along-track SSH measurements. The surface currents are obtained by interpolation of all these pseudo-observations as shown in Fig. 3(a). Result of the classical interpolation and surface current retrieval from true available along-track SSH is given in Fig. 3(b).

Fig. 2. From a single snapshot of observed SST the 20^{th} of February 2004 in the Agulhas current (42°E to 46°E and 36°S to 40°S), we identify $k = 8$ analog situations of SST with their corresponding along-track SSH measurements, we finally interpolate them to retrieve the surface currents.

The results indicate an overall good agreement between the two interpolation methods. The 15 years of historical data seems large enough to encompass most of the prominent dynamics of the system. To validate the procedure, we also compare to the surface current climatology, corresponding to the mean of the 15 years of altimetric data for the month of September, given in Fig. 3(c). When comparing to the true SSH observations, the root mean squared error is reduced from 0.18 meters for the climatology to 0.12 for the proposed method. These encouraging results open new possibilities of using collocated SST/SSH information to improve the surface currents retrieval or replace the along-track SSH measurements when they are not available. This method could be particularly useful to estimate the mesoscale circulation in large current systems during the period 1981-1997 where only SST measurements were available.

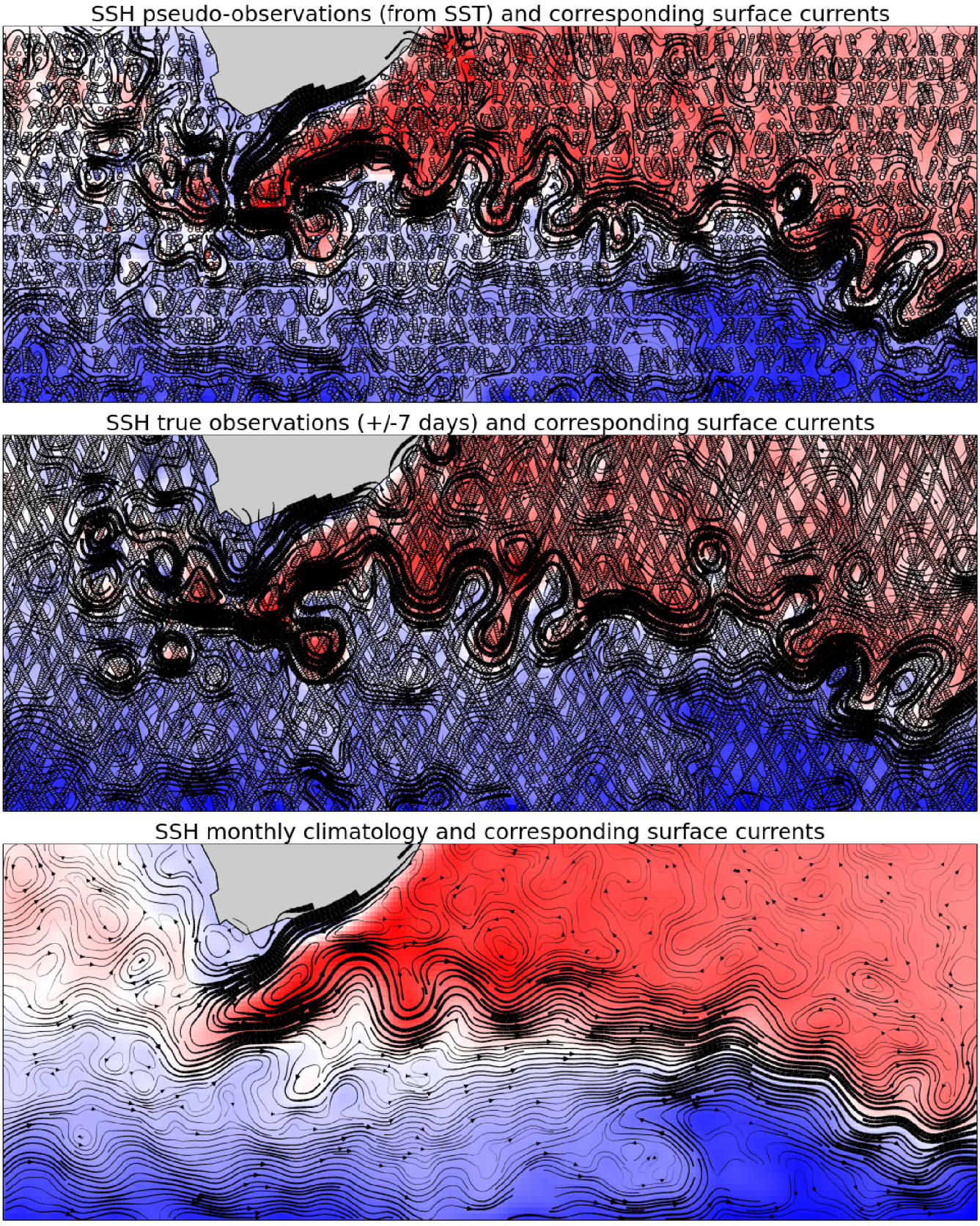


Fig. 3. (a) Proposed surface current retrieval from combination of historical SST/SSH data, (b) classical surface current retrieval from near past and future along-track SSH measurements and (c) monthly mean surface current climatology, the 1^{st} of September 2004 in the Agulhas current.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] M. M. Bowen, W. J. Emery, J. L. Wilkin, P. C. Tildesley, I. J. Barton, and R. Knewtson, "Extracting multiyear surface currents from sequential thermal imagery using the maximum cross-correlation technique," *Journal of Atmospheric and Oceanic Technology*, vol. 19, no. 10, pp. 1665–1676, 2002.

[2] J. Isern-Fontanet, B. Chapron, G. Lapeyre, and P. Klein, "Potential use of microwave sea surface temperatures for the estimation of ocean currents," *Geophysical research letters*, vol. 33, no. 24, 2006.

[3] P. Tandeo, B. Chapron, S. Ba, E. Autret, and R. Fablet, "Segmentation of mesoscale ocean surface dynamics using satellite sst and ssh observations," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 7, pp. 4227–4235, 2014.

[4] A. Atencia and I. Zawadzki, "A comparison of two techniques for generating nowcasting ensembles. part ii: Analogs selection and comparison of techniques," *Monthly Weather Review*, vol. 143, no. 7, pp. 2890–2908, 2015.

[5] A. Pascual, Y. Faugère, G. Larnicol, and P.-Y. Le Traon, "Improved description of the ocean mesoscale variability by combining four satellite altimeters," *Geophysical Research Letters*, vol. 33, no. 2, 2006.

[6] C. Le Goff, R. Fablet, P. Tandeo, E. Autret, and B. Chapron, "Spatio-temporal decomposition of satellite-derived sst-ssh fields: links between surface data and ocean interior dynamics in the agulhas region," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2016, in press.

[7] C. González-Haro and J. Isern-Fontanet, "Global ocean current reconstruction from altimetric and microwave sst measurements," *Journal of Geophysical Research: Oceans*, vol. 119, no. 6, pp. 3378–3391, 2014.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

PREDICTING EXECUTION TIME OF CLIMATE-DRIVEN ECOLOGICAL FORECASTING MODELS

Scott Farley[1] and John W. Williams[1,2]

***Abstract*—Species distribution models are climate-driven ecological forecasting tools that are widely used to predict species range shifts and ecological responses to 21st century climate change. As modern and fossil biodiversity databases improve and statistical methods become more computationally intensive, choosing the correct computing configuration on which to run these models becomes more important. We present a predictive model for estimating species distribution model execution time based on algorithm inputs and computing hardware. The model shows considerable predictive skill and can inform future resource provisioning strategies. We also demonstrate a technique for predicting model accuracy that suggests that inclusion of training data from the fossil record can enhance the accuracy of distribution models.**

I. INTRODUCTION

21st century climate change is expected to significantly alter species distributions, both at global and regional scales. Species Distribution Models (SDMs) are statistical methods that estimate species-specific responses to climatic gradients, and are widely used to to predict species presence under future climate scenarios [1]. While the models are widespread in the literature, a thorough understanding of algorithm execution time and accuracy produced by different input datasets and on different computing hardware has not yet been established. Here we discuss models for predicting the accuracy and run time of three SDMs given these factors. Execution time and accuracy models can improve computing resource utilization and identify performance bottlenecks in popular SDM code repositories.

SDMs can be fit with both modern- and paleo-climate training data, the scale, size, and resolution of which has increased rapidly over the last several years. Emerging databases, such as the Neotoma Paleoecological Database (http://neotomadb.org) and the Global Biodiversity Information Facility (GBIF, http://gbif.org), provide biogeographical data for millions of species worldwide, both in the recent fossil record and for the modern era. Environmental covariates to species presences are obtained from widely-available climate model output, which can provide decadal or sub-decadal temporal resolution for the last 21,000 years. Downscaling techniques can improve the spatial resolution of gridded data to scales suitable for regional and sub-regional study.

While the size and resolution of climatic and biodiversity data used to train SDMs increase, the methods used to learn species' responses to climatic gradients are becoming more computationally expensive. Most competitive SDMs use statistical learning procedures to estimate the functional relationship between species presence and climatic patterns. Novel techniques, such as Bayesian learning have also demonstrated high accuracy in this setting [2]. Moreover, many researchers now model hundreds of species in a single study (e.g., [3]), or use joint modeling techniques to capture inter-species interactions [4], resulting in larger modeling workloads. More powerful computing hardware has the potential to reduce the execution time of SDMs, particularly those with high dimensionality, large training sets, and/or wide spatiotemporal extents. While work has been done to assess the characteristic complexity of machine learning models (e.g. big-O notation) [5], less has been done to characterize the differences in model execution time of SDM techniques due to different computing hardware configurations and algorithm inputs. Though internal variations in memory management make it is difficult to exactly define model runtime as a function of hardware [6], models of computer performance that consider input data and static hardware configuration may be capable of capturing high-level trends [7], [8].

Here we model algorithm speed using two static

Corresponding author: S Farley, sfarley2@wisc.edu [1]University of Wisconsin-Madison, Department of Geography, Madison, WI 53706 [2]University of Wisconsin-Madison, Center for Climatic Research, Madison, WI 53706

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

hardware components capable of improving performance: (1) main memory size (i.e., RAM) and (2) the number computing cores, and two algorithm inputs: (1) the size of the training data used to fit the model and (2) the spatial resolution of the output. While different learning techniques may have implementation- or algorithm-specific differences (i.e., tuning parameters, language differences) that may influence model execution time, we test several popular R implementations with experimental variables that extend across model classes.

We also examine the predictive accuracy gains made by fitting SDMs with different training data sizes. Recent studies have examined the best practices for using small numbers of training examples ($n < 300$), such as for rare species [9]. However, while very large training datasets ($n > 100,000$) are unlikely in the ecological domain, the fossil record can be used to fit SDMs with a larger set of training data than the modern era alone [10], perhaps by several times for some species. While there is a greater degree of uncertainty associated with fossil occurrences, their utilization may significantly enhance SDM skill when included in the fitting process.

II. METHOD

We systematically tested the accuracy and execution time of three popular species distribution modeling algorithms on four different training set sizes and four spatial resolutions on 44 computing configurations (4 x CPU, 11 x RAM). All experiments were done using the R programming language on virtual machines hosted on the Google Cloud Compute platform. Ten replicates of each combination of hardware and algorithm inputs were completed to improve understanding system-induced variance.

Fossil occurrences for the *Picea* (spruce) genus over the last 21,000 years were obtained from the Neotoma Paleoecological Database (http://neotomadb.org). Decadally-averaged climatic covariates for each fossil occurrence were extracted from 0.5 degree spatial resolution debiased and downscaled CCSM3 climate model output for North America [11], and used to fit the SDMs.

Three SDM algorithms that have shown competitive predictive skill in the literature were evaluated: (a) boosted regression trees (GBM-BRT) [12], (b) multivariate adaptive regression splines (MARS) [13], (c) generalized additive models (GAM) [14]. All models were fit using a randomized training data subset of a pre-specified size, and then projected onto a climatic grid for the year 2100. The output grid resolution was

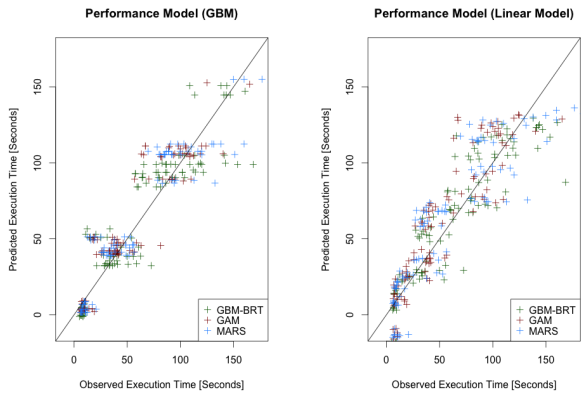


Fig. 1. Gradient Boosted Regression Tree (left) and Linear (Right) models of three different SDM modeling algorithms execution time under different algorithm and hardware parameters.

varied between 0.1 and 1 degree of latitude. SDM accuracy skill was evaluated using an independent testing set of 20% of the available data using the area under the receiver operator curve (AUC) statistic.

To predict model execution time, two predictive models were built for each SDM technique: (a) a multiple linear regression and (b) a gradient boosted regression tree model. Models were fit from the set of experiments ($n = 20,583$) using the number of training examples, CPU cores, memory, and spatial resolution as predictors of execution time. Model were evaluated using ANOVA and partial dependence plots, and skill was estimated using observed-to-predicted correlation metrics. Predictive models of SDM AUC score were also developed and fit using a boosted regression tree approach.

III. EVALUATION

Predictive models of SDM execution time demonstrate considerable skill when evaluated against a hold-out testing set of observed values. In general, the boosted regression tree model approach significantly outperformed the linear models. Regression trees are able better capture the potential non-linearities of the experimental dataset and can remove the negative predictions forecast by the linear model. However, both sets of models consistently showed $r^2 > 0.8$ correlation between observed and predicted values with a mean prediction error of less than 4 seconds.

Of all six execution time models (2 models x 3 SDMs), the regression tree prediction model of the MARS SDM performed the best, with a mean error of -0.457 ± 1.895 seconds and an r^2 value of 0.936. The regression tree models for GAM and GBM-BRT

SDMs both performed well with r^2 values of 0.892 and 0.880, respectively. The linear models all showed lower r^2 correlation values and had larger prediction variance and mean prediction errors than their decision tree counterparts. The best performing linear model was again for the MARS SDM, with an r^2 correlation of 0.876, with a significantly larger mean error of 2.17 ± 1.73 seconds. Figure 1 shows the observed and predicted values for each SDM for both of the prediction models.

Model interrogation using ANOVA (linear model) and partial dependency plots (GBM model) reveals that model execution time depends strongly on the number of training examples used to fit the SDM. In all cases, the number of training examples and spatial resolution of the output were shown to be highly significant ($p < 0.001$). Computer hardware variables were not shown to be significant predictors of execution time for these SDMs. In some cases, additional memory was shown to reduce model speed, perhaps due to increased overhead of memory management. Runtime logs indicate that model execution was bounded by CPU processing capability, rather than main memory capacity, suggesting that SDM workflows could be improved if the algorithms were written to run in parallel, rather than sequentially.

Models of SDM accuracy suggest that significant accuracy gains can be achieved by fitting the models with more than 2000 training examples. All three SDM algorithms showed a similar pattern of increasing accuracy as the number of input training examples increased, though the increase was not linear. Figure 2 demonstrates the accuracy of a GBM-BRT model with up to 9000 training training examples. The accuracy prediction model shows an observed-to-predicted r^2 of 0.900 and a mean prediction error of 0.001 ± 0.002 AUC. The model strongly suggests that use of the additional training data available in the fossil record can significantly enhance SDM accuracy.

Future work will be directed towards larger and more complex models of climate-species dynamics. Additional research should also investigate explicitly parallel machine learning techniques and their feasibility for SDM studies, as our results show that execution time is strongly limited by CPU-bound serial learning techniques.

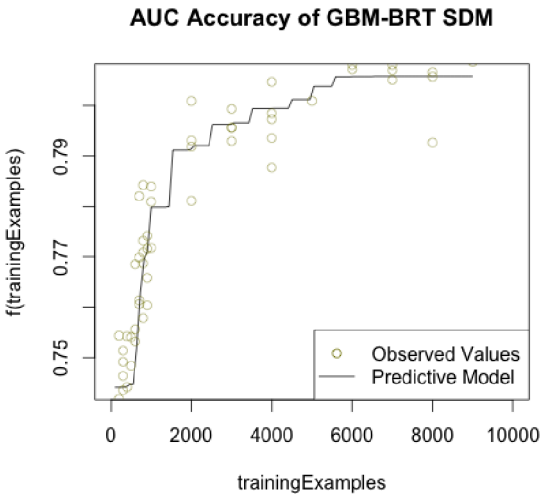


Fig. 2. Predictive model of GBM-BRT species distribution model achieved when using different input sizes.

ACKNOWLEDGMENTS

Funding for the authors was provided by University of Wisconsin-Madison Geography Department's Trewartha Research Award, the University of Wisconsin-Madison Vilas Research Trust, and the National Science Foundation (EAR-1550707).

REFERENCES

[1] J. Franklin, *Mapping species distributions: spatial inference and prediction*. Cambridge University Press, 2010.

[2] N. Golding and B. V. Purse, "Fast and flexible bayesian species distribution modelling using gaussian processes," *Methods in Ecology and Evolution*, vol. 7, no. 5, pp. 598–608, 2016.

[3] J. Elith, C. Graham, R. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J Hijmans, F. Huettmann, J. R Leathwick, A. Lehmann, *et al.*, "Novel methods improve prediction of species distributions from occurrence data," *Ecography*, vol. 29, no. 2, pp. 129–151, 2006.

[4] J. S. Clark, A. E. Gelfand, C. W. Woodall, and K. Zhu, "More than the sum of the parts: forest climate response from joint species distribution models," *Ecological Applications*, vol. 24, no. 5, pp. 990–999, 2014.

[5] T. Hastie, J. Friedman, and R. Tibshirani, "Additive models, trees, and related methods," in *The Elements of Statistical Learning*, pp. 257–298, Springer, 2001.

[6] D. J. Lilja, *Measuring computer performance: a practitioner's guide*. Cambridge university press, 2005.

[7] Q. Wu and V. V. Datla, "On performance modeling and prediction in support of scientific workflow optimization," in *2011 IEEE World Congress on Services*, pp. 161–168, IEEE, 2011.

[8] B. C. Lee, D. M. Brooks, B. R. de Supinski, M. Schulz, K. Singh, and S. A. McKee, "Methods of inference and learning for performance modeling of parallel applications," in *Proceedings of the 12th ACM SIGPLAN symposium on Principles and practice of parallel programming*, pp. 249–258, ACM, 2007.

[9] M. S. Wisz, R. Hijmans, J. Li, A. T. Peterson, C. Graham, and A. Guisan, "Effects of sample size on the performance of species distribution models," *Diversity and Distributions*, vol. 14, no. 5, pp. 763–773, 2008.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

[10] K. C. Maguire, D. Nieto-Lugilde, M. C. Fitzpatrick, J. W. Williams, and J. L. Blois, "Modeling species and community responses to past, present, and future episodes of climatic and ecological change," *Annual Review of Ecology, Evolution, and Systematics*, vol. 46, pp. 343–368, 2015.

[11] D. Lorenz, D. Nieto-Lugilde, J. Blois, M. Fitzpatrick, and J. Williams, "Data from: Downscaled and debiased climate simulations for north america from 21,000 years ago to 2100ad," 2016.

[12] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.

[13] J. Leathwick, J. Elith, and T. Hastie, "Comparative performance of generalized additive models and multivariate adaptive regression splines for statistical modelling of species distributions," *Ecological modelling*, vol. 199, no. 2, pp. 188–196, 2006.

[14] A. Guisan, T. C. Edwards, and T. Hastie, "Generalized linear and generalized additive models in studies of species distributions: setting the scene," *Ecological modelling*, vol. 157, no. 2, pp. 89–100, 2002.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

SPATIOTEMPORAL ANALYSIS OF SEASONAL PRECIPITATION OVER US USING CO-CLUSTERING

Mohammad Gorji–Sefidmazgi[1], Clayton T. Morrison[1]

***Abstract*—Clustering techniques are commonly used in climate science for clustering in time or space. In this study, we apply a co-clustering technique for simultaneous analysis of spatial and temporal patterns in precipitation data collected at stations across the continental US. This approach reveals clusters of stations with similar precipitation patterns and also clusters of time where the values of precipitation are similar relative to location.**

I. MOTIVATION

Detecting regions of homogeneous precipitation patterns is essential for modeling the hydrological cycle and the design and management of water resource systems. A common requirement by these tasks is the need to group measurements into meaningful categories based on historical observations. However, patterns of precipitation change over time due to internal climate dynamics and external forces [1]. In fact, the time and space domains of the climate system are highly correlated with each other. It is therefore desirable to find spatial groupings while taking into account related temporal structure of precipitation measurements.

Clustering techniques have been developed to find groups of elements that are similar to each other, but dissimilar to the elements of other groups. Current clustering techniques used in climate science are typically only used within time [2], [3] or space domains [4], [5], [6], but not both simultaneously. In time-clustering techniques, segments of time are detected where the values of their time series are similar to each other. Spatial-clustering approaches use the average (or trend) of measurements in each station to identify the stations with similar measurement records. Finally, some spatiotemporal approaches first analyze the temporal domain (e.g., at each station) and then analyze the spatial domain (e.g., use GIS data to find spatial patterns in climate data) [7], [8].

Corresponding author: M. Gorji-Sefidmazgi,
Emails: {mgorjise,claytonm}@email.arizona.edu,
[1]School of Information, The University of Arizona

Co-clustering is a variant of clustering that introduces the ability to simultaneously cluster along two variables. Here we seek to simultaneously cluster stations while also clustering precipitation values over time, producing co-clusters as subsets of stations that have similar precipitation patterns over time. By co-clustering, it is not necessary to calculate the average (or trend) of precipitation in each station before spatial clustering. Instead, it is possible to detect clusters of stations with similar precipitation temporal patterns. Additionally for each cluster of stations, we will find the clusters of time where the values of precipitations are similar to each other.

In this paper, we present the results of applying a co-clustering algorithm to monthly precipitation data of 1218 stations in the continental United States. Our results suggest the method is capable of identifying interesting spatial patterns in seasonal precipitation.

II. DATA

The dataset used in this study is the bias-adjusted monthly precipitation time series over 1218 stations within the continental US derived from *United States Historical Climatology Network (USHCN)* database [9]. This dataset can be downloaded from http://cdiac.ornl.gov/epubs/ndp/ushcn/ushcn.html. The precipitation amounts are in hundredths of inches (e.g., 468 indicates 4.68”). In this analysis, we need data for the same time period for each station. We selected the 64-year time period of 1951 to 2014 and calculated the average precipitation across each month. We then partitioned the data by season (January–February–March (JFM), April–May–June (AMJ), July–August–September (JAS) and October–November–December (OND)). This results in 4 data sets, one for each season, each represented by a matrix of 1218 rows (for each station) by 192 columns (for the three months for the season, across 64 years).

III. METHOD

In this work, use the *Spectral Co-clustering* algorithm, originally developed for clustering of gene ex-

pression levels in thousands of genes [10]. This method has efficient performance, suitable for working with large matrices.

Let a dataset of samples be represented along two variable dimensions, M and N (each of m and n possible variable states, respectively), as a matrix $A = (a_{ij})_{m\times n}$. Each value in the matrix, a_{ij}, is indexed by the i-th state of variable M and j-th state of variable N. Co-clustering then involves simultaneously partitioning the m states of variable M into K distinct row clusters, and partitioning the n states of variable N into L distinct column clusters.

Suppose that the matrix A is co-clustered into $(K \times L)$ sub-matrices. Let $P_{(k,l)}$ be the sub-matrix of k-th row cluster and l-th column cluster, and denote the average of all values of this sub-matrix as $\mu_{(k,l)}$. The quality of this co-cluster is defined as:

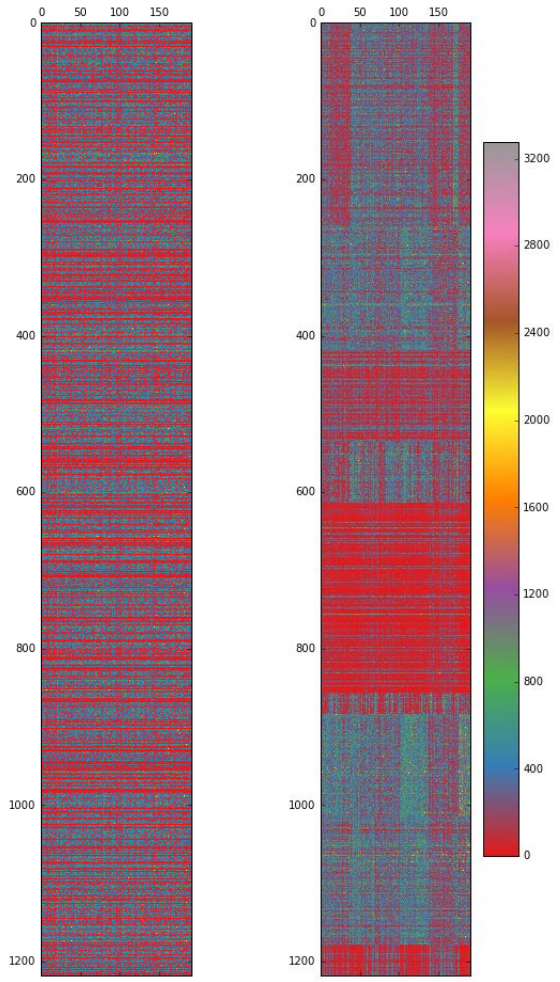
$$e_{(k,l)} = \sum_{a_{ij}\in P_{(k,l)}} [a_{ij} - \mu_{(k,l)}]^2. \quad (1)$$

We then define the error, E, of a co-clustering as the sum of $e_{(k,l)}$ for all of $(K \times L)$ sub-matrices. The goal of a co-clustering algorithm is to find a partition that minimize E. However, note that a co-clustering solution with single-row and single-column sub-matrices is a trivial solution with $E = 0$. Thus, it is necessary to fix the values of K and L and then determine the members of those clusters that minimize E.

The problem remains as to how to choose K and L. This is still an open model-selection problem, but here we adopt the method proposed in [11]. First, note that our error score E will decrease as we increase K and L no matter how we associate the values in the cells of matrix A with rows and columns. That is, this will be true if we choose any random matrix $\hat{A}$ whose individual elements are the same as A but whose locations in the matrix are randomized with respect to rows and columns. We search for K and L such that E for the clustering over A is as low as possible but not lower than the expected value of E across random $\hat{A}$'s. To incorporate this constraint, we perform a grid search over different values of (K, L) for A and along with R randomly sorted matrices $\hat{A}_r$, and seek to minimize:

$$\sum_{i=1}^{N}(E - \mathbb{E}\{\hat{E}_r\})^2 \quad (2)$$

For spatiotemporal analysis of precipitation, we seek row clusters that represent sets of stations that record similar precipitation and the column clusters that represent the periods of time where the values of precipitations are similar to each other. We applied the spectral co-clustering approach to the four matrices of seasonal precipitation for $K \in \{2, \ldots, 15\}$ and $L \in \{2, \ldots, 10\}$. For each (K, L), the algorithm was also applied to 100 randomly row- and column-permuted versions of each matrix to calculate the cost of Equation 2. In this way, we can find the optimal values of (K, L) for each of four matrices.



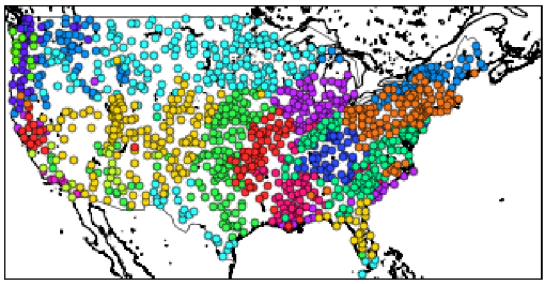
(a) (b)

Fig. 1: (a) Heatmap of summer precipitation for 1218 stations and 192 time points, prior to clustering. (b) Heatmap after permuting the rows and columns to group by clusters, with optimal values of $K = 9$ and $L = 7$. (Note that the row and column indices for (b) do not correspond to those in (a).) The checkerboard structure of the sorted matrix (b) can be seen across the clusters.

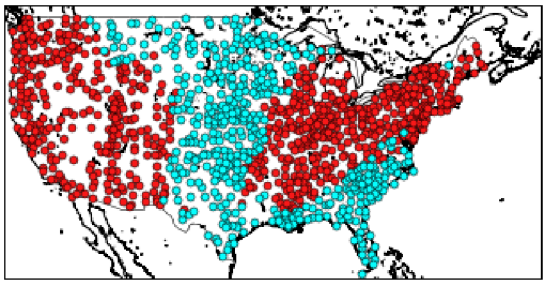
IV. RESULTS

Heatmaps are a common visualization tools for representing the quality of co-clustering. The color of

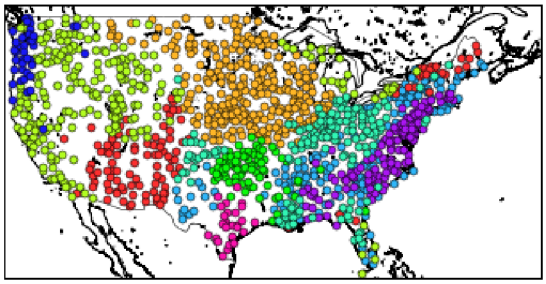
6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO



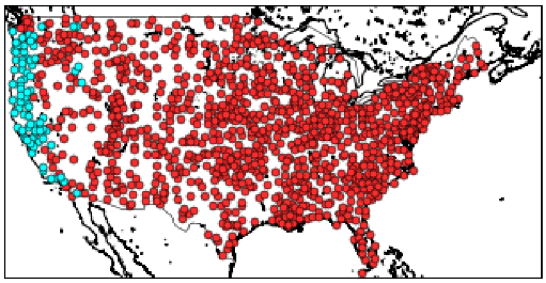
(a) JFM



(b) AMJ



(c) JAS



(d) OND

Fig. 2: Space clusters in seasonal precipitation over US 1951-2014. (a) Winter $K = 14$, $L = 6$. (b) Spring $K = 2$, $L = 2$. (c) Summer $K = 9$, $L = 7$. (d) Autumn $K = 2$, $L = 2$.

pixels of a rectangular grid represent the magnitude of the a_{ij} values. After the co-clustering algorithm is used to cluster row and column indices, the rows and columns are reordered to group by cluster. In general, a co-clustering is qualitatively more successful to the extent that the reordering by clusters results in sub-

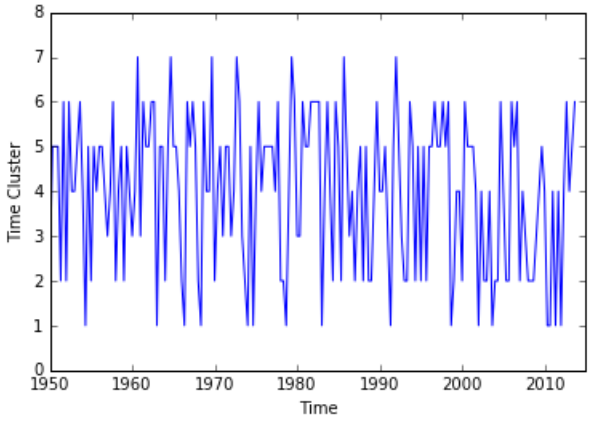


Fig. 3: Indices of 7 time-clusters in a station located Eureka, CA. Each Time Cluster (indexed along the y-axis by numbers $1 - 7$) has a different pattern of precipitation. Note that the Time Cluster indices are categorical indices for naming the cluster, but here we have also ordered the indices so that they correspond to the order of the precipitation values (i.e., index 1 is the lowest precipitation mean (86), index 2 the next lowest (91), and so on up to index 7, which has the highest precipitation mean (162).)

matrices whose within-cluster values are more similar than between clusters, resulting in a "checker board" pattern [12]. Figure 1 shows the results of applying the spectral co-clustering algorithm to the summer precipitation data.

Figure 2 shows the spatial patterns of precipitation across the four seasons. It can be seen that there are higher variabilities for precipitations of summer and winter. Also, the spatial patterns of summer and winter are similar to each other, although variability increases toward the east coast during winter, with additional clusters.

For each one of these spatial clusters there are also associated temporal clusters. These are periods of time where the values of precipitation are similar to each other. As an example, Figure 3 shows the indices of detected temporal clusters of precipitation for the station located at Eureka, CA. Each of these 7 time clusters represents a pattern of precipitation. For the Eureka, CA station, the average of precipitation in these 7 time clusters are $\{86, 91, 99, 100, 117, 136, 162\}$.

ACKNOWLEDGMENTS

This work was supported in part by the DARPA Big Mechanism program under ARO contract W911NF-14-1-0395.

REFERENCES

[1] P. C. D. Milly, J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer, "Stationarity Is Dead: Whither Water Management?," *Science*, vol. 319, no. 5863, pp. 573–574, 2008.

[2] M. Gorji Sefidmazgi, M. Sayemuzzaman, A. Homaifar, M. K. Jha, and S. Liess, "Trend analysis using non-stationary time series clustering based on the finite element method," *Nonlinear Processes in Geophysics*, vol. 21, no. 3, pp. 605–615, 2014.

[3] H. Wan, X. Zhang, F. Zwiers, and S.-K. Min, "Attributing northern high-latitude precipitation change over the period 1966–2005 to human influence," *Climate Dynamics*, vol. 45, no. 7, pp. 1713–1726, 2015.

[4] A. Modaresi Rad and D. Khalili, "Appropriateness of Clustered Raingauge Stations for Spatio-Temporal Meteorological Drought Applications," *Water Resources Management*, vol. 29, no. 11, pp. 4157–4171, 2015.

[5] Y. Zhang, S. Moges, and P. Block, "Optimal Cluster Analysis for Objective Regionalization of Seasonal Precipitation in Regions of High Spatial–Temporal Variability: Application to Western Ethiopia," *Journal of Climate*, vol. 29, no. 10, pp. 3697–3717, 2016.

[6] X. Huang, V. Lyubchich, A. Brenning, and Y. R. Gel, "Analysis of dynamic trend-based clustering on Central Germany precipitation," in *Fifth International Workshop on Climate Informatics*, (Boulder), 2015.

[7] S. Irwin, R. K. Srivastav, S. P. Simonovic, and D. H. Burn, "Delineation of Precipitation Regions using Location and Atmospheric Variables in Two Canadian Climate Regions: The role of attribute selection," *Hydrological Sciences Journal*, vol. 0.

[8] M. R. Rahman and H. Lateh, "Spatio-temporal analysis of warming in Bangladesh using recent observed temperature data and GIS," *Climate Dynamics*, vol. 46, no. 9, pp. 2943–2960, 2016.

[9] M. J. Menne, C. N. Williams, and R. S. Vose, "The U.S. Historical Climatology Network Monthly Temperature Data, Version 2," *Bulletin of the American Meteorological Society*, vol. 90, no. 7, pp. 993–1007, 2009.

[10] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein, "Spectral biclustering of microarray data: Coclustering genes and conditions," 2003.

[11] B. Percha and R. B. Altman, "Learning the Structure of Biomedical Relationships from Unstructured Text.," *PLoS computational biology*, vol. 11, no. 7, p. e1004216, 2015.

[12] S. Busygin, O. Prokopyev, and P. M. Pardalos, "Biclustering in data mining," *Computers and Operations Research*, vol. 35, no. 9, pp. 2964–2987, 2008.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

PREDICTION OF EXTREME RAINFALL USING HYBRID CONVOLUTIONAL-LONG SHORT TERM MEMORY NETWORKS

Sulagna Gope[1], Sudeshna Sarkar[1], Pabitra Mitra[1]

***Abstract*—A hybrid convolutional-Long Short Term Memory (LSTM) neural network model has been proposed as a data driven model to predict the amount of rainfall. The convolutional model takes care of the spatial features responsible for rainfall and the temporal variation in the features leading to rainfall is captured by the LSTM model. We demonstrate the effectiveness of the model in rainfall prediction tasks in India. Our model has been found to work better when compared with some of the standard machine learning models that has been used for the task.**

I. MOTIVATION

Rainfall is a very important natural phenomena which greatly affects the life and livelihood of people in places like India. Though heavy rainfall creates a lot of hazards and losses, little or no rainfall affects cultivation, dries up rivers and lakes, leading to a shortage of food and fresh water. Thus rainfall forecasting remains an important problem. Accurate rainfall prediction in complex climatic regions like India has been a difficult task. A lot of effort has been made to identify the true factors responsible for it, however the uncertainty in prediction is still persistent [1][2]. This is because the true mechanism of rainfall is not thoroughly known. Different atmospheric factors interact to bring about rainfall at a certain place at a certain time. This is a very complex mechanism which has not yet been solved with the known equations. Besides, accurate rainfall prediction models need to capture the local relief and local phenomena. Numerical weather prediction (NWP) models, which are mathematical models based on the laws of physics, atmospheric science and fluid motion, have failed to predict rainfall accurately in certain regions of the world. Statistical and machine learning based models have been used to overcome the shortcomings of the NWP models. Their advantage is that they are portable and can be learned for different regions of the world. Among statistical models analog methods are quite popular for precipitation forecasting [3]. When the analog methods are compared with deterministic models like MM5 [4] it is found that the deterministic models overestimates the rainfall whereas the analog models underestimate it. This calls for further improvement of the technique. Some work have been carried out on precipitation prediction in India such as [5], [6], most of which have tried to relate extreme rainfall with anomalous weather behavior. Though these models could predict rainfall in general, they have failed to predict in advance extreme rainfall events.

Recently many advanced machine learning techniques like deep learning have been found to work well in solving complex tasks. Nayak et al.[7] used support vector machines to predict extreme or non-extreme rainfall events. JNK Liu [8] developed a deep neural network model to predict temperature, dew point, mean sea level pressure and wind speed in the next few hours. In [9], [10] deep learning models have used for different weather prediction tasks. Deep learning models have been found to work very well in complex tasks. For instance convolutional neural network [11] beats the state of the art methods in image recognition tasks. Convolutional neural network (CNN) works very well on spatial data where localized features play an important role. On the other hand recurrent neural network (RNN) models [12] are effective in dealing with temporal data and in various sequence generation tasks. RNN has been modified to use new architectures like LSTM [13] which overcome certain limitations of RNN.

In this paper we have built a model for rainfall prediction where the spatial aspect of weather parameters is taken care of by a CNN model and the variation in the behavior of the features over time is captured by a LSTM model.

Corresponding author: S Gope, sulagna.student12@gmail.com
[1]Indian Institute of Technology Kharagpur, India

II. DATA

Atmospheric features over the entire Indian subcontinent have been used as predictors for rainfall. These features include temperature, mean sea level pressure, precipitable water, relative humidity, U-wind and V-wind at the surface level. Atmospheric parameters like temperature, vertical wind velocity (omega), relative humidity, u-wind and v-wind are also considered at the 850-, 600-, 400-hPa pressure levels. All these parameters have been found to be important factors responsible for rainfall and are generally used for rainfall prediction task. The data is collected from the National Centers for Environmental Prediction/National Center for Atmospheric Research (NCEP/NCAR) reanalysis data for the Indian subcontinent ranging from 5 degrees to 40 degrees north latitude and 65 degrees to 100 degrees east longitude. The extreme rainfall data has been obtained for Mumbai from Indian Meteorology Department (IMD). The rest of the rainfall data required for the experiment has been obtained from APHRODITE. For training the model, we have used the data from 1969-1999 for the four Indian Summer Monsoon months. The data from 2000-2008 has been used for testing. Though the model is used for short-term prediction only, still a huge amount of historic data is needed for training, so that the model can learn the underlying physical process solely from the data, without any extra domain knowledge. This enables the model to make better prediction.

III. METHOD

In this work we have first tried to identify the most significant spatial features from the large set of input features using a convolutional neural network or CNN [11]. The significant features identified are then fed to a LSTM model to solve the rainfall regression problem.

We have first constructed a multi-layered CNN for a classification task. There are about 21 weather features over a $15 * 15$ grid spatial region, where each grid has a dimension of $2.5^\circ * 2.5^\circ$. Our CNN architecture consists of three convolutional layer with two pooling layers in between and finally two simple neural network layers. We have performed a number of experiments with different number of layers and found that the above mentioned configuration gives the best result. Each convolutional layer transforms one set of feature maps to another set by convolution with a set of filters. The feature map h^k for a given layer with filters having weight W^k and bias b^k is given by

$$h_{ij}^k = tanh((W^k * x)_{ij} + b^k) \quad (1)$$

where $*$ is the convolution operator and x is the input from the previous layer. The feature maps are subsampled using max pooling. The final feature map is then connected to a sparse neural network layer, which is then connected to a fully connected NN layer. The fully connected layer is mapped to the output. In our case the output for the CNN is binary, indicating extreme or non-extreme rainfall.In [7], rainfall is classified as extreme or non-extreme event based on a threshold amount which is dependent on the region considered. For example, Mumbai receives very heavy rain every year during monsoon, thus the threshold for Mumbai is taken as 75mm (which is quite high) of accumulated rain. The same has been done in this work.

The features of each day is fed to the CNN model. The feature map obtained from the last convolutional net layer is extracted for each day. The new feature maps for the past three days is then fed to a single layer LSTM model. A block diagram of the hybrid CNN-LSTM model is shown in Fig. 1. We have also

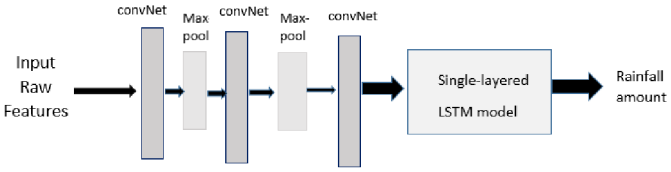


Fig. 1. Block diagram of our approach

used AFM for feature extraction followed by lstm for rainfall estimation. The AFM method finds out the features which consistently show anomalous behavior corresponding to extreme rainfall events. Only those features, which have shown anomalous behaviour for most of the extreme events, are used out of the full input feature set is used as input to the lstm for rainfall estimation.

LSTM is a memory based neural network which is similar to RNN with a few gates. The LSTM architecture includes a input gate, forget gate, output gate and a cell which helps it to selectively retain useful past information and forget the rest of the unnecessary information. The equations of the gates are as follows:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \quad (2)$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \quad (3)$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \quad (4)$$

$$c_t = f_t * c_{t-1} + i_t * \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (5)$$

$$h_t = o_t * \tanh(c_t) \quad (6)$$

The input i_t, output o_t, forget f_t gates and the cell c_t follow the above equations with different parameters

and uses sigmoid activation function. They receive the present input and the previous hidden state as input and generate an output just like a RNN. The memory cell c_t is responsible for combining the previous memory with the newly generated hidden state. Finally the current hidden state h_t is calculated using the cell state and the output from the output gate. In our model the feature maps of the past few days are fed as input and the amount of rainfall for the next 3 days are obtained. The model is trained using stochastic gradient descent.

IV. EVALUATION

Our model can predict rainfall in a particular region based on the weather parameters of that region and its surroundings. The model performs well in predicting the total rainfall that will take place in the next 1, 2 and 3 days based on all the weather parameters of the previous three to five days. We have tested our model for predicting rainfall during the Indian Summer Monsoon, over Mumbai, India, since every year this region receives high rainfall. The 21 weather features observed over the 2-dimensional space spread over the Indian sub-continent has been fed to the convolutional network. We have used daily observations of the features as input. The feature map of the last convolutional layer is used as input to the LSTM network. Our hybrid model has been compared to a CNN, a hybrid stacked-autoencoder followed by LSTM (SAE+LSTM) model and anomaly frequency method [7] of feature extraction followed by LSTM (AFM+LSTM). The plots of the true rainfall and the predicted rainfall using AFM-LSTM and CNN-LSTM hybrid model is shown in Fig.2 and Fig.3 respectively. In the figures, the red line indicates the predicted output and blue line indicates the true output. Due to lack of space we have only included the plots of results for prediction of rainfall with 1 day lag.

The performance of the different methods used has been shown in the Table I in the form of mean square error and symmetric mean absolute percentage error (SMAPE) which is calculated as follows:

$$\frac{1}{n}\sum_{1}^{n}\frac{|F_t - A_t|}{|F_t| + |A_t|} * 100 \qquad (7)$$

where F_t is the predicted value and A_t is the true value. The mean square error (mse) does not give a good measure of curve fitting. We thus use the SMAPE measure which gives a better indication of fitting. Since the rainfall values are normalized between 0 and 1, the mean square error is low in all the cases. The AFM followed by LSTM and the CNN followed by

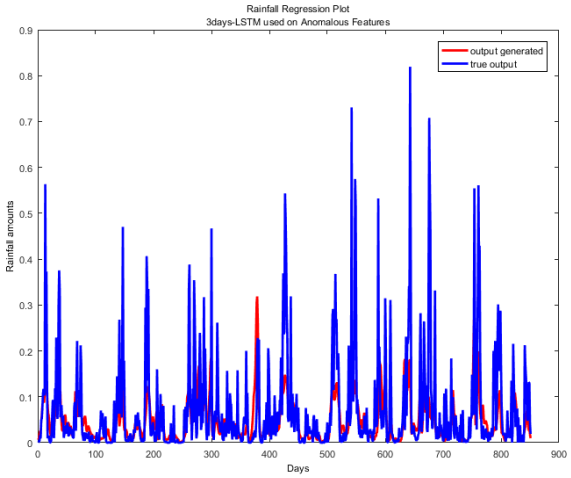


Fig. 2. Plot of true and predicted rainfall using AFM-LSTM model with 1 day lag

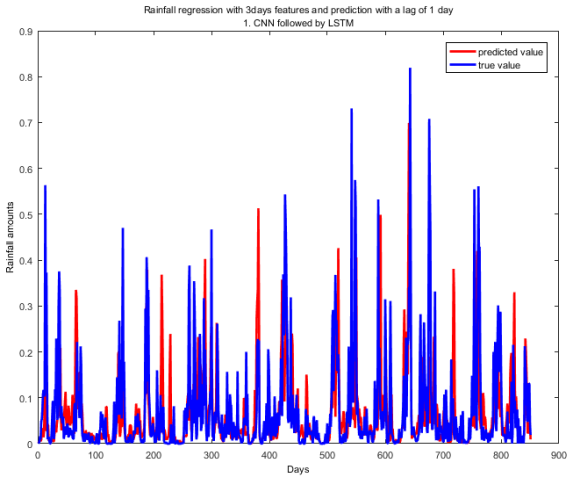


Fig. 3. Plot of true and predicted rainfall using CNN-LSTM model with 1 day lag

TABLE I
COMPARISON OF ERRORS FOR REGRESSION PLOTS OF DIFFERENT METHODS

	1 Day lag		2 Days lag		3 Days lag	
Method	mse	SMAPE	mse	SMAPE	mse	SMAPE
AFM+SVR	0.063	75.50	0.069	77.90	0.08	86.01
SAE + LSTM	0.011	50.66	0.017	51.55	0.029	54.00
CNN	0.016	83.00	0.021	83.571	0.022	84.01
AFM + LSTM	**0.008**	**42.14**	**0.009**	**42.92**	**0.010**	**46.73**
CNN + LSTM	**0.01**	**42.48**	**0.009**	**44.55**	**0.009**	**44.67**

LSTM gives the least mse and SMAPE. Low value of SMAPE indicates a better fit. AFM+LSTM method shows a good fit for low and medium rainfall and has less false positives. This is probably the reason why SMAPE has lower value for AFM+LSTM method than CNN+LSTM. However for high rainfall cases CNN+LSTM performs better. Thus both the models can be implemented to get an overall idea of the amount of rainfall that may occur. Also we find that the SMAPE of AFM+LSTM exceeds that of CNN+LSTM as the time lag increases which indicates that CNN+LSTM model is capable of predicting rainfall in much advance. Fig 3 also shows a few false positive cases, where our model predicts high rainfall but actually there is much less rain. We would like to improve our method, concentrating more on the feature selection part to get better results. We have tried the same experiment using an end-to-end model, however we did not get good result with it. In future we would like to make a better end-to-end model, with proper parameter tuning to address the problem of rainfall prediction and other severe weather events in advance.

ACKNOWLEDGEMENTS

This research was supported and funded by Indian Institute of Technology Kharagpur, India and MHRD, India, under the project named ”Feature Extraction and Data Mining from Climate Data (FAD)”.

REFERENCES

[1] S.-Y. Hong and J.-W. Lee, “Assessment of the wrf model in reproducing a flash-flood heavy rainfall event over korea,” *Atmospheric Research*, vol. 93, no. 4, pp. 818–831, 2009.

[2] R. Khaladkar, S. Narkhedkar, and P. Mahajan, *Performance of NCMRWF Models in Predicting High Rainfall Spells During SW Monsoon Season: A Study for Some Cases in July 2004*. Indian Institute of Tropical Meteorology, 2007.

[3] A. Ben Daoud, E. Sauquet, M. Lang, G. Bontron, and C. Obled, “Precipitation forecasting through an analog sorting technique: a comparative study,” *Advances in Geosciences*, vol. 29, no. 29, pp. 103–107, 2011.

[4] V. Altava-Ortiz, A. Barrera, M. Llasat, O. P. Prat, J. Gibergans-Báguena, and M. Barnolas, “Application of the MM5 and the analogous method to heavy rainfall event, the case of 16?18 October 2003 in Catalonia (NE Spain),” *Advances in Geosciences*, vol. 7, pp. 313–319, Apr. 2006.

[5] S. Roy Bhowmik and V. Durai, “Application of multimodel ensemble techniques for real time district level rainfall forecasts in short range time scale over indian region,” *Meteorology and Atmospheric Physics*, vol. 106, no. 1, pp. 19–35, 2010.

[6] A. Sahai, M. Soman, and V. Satyan, “All india summer monsoon rainfall prediction using an artificial neural network,” *Climate dynamics*, vol. 16, no. 4, pp. 291–302, 2000.

[7] M. A. Nayak and S. Ghosh, “Prediction of extreme rainfall event using weather pattern recognition and support vector machine classifier,” *Theoretical and applied climatology*, vol. 114, no. 3-4, pp. 583–603, 2013.

[8] J. N. Liu, Y. Hu, J. J. You, and P. W. Chan, “Deep neural network based feature representation for weather forecasting,” in *Proceedings on the International Conference on Artificial Intelligence (ICAI)*, p. 1, The Steering Committee of The World Congress in Computer Science, Computer Engineering and Applied Computing (WorldComp), 2014.

[9] M. M. Kordmahalleh, M. G. Sefidmazgi, A. Homaifar, and S. Liess, “Hurricane trajectory prediction via a sparse recurrent neural network,”

[10] C. Anderson, I. Ebert-Uphoff, Y. Deng, and M. Ryan, “Discovering spatial and temporal patterns in climate data using deep learning,” in *5th International Workshop on Climate Informatics, NCAR Mesa lab, Boulder, CO*, 2015.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[12] A. Graves, “Generating sequences with recurrent neural networks,” *arXiv preprint arXiv:1308.0850*, 2013.

[13] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

SPATIOTEMPORAL PATTERN EXTRACTION WITH DATA-DRIVEN KOOPMAN OPERATORS FOR CONVECTIVELY COUPLED EQUATORIAL WAVES

Joanna Slawinska[1], Dimitrios Giannakis[2]

***Abstract*—We study spatiotemporal patterns of convective organization using a recently developed technique for feature extraction and mode decomposition of spatiotemporal data generated by ergodic dynamical systems. The method relies on constructing low-dimensional representations (feature maps) of spatiotemporal signals using eigenfunctions of the Koopman operator governing the evolution of observables in ergodic dynamical systems. This operator is estimated from time-ordered data through a Galerkin scheme applied to basis functions computed via the diffusion maps algorithm. We apply this technique to brightness temperature data from the CLAUS archive and extract a multiscale hierarchy of spatiotemporal patterns on timescales spanning years to days. In particular, we detect for the first time without prefiltering the input data traveling waves on temporal and spatial scales characteristic of convectively coupled equatorial waves (CCEWs). We discuss the salient properties of waves in this hierarchy and find that the activity of certain types of CCEWs is modulated by lower-frequency signals.**

I. MOTIVATION

Clouds are omnipresent throughout the Earth's atmosphere. They constitute an important part of the climate system, having potentially large yet uncertain impact on climate change [1], [2], [3], [4], [5]. Convection occurs on relatively small and short scales, on the order of a few tens of kilometers and a few hours at most, but interacts strongly with other scales and is coupled with large-scale circulation and moisture [6]. Moreover, clouds organize themselves in a number of distinctive mesoscale and synoptic scale convective systems, oftenmost embedded within each other and propagating throughout the tropics [7], [8], [9], [6]. Despite the considerable economic impact of these systems (e.g., they provide the majority of precipitation in the tropics [10], [11]), their predictability with operational models is limited [12], [13], [14], [15], due to their poorly understood and multiscale nonlinear nature.

Convectively coupled equatorial waves (CCEWs, [16]) are typically categorized either as Kelvin, mixed Rossby-gravity (MRG), east/west inertio-gravity (EIG/WIG), and equatorial Rossby (ER) waves [16]. This classification relies on theoretical solutions of dry-wave linear theory and filtering over the corresponding frequency-wavelength band. However, since the observed spatiotemporal signal corresponds to propagating organized convection coupled nonlinearly with moist multiscale tropical dynamics, these idealized solutions have a number of discrepancies with the wave types observed in nature [16], [17], and more advanced methods for their detection and tracking are being sought [8], [18]. Wavelet transforms can succesfully decompose such multiscale structures [19], [20], [21], yet they do not reduce the dimensionality of the signal and are not well-suited for objective classification and nonparameteric modeling of a finite number of spatiotemporal modes.

II. KOOPMAN OPERATOR APPROACH

Here, we demonstrate the potential of a new machine learning technique for extraction of spatiotemporal features (such as CCEWs) that are defined by eigenvectors of data-driven Koopman operators. This approach has been introduced recently in [22], where a more detailed analysis can be found, and here we summarize this reference briefly. The approach utilizes the framework of ergodic theory, and in particular, it relies on the notion that the temporal sequence of states $a_0, a_1, \ldots$ of the system (here, of the atmosphere system) is the outcome of measure-preserving ergodic dynamics. The objective is to derive observables (temporal patterns) that are functions of these states, $f(a_i)$, and can be associated with the particular phenomena of interest

Corresponding author: J. Slawinska, joanna.slawinska@nyu.edu
[1]Center for Environmental Prediction, Rutgers, The State University of New Jersey, New Brunswick, NJ, USA. [2]Courant Institute of Mathematical Sciences, New York University, New York, NY, USA.

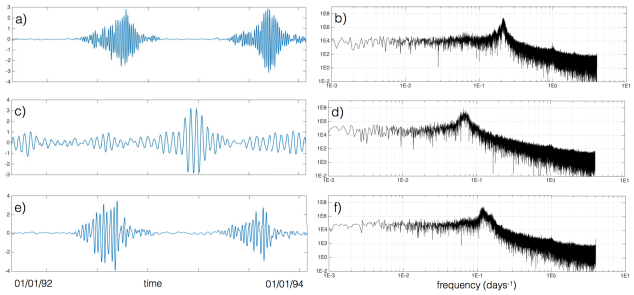


Fig. 1. Temporal patterns (left) and their frequency spectra (right) for three examples of CCEW-like patterns obtained here as eigenfunctions of data-driven Koopman operators.

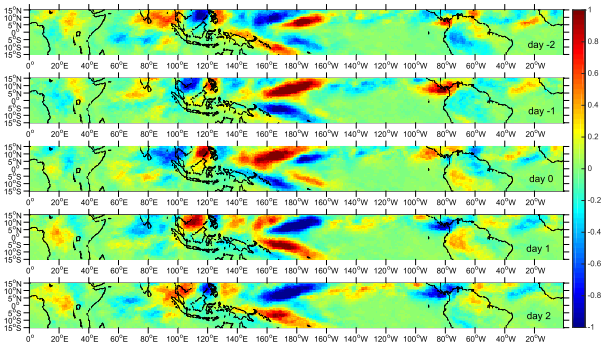


Fig. 2. 5-day evolution of the T_b anomalies (arbitrary units) associated with the 5-day westward MRG wave shown in Fig. 1(a,b).

(here, CCEWs). In order to achieve that, the Koopman operators that govern the temporal evolution of the observables are being considered. The Koopman operator U of a dynamical system (sampled discretely in time) acts on observables by time shifts, i.e., $g = U(f)$ with $g(a_i) = f(a_{i+1})$. Note that even if the dynamics is nonlinear, U is a linear operator acting on a (generally, infinite-dimensional) space of observables. Historically, this operator-theoretic approach was introduced in the 1930s in the context of ergodic theory [23], and has been employed more recently in dynamic mode decomposition techniques [24], [25].

The method in [22] uses kernel methods [26] to represent the Koopman operator U in a smooth orthonormal basis $\{\phi_i\}$ of kernel eigenfunctions learned from time-ordered observations of the system. Dynamical temporal patterns, ψ_k, are then determined by solving the Koopman eigenvalue problem, $U\psi_k = \lambda_k\psi_k$, in this basis. Numerically, this is a Galerkin method involving the solution of the matrix eigenvalue problem $\mathbf{A}\vec{c}_k = \lambda_k\vec{c}_k$, where $A_{mn} = \phi_m^\top U(\phi_n)$ are the matrix elements of the Koopman operator, and the column vectors $\vec{c}_k$ store the expansion coefficients of ψ_k in the $\{\phi_i\}$ basis. The corresponding eigenvalues λ_k are complex, $\lambda_k = e^{\gamma_k + i\omega_k}$, and capture growth rates (γ_k) and oscillatory frequencies (ω_k). As a result, the input signal is decomposed into quasi-oscillatory patterns with distinctive timescale separation. These patterns are intrinsic to the dynamical system in the sense that they are invariant under invertible nonlinear transformations of the input data. Moreover, spatiotemporal patterns associated with the ψ_k can be obtained via projections of the input data; a procedure familiar from PCA.

As with any Galerkin method, the efficacy of this scheme depends strongly on the choice of basis $\{\phi_i\}$, and hence the choice of kernel. In [22], the kernel is carefully constructed to ensure that the ϕ_i are orthonormal with respect to the invariant measure of the dynamics (see also [27]). A validation of the method can be found in [28], where extraction of traveling waves is demonstrated for a synthetic dataset [19] consisting of two superposed waves of different wavelength and with amplitude modulation by a background state. Reference [28] also employs kernels based on Takens delay embeddings [29] to extract modes of convective organization (including CCEW signals) from equatorially averaged brightness temperature (T_b) data. The same kernels have previously been employed for extraction and prediction of intraseasonal and interannual modes via so-called NLSA algorithms [30], [31], [32], [33], but the higher-frequency intermittent CCEW signals of interest here could not be recovered by NLSA.

III. RESULTS

We apply the Koopman operator approach described in section II to extract CCEW patterns from two-dimensional (2D) T_b data. In this study, T_b data from the CLAUS multi-satellite archive [34] were sampled on a uniform longitude-latitude grid of 0.5° resolution, and observed every 3 hours for the period July 1, 1983 to June 30, 2009. Since knowledge of the T_b field at a given time is not sufficient to uniquely determine its evolution at a later time, we embed the input data in a higher dimensional space via delay-coordinate maps [35]. In particular, selecting an integer parameter q, we construct the time series $u(t) = (T_b(t), T_b(t-1), \ldots, T_b(t-q+1))$, where $T_b(t)$ denotes the sampled 2D T_b field at time t. Due to a theorem of Takens [36], the signal $u(t)$ is expected to be more Markovian than the individual T_b snapshots. Here, following [30], [31], [28], [32], we set $q = 512$, corresponding to a time interval of 64 days for our 3-hour sampling

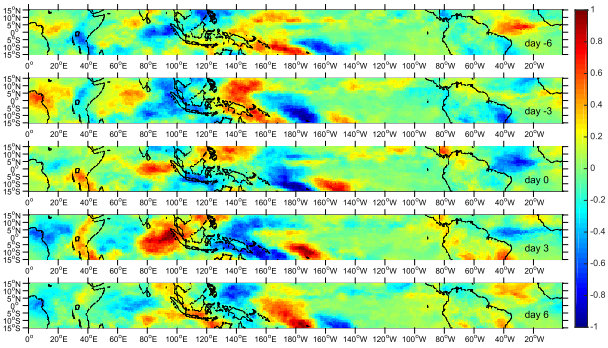


Fig. 3. 15-day evolution of the T_b anomalies (arbitrary units) associated with the 15-day westward ER wave shown in Fig. 1(c,d).

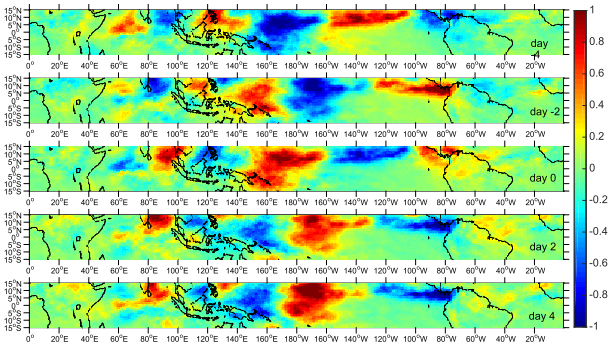


Fig. 4. 10-day evolution of the T_b anomalies (arbitrary units) associated with the 8-day eastward Kelvin wave shown in Fig. 1(e,f).

interval. After delay-coordinate mapping, the number of samples available for analysis is $n = 66{,}693$ and the ambient space dimension is $q \times d = 5{,}713{,}920$, where $d = 11,160$ is the number of sampled spatial points covering the whole equatorial belt. The kernel eigenfunctions ϕ_i are computed as described in [28]. The Koopman eigenfunctions ψ_k are subsequently computed using the leading $l = 110$ kernel eigenfunctions as a basis for the Galerkin approximation space.

The resulting eigenfunctions exhibit a broad range of timescales, ranging from interannual to diurnal timescales. Among them are periodic patterns representing the seasonal and diurnal cycles and their harmonics. Other patterns have the structure of intermittent, amplitude-modulated traveling waves. These waves include intraseasonal oscillations (ISOs), but also higher-frequency CCEWs. Representative eigenfunctions of the latter class are displayed in Fig. 1. There, the strong temporal intermittency of these modes, characterized by periods of energetic, yet coherent, activity interspersed between periods of almost complete quiescence is clearly evident. The modulating envelopes of the CCEW modes evolves on intraseasonal timescales, suggesting possible interactions with larger-scale ISOs. To our knowledge, this is the first time that such CCEW patterns have been recovered from 2D T_b data via an objective eigendecomposition technique.

The spatiotemporal patterns associated with the Koopman eigenfunctions in Fig. 1 are displayed in Figs. 2–4. There, various types of eastward and westward traveling convective organization can be seen. The modes are in good qualitative agreement with the dispersion characteristics and spatial structures obtained by dry linear CCEW theory (see, e.g., Figs. 2–4 in [16]). In particular, Fig. 2 shows a 5-day westward-propagating mixed Rossby-gravity (MRG) wave, Fig. 3 a 15-day westward-propagating equatorial Rossby wave, and Fig. 4 an 8-day eastward-propagating Kelvin wave. It is important to note that despite the qualitative agreement with linear dry theory, the patterns in Figs. 2–4 have notable differences from the idealized structures. For example, a prominent feature of the MRG in Fig. 2 is a T_b anomaly tilt with latitude which is not consistent with the theoretical MRG solutions of the shallow water equations. This tilted structure was also identified in [37] from filtered OLR data, but was recovered here from unfiltered data through eigenfunctions of an operator governing the time evolution of observables of the atmosphere system.

IV. SUMMARY AND OUTLOOK

In this work, we have demonstrated the potential of data-driven Koopman operator techniques for extraction of spatiotemporal patterns from high-dimensional multiscale timeseries generated by nonlinear dynamical systems. In particular, applying the method developed in [22] to 2D brightness temperature data over the tropics, we identified several propagating patterns corresponding to CCEWs. To our knowledge, recovery of such patterns from brightness temperature data has previously not been possible via objective eigendecomposition techniques. This provides an opportunity to improve the understanding of CCEW structures and their interactions with lower-frequency modes [9], [6].

ACKNOWLEDGMENTS

The research of D. Giannakis is supported by ONR Grant N00014-14-0150, ONR MURI Grant 25-74200-F7112, and NSF Grant DMS-1521775.

REFERENCES

[1] J. Slawinska, W. W. Grabowski, H. Pawlowska, and A. A. Wyszogrodzki, "Optical Properties of Shallow Convective Clouds Diagnosed from a Bulk-Microphysics Large-Eddy Simulation," *J. Climate*, vol. 21, no. 7, pp. 1639–1647, 2008.

[2] J. Slawinska, W. W. Grabowski, H. Pawlowska, and H. Morrison, "Droplet Activation and Mixing in Large-Eddy Simulation of a Shallow Cumulus Field," *J. Atmos. Sci*, vol. 69, no. 2, pp. 444–462, 2012.

[3] S. C. Sherwood, S. Bony, and J.-L. Dufresne, "Spread in model climate sensitivity traced to atmospheric convective mixing," *Nature*, vol. 505, no. 7481, pp. 37–42, 2014.

[4] S. Bony, B. Stevens, D. M. W. Frierson, C. Jakob, M. Kageyama, R. Pincus, T. G. Shepherd, S. C. Sherwood, A. P. Siebesma, A. H. Sobel, M. Watanabe, and M. J. Webb, "Clouds, circulation and climate sensitivity," *Nature Geosci*, vol. 8, pp. 261–268, 2015.

[5] B. Stevens, "Rethinking the lower bound on aerosol radiative forcing," *J. Climate*, vol. 28, no. 12, pp. 4794–4819, 2015.

[6] J. Slawinska, O. Pauluis, A. J. Majda, and W. W. Grabowski, "Multiscale Interactions in an Idealized Walker Circulation: Mean Circulation and Intraseasonal Variability," *J. Atmos. Sci.*, vol. 71, no. 3, pp. 953–971, 2014.

[7] B. Mapes, S. Tulich, J. Lin, and P. Zuidema, "The mesoscale convection life cycle: Building block or prototype for large-scale tropical waves?," *Dyn. Atmos. Oceans*, vol. 42, 2006.

[8] J. Dias, S. N. Tulich, and G. N. Kiladis, "An Object-Based Approach to Assessing the Organization of Tropical Convection," *J. Atmos. Sci.*, vol. 69, no. 8, pp. 2488–2504, 2012.

[9] J. Dias, S. Leroux, S. N. Tulich, and G. N. Kiladis, "How Systematic is Organized Tropical Convection within the MJO?," *Geophys. Res. Lett.*, vol. 40, pp. 1420–1425, 2013.

[10] A. D. D. Genio and W. Kovari, "Climatic Properties of Tropical Precipitating Convection under Varying Environmental Conditions," *J. Climate*, vol. 15, no. 18, pp. 2597–2615, 2002.

[11] J. Slawinska, W. W. Grabowski, and H. Morrison, "The impact of atmospheric aerosols on precipitation from deep organized convection: A prescribed-flow model study using double-moment bulk microphysics," *Q. J. Roy. Meteorol. Soc.*, vol. 135, no. 644, pp. 1906–1913, 2009.

[12] A. D. Del Genio, "Representing the Sensitivity of Convective Cloud Systems to Tropospheric Humidity in General Circulation Models," *Surv. Geophys.*, vol. 33, pp. 637–656, 2012.

[13] M. W. Moncrieff, D. E. Waliser, M. J. Miller, M. A. Shapiro, G. R. Asrar, and J. Caughey, "Multiscale Convective Organization and the YOTC Virtual Global Field Campaign," *Bull. Amer. Meteor. Soc.*, vol. 93, no. 8, pp. 1171–1187, 2012.

[14] I. Tobin, S. Bony, C. E. Holloway, J.-Y. Grandpeix, G. Sèze, D. Coppin, S. J. Woolnough, and R. Roca, "Does convective aggregation need to be represented in cumulus parameterizations?," *J. Adv. Model. Earth Syst.*, vol. 5, pp. 692–703, 2013.

[15] J. Slawinska, O. Pauluis, A. J. Majda, and W. W. Grabowski, "Multiscale Interactions in an Idealized Walker Cell: Simulations with Sparse Space-Time Superparameterization," *Mon. Wea. Rev.*, vol. 143, no. 2, pp. 563–580, 2015.

[16] G. N. Kiladis, M. C. Wheeler, P. T. Haertel, K. H. Straub, and P. E. Roundy, "Convectively coupled equatorial waves," *Rev. Geophys.*, vol. 47, no. 2, 2009. RG2003.

[17] G. N. Kiladis, J. Dias, and M. Gehne, "The relationship between Equatorial Mixed Rossby-Gravity and Eastward Inertio-Gravity Waves. Part I," *J. Atmos. Sci.*, pp. 2123–2145, 2016.

[18] H. R. Ogrosky and S. N. Stechmann, "Identifying Convectively Coupled Equatorial Waves Using Theoretical Wave Eigenvectors," *Mon. Wea. Rev.*, vol. 144, pp. 2235–2264, 2016.

[19] K. Kikuchi and B. Wang, "Spatiotemporal Wavelet Transform and the Multiscale Behavior of the Madden-Julian Oscillation," *J. Climate*, vol. 23, pp. 3814–3834, 2010.

[20] K. Kikuchi, "An introduction to combined Fourier–wavelet transform and its application to convectively coupled equatorial waves," *Climate Dyn*, no. 5, pp. 1339–1356, 2014.

[21] J. Slawinska, O. Pauluis, A. J. Majda, and W. W. Grabowski, "Multiscale Interactions in an Idealized Walker Cell: Analysis with Isentropic Streamfunctions," *J. Atmos. Sci.*, vol. 73, no. 3, pp. 1187–1203, 2016.

[22] D. Giannakis, "Data-driven spectral decomposition and forecasting of ergodic dynamical systems," *Appl. Comput. Harmon. Anal.*, 2016. Submitted.

[23] O. Koopman, B., "Hamiltonian Systems and Transformation in Hilbert Space," *Proc. Natl. Acad. Sci.*, vol. 17, no. 5, pp. 315–318, 1931.

[24] I. Mezić, "Spectral properties of dynamical systems, model reduction and decompositions," *Nonlinear Dynamics*, vol. 41, no. 1, pp. 309–325, 2005.

[25] M. Budišić, R. Mohr, and I. Mezić, "Applied Koopmanism," *Chaos*, vol. 22, no. 4, 2012.

[26] R. R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal.*, vol. 21, pp. 5–30, 2006.

[27] T. Berry, D. Giannakis, and J. Harlim, "Nonparametric forecasting of low-dimensional dynamical systems," *Phys. Rev. E*, vol. 91, p. 032915, Mar 2015.

[28] D. Giannakis, J. Slawinska, and Z. Zhao, "Spatiotemporal Feature Extraction with Data-Driven Koopman Operators," *Journal of Machine Learning Research, Proceedings of the 1st International Workshop on 'Feature Extraction: Modern Questions and Challenges' and NIPS Conference*, vol. 44, pp. 103–115, 2015.

[29] D. Giannakis and A. J. Majda, "Nonlinear Laplacian spectral analysis: capturing intermittent and low-frequency spatiotemporal patterns in high-dimensional data," *Statistical Analysis and Data Mining*, vol. 6, no. 3, pp. 180–194, 2013.

[30] D. Giannakis, W.-w. Tung, and A. J. Majda, "Hierarchical structure of the Madden-Julian oscillation in infrared brightness temperature revealed through nonlinear Laplacian spectral analysis," in *2012 Conference on Intelligent Data Understanding (CIDU)*, (Boulder, Colorado), pp. 55–62, 2012.

[31] W.-w. Tung, D. Giannakis, and A. J. Majda, "Symmetric and Antisymmetric Signals in MJO Deep Convection. Part I: Basic modes in infrared brightness temperature," *J. Atmos. Sci.*, vol. 71, pp. 3302–3326, 2014.

[32] E. Székely, D. Giannakis, and A. J. Majda, "Extraction and predictability of coherent intraseasonal signals in infrared brightness temperature data," *Climate Dyn.*, vol. 46, no. 5, pp. 1473–1502, 2016.

[33] R. Alexander, Z. Zhao, E. Székely, and D. Giannakis, "Kernel analog forecasting of tropical intraseasonal oscillations," *J. Atmos. Sci.*, 2016. In revision.

[34] K. Hodges, D. Chappell, G. Robinson, and G. Yang, "An Improved Algorithm for Generating Global Window Brightness Temperatures from Multiple Satellite Infrared Imagery," *J. Atmos. Oceanic Technol.*, vol. 17, pp. 1296–1312, 2000.

[35] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *J . Stat. Phys.*, vol. 65, pp. 579–616, 1991.

[36] F. Takens, "Detecting strange attractors in turbulence," *Dynamical Systems and Turbulence, Warwick 1980, Lecture Notes in Mathematics, Vol. 898, Springer, Berlin*, pp. 366–381, 1981.

[37] H. Hendon and B. Liebmann, "The Structure and Annual Variation of Antisymmetric Fluctuations of Tropical Convection and Their Association with Rossby-Gravity Waves," *J. Atmos. Sci.*, vol. 48, no. 19, pp. 2127–2140, 1991.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

COVARIANCE STRUCTURE ANALYSIS OF CLIMATE MODEL OUTPUT

Chintan Dalal[1],]

***Abstract*—To understand future climate different Earth system models from groups wo simulate projections of future climates. However from these simulations are computationally very sive, often requiring several months on a superco In this paper, we provide a new statistical en method that may allow a realization of future projections within a day rather than several Specifically, we analyze the structure of several outputs from various climate models on a man covariance matrices. The manifold covariance st provides a method to compare existing climate outputs, as well as to sample a new realization o climate projections. We validated our climate output comparison method using known depen between various climate models. Additiona showed, using semi-variogram plots, that the distribution of our realizations lie within the distribution of existing climate model outputs. The proposed statistical emulator could find its use in future climate impact assessment.**

I. INTRODUCTION

Our understanding of future climate changes can improve by analyzing various plausible realizations of future climate projections. However, generating a climate simulation from an Earth System model is computationally very expensive since the model captures the complex interactions among the many components of the Earth's climate system (see [1]). The Coupled Model Inter-comparison Project (CMIP [2]) coordinates efforts between various groups developing Earth system models to create a database of multi-model ensembles of climate simulations. For example, Fig. 1 shows changes in precipitation for North America from two separate Earth system models that are part of the CMIP Phase 5 (CMIP5) multi-model ensemble. Both of these models show a plausible, yet, different view of future climate changes. Hence, a thorough assessment

Corresponding author: C. Dalal, chintan.dalal@rutgers.edu, [1] Department of Computer Science, Rutgers University, NJ, [2]National Center for Atmospheric Research, Boulder, CO, [3]Climate Central, NJ

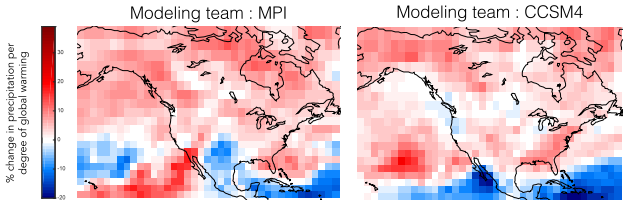


Fig. 1. Projections (2090) of percent change in precipitation per degree of change in the global mean temperature for North America from the CMIP5 multi-model ensemble. Shown here are projections from the Max Planck Inst. (MPI, Germany) and Community Earth System Model (CCSM4, USA).

of future climate impact requires a framework that can capture the variability across all climate model outputs.

An overview of methodologies that can capture the variability among climate model outputs is given in [3], along with the limitations of these approaches. For example, some of the climate models share common physical representation and numerical methods, and, thereby, cannot be considered as independent simulations. Additionally, the dependencies in climate models reduces the spread of future climate projections. To address the inter-model dependency issue, a Bayesian hierarchical framework has been suggested by [4], [5], [6], [7]. However, the proposed Bayesian framework faces difficulties in robustly modeling the inter-dependencies because of its sensitivity to prior assumptions.

Recent work by [8] shares similar methodological goals as ours in that the authors address issues of model dependencies and sampling in a non-parameteric set-up. The authors use a standard Euclidean metric on a low dimensional space by fixing the modes of variance within the available ensemble. Thus, limiting the amount of variability information that is present in the climate model outputs.

This paper presents an approach that allows for the variability information from the climate model outputs to be estimated. Specifically, we assume that the ensemble of well fitted covariance matrices provides sufficient information to characterize a distance measure. One application is to sample new realizations from an existing ensemble of climate model outputs.

II.

To capture the variabi
tions at locations arour
ensemble members, one
framework. If $\tilde{\mathbf{y}}$ is a n
then the standard multiva
(sMVN) is given by

$$\tilde{\mathbf{y}} =$$

where $\tilde{\mathbf{y}}$ is the new mu
the future climate project
Σ is the ensemble covari
is the standard normal ra

In sMVN, the estimat
distribution of the climat
of the forms

$$\Sigma(\theta)(i.e.\ \hat{\Sigma}) = \psi\mathbf{I} + c$$

In this preliminary study, H is selected as a stationary anisotropic matérn covariance function. Here, stationarity is selected for simplicity, and the anisotropic matérn covariance function is a standard choice in geostatistics. Finally, $\hat{\mu}$ is estimated as an equally weighted average.

The parameters $\theta = \{\phi, \sigma, \psi\}$ are also known as range, sill, and nugget, resp., in the geostatistics literature. They are estimated by maximizing the likelihood function, which is of the form $\ell(\theta|\mathbf{y}_1, \ldots, \mathbf{y}_\mathrm{N}) \propto |\Sigma(\theta)|^{-\frac{\mathrm{N}}{2}} \prod_{i=1}^{\mathrm{N}} \exp(-\frac{1}{2}(\mathbf{y_i} - \hat{\mu})^\mathsf{T}\Sigma(\theta)^{-1}(\mathbf{y_i} - \hat{\mu})))$, where N is the number of ensemble members, and $\mathbf{y}_i$ is a vector field of climate model outputs.

Our statistical emulation method, which we call the information geometric multivariate normal sampling method (igMVN), is depicted in Fig. 2. In igMSN, the estimate of the parameters Σ and μ is of the form

$$\hat{\Sigma} = \bar{\Sigma} + \Lambda^{\frac{1}{2}}\epsilon, \quad \hat{\mu} = \sum_{j=1}^{n}\sum_{k=1}^{m_j}\frac{1}{nm_j}\mathbf{y}_{j,k}, \qquad (3)$$

where $\hat{\Sigma}$ is sampled from a normal distribution on a manifold of covariance matrices, i.e., $\Sigma(\theta) \sim \mathcal{N}(\bar{\Sigma}, \Lambda|\Sigma(\theta_1), \ldots, \Sigma(\theta_\mathrm{N}))$. Finally, $\hat{\mu}$ is estimated as a weighted average. Here, n is the number of clusters of covariance matrices, and m_j is the number of covariance matrices in each cluster.

The parameters $\bar{\Sigma}$ and Λ are the mean and variance, resp., of the ensemble of covariance matrices. Each $\Sigma(\theta_i)$(i.e. Σ_i) is a covariance matrix of individual ensemble members, and θ_i is learned by maximizing a likelihood function.

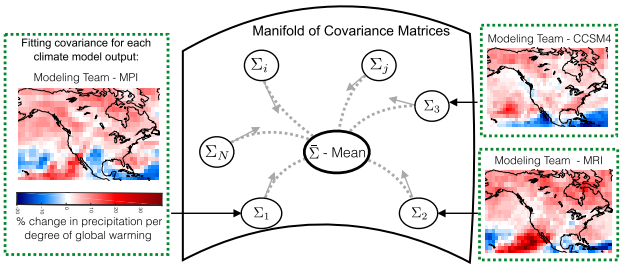


Fig. 2. Our inter-model comparison and sampling method: A Manifold view of the covariance structure of climate model outputs from various modeling teams (e.g. MPI, MRI, CCSM4)

A theoretical background for statistical distributions of symmetric positive definite matrices on a manifold can be found in [9], and the computational form to estimate $\bar{\Sigma}$ and Λ is given in [10], [11], [12], [13].

In order to estimate $\hat{\mu}$ using the weighted average, we first cluster the covariance matrices on a manifold using a standard hierarchical clustering method. The criteria for a cluster is $\max\{D(\Sigma_1, \Sigma_2) : \Sigma_1 \in S_1(\Sigma_i), \Sigma_2 \in S_2(\Sigma_i)\} < \text{threshold}$. The choice of the clustering method and the criteria for clustering are chosen for simplicity. The threshold is empirically chosen as 2 in our experiments, S_1 and S_2 are two sets of clusters of Σ_i's, and the distance metric (geodesic) on the manifold of covariance matrices is of the form $\mathrm{D}^2(\Sigma_1, \Sigma_2) = \frac{1}{2}\mathrm{Tr}(\log^2(\Sigma_1^{-\frac{1}{2}}\Sigma_2\Sigma_1^{-\frac{1}{2}}))$.

The estimates of $\hat{\mu}$ and $\hat{\Sigma}$ in igMVN incorporate extra information about the structure of covariance matrices that the sMVN fails to consider. This extra information is enabled using statistics on the structure of the covariance matrices in order to detect dependencies in climate model outputs and, thereby, incorporate known limitations in the ensemble members.

III. EVALUATION

To gain insight into the applicability of our proposed statistical emulation method, we used the ensemble of climate model outputs from CMIP5 experiments of future projections under RCP scenarios (see [1]). In order to test our method against various patterns in climate model outputs, we selected the climate variable of percent change in precipitation per degree of change in the global mean temperature.

In this paper, we restrict our study to the spatial dataset of the North American region in order to analyze the regional spatial variability aspect of the climate model outputs. Additionally, we have included single simulation runs from each of the Earth System Models (ESMs), rather than multiple simulation runs, in order to reduce biases in the ensemble.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

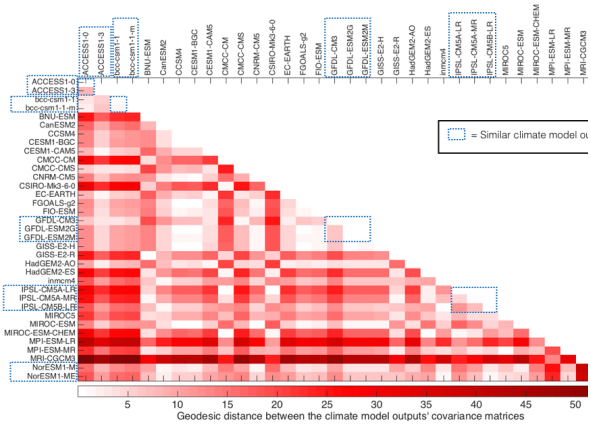


Fig. 3. A representation of the similarity measure between cl models outputs of the CMIP5 ensemble members. The simi measure that we designed is a geodesic distance between the covariance matrix of individual climate model outputs. Row columns of the above plot represent various climate model ou lighter shades of red represent higher similarity between m and boxes represent climate models that are validated to have inter-model dependencies.

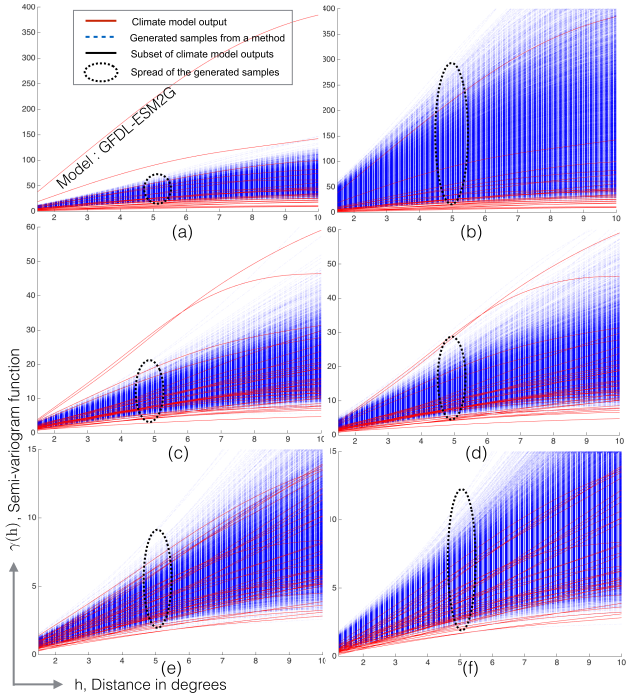


Fig. 4. Diagnostic plots showing the experimental semi-variogram function for various climate model outputs from CMIP5 ensembles (red lines) and the statistically generated samples (blue lines) from the standard multi-variate normal sampling method (sMVN) for the (a) RCP2.6 ensemble, (c) RCP4.5 ensemble, and (e) RCP 8.5 ensemble. Realizations from our sampling method (igMVN) are shown for the (b) RCP2.6 ensemble, (d) RCP4.5 ensemble, and (f) RCP8.5 ensemble. The ellipse in each plot focuses on the spread of the generated samples from each sampling method.

Fig. 3 shows the values of our distance m between each of the ensemble members. In this figure, lighter shades of red represent higher similarity in the covariance matrices of the climate model outputs, and, in turn, imply higher dependencies between the models. The climate model outputs from the same Earth system modeling group are highlighted by the blue boxes and are known to have high inter-model dependencies for reasons that include code and data sharing (see [14]). The highlighted blue boxes show lighter shades of red, and, in turn, demonstrate that the chosen geodesic distance metric can be used to compare and cluster climate model outputs in a non-parametric fashion.

Fig. 4 shows the experimental semi-variogram plots of climate model outputs and statistically generated samples from a number of methods. Given the semi-variogram function, one can estimate the parameters (range, sill, and nugget) of the covariance function. Hence, semi-variogram plot, explained in detail in [15], is a good tool in spatial statistics to visualize the differences in covariance matrices.

The climate model outputs (as shown by the red lines in Fig.4)) in the RCP2.6 and 4.5 ensembles (Fig.4) (a), (b), (c), (d)) has higher inter-model variability in its semi-variogram plots than the RCP8.5 ensemble (Fig.4) (e) and (f)). Hence, the spread of the climate model outputs realizations (as shown by the blue lines in Fig. 4) using the igMVN method (Fig.4) (b) and (d)) is better than the sMVN method (Fig.4) (a) and (b)) in representing the underlying spread of the climate model outputs. The wide spread in the igMVN samples could be attributed to the sampling of the covariance matrices from a manifold.

Fig.5 (a) shows the spatial field of the GFDL-ESM2G model output (a member in the CMIP5-RCP2.6 ensemble) overlaying the North American region. From the semi-variogram plots in Fig.4 (a) and (b) we see that the realizations from the sMSV method does not emulate the climate data well, when compared to the igMVN method, for the GFDL-ESM2G model. Similarly, in Fig.5 (b) and (c) we see that there are more matching pixels (as shown by the grey colored boxes) in the realizations from the igMVN method (c) than the sMVN method (a). Therefore, the igMVN method may have some advantages over more traditional approaches; hence, it would be worth pursuing this method to compare and sample climate model outputs.

IV. DISCUSSION

In this paper, we have shown a non-parametric statistical emulator that can potentially mimic the existing ensemble of climate model outputs for projections of precipitation changes over North America and under the RCP scenarios. Additionally, we have provided

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

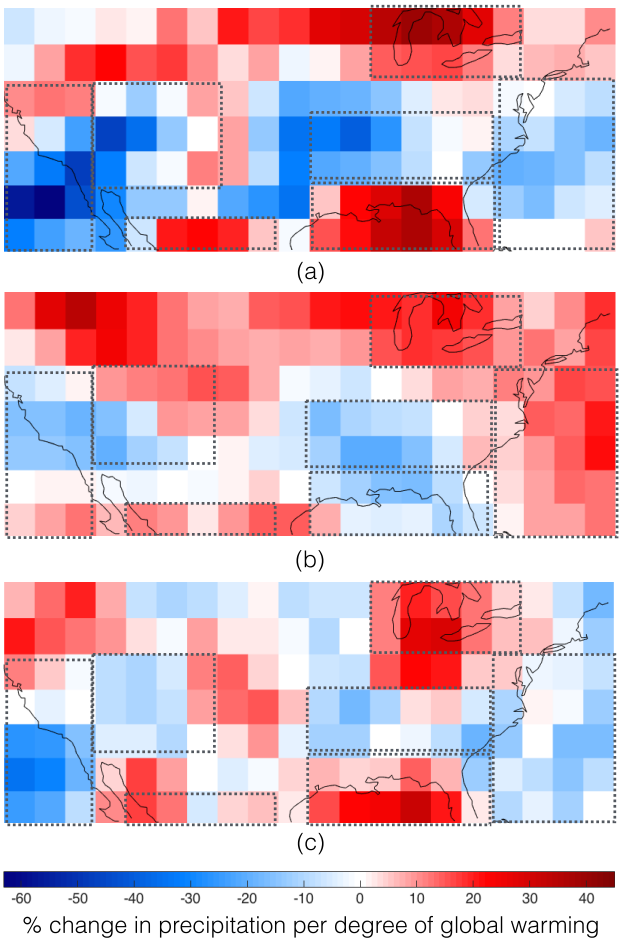


Fig. 5. Diagnostic plots showing the spatial field of climate variables from the Geophysical Fluid Dynamics Laboratorys climate model output of GFDL-ESM2G (a CMIP5-RCP2.6 ensemble member). The spatial field shown here is restricted to the North American region. (a) shows the climate model output, (b) shows one of closest realization (from Fig. 4(a)) using the sMVN method, and (c) shows one of closest realization (from Fig. 4(b)) using the igMVN method. The coast is represented by black lines, and the boxes represents patterns of similarity between the realizations and the climate model output.

a method to compare climate model outputs, which can be potentially used to investigate multi-model interdependencies in the CMIP5 ensembles.

By providing an emulator and a method for inter-model comparison, we can make the uncertainty in future climate projections more comprehensive and robust.

ACKNOWLEDGMENTS

This work is supported by the U.S. Department of Homeland Security under Grant Award Number 2012-ST-104-000044. We thank Vladimir Pavlovic, Dimitris Metaxas, Benjamin Sanderson, Matthew Edwards, Netta Gurari, and the anonymous reviewers for their feedback.

REFERENCES

[1] IPCC, "The physical science basis: Working group 1 contribution to the fifth assessment report of the intergovernmental panel on climate change," *New York: Cambridge University Press*, vol. 1, pp. 535–1, 2013.

[2] K. E. Taylor, R. J. Stouffer, and G. A. Meehl, "An overview of cmip5 and the experiment design," *Bulletin of the American Meteorological Society*, vol. 93, no. 4, p. 485, 2012.

[3] C. Tebaldi and R. Knutti, "The use of the multi-model ensemble in probabilistic climate projections," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1857, pp. 2053–2075, 2007.

[4] C. Tebaldi, R. L. Smith, D. Nychka, and L. O. Mearns, "Quantifying uncertainty in projections of regional climate change: A bayesian approach to the analysis of multimodel ensembles," *Journal of Climate*, vol. 18, no. 10, pp. 1524–1540, 2005.

[5] N. A. Leith and R. E. Chandler, "A framework for interpreting climate model outputs," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 59, no. 2, pp. 279–296, 2010.

[6] C. Tebaldi and B. Sansó, "Joint projections of temperature and precipitation change from multiple climate models: a hierarchical bayesian approach," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, vol. 172, no. 1, pp. 83–106, 2009.

[7] R. Furrer, S. R. Sain, D. Nychka, and G. A. Meehl, "Multivariate bayesian analysis of atmosphere–ocean general circulation models," *Environmental and ecological statistics*, vol. 14, no. 3, pp. 249–266, 2007.

[8] B. M. Sanderson, R. Knutti, and P. Caldwell, "Addressing interdependency in a multimodel ensemble by interpolation of model properties," *Journal of Climate*, vol. 28, no. 13, pp. 5150–5170, 2015.

[9] S. Amari, *Differential geometry in statistical inference*, vol. 10. Inst. of mathematical statistic, 1987.

[10] S.-I. Amari, "Information geometry on hierarchy of probability distributions," *IEEE transactions on information theory*, vol. 47, no. 5, pp. 1701–1711, 2001.

[11] X. Pennec, "Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements," *Journal of mathematical imaging and vision*, vol. 25, no. 1, pp. 127–154, 2006.

[12] T. Imai, A. Takaesu, and M. Wakayama, "Remarks on geodesics for multivariate normal models," *2011B-6*, 2011.

[13] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Communications on pure and applied mathematics*, vol. 30, no. 5, pp. 509–541, 1977.

[14] R. Knutti, D. Masson, and A. Gettelman, "Climate model genealogy: Generation cmip5 and how we got there," *Geophysical Research Letters*, vol. 40, no. 6, pp. 1194–1199, 2013.

[15] N. Cressie, "Statistics for spatial data: Wiley series in probability and statistics," *Wiley-Interscience, New York*, vol. 15, pp. 105–209, 1993.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

SIMPLE AND EFFICIENT TENSOR REGRESSION FOR SPATIO-TEMPORAL FORECASTING

Rose Yu, Yan Liu

Computer Science Department, University of Southern California

***Abstract*—Forecasting, a classic problem in climate science, has gained much improved performance by incorporating spatio-temporal correlations. Tensor regression provides an efficient framework for spatio-temoral forecasting. However, many of the existing algorithms for tensor regression suffer from memory bottleneck. In this paper, we develop the tensor projected gradient (TPG) algorithm, whose memory requirement is linear in the problem size. We demonstrate that our algorithm obtains comparable prediction accuracy with significant speed-up and memory budge on climate measurements.**

I. MOTIVATION

The increasing capabilities of climate measurement devices have lead to large amount of data with spatial and temporal information. It is critical yet challenging to incorporate the spatio-temporal correlations when performing data analysis. One classic problem in climate science is spatio-temporal foreacasting [12], i.e., predicting climate variables at different locations and time using their historical measurements. We note that spatio-temporal data can be naturally represented as a tensor (time $\times$ location $\times$ climate variables). Many work (e.g.[1], [2]) have shown that spatio-temporal forecasting can be formulated as a low-rank tensor regression problem, which provides us with a concise way of modeling complex structures in climate data.

Tensor regression assumes that the model parameters form a high order tensor and there exists a low-dimensional factorization for the model tensor. Existing tensor regression algorithms (e.g.[3], [4], [5], [6], [7]) fall into two categories : (1) alternating least square (ALS) sequentially finds the factor that minimizes the loss while keeping others fixed; (2) spectral regularization approximates the original non-convex problem with a convex surrogate loss, such as the nuclear norm of the unfolded tensor.

Corresponding author: Rose Yu, qiyu@usc.edu

A clear drawback of all the algorithms mentioned above is high computational cost. ALS displays unstable convergence properties and outputs sub-optimal solutions [8]. Trace-norm minimization suffers from slow convergence [9]. Moreover, those methods face the memory bottleneck when dealing with large-scale datasets. For example, the greedy algorithm [1] follows the Orthogonal Matching Pursuit (OMP) scheme. Though significantly faster, it requires the matricization of the data tensor, and thus would face memory bottleneck when dealing with large sample size.

In this paper, we introduce subsampled Tensor Projected Gradient (TPG), a simple and fast solution. Our algorithm is based upon projected gradient descent [10] and can also be seen as a tensor generalization of iterative hard thresholding algorithm [11]. The algorithm only needs a fixed number of iterations, depending solely on the logarithm of signal to noise ratio. The memory requirement grows linearly with the size of the problem. We demonstrate the empirical performance on multivariate spatio-temporal forecasting for climate measurements. Experiment results show that the proposed algorithm significantly outperforms existing approaches in both prediction accuracy and speed.

II. METHOD

A. Spatio-Temporal Forecasting

Suppose we are given access to measurements $\mathcal{X} \in \mathcal{R}^{T\times P\times M}$ of T timestamps of M variables over P locations as well as the geographical coordinates of P locations. We can model the time series with a Vector Auto-regressive (VAR) model of lag L, where we assume the generative process as $\mathcal{X}_{t,:,m} = \mathbf{X}_{t,m}\mathcal{W}_{:,:,m} + \mathcal{E}_{t,:,m}$, for $m = 1,\ldots,M$ and $t = L+1,\ldots,T$. Here $\mathbf{X}_{t,m} = [\mathcal{X}^\top_{t-1,:,m},\ldots,\mathcal{X}^\top_{t-L,:,m}]$ denotes the concatenation of L-lag historical data before time t. We learn a model coefficient tensor $\mathcal{W} \in \mathbb{R}^{PL\times P\times M}$ to forecast

multiple variables simultaneously. The forecasting task can be formulated as follows:

$$\widehat{\mathcal{W}} = \underset{\mathcal{W}}{\operatorname{argmin}} \left\{ \|\widehat{\mathcal{X}} - \mathcal{X}\|_F^2 + \mu \sum_{m=1}^{M} \operatorname{tr}(\widehat{\mathcal{X}}_{:,:,m}^\top \mathbf{L} \widehat{\mathcal{X}}_{:,:,m}) \right\}$$
$$\text{s.t. } \widehat{\mathcal{X}} = \mathbf{X}_{t,m} \mathcal{W}_{:,:,m}, \text{ s.t. } \operatorname{rank}(\mathcal{W}) \leq R \quad (1)$$

where rank constraint imposes structures such as spatial clustering and temporal periodicity on the model. The Laplacian regularizer $\mathbf{L}$ is constructed from the kernel using the geographical information, which accounts for the spatial proximity of observations. With simple change of variables, spatio-temporal forecasting can be shown as a special case of the following tensor regression problem [1].

B. Tensor Regression

Given a predictor tensor $\mathcal{X}$ and a response tensor $\mathcal{Y}$, tensor regression targets at the following problem:

$$\mathcal{W}^\star = \underset{\mathcal{W}}{\operatorname{argmin}} \mathcal{L}(\mathcal{W}; \mathcal{X}, \mathcal{Y})$$
$$\text{s.t. } \operatorname{rank}(\mathcal{W}) \leq R \quad (2)$$

The problem aims to estimate a model tensor $\mathcal{W} \in \mathbb{R}^{D_1 \times D_2 \times D_3}$ that minimizes the empirical loss $\mathcal{L}$, subject to the constraint that the Tucker rank of $\mathcal{W}$ is at most R. Equivalent, the model tensor $\mathcal{W}$ has a low-dimensional factorization $\mathcal{W} = \mathcal{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$ with core $\mathcal{S} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$ and orthonormal projection matrices $\{\mathbf{U}_n \in \mathbb{R}^{D_n \times R_n}\}$. The dimensionality of $\mathcal{S}$ is at most R. The reason we favor Tucker rank over others is due to the fact that it is a high order generalization of matrix SVD, thus is computational tractable, and carries nice properties that we later would utilize.

Algorithm 1 Subsampled Tensor Projected Gradient

Input: predictor $\mathcal{X}$, response $\mathcal{Y}$, rank R
Output: model tensor $\mathcal{W} \in \mathbb{R}^{D_1 \times D_2 \times D_3}$
Compute count sketch $\mathbf{S}$
Sketch $\tilde{\mathcal{Y}} \leftarrow \mathcal{Y} \times_1 \mathbf{S}$, $\tilde{\mathcal{X}} \leftarrow \mathcal{X} \times_1 \mathbf{S}$
Initialize $\mathcal{W}^0$ as zero tensor
repeat
  $\tilde{\mathcal{W}}^{k+1} = \mathcal{W}^k - \eta \nabla \mathcal{L}(\mathcal{W}^k; \tilde{\mathcal{X}}, \tilde{\mathcal{Y}})$
  $\mathcal{W}^{k+1} = \text{ITP}(\tilde{\mathcal{W}}^{k+1})$
until Converge

As shown in Algorithm 1, subsampled Tensor Projected Gradient (TPG) combines a gradient step with a proximal point projection step [14]. The gradient step treats (2) as an unconstrained optimization of $\mathcal{W}$. As long as the loss function is differentiable in a neighborhood of current solution, standard gradient descent methods can be applied. For our case, computing the gradient under linear model is trivial: $\nabla \mathcal{L}(\mathcal{W}; \mathcal{X}, \mathcal{Y}) = \langle \mathcal{X}^T, \mathcal{Y} - \langle \mathcal{X}, \mathcal{W} \rangle \rangle$. After the gradient step, the subsequent proximal point step aims to find a projection $\mathcal{P}_R(\mathcal{W}) : \mathbb{R}^{D_1 \times D_2 \times D_3} \to \mathbb{R}^{D_1 \times D_2 \times D_3}$ satisfying:

$$\mathcal{P}_R(\mathcal{W}^k) = \underset{\mathcal{W}}{\operatorname{argmin}} (\|\mathcal{W}^k - \mathcal{W}\|_F^2)$$
$$\text{s.t. } \mathcal{W} \in \mathcal{C}(R) = \{\mathcal{W} : \operatorname{rank}(\mathcal{W}) \leq R\} \quad (3)$$

Algorithm 2 Iterative Tensor Projection (ITP)

Input: model $\tilde{\mathcal{W}}$, predictor $\mathcal{X}$, response $\mathcal{Y}$, rank R
Output: projection $\mathcal{W} \in \mathbb{R}^{D_1 \times D_2 \times D_3}$
Initialize $\{\mathbf{U}_n^0\}$ with R left singular vectors of $\mathcal{W}_{(n)}$
while $i \leq R$ **do**
  repeat
    $\mathbf{u}_1^{k+1} \leftarrow \tilde{\mathcal{W}} \times_2 \mathbf{u}_2^{kT} \times_3 \mathbf{u}_3^{kT}$
    $\mathbf{u}_2^{k+1} \leftarrow \tilde{\mathcal{W}} \times_1 \mathbf{u}_1^{kT} \times_3 \mathbf{u}_3^{kT}$
    $\mathbf{u}_3^{k+1} \leftarrow \tilde{\mathcal{W}} \times_1 \mathbf{u}_1^{kT} \times_2 \mathbf{u}_2^{kT}$
  until Converge to $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$
  Update $\{\mathbf{U}_n\}$ with $\{\mathbf{u}_n\}$
  $\mathcal{W} \leftarrow \tilde{\mathcal{W}} \times_1 \mathbf{U}_1 \mathbf{U}_1^T \times_2 \mathbf{U}_2 \mathbf{U}_2^T \times_3 \mathbf{U}_3 \mathbf{U}_3^T$
  if $\mathcal{L}(\mathcal{W}; \mathcal{X}, \mathcal{Y}) \leq \epsilon$ **then**
    RETURN
  end if
end while

The difficulty of solving the above problem mainly comes from the non-convexity of the set of low-rank tensors. Common approaches based on spectral approximation requires a full SVD for each unfolding of the tensor. Iterative hard thresholding, on the other hand, takes advantage of the general Eckart-Young-Mirsky theorem [15] for matrices, which allows the Euclidean projection to be efficiently computed with thin SVD. Unfortunately, Eckart-Young-Mirsky theorem does not apply to higher order tensors [16]. Therefore, computing high-order singular value decomposition (HOSVD) [17] and discarding small singular values do not guarantee optimality of the projection.

To address the challenge, we note that for tensor Tucker model, we have : $\mathcal{W} = \mathcal{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3$. And the projection matrices $\{\mathbf{U}_n\}$ happen to be the left singular vectors of the unfolded tensor, i.e., $\mathbf{U}_n \mathbf{\Sigma}_n \mathbf{V}_n^T = \mathcal{W}_{(n)}$. This property allows us to compute each projection matrix efficiently with thin SVD. By iterating over all factors, we can obtain a local optimal solution that is guaranteed to have rank at most R. We want to emphasize that there is no known algorithm

that can guarantee the convergence to the global optimal solution. However, in the Tucker model, different local optimas are highly concentrated, thus the choice of local optima does not really matter [18].

When the model parameter tensor $\mathcal{W}$ is very large, performing thin SVD itself can be expensive. In our problem, the dimensionality of the model is usually much larger than its rank. With this observation, we utilize another property of Tucker model $\mathbf{U}_n = \mathcal{W} \times_1 \cdots \times_{n-1} \mathbf{U}_{n-1}^T \times_{n+1} \mathbf{U}_{n+1}^T \cdots \times_N \mathbf{U}_N$. This property implies that instead of performing thin SVD on the original tensor, we can trade cheap tensor matrix multiplication to avoid expensive large matrix SVD. This leads to the Iterative Tensor Projection (ITP) procedure as described in Algorithm 2. Denote $\{\mathbf{u}_n\}$ as row vectors of $\{\mathbf{U}_n\}$, ITP uses power iteration to find one leading singular vector at a time. The algorithm stops either when hitting the rank upper bound R or when the loss function value decreases below a threshold ϵ.

ITP is significantly faster especially when the model is low-rank. If we initialize our solution with the top R left singular vectors of tensor unfoldings, the projection iteration can start from a close neighborhood of the stationary point, thus leading to faster convergence. In tensor regression, our main focus is to minimize the empirical loss. Sequentially finding the rank-1 subspace allows us to evaluate the performance as the algorithm proceeds. The decrease of empirical loss would call for early stop of the thin SVD procedure. Another acceleration trick we employ is randomized sketching. This trick is particularly useful when we are encountered with ultra high sample size or extremely sparse data. In practice, we find count sketch works well with TPG, even when the sample size is very small.

III. EVALUATION

To evaluate the performance of our framework, we experiment with the U.S. Historical Climatology Network (USHCN) daily [1]. The data set contains daily measurements for 5 climate variables (temperature max, temperature min, precipitation, snow fall and snow depth) for more than 100 years. The records were collected across more than $1,200$ locations and spans over $45,384$ time stamps.

We split the data along the temporal dimension into 80% training set and 20% testing set. We choose VAR (3) model and use 5-fold cross-validation to select the rank during the training phase. For both datasets, we normalize each individual time series by removing the mean and dividing by standard deviation. Due to the memory constraint of the Greedy algorithm, evaluations are conducted on down-sampled datasets.

[1]http://cdiac.ornl.gov/ftp/ushcn_daily/

TABLE I
FORECASTING RMSE AND RUN TIME ON USHCN DAILY MEASUREMENT FOR VAR PROCESS WITH 3 LAGS

	TPG	OLS	THOSVD	GREEDY	ADMM
RMSE	**0.3872**	1.4265	0.7224	0.4389	0.5893
RUNTIME	144.43	23.69	46.26	410.38	6786

Table I presents the best forecasting performance (w.r.t sketching size) and the corresponding run time for each of the methods. TPG outperforms baseline methods with higher accuracy. Greedy shows similar accuracy, but TPG converges in very few iterations. For USHCN, TPG achieves much higher accuracy with significantly shorter run time. Those results demonstrate the efficiency of our proposed algorithm for spatio-temporal forecasting tasks.

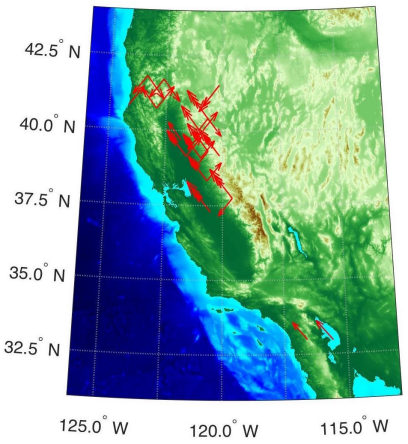


Fig. 1. Velocity vectors plot of spatial-temporal dependency graph obtained via TPG. Results are averaged across all five different climate variables.

We further investigate the learned structure of TPG algorithm from USHCN data. Figure 1 shows the spatial-temporal dependency graph on the terrain of California. Each velocity vector represents the aggregated weight learned by TPG from one location to the other. The graph provides an interesting illustration of atmospheric circulation. For example, near Shasta-Trinity National forecast in northern California, the air flow into the forecasts. On the east side along Rocky mountain area, there is a strong atmospheric pressure, leading to wind moving from south east to north west passing the bay area. Another notable atmospheric circulation happens near Salton sea at the border of Utah, caused mainly by the evaporation of the sea.

ACKNOWLEDGMENTS

This work is supported in part by the U. S. Army Research Office under grant number W911NF-15-1-0491, NSF Research Grant IIS- 1254206 and IIS-1134990. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agency, or the U.S. Government.

REFERENCES

[1] R. Yu, M. T. Bahadori, and Y. Liu, "Fast multivariate spatio-temporal analysis via low rank tensor learning," *NIPS*, 2014.

[2] R. Yu, D. Cheng, and Y. Liu, "Accelerated online low rank tensor learning for multivariate spatiotemporal streams," in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pp. 238–247, 2015.

[3] Q. Zhao, C. F. Caiafa, D. P. Mandic, L. Zhang, T. Ball, A. Schulze-Bonhage, and A. Cichocki, "Multilinear subspace regression: An orthogonal tensor decomposition approach.," in *NIPS*, vol. 2011, pp. 1269–1277, 2011.

[4] H. Zhou, L. Li, and H. Zhu, "Tensor regression with applications in neuroimaging data analysis," *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 540–552, 2013.

[5] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze, and M. Pontil, "Multilinear multitask learning," in *Proceedings of the 30th International Conference on Machine Learning*, pp. 1444–1452, 2013.

[6] K. Wimalawarne, M. Sugiyama, and R. Tomioka, "Multitask learning meets tensor factorization: task imputation via convex optimization," in *Advances in Neural Information Processing Systems*, pp. 2825–2833, 2014.

[7] M. Signoretto, Q. T. Dinh, L. De Lathauwer, and J. A. Suykens, "Learning with tensors: a framework based on convex optimization and spectral regularization," *Machine Learning*, vol. 94, no. 3, pp. 303–351, 2014.

[8] A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari, *Non-negative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[9] S. Gandy, B. Recht, and I. Yamada, "Tensor completion and low-n-rank tensor recovery via convex optimization," *Inverse Problems*, vol. 27, no. 2, p. 025010, 2011.

[10] P. H. Calamai and J. J. Moré, "Projected gradient methods for linearly constrained problems," *Mathematical programming*, vol. 39, no. 1, pp. 93–116, 1987.

[11] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[12] N. Cressie and C. K. Wikle, *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

[13] K. L. Clarkson and D. P. Woodruff, "Low rank approximation and regression in input sparsity time," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 81–90, ACM, 2013.

[14] R. T. Rockafellar, "Monotone operators and the proximal point algorithm," *SIAM journal on control and optimization*, vol. 14, no. 5, pp. 877–898, 1976.

[15] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[16] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[17] L. De Lathauwer, B. De Moor, and J. Vandewalle, "A multilinear singular value decomposition," *SIAM journal on Matrix Analysis and Applications*, vol. 21, no. 4, pp. 1253–1278, 2000.

[18] M. Ishteva, P.-A. Absil, S. Van Huffel, and L. De Lathauwer, "Tucker compression and local optima," *Chemometrics and Intelligent Laboratory Systems*, vol. 106, no. 1, pp. 57–64, 2011.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

TRACKING OF TROPICAL INTRASEASONAL CONVECTIVE ANOMALIES

Bohar Singh[2], James L. Kinter[1,2]

Abstract— **A new algorithm based upon a multiple object tracking method is developed to identify, tracks and classify Tropical intraseasonal oscillations (TISO) on the basis of their direction of propagation. Daily NOAA Outgoing Longwave Radiation anomalies (OLRA) from 1979-2013 are Lanczos bandpass filtered for the intraseasonal time scale (20-100 days) and spatially averaged with eight neighboring point to get large spatial scales (~10^5 km^2). Tracking of TISO is performed by using a two-stage Kalman filter predictor-corrector. Two dominant components of the TISO (Eastward propagating and Northward propagating) are classified, and it is found that TISO remains active throughout the year. Eastward propagation of the TISO occurs from November to April with phase speed of ~4 m/s and Northward propagation of the TISO occurs from May to October at ~2 m/s in both Indian and Pacific Ocean basins.**

I. MOTIVATION

The Tropical Intraseasonal Oscillation (TISO) refers to variability on the time scale of 20-100 days, intermediate between the time scales traditionally associated with weather and climate. Physical understanding of TISO is a very important and challenging aspect of making predictions beyond the limit of instantaneous weather. TISO can be classified into two dominant components on the basis seasonality: (a) Madden Julian Oscillations (MJO); and (b) Monsoon Intra-seasonal Oscillations (MISO). The prevailing view of the dynamics of the MJO is that it is governed by the coupling of a moist Kelvin-Rossby wave to convective heating by boundary layer moisture convergence [9] although there is debate [7].

Corresponding author: Bohar Singh, bsingh5@gmu.edu
George Mason University, Fairfax, VA.

In contrast, the dynamics of the MISO is governed primarily by barotropic cyclonic vorticity and easterly wind shear [5]. MJO and MISO occur in different seasons, at different latitudes and are governed by different mechanisms. MJO and MISO also have a different phase speed and direction of propagation. TISO has great importance because of its influence on the highly populated and largely agrarian economies of the tropics, where it regulates wet and dry spells of rainfall ([3]; [2]; [4]), which has a direct relationship with crop production. MJO also affects the tropical cyclone activity in all the ocean basins ([6]; [8]), El Nino Southern Oscillation (ENSO) as in [12] and extra-tropical weather and climate. Understanding of TISO is very crucial to realize the dream of seamless perdition. Most of the commonly used diagnostics to understand TISO consider dimensional reduction by seasonal and spatial averages or empirical orthogonal function decomposition. This approach may miss some information regarding direction of propagation, location and phase speed.

As an alternative, we examine TISO by tracking each event and compositing events on the basis of direction of propagation. The characterstics of TISO such as prefered geographical loaction of propagation, phase speed, life span, regions of initiation and dissipation and their seasonal and intraannual variability are analyzed in this study.

II. DATA AND METHOD

To identify and track TISOs, 34 years (1979-2013) of daily outgoing long-wave radiation (OLR) data from the National Oceanic and Atmospheric Administration (NOAA; 2.5° X 2.5° grid) is used. Anomalous OLR is considered a proxy for large-scale tropical convective anomalies [1] and [10], because negative OLR anomalies (OLRA) are well correlated with convective clouds. Daily OLR anomalies are obtained by removing the first four annual harmonics from the data at each gridpoint. The data is band pass filtered using a Lanczos filter in the 20-100 day band to obtain intra-

seasonal anomalies. Finally OLRA are spatially smoothed using a 9-point weighted average. An event is classified as TISO if it satisfies the following three criteria, which are quite similar to [9]: (a) Life span at least 20 days; (b) During its lifetime, the zonal dimension exceeds 30° longitude and the mean OLRA remains less than -15 W/m^2; (c) At the strongest stage, the zonal dimension exceeds 50° longitude and the central intensity is less than -25 W/m^2.

Sr. No	Class	Events	Ave. Speed	Ave. Life Span
1	Eastward	71	4.04	33
2	Northward	96	2.28	24

Table 1: Characteristics of Tropical intraseasonal oscillation

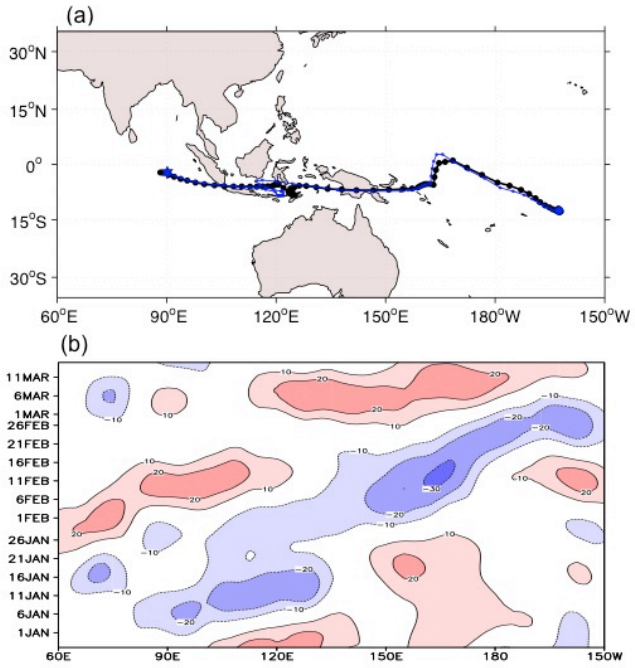


Figure (1): (a) Track of eastward propagating TISO identified visually (black) and identified by algorithm (blue), (b) Hovmoller diagram averaged between 15°S-15°N for the same dates as in (a)

A systematic framework using a motion-based multiple-object tracking algorithm, as given in [11], has been developed to identify, track and classify tropical intra-seasonal oscillations. The method is applied to intra-seasonally filtered daily and spatially smoothed OLRA data from 35°S to 35°N to track every individual TISO event.

The steps involved in tracking are as follows:

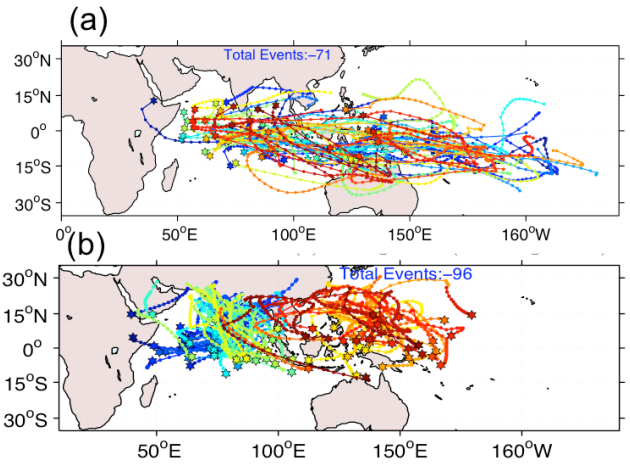


Figure (2): (a) Tracks of eastward propagating and (b) northward propagating TISOs identified by the algorithm

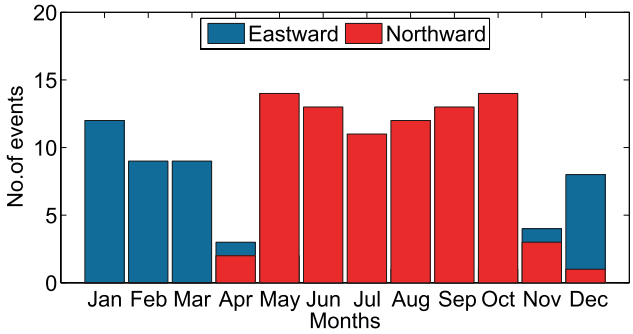


Figure (3): Frequency of occurrence of the eastward and northward propagating TISOs

1) To make this process autonomous, each daily frame is considered as an image from a static camera.
2) Group of connected pixels in each image is detected by using blob analysis after background subtraction, which is already performed during anomaly calculations. Group of connected pixels is considered as an object (clouds cluster)
3) The smallest size of cloud cluster (number of pixel), which can be tracked by algorithm, is controlled by a size threshold parameter currently set at size of 15 pixels.
4) Tracks are initialized for each region and properties like area, centroid, mean intensity (OLRA) and date are stored.
5) Cloud clusters can change size and shape, vanish or originate form frame to frame So after detection, the position and velocity of each centroid is predicted in the next frame using a

Kalman filter with constant velocity dynamical model.

6) To associate the position of a given cloud cluster to a cloud cluster in next frame, the assignment problem is solved by calculating the minimum distance between the predicted cloud cluster and all other cloud cluster in that frame.
7) Distance of association, is a maximum threshold up to which two cloud cluster can be associated. We are using 6 grid points in the algorithm
8) A correction is applied on the basis of measurement and prediction variance, and re-labeling is performed.
9) A cloud cluster is presumed to be dead if none of the cloud clusters from the current frame is assigned (not found within the distance of association) to a cloud cluster from the previous frame.
10) A cloud system is called newly originated, if it is not assigned to any track and a new track is initialized.
11) For all the cloud systems in the current frame, the position of each is again predicted and the algorithm repeats.

After obtaining all the tracks, they can be classified according to the objective criteria like propagation direction, lifetime, and intensity. Each class of tracks can be used to develop climatology for that class. The climatology of each class can used to investigate the dynamical characteristics of that class by compositing it with other variables.

III. EVALUATION

In this section we summarize the results of applying the objective-tracking algorithm to daily NOAA OLR data from 1979-2013. As shown in Table 1, the objective method identifies 71 eastward-moving and 96 northward-moving TISOs. The average speed for eastward-moving TISOs is 4.0 m/s, with duration of about 33 days. Northward-moving TISOs propagate at about 2.0 m/s and last about 24 days before dying in the high latitudes. Both the speed and average life span of events detected by the objective-tracking algorithm are in good agreement with results reported in previous studies [9] and [4]. Fig (1a) is showing an eastward propagating event identified by visual analysis and than algorithm is ran for those dates. As we can see that algorithm successfully identified the same event. The very same event can also be confirmed with conventional hovemoller diagram in Fig (1b). Advantage of detection with algorithm is that we can save more information about the event like start date, end date, exact geographical location of the track, velocity, intensity and size of the cloud cluster on each day as compared to hovemoller identification. This information about each TISO event may help us to understand better about it in further analysis. Fig. (2) shows the tracks of eastward-moving TISOs from objective-tracking algorithm. Tracks from the method look similar in term of geographical location of occurrence, initiation and dissipation as in previous [5] and [2]. Most eastward-moving tracks are found south of the equator between 0° and 15° S, beginning in the western to central Indian Ocean and propagating eastward to the South Pacific Convergence Zone (SPCZ). Northward propagation can be seen in both of the ocean basins, while initiation more often occurs in the Indian Ocean. Two types of northward propagation happen in the Indian Ocean sector: (a) Moving northward immediately after initiation; and (b) Moving eastward at first, then turning northward. TISOs originating in the Pacific Ocean sector propagate directly northward only. After initiation, northward propagating TISOs die after crossing 20°N. Eastward-moving TISOs occur more often in boreal winter (NDJFMA), but they can occur throughout the year (Fig. 2). Northward propagation is more common in boreal summer (MJJASO), but sometimes it occurs in November and December.

In this method of tracking, no assumption is made regarding seasonality of TISO. Events are solely classified on the basis of direction of propagation, and therefore seasonality is confirmed naturally. We are not missing any event that happens outside the predefined season as opposed to other conventional methods. Advantage of this method over hovemoller and any index identification method (MJO identified using any area averaged index) is that it can give us actual track of a MJO event with daily position of centroid (center of mass of cloud system), mean intensity (OLR), minimum intensity (OLR), daily phase speed, size of tracked cloud cluster (number of grid points), positions of initiations and dissipations, actual date occurrence and number of days that event remains active for, which is not possible in both of above mentioned methods. This method has some limitations; such as it cannot track an event that bifurcates into two tracks, which are possible in some TISO cases. Sometimes algorithm can lose track of TISO when it becomes weak while crossing maritime continent.

ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the National Science Foundation (ASG-1338427) for this study.

REFERENCES

[1] Lau, K. M., and P. H. Chen, "Aspects of the 40-50 day oscillation during northern summer as inferred from outgoing longwave radiation", *Mon. Wea. Rev.*, vol. 114, pp. 1354–1367, 1986.

[2] Lawrence, D. M., and P. J. Webster, "The boreal summer intraseasonal oscillation and the south Asian monsoon". J. Atmos. Sci., 59, pp 1593–1606, 2001.

[3] D R Sikka and Sulochana Gadgil, "On the maximum cloud zone and the ITCZ over India longitude during the Southwest monsoon", *Mon. Weather Rev.*, Vol.108, pp.1840–1853, 1980.

[4] Jones, C., L. M. V. Carvalho, R.W. Higgins, D. E. Waliser, J.K.E. Schemm: "Climatology of tropical intra-seasonal convective anomalies 1979-2002," *J. Climate*, vol. 17, pp. 523–539, 2004a.

[5] Jiang, X., T. Li, and Bin Wang, "Structure and Mechanisms of the northward-propagating intraseasonal oscillations," *J. Climate*, vol. 17, pp. 1022–1039, 2004.

[6] Maloney, E. D., and A. H. Sobel, "Surface fluxes and ocean coupling in the tropical intraseasonal oscillation", J. Climate, 17, pp. 4368– 4386, 2000.

[7] Pritchard, M. S. and D. Yang, "Response of superparmetrized Madden Julian ocscillation to extreme climate basic state variation challenges moisture mode view," *J. Climate*, vol. 29, pp. 4995–5008, 2016.

[8] Philip J. Klotzbach, " On the Madden–Julian Oscillation–Atlantic Hurricane Relationship", *J. Climate,* **23**, pp 282–293, doi: 10.1175/2009JCLI2978.1,2010.

[9] Wang, B., and H. Rui, "Synoptic climatology of transient tropical intra-seasonal convection anomalies," *Meteor. Atmos. Phys.*, vol. 44, pp. 44–61, 1989.

[10] Waliser, D.E., N. E. Graham and C. Gautier, "Comparison of highly reflective cloud and outgoing longwave dataset for use in estimating tropical deep convection," *J. Climate*, vol. 6, pp. 331–353, 1993.

[11] Yilmaz, A.. O. Javed, and M. Shah, "Object tracking: A survey," *ACM Computing Surveys: Atmospheres,* vol. 38, no. 4, pp. 1–45, 2006.

[12] Zhang, C., and J. Gottschalck, "SST anomalies of ENSO and the Madden–Julian oscillation in the equatorial Pacific", *J. Climate*, **15, pp** 2429–2445, 2002.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

ANALYSIS OF AMAZONIA DROUGHTS USING SUPERVISED KERNEL PRINCIPAL COMPONENT ANALYSIS

Carlos H. R. Lima[1,2], Amir AghaKouchak[2]

***Abstract*—Extreme droughts in Amazonia seem to become more frequent in the last years and have been associated with local and global impacts on society and the ecosystem. Here we try to better understand the dynamics and causes of Amazonia droughts by analyzing the moisture and heat fluxes that cross the region. Particularly, we decompose the high-dimensional moisture fluxes on the boundaries of the Amazonia region into a low-dimensional space using supervised kernel principal component analysis, where the side information is provided by the gridded PDSI drought index over Amazonia. Subsequently, we apply K-means to cluster the first two modes into three groups. The distribution of drought indexes (PDSI, SPI, SSI and MSDI), temperature and rainfall over Amazonia associated with each cluster is then analyzed. The results reveal at least three distinguished patterns in the moisture and heat fluxes crossing Amazonia that are associated with extreme drought conditions. These findings could not be obtained using standard PCA or from first clustering the response variable (i.e. drought indices). Furthermore, the manuscript offers insights into the dynamics and causes of Amazonia droughts.**

I. MOTIVATION

It is well recognized that the dynamics of Amazonia ecosystem plays a significant role on biogeochemical cycles [1], moisture transport [2] and on the regional climate of distant regions [3]. Extreme droughts in the Amazonia seem to become more frequent in the last years [4] and have the potential to trigger a large number of fires and cause extensive impacts on society and the ecosystem [5], such as the events of 2005 and 2010.

The major cause of droughts in Amazonia is related to the El Niño-Southern Oscillation (ENSO) [4] and to a minor extent to the sea surface temperature (SST) variability in the Tropical North Atlantic. Warm SST anomalies in the eastern Tropical Pacific shifts the descending branch of the Walker circulation over Amazonia and inhibits precipitation during the austral summer rainfall season [6]. A warmer tropical north Atlantic will displace north the Inter-Tropical Convergence Zone from its climatological position and therefore the ascending branch of the Hadley cell and reduce convection and precipitation over Amazonia.

However, the role of moisture and heat fluxes on Amazonia droughts has not been extensively explored in the literature. This work is carried out to better understand how these fluxes affect the temperature, rainfall and drought indices over Amazonia. This is accomplished using the supervised kernel principal component analysis (SKPCA), which is a variant of PCA to deal with nonlinearities and include side information.

Corresponding author: C Lima, University of Brasilia, chrlima@unb.br [1] Department of Civil and Environmental Engineering, University of Brasilia [2] Department of Civil and Environmental Engineering, University of California, Irvine

II. DATA

A. Moisture and Heat Fluxes

Vertically integrated moisture and heat fluxes are obtained from the ERA-Interim reanalysis data [7]. It covers the period January/1980 – December/2013 and are retrieved for the Amazonia boundaries (north, south, east and west edges) as defined in Figure 1 (dashed line). The moisture fluxes on each edge are combined to form the input matrix X (size 408 data points versus 340 dimensions) to the SKPCA model.

B. Drought Indices

Drought indices are derived in order to reveal specific features of drought conditions. Here we use four gridded (over the Amazonia region defined in Fig. 1) drought indices: Palmer Drought Severity Index (PDSI [8]); Standardized Precipitation Index (SPI [9]); Standardized Soil Moisture Index (SSI [10]) and Multivariate Standardized Drought Index (MSDI [11]).

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

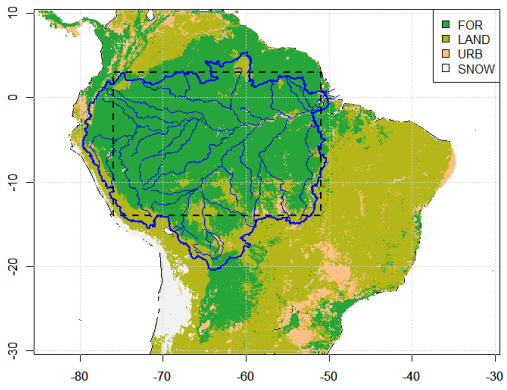


Fig. 1. Grid domain (dashed line) defining the Amazonia region for this study. The Amazonia watershed is defined by the blue line. The different land cover types follow the NASA land cover maps.

PDSI data is provided by the NOAA/OAR/ESRL PSD, Boulder, at http://www.esrl.noaa.gov/psd/. SPI, SSI and MSDI data are available at the Global Integrated Drought Monitoring and Prediction System (GIDMaPS, http://drought.eng.uci.edu/). We refer the reader to the references cited above for more details regarding each drought index. The gridded PDSI (matrix size: 408 x 70) is used as side information for the SKPCA model. For the subsequent analysis, all the indices are averaged over the Amazonia area as defined in Figure 1.

C. Rainfall and Temperature

Monthly gridded temperature and rainfall data for the period 1980–2013 are provided by [12]. These data consist of interpolated daily rainfall and temperature observations from 3625 rainfall gauges and 735 weather stations across Brazil available from different institutions (INMET, ANA and DAEE). The interpolation schemes and validation procedures are described in [12]. The geographical region delimited by the dataset is the Amazonia boundary as shown in Figure 1 and the spatial average is used in the subsequent analysis.

III. Technical Approach

The kernel PCA is an extension of PCA [13] designed for dealing with nonlinear data through the use of kernels [14]. It consists of a nonlinear mapping of the input data onto a linear space, in which PCA can be freely applied. The *kernel trick* is employed in order to avoid the explicit mapping of the input data coordinates onto the feature space. We use the Gaussian (RBF) kernel, which is defined for two points $\mathbf{x}$ and $\mathbf{x}'$ of moisture flux data as:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\sigma^2}\right), \qquad (1)$$

where $||\cdot||$ is the Euclidean norm and σ is a parameter often called the width of the kernel, which is arbitrarily set here as the sample standard deviation of the input data.

In order to highlight the main modes of the moisture fluxes that are most associated with Amazonia droughts, we use the so called supervised PCA [15], where the PCs are obtained considering the maximum dependence on the response variable (side information). This has a similar interpretation of canonical correlation analysis, but has more advantages, particularly when the number of variables exceeds the number of observations [15]. Essentially, the procedure to obtain the supervised PCs consists of an eigen-decomposition of a product matrix of input and side information data. We refer the reader to [15] for the mathematical details. Here we use the gridded PDSI as side information and the code developed by [15].

Once the SKPCA modes are obtained, we apply K-means [13] to the first two modes to find clusters in the reduced space. The distribution of moisture and heat fluxes, drought indices, rainfall and temperature associated with each cluster are then analyzed.

IV. Evaluation

A. Clustering, Moisture and Heat Fluxes

The K-means clustering of the first and second modes of the moisture fluxes considering the gridded PDSI as side information is displayed in Figure 2. There is a linear structure between the two modes and a continuous density of points partitioned into three clusters. Hereafter we will refer to them as clusters 1 (red dots), 2 (blue dots) and 3 (green dots). The vertically integrated northward fluxes of moisture and heat averaged over the months correspondent to each of these clusters are shown in Figure 3. There is a clear separation among the clusters. A stronger moisture and heat inflow is observed for cluster 1 (red line in top panels of Fig. 3) while a stronger outflow occurs in clusters 2 and 3 (green and blue lines in bottom panels of Fig. 3). The vertically integrated eastward fluxes of moisture and heat are displayed in Figure 4. The most intense westward moisture flux on both east and west boundaries is observed for cluster 3, followed by clusters 2 and 1. A similar pattern is observed for the heat fluxes (bottom panels in Fig. 4) up to 7°S, where the patterns are then reversed.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

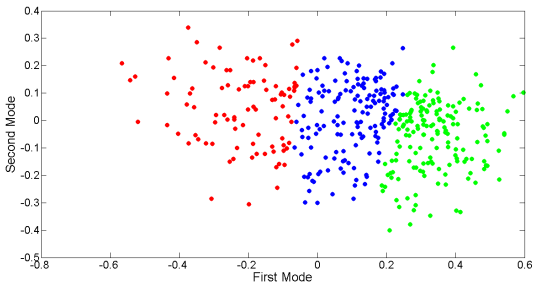


Fig. 2. K-means clustering of first and second modes after applying SKPCA to the moisture fluxes over Amazonia.

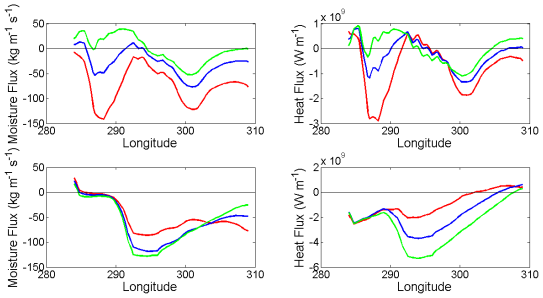


Fig. 3. Vertically integrated northward fluxes of moisture and heat on the north (top panels) and south (bottom panels) Amazonia boundaries as defined in Fig. 1. The colors represent the clusters showed in Fig. 2.

B. *Drought Indices, Rainfall and Temperature*

The distribution of the drought indices, rainfall and temperature associated with each cluster of Figure 2 is shown in Figure 5. In general, clusters 1 to 3 are associated with wet to dry conditions as per the drought indices (high to low values), and more significant with high (low) to low (high) rainfall (temperature). Essentially, the remarkable patterns of moisture and heat fluxes as displayed in Figures 3 and 4 for cluster 3 - reduced inflow on the north boundary, stronger outflow on the south boundary and stronger inflow on the east boundary - are associated with more intense drought conditions: low rainfall and high temperatures.

This conclusion is more evident when we analyze

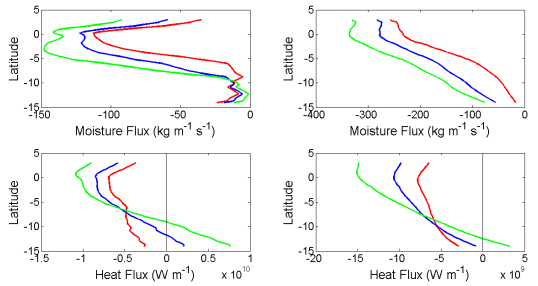


Fig. 4. As in Fig. 3, but for the vertically integrated eastward fluxes of moisture and heat on the east (right panels) and west (left panels) Amazonia boundaries.

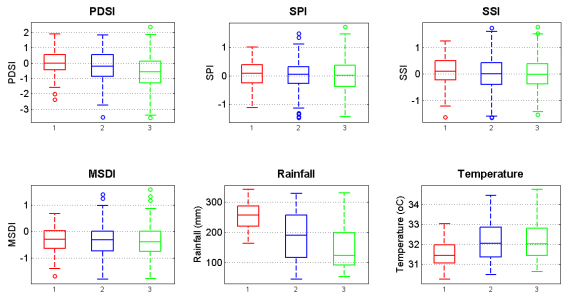


Fig. 5. Distribution (period 1980–2013) of PDSI, SPI, SSI, MDSI, rainfall and temperature (averaged over Amazonia) according to the clusters showed in Fig. 2.

such extreme events. In figure 6 we take the 10% (monthly data) most extreme events of each variable associated with drought conditions (rainfall and drought indices in the lower tail of the distribution, i.e., values below the 10th quantile; temperature in the upper tail of the distribution, i.e., values above the 90th quantile) and calculate the frequency of these extreme events in each of the clusters. Clearly, clusters 3 and 2 are associated with the most extreme droughts as measured by all variables analyzed. A hypothesis test (assuming independent events) reveals that the moisture and heat patterns of cluster 3 are statistically associated with the most extreme events of PDSI, MSDI, rainfall and temperature over Amazonia.

Finally, if we follow a baseline approach and first apply $K-$means to cluster the spatially averaged PDSI, we are able to find distinguished clusters for this variable as well as for rainfall and temperature (Fig. 7), but there is no clear separation among the clusters associated with the moisture and heat fluxes. Hence, when averaging across the associated events (dry, wet and average PDSI), this baseline approach does not find any particular moisture and heat flux pattern associated with Amazonia moisture conditions, suggesting a minor or no role of these variables and that other factors are responsible for Amazonia droughts. On the other hand, the SKPCA method adopted here was able to show the most likely moisture and heat flux patterns connected with droughts, highlighting then its relevance as technical approach and revealing that such fluxes certainly play a significant role on Amazonia droughts, particularly the extreme ones. This new understanding will also help future studies in finding the other climate and ecosystem factors behind Amazonia droughts.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

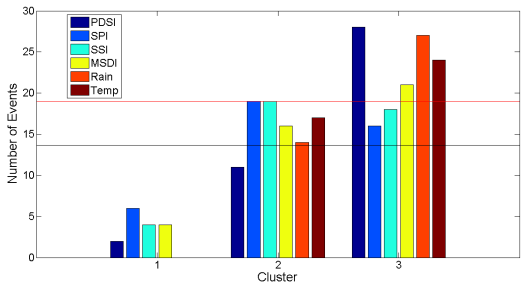


Fig. 6. Frequency of the top 10% extreme events for PDSI, SPI, SSI, MDSI, rainfall and temperature for each cluster. The black and red lines show, respectively, the 50% and 95% quantiles assuming a multinomial distribution with equally probable clusters.

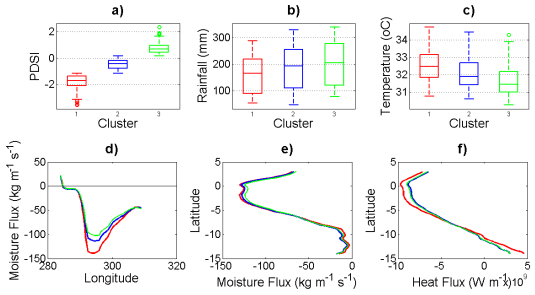


Fig. 7. a) K-means clustering applied to PDSI and associated clusters for b) rainfall; c) temperature; d) northward moisture flux on south edge; e) eastward moisture flux on east edge and f) eastward heat flux on east edge.

ACKNOWLEDGMENTS

We thank all agencies and authors that provide dataset and codes. The first author acknowledges a Postdoctoral Fellowship from the Brazilian Government Agency CNPq during part of this work.

REFERENCES

[1] M. E. McClain, R. Victoria, and J. E. Richey, *The Biogeochemistry of the Amazon Basin*. Oxford University Press, 2001.

[2] A. Drumond, R. Nieto, L. Gimeno, and T. Ambrizzi, "A Lagrangian identification of major sources of moisture over Central Brazil and La Plata Basin," *Journal of Geophysical Research*, vol. 113, p. D14128, July 2008.

[3] C. A. Nobre, P. J. Sellers, and J. Shukla, "Amazonian deforestation and regional climate change," *Journal of Climate*, vol. 4, no. 10, pp. 957–988, 1991.

[4] J. A. Marengo and J. C. Espinoza, "Extreme seasonal droughts and floods in Amazonia: causes, trends and impacts," *Int. J. Climat.*, vol. 36, pp. 1033–1050, 2016.

[5] M. A. Cochrane, "Fire science for rainforests," *Nature*, vol. 421, pp. 913–919, 2003.

[6] C. Ropelewski and M. Halpert, "Global and Regional Scale Precipitation Patterns Associated with the El Niño/Southern Oscillation," *Mon. Wea. Rev.*, vol. 115, pp. 1606–1626, 1987.

[7] D. P. Dee, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Holm, L. Isaksen, P. Kallberg, M. Kohler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thepaut, and F. Vitart, "The era-interim reanalysis: configuration and performance of the data assimilation system," *Quarterly Journal of the Royal Meteorological Society*, vol. 137, no. 656, pp. 553–597, 2011.

[8] A. Dai, K. E. Trenberth, and T. Qian, "A global data set of Palmer Drought Severity Index for 1870-2002: Relationship with soil moisture and effects of surface warming," *J. Hydrometeorology*, vol. 5, pp. 1117–1130, 2004.

[9] T. B. McKee, N. J. Doesken, and J. Kleist, "The relationship of drought frequency and duration of time scales," in *Eighth Conference on Applied Climatology*, (Anaheim CA), pp. 179–186, American Meteorological Society, 1993.

[10] Z. Hao and A. AghaKouchak, "Multivariate Standardized Drought Index: A Parametric Multi-Index Model," *Advances in Water Resources*, vol. 57, pp. 12–18, 2013.

[11] Z. Hao and A. AghaKouchak, "A Nonparametric Multivariate Multi-Index Drought Monitoring Framework," *Journal of Hydrometeorology*, vol. 15, no. 1, pp. 89–101, 2014.

[12] A. C. Xavier, C. W. King, and B. R. Scanlonc, "Daily gridded meteorological variables in Brazil (1980–2013)," *Int. J. Climatol.*, vol. 36, pp. 2644–2659, 2016.

[13] T. Hastie, R.Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2001.

[14] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299 – 1319, 1998.

[15] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Z. Jahromi, "Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds," *Pattern Recognition*, vol. 44, no. 7, pp. 1357 – 1371, 2011.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

A BAYESIAN PREDICTIVE ANALYSIS OF DAILY PRECIPITATION DATA

Sai K. Popuri[1], Nagaraj K. Neerchal[1], Amita Mehta[2]

***Abstract*—We develop a Bayesian predictive model for data that features a point mass at zero (semi-continuous data), where the predictions are also semi-continuous. The procedure is illustrated for a Tobit modeling algorithm using data augmentation, and Gibbs sampling. We apply the procedure to the daily precipitation data at a location in the upper Missouri River Basin (MRB) region to generate predictions of daily rainfall. Historical simulated data by MIROC5, a Global Climate Model (GCM), is used as a covariate in our model. We further compare the accuracy of our predictions with a few frequentist methods using a criteria suitable for semi-continuous data.**

I. MOTIVATION

Predictions of daily precipitation are often required as an input to hydrological modeling tools (ex.: Soil and Water Assessment Tool (SWAT)[1]) for regional hydrological assessment studies. One of the methods to predict precipitation is to use simulated historical data provided by GCMs as covariates in regression models with the observed precipitation as the response. A common approach is to forecast precipitation at the monthly level, and use a 'weather generator' (ex.:[1],[2]) to simulate daily precipitation in a manner consistent with the monthly forecasts. Here we are interested in predicting at the daily level instead. This introduces challenges because the data is semi-continuous, and we seek predictions that also feature a point mass at 0.

One simple way to predict data with a point mass at 0 is to ignore the semi-continuous nature of the data, and fit a linear regression model with normal errors. Predictions can then be made by treating the negative predictions (after plugging in the estimated parameters) as 0 values. A better approach would be is to account for the point mass, and accordingly assume a parametric model, for example, a Tobit model ([3]), and maximize the resulting likelihood function. Predictions can then be made by treating the negative latent predictions as 0 values. An alternate approach is to minimize an L_1 loss (Least Absolute Deviations) for the Tobit model ([4]), and make predictions in a similar fashion by thresholding at zero. While thresholding at 0 results in semi-continuous predictions, it is adhoc. Also, the distributional properties of the resulting predictions are not clear. In this paper we propose a more constructive method to generate semi-continuous predictions by taking a Bayesian approach. As a result, predictions are semi-continuous by design and a post-estimation adhoc thresholding can be avoided. We illustrate the implementation using data augmentation, and Gibbs sampling to estimate the posterior predictive distribution for a Tobit model. We apply the method to the daily precipitation data, and compare the accuracy of our predictions with the frequentist methods mentioned above.

Corresponding author: Sai Popuri, saiku1@umbc.edu
[1]Department of Mathematics and Statistics, University of Maryland, Baltimore County [2]Joint Center for Earth Systems Technology, Baltimore, MD

II. METHOD

Consider a random variable Y with support $\{0\} \cup (0, \infty)$, that is, it takes the value 0 with a positive probability, and follows a continuous probability distribution on $(0, \infty)$, if greater than 0. We call such a random variable a semi-continuous random variable. Let $\mathbf{x} = (x_1, x_2, .., x_p)$, a p-dimensional vector of covariates. The density of Y given the covariate $\mathbf{x}$ can be written as:

$$f(y \mid \mathbf{x}, \theta) = \delta(y)\gamma(\mathbf{x}, \theta) + \delta^*(y)[1 - \gamma(\mathbf{x}, \theta)]g_{\eta(\mathbf{x},\theta)}(y), \quad (1)$$

where $\delta(y)$ is the indicator function that is 1 if $y = 0$, $\delta^*(y) = 1 - \delta(y)$, and g is the density of a continuous random variable with support on $(0, \infty)$ governed by the parameter θ.

Let $\mathbf{y} = (y_1, .., y_n)$ be an i.i.d. sample of size n of semi-continuous data, $X = [\mathbf{x}_1\mathbf{x}_2..\mathbf{x}_n]^T$ (treated as fixed), the $n \times p$ matrix of covariates, $\tilde{y}$ a new observation, and $\tilde{\mathbf{x}}$ the corresponding new covariate. Let the parameter space of θ be Θ. Let $p(\theta)$ be a prior on θ, and $p(\theta \mid \mathbf{y}, X)$ the posterior of θ. If the posterior of θ is proper, then the posterior predictive density, defined by, $f_{\tilde{Y}|\tilde{x},\mathbf{y},\mathbf{X}}(\tilde{y}) = \int_\Theta f(\tilde{y} \mid \tilde{x}, \theta)p(\theta \mid \mathbf{y}, X)d\theta$, is also

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

TABLE I: Goodness of fit criteria

	Pred. $\hat{y}=0$	**Pred.** $\hat{y}>0$
Obs. $y=0$	$\frac{n_{00}}{n_0}$	$\frac{1}{n_0-n_{00}}\sum_{y_i=0,\hat{y}_i>0}\hat{y}_i$
Obs. $y>0$	$\frac{1}{n_1-n_{11}}\sum_{y_i>0,\hat{y}_i=0}y_i$	$\frac{n_{11}}{n_1}, MSPE, MAD$

semi-continuous, and can be written as:

$$\delta(\tilde{y})\gamma(\tilde{x},\mathbf{y},\mathbf{X})+\delta^*(\tilde{y})(1-\gamma(\tilde{x},\mathbf{y},\mathbf{X}))g_{\tilde{Y}|\tilde{x},\mathbf{y},\mathbf{X}}(\tilde{y}), \quad (2)$$

where

$$\gamma(\tilde{x},\mathbf{y},\mathbf{X})=\int_\Theta \gamma(\tilde{x},\theta)p(\theta\mid\mathbf{y},\mathbf{X})d\theta$$

is a function with range in $(0,1)$, and

$$g_{\tilde{Y}|\tilde{x},\mathbf{y},\mathbf{X}}(\tilde{y})=\frac{1}{1-\gamma(\tilde{x},\mathbf{y},\mathbf{X})}\int_\Theta(1-\gamma(\tilde{x},\theta))g_{\eta(\tilde{x},\theta)}(\tilde{y})p(\theta\mid\mathbf{y},\mathbf{X})d\theta$$

is the density of a continuous random variable with support on $(0,\infty)$.

Due to the discrete/continuous mixed nature of the semi-continuous data, goodness-of-fit criteria to measure accuracy of predictions becomes multi-faceted as depicted in Table I. Let $\hat{y}$ be a semi-continuous prediction of y_i, a semi-continuous random variable, $i=1,..,n$. Let n_0, and n_1 be the number of zero, and positive observations in the sample respectively. Let n_{00} be the number of observations with value 0 and whose predictions are also 0, and n_{11} be the number of positive observations whose predicted values are also positive. In Table I, Mean Squared Prediction Error (MSPE) is $\frac{1}{n}\sum_{Y_i>0,\hat{Y}_i>0}(Y_i-\hat{Y}_i)^2$, and Mean Absolute Deviations (MAD) is $\frac{1}{n}\sum_{Y_i>0,\hat{Y}_i>0}\mid Y_i-\hat{Y}_i\mid$. We use the criteria in the Table for our analysis.

Predictive density for the Tobit model

One way to realize semi-continuous data is to assume that the observed y_i depends on $\mathbf{x}_i$ via a latent random variable y_i^*, which is assumed to be normal with mean $\beta\mathbf{x}_i$, where β is the $p-$dimensional vector of regression coefficients, and variance σ^2 as shown in equation (3). This model for y_i is know as the Tobit model in Econometrics ([3]).

$$\begin{aligned} y_i^* &= \beta\mathbf{x}_i+u_i \\ y_i &= \begin{cases} y_i^*, & \text{if } y*_i>0 \\ 0, & \text{if } y_i^*\le 0, \end{cases} \end{aligned} \quad (3)$$

where u_i is i.i.d. normal with mean 0, and variance σ^2. Under this model, the density of $y_i\mid\mathbf{x}_i,\theta$ is given by equation (1) with $\theta=(\beta,\sigma^2)$, $\gamma(\mathbf{x}_i,\theta)=1-\Phi\left(\frac{\beta\mathbf{x}_i}{\sigma}\right)$, and $g_{\eta(\mathbf{x}_i,\theta)}$ as truncated (>0) normal with mean $\beta\mathbf{x}_i$, and variance σ^2.

We extend the Bayesian analysis of the Tobit model described in [5] with a numerical approximation to the posterior predictive distribution. Let $\mathbf{y}^+=\{y_i: i\ni y_i>0\}$, C be the index set of zero valued observations. Let $\mathbf{z}=\{z_j: j\in C\}$ be the latent parameters representing the unobserved y_j^* for $y_j=0$. Therefore $\mathbf{y}^*=(\mathbf{z},\mathbf{y}^+)$ is the augmented complete data. Assuming a non-informative prior $p(\beta,\sigma^2)\propto\sigma^{-2}$, $\mathbf{X}^T\mathbf{X}$ is non-singular, and given $\mathbf{z}$, the posterior of (β,σ^2) can be obtained as ([6]):

$$\begin{aligned} \beta\mid\sigma^2,\mathbf{y}^*,\mathbf{X} &\sim N(\hat{\beta},\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}) \\ \sigma^2\mid\mathbf{y}^*,\mathbf{X} &\sim Inv-Gamma(\frac{n-p}{2},\frac{(\mathbf{y}-\mathbf{X}\hat{\beta})^T(\mathbf{y}-\mathbf{X}\hat{\beta})}{2}), \end{aligned} \quad (4)$$

where $\hat{\beta}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}^*$. If the conditional distributions of individual parameters given rest of the parameters, and the data are available, and since $n>p$, and $\mathbf{X}$ is full rank, Gibbs sampling ([6]) in Algorithm 1 enables us to draw approximate samples from the joint distribution of $(\mathbf{z},\theta)$. In Algorithm 1, let $\theta^{(0)}$ be the initial value of θ, and $\mathbf{z}^{(0)}$ are initialized to samples from truncated (<0) normal with parameters $\theta^{(0)}$.

Algorithm 1 Gibbs sampler

Step 1: Update $\mathbf{z}^{(k+1)}$ by drawing from truncated (<0) normal with mean $\beta^{(k)}\mathbf{x}_j$, and variance $\sigma^{2^{(k)}}$, for each $j\in C$.
Step 2: Update $\sigma^{2^{(k+1)}}$ by drawing from the conditional (given $\mathbf{z}^{(k)},\mathbf{y}^+$) in equation (4).
Step 3: Update $\beta^{(k+1)}$ by drawing from the conditional (given $\sigma^{2^{(k+1)}},\mathbf{z}^{(k)},\mathbf{y}^+$) in equation (4).

Based on L independent samples from the posterior of $\theta=(\beta,\sigma^2)$, we estimate the components of the posterior predicitive distribution in equation (2) as:

$$\begin{aligned} \hat{\gamma}(\tilde{\mathbf{x}},\mathbf{y},\mathbf{X}) &= L^{-1}\sum_{j=1}^{L}\gamma(\tilde{\mathbf{x}},\theta_j) \\ g_{\tilde{Y}|\mathbf{x},\mathbf{y},\mathbf{X}} &= \frac{1}{1-\hat{\gamma}(\tilde{\mathbf{x}},\mathbf{y},\mathbf{X})}L^{-1}\sum_{j=1}^{L}(1-\gamma(\tilde{\mathbf{x}},\theta_j))g_{\eta(\tilde{\mathbf{x}},\theta_j)}, \end{aligned} \quad (5)$$

where $\gamma(\tilde{\mathbf{x}},\theta_j)=1-\Phi\left(\frac{\beta_j\tilde{\mathbf{x}}}{\sigma_j}\right)$, and $g_{\eta(\tilde{\mathbf{x}},\theta_j)}$ is truncated (>0) normal with mean $\beta_j\tilde{\mathbf{x}}$, and variance σ^2. Since semi-continuous predictions are desired, we use the median of simulations from the (approximate) predictive density as predictions.

We fitted the procedure described above on datasets simulated from a few Tobit models (results are not shown here). We found that the maximum likelihood estimates (MLE) of the parameters in the Tobit model are covered around the modes of the corresponding posteriors. Also, the predictions from the Bayesian method agreed well with the Tobit MLE plug-in predictions based on the criteria in Table I.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

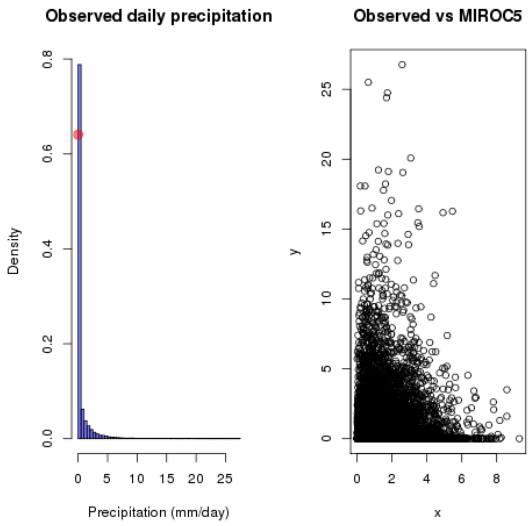


Fig. 1: Observed vs MIROC5

III. EVALUATION

We analyze the daily precipitation data at $-109.8125°$ W, $41.4375°$ N (Rock Springs, WY) using the Bayesian procedure for the Tobit model described in section II. This data was analyzed in [7] using Multiple Linear Regression (MLR), and the Tobit maximum likelihood estimation (MLE). We compare these methods with the Bayesian method for predictive accuracy using the criteria in Table I. In addition, we also fit a regression model using the Least Absolute Deviation (LAD) criterion ([4]). Observed daily precipitation (measured in mm/day) is provided by [8], and has a temporal coverage of $1949 - 2005$. Data from $1949 - 2000$ is used for model fitting, and $2001 - 2005$ for evaluation. The covariate daily precipitation data provided by MIROC5 has the same temporal coverage. Figure 1 shows the histogram of the observed daily precipitation, and a scatter plot of the observed data against the covariate data from MIROC5. The observed data has approximately 65% (indicated by the red dot in the histogram) of zeroes, and is heavy tailed with a few very large intensities of rainfall. The scatter plot suggests a complex relationship between the observed, and the covariate data. Despite the suggestion of a possible non-linear relationship, we fit a Tobit model primarily to illustrate the proposed methodology.

Let y_t be the observed precipitation for t^{th} day starting from $01/01/1949$. Let x_t be the corresponding precipitation provided by MIROC5. The regression model considered in [7] is

$$y_t = \beta^0 + \beta^1 x_t + \sum_{k=2}^{12} \alpha_k m_{tk} + \sum_{k=2}^{12} \gamma_k m_{tk} x_t + u_t \qquad (6)$$

where the dummy variables m_{tk}, $k = 1, 2, .., 11$ represent the month effects and the errors $\{u_t\}$ are assumed to be i.i.d $N(0, \sigma^2)$. The Tobit model (equation 3) has the mean term from equation (6) for the latent process. The LAD model is similar to the MLR model in equation (6) with the mean absolute deviations as the risk criterion.

Let $\theta = (\beta_0, \beta_1, \alpha_2, .., \alpha_{12}, \gamma_2, .., \gamma_{12})$ be the vector of parameters in equation (6), and $\hat{\theta}$ the point estimate from MLR, Tobit MLE, or LAD models. Then the forecast of y at a future time f based on the MLR, Tobit MLE, or the LAD model is

$$\hat{y}_f = y_f(\hat{\theta}, x_f) I(y_f(\hat{\theta}, x_f) > 0), \qquad (7)$$

where x_f is the covariate at time f, and $y_f(\hat{\theta}, x_f)$ is the mean term in equation (6) evaluated at $\hat{\theta}$.

The Bayesian procedure for the Tobit model described in section II was fitted using Gibbs sampling. Twenty chains were run for $3,000$ iterations each with initial values for (β, σ^2) set at disparate locations in the parameter space. We burn in 500 iterations in each chain and thin every 5^{th} simulation to collect a posterior sample of size $10,000$ in total. Figure 2 shows the trace plots, histograms, and auto-correlation plots for β^0, β^1, and σ^2 from one of the ten chains. It suggests reasonable convergence rates, and the posterior samples seem approximately independent. Diagnostics for other parameters, and from other chains look similar.

Predictions based on the posterior samples could be made in several ways. One could perform forward simulation of the latent process y^* in the Tobit model in equation (3), and predict by thresholding at 0. Our predictions are the median values of samples from the approximate posterior predictive density in equation (5) over the posterior sample. We found little difference in terms of computational efficiency or accuracy between these two approaches. However, an advantage of the analytical form of the posterior predictive density is that it could provide useful insights into predictive uncertainty ([6]).

Table II shows the proportions of matched dry days ($\frac{n_{00}}{n_0}$), and matched wet days ($\frac{n_{11}}{n_1}$), which indicate the accuracy of various methods in terms of prediction of dry, and wet days. Table III shows the details of the prediction error when a. the true observed, and predicted values are positive (MSPE, and MAD), b. the true value is 0, but the prediction is positive (3^{rd} column shows the average of predicted rain intensities

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

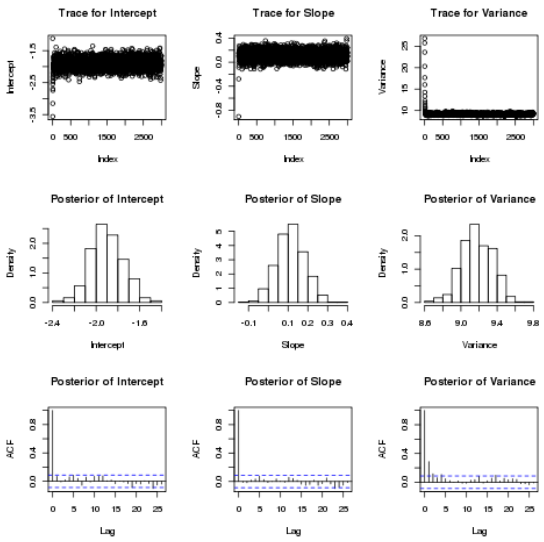


Fig. 2: Diagnostics from an MCMC chain

TABLE II: Proportion of Matched Dry and Wet Days.

Method	% of dry day matches	% of rainy day matches
MLR	0	1
Tobit	1	0
LAD	0.98	0.03
Bayesian	1	0

for true dry days), and c. the true value is positive, but the prediction is 0 (4^{th} column is the average of true rain intensities for dry predicted days). For the data considered, the MLR method produced only positive predictions. This is because MLR does not account for the point mass at 0. As a result, the regression parameter estimates tend to have a strong positive bias. To the other extreme, the Tobit MLE method produced only 0 values as predictions. A possible reason is the model misspecification as the scatter plot in Figure 1 suggests. As expected, predictions from the Bayesian method are similar to those from the Tobit MLE. Since the LAD method is a more robust alternative to Tobit MLE, we expect it to perform better as Tables II, and III indicate. As suggested by one of the reviewers, more flexible distributional forms like zero-adjusted gamma distribution to model the semi-continuous observed data could yield better results. Furthermore, statistical methodologies to incorporate time, and space structure of the data need to be developed.

ACKNOWLEDGMENTS

The first author is grateful to the Joint Center for Earth Systems Technology (JCET), UMBC, for funding. The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF).

TABLE III: Errors in the Predicted Intensities of Rain.

Method	MSPE	MAD	Avg. of DR	Avg. of RD
MLR	2.07	0.43	0.34	0
Tobit	0	0	0	0.53
LAD	0.05	0.01	0	0.52
Bayesian	0	0	0	0.53

REFERENCES

[1] P. W. Gassman, M. R. Reyes, C. H. Green, and J. G. Arnold, "The Soil and Water Assessment Tool: Historical development, applications, and future research directions," 2007.

[2] C. Kilsby, P. Jones, A. Burton, A. Ford, H. Fowler, C. Harpham, P. James, A. Smith, and R. Wilby, "A daily weather generator for use in climate change studies," *Environ. Modelling and Software*, 2007.

[3] A. Takeshi, "Advanced econometrics," 1985.

[4] J. L. Powell, "Least absolute deviations estimation for the censored regression model," *Journal of Econometrics*, vol. 25, pp. 303–325, 1984.

[5] S. Chib, "Bayes inference in the tobit censored regression model," *Journal of Econometrics*, vol. 51, pp. 79–99, 1992.

[6] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*. Chapman and Hall/CRC, 2nd ed., 2003.

[7] S. K. Popuri, N. K. Neerchal, and A. Mehta, "Comparison of Linear and Tobit Modeling of Downscaled Daily Precipitation over Missouri River Basin using MIROC5," in *Machine Learning and Data Mining Approaches to Climate Science. Proc. of the 4th Inter. Workshop on Climate Informatics*, pp. 35–49, Springer, 2015.

[8] E. P. Maurer, A. W. Wood, J. C. Adam, and D. P. Lettenmaier, "A long-term hydrologically based dataset of land surface fluxes and states for the conterminous united states," *Journal of Climate*, vol. 15, no. 22, pp. 3237–3251, 2002.

CI 2016 6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

INCORPORATING PRIOR KNOWLEDGE IN SPATIO-TEMPORAL NEURAL NETWORK FOR CLIMATIC DATA

Arthur Pajot[1], Ali Ziat[1,2], Ludovic Denoyer[1], Patrick Gallinari[1]

***Abstract*—We introduce a methodology for incorporating prior knowledge in spatio-temporal statistical models and develop this idea for designing a spatio-temporal neural network. More specifically, starting from an analytic description of a physical phenomenon using partial differential equations, we derive functional dependencies for the variables of the neural network. The latter is then trained on data gathered from the observed phenomenon so as to reproduce its underlying dynamics. The method is illustrated via preliminary experiments performed on two simple but representative datasets.**

I. STATISTICAL MOTIVATION

Spatio-temporal statistical models are increasingly being used across a wide variety of scientific disciplines to describe and predict spatial processes that evolve over time. They are efficient at modeling climatic phenomena described by the evolution of specific quantities over time. A key difficulty is that modeling spatio-temporal dependencies at large scale rapidly becomes prohibitive in terms of model complexity. It can be useful then to exploit prior knowledge on the physical phenomenon, in order to constrain the dependencies between the model variables thus reducing its complexity and making training easier and more accurate. We propose here to use prior knowledge developed by physicists under the form of partial differential equation (PDE) for designing our statistical models. This methodology has been advocated in spatio-temporal statistics for example by Wikle and al. [1] using a Bayesian framework. Our contribution is (i) to adapt this idea to the recent field of Deep Learning (ii) to develop a spatio-temporal neural network along these lines for modeling general spatio-temporal processes. The family of PDEs we have in mind is that of reaction-diffusion equations which are used in several fields of natural science, such as physics, ecology and biology.

Corresponding author: arthur.pajot@lip6.fr [1]Sorbonne Universities, UPMC Univ Paris 06, UMR 7606, LIP6, [2]Vedecom Institute, Eco-Mobility Department, 77000, Versailles, France

For this contribution, we will illustrate the approach with simple diffusion equations.

Let us first illustrate the ideas using as an example the one-dimensional heat equation:

$$\frac{\partial u}{\partial t} = a\left(\frac{\partial^2 u}{\partial x^2}\right)$$

Where u is the heat measurement and a is a diffusion coefficient. This equation expresses the temporal evolution of quantity u on the real axis. Discretizing time and space to approximate the PDE allows us to describe a spatio-temporal process as an ensemble of time series, where each time step represents a spatial state $\{X_t; t = 1, \dots\}$ with X_t a multidimensional spatial vector.

Using the finite difference method the heat equation can be discretized as:

$$u_{t+1}^i = u_t^i + a\Delta t\left(\frac{u_t^{i+1} - 2u_t^i + u_t^{i-1}}{\Delta x^2}\right)$$

which can be rewritten as:

$$u_{t+1}^i = \theta_1 u_t^{i-1} + \theta_2 u_t^i + \theta_1 u_t^{i+1}$$

Where θ_1 and θ_2 are parameters depending on a, Δx and Δt, to be estimated. This parameterization allows us to capture the structure of a heat diffusion phenomena and gives the following first order dynamics:

$$\begin{bmatrix} X(x_1; t+1) \\ X(x_2; t+1) \\ \dots \\ X(x_n; t+1) \end{bmatrix} = \begin{bmatrix} \theta_2 & \theta_1 & \dots & 0 \\ \theta_1 & \theta_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \theta_2 \end{bmatrix} \begin{bmatrix} X(x_1; t) \\ X(x_2; t) \\ \dots \\ X(x_n; t) \end{bmatrix} \quad (1)$$

where $X(x_i; t)$ is the modeled quantity at time t and location i.

As shown in [2], finite-difference discretizations of PDEs imply a lagged nearest-neighbor parameterization of the dynamic function (i.e.sparse, with tri-diagonal

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

structure in the example above), where the parameterization is controlled by the problem parameters. Several other discretization choices can be applied to this problem: such as finite-difference, and Runge-Kutta schemes. Those different schemes and order of discretizations can motivate alternative spatio-temporal statistical models.

II. Representation Learning

Representation learning approaches and particularly Deep Learning are today state of the art methods in many fields such as computer vision or speech processing ([3], [4]). For temporal data, several successful models based on recurrent neural networks (RNNs) have been recently proposed for complex sequence or series analysis tasks ([5], [6], [7]). RNNs attempt to capture characteristics of the sequential process underlying temporal data dynamics. To the best of our knowledge the spatial dimension has not been considered in these temporal models but very recently [8].

Our NN model is an extension of [8]. It has two components: a decoding component which computes a prediction of future observations given a latent representation of this future observation and a spatio-temporal component which captures the dynamic of the data series in the latent space. Let us consider multivariate time series $X = \{X_t; t = 1, \dots, T\}$ with $X_t^i \in R$ the value of the i-th spatial component of the series at time t. Each series will have its own latent representation at each time step. Let Z_t^i be the latent factor of series i at time t, Z_t is thus a $n \times N$ matrix, N being the dimension of the latent space and n the size of the temporal dimension. If we consider for simplicity a first order dynamic in the latent space, the model writes:

$$\hat{X}_t = d(Z_t, \Gamma) \text{ with } Z_t = h(Z_{t-1}; \Theta) \tag{2}$$

Where $\hat{X}_t \in R^m$ is the predicted value of the multivariate series $X_t \in R^m$ at time t. d is the encoding function and h is the dynamic model in the latent space. The aim of statistical modeling is to estimate the parameters Γ and Θ. For a linear model Γ and Θ will be linear operators (weight matrices), and for a non-linear a neural network, for example.

With the above notations, X corresponds to a $\mathbb{R}^{T \times m}$ multidimensional matrix. The latent representation Z_{t+1}^i of any point i of a series at time $t+1$ will be dependent on both the latent representation of this point at time t, and of the representations of the other points. The shape of the h function will depend on the process one wants to describe, and will be defined based on the underlying differential equations that model this process. Learning the parameters is apprehended as a minimization problem defined as:

$$\begin{aligned} \min_{Z,\Theta;\Gamma} \frac{1}{T} \sum_t ||d(Z_t, \Gamma) - X_t|| \\ + \lambda \frac{1}{T} \sum_t ||Z_{t+1} - h(Z_t, \Theta)|| \end{aligned} \tag{3}$$

Where λ is a hyper-parameter set by cross-validation which corresponds to the strength of the dynamic constraint. Learning is then performed through Stochastic Gradient Descent. Let us now describe how h can be defined for two simple dynamic processes.

A. One Dimensional Heat Equation

The equation 2 models the successive observation vector as an auto-regressive process. In our model 3 the dynamic is modeled in the latent (Z)-space. In this context, the heat equation described in section I can be rewritten in the latent space under the following form:

$$Z_{t+1}^i = \theta_1 Z_t^{i-1} + \theta_2 Z_t^i + \theta_1 Z_t^{i+1}$$

This equation is a parameterization of the dynamic function h.

B. Two-dimensional Heat equation

We can follow a similar methodology for the 2-dimensional heat equation:

$$\frac{\partial u}{\partial t} = a \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$$

In the latent space the parameterization gives:

$$Z_{t+1}^{ij} = \theta_1 (Z_t^{i-1,j} + Z_t^{i+1,j}) + \theta_2 Z_t^{i,j} + (Z_t^{i,j-1} + Z_t^{i,j+1})$$

This parameterization seems natural, as the 2D heat equation describes a diffusion phenomenon is two directions: horizontal and vertical.

III. Experiment

Preliminary experiments have been performed on two simple spatio-temporal forecasting problems. In the following, we will consider prediction (forecasting) at different horizons $T+1$, $T+2$ up to $T+100$, with $T = 100$. This means that all observations until $T = 100$ are used for training through equation (3), and prediction is performed and evaluated on the next 100 observations. We use as a comparison baseline a state of the art RNN with Gated Recurrent Unit called GRU [7]. The addition of a tanh non-linearity to the h dynamic function improves the performance.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

1) A Toy Example, the Heat Equation: The heat equation simulation is used here as a canonical example allowing us to demonstrate the capacity of our model to accurately capture simple physical processes. It has been simulated with a finite-difference discretization of this equation.

We show in Fig 1 the Mean Squares Error (MSE) achieved by our model and by the neural-network baseline on the test sequence. Remember that the models are trained on the first 100 observations of the phenomenon and evaluated on the next 100. The spatial location has one dimension of size 200. The initial conditions are a heat of 1 for positions 70 to 130 and -1 for the rest.

As can be seen, recurrent networks perform slightly better for the first few iterations, but then quickly diverge while the proposed model is able to keep the predicted MSE at a reasonable level.

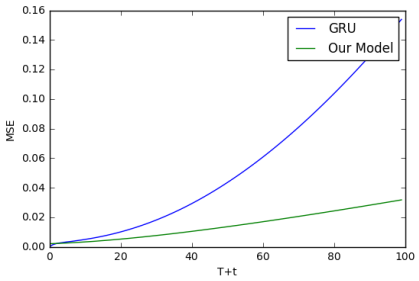


Fig. 1. MSE of the heat equation at time T+1 to T+100

The behavior of our model and its ability to capture the diffusion process is illustrated on figure 2.

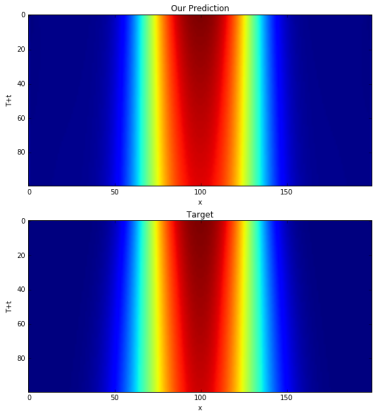


Fig. 2. Heat equation at time T+1 to T+100

2) Pacific Sea Surface Temperature: Tropical Pacific Sea Surface Temperature (SST) exhibits strongly structured variability on multiple spatial and temporal scales. More particularly, the El Nino phenomenon is strong in the Pacific and is known to be one of the most important sources of climate variability. The effects of such variability influence ecological systems on a very large scale, suggesting that it is useful to accurately forecast such effects many months in advance.

We used monthly SST data from the Pacific Ocean (fig. 3) across 2261 gridded spatial locations at 2×2 resolution from January 1970 to March 1998 as described in [9]. Our model is trained on the first 100 months and tested on the following 100 months.

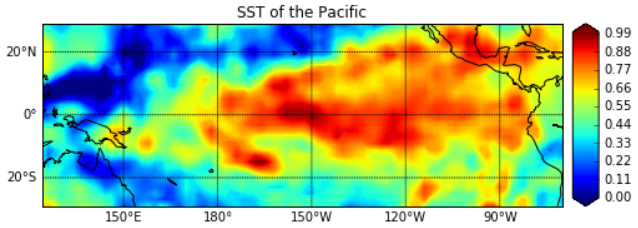


Fig. 3. Pacific SST at time 100

For this preliminary experiments, the statistical parameterization chosen was a heat diffusion equation in two dimensions, as described in section II-B.

Figure 4 shows the MSE of the forecasting for our method and the baseline GRU. This data is more complex than the artificially generated data from the previous experiment. The GRU network works better for several iterations of the dynamics corresponding to a horizon up to 15 steps, but fails to forecast the long term phenomena while or model remains relatively stable.

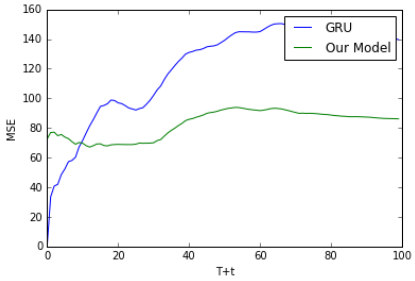


Fig. 4. MSE of the SST at time T+1 to T+100

IV. CONCLUSION

We have proposed (i) a method to incorporate prior physical knowledge into a statistical model that is easily transferable to representation learning models (ii) a neural network model inspired from these principles, able to model spatio-temporal dynamics. We have conducted preliminary experiments and shown that our

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

method outperforms a state-of-the-art RNN for long-term forecasting. Future work will evaluate the model on more complex datasets and examine more general types of reaction-diffusion equations.

REFERENCES

[1] C. K. Wikle and M. B. Hooten, "A general science-based framework for dynamical spatio-temporal models," *Test*, vol. 19, no. 3, pp. 417–451, 2010.

[2] C. K. Wikle, "Hierarchical bayesian models for predicting the spread of ecological processes," *Ecology*, vol. 84, no. 6, pp. 1382–1394, 2003.

[3] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] A. Graves, A. Mohamed, and G. E. Hinton, "Speech recognition with deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing,ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pp. 6645–6649, 2013.

[7] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[8] A. Ziat, G. Contardo, N. Baskiotis, and L. Denoyer, "Learning embeddings for completion and prediction of relationnal multivariate time-series," *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning - ESANN*, 2016.

[9] N. Cressie and C. K. Wikle, *Statistics for spatio-temporal data*. John Wiley & Sons, 2015.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

DIMENSIONALITY-REDUCTION OF CLIMATE DATA USING DEEP AUTOENCODERS

Juan A. Saenz[1] , Nicholas Lubbers[1,2], Nathan M. Urban[1]

***Abstract*—We explore the use of deep neural networks for nonlinear dimensionality reduction in climate applications. We train convolutional autoencoders (CAEs) to encode two temperature field datasets from pre-industrial control runs in the CMIP5 first ensemble, obtained with the CCSM4 model and the IPSL-CM5A-LR model, respectively. With the later dataset, consisting of 36500 96×96 surface temperature fields, the CAE out-performs PCA in terms of mean squared error of the reconstruction from a 40 dimensional encoding. Moreover, the noise in the filters of the convolutional layers in the autoencoders suggests that the CAE can be trained to produce better results. Our results indicate that convolutional autoencoders may provide an effective platform for the construction of surrogate climate models.**

I. INTRODUCTION

Uncertainty quantification of the response of the Earth system to greenhouse-gas emission scenarios is important for evaluating the impacts of climate change on infrastructure, agriculture, and the environment, among other areas. However, simulations using global, coupled earth system models are computationally expensive, making it impossible to produce large ensembles needed for statistical uncertainty quantification. To overcome this, surrogate models–simplified models that emulate more complex climate models–are built and trained [1], [2]. The computational cost of running these climate model emulators is much lower than their full complexity counterparts. As a result, large ensembles of simulations (order tens-hundreds of thousands) can be produced and used to carry out uncertainty quantification.

Emulators can be devised to dynamically evolve the state of the climate on a dimensionally reduced manifold [3]. An important requirement of such emulators is that the state in physical dimensions can be recovered. Linear dimensionality reduction via principal component analysis (PCA) is well-known in the climate science community, however, nonlinear methods have not been fully explored. Ross [3] investigated nonlinear dimensionality reduction methods for a different climate aplication: identifying low-dimensional nonlinear dynamics in El Nino variability. Errors from reconstructions using nonlinear methods were not significantly better than using linear PCA. Methods include nonlinear PCA (autoencoders), Isomap, and Hessian locally linear embedding.

Here, motivated by the success of deep autoencoders for dimensionality reduction [4] and convolutional neural networks for image processing [5], [6], we present work in progress using convolutional autoencoders [7] to reduce the dimensionality of data from climate models. In section II we describe our methods, and in section III we present results on two pre-industrial climate model simulation datasets: the CCSM4-T31 temperature at the surface dataset, and the IPSL-CM5A-LR temperature at the surface data. We end with a brief discussion in section IV.

II. METHODS

A. Principal Components Analysis

Consider a dataset of dimensionality M with N datapoints collected into a $N \times M$ data matrix $\mathbf{X}$. PCA constructs a rank m reduced matrix $\hat{\mathbf{X}}_{pca}$ by projecting $\tilde{\mathbf{X}}$ (obtained by centering and normalizing $\mathbf{X}$ using the global mean and standard deviation) onto the first m principal components which maximizes the covariance of the data, thus minimizing the mean reconstruction error

$$MSE = \frac{1}{NM}\sum_n (\mathbf{x}_n - \hat{\mathbf{x}}_n)^2 = \frac{1}{NM}\|\mathbf{X} - \hat{\mathbf{X}}\|_2^2. \quad (1)$$

This can be obtained by singular value decomposition of $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^T$ with $\mathbf{\Sigma}$ the diagonal matrix of singular values. The data covariance is given by $\mathbf{X}^T\mathbf{X} = \mathbf{W}\mathbf{\Sigma}^2\mathbf{W}^T$, and so the first m principal component vectors are the columns of $\mathbf{W}$ associated with the m largest singular values.

Corresponding author: J.A. Saenz, juan.saenz@lanl.gov; N. Lubbers, nlubbers@bu.edu; N.M. Urban, nurban@lanl.gov.
[1]Los Alamos National Laboratory, [2]Boston University.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

B. Convolutional Autoencoder

The autoencoder provides an alternative method for dimensionality reduction of $\mathbf{X}$. Data is fed though a series of neural network layers, i.e. an affine transform followed by an elementwise nonlinearity f, to create activations $\mathbf{X}_l$ at layer l:

$$\mathbf{X}_l = f(\mathbf{W}_l \mathbf{X}_{l-1} + \mathbf{b}_l). \tag{2}$$

The trainable parameters of the network are the weights $\mathbf{W}_l$ and biases $\mathbf{b}_l$ of each layer. The waist layer of the autoencoder is constrained to a number of neurons m, so that the activations at that layer can be used as an m length code for each image. This is followed by decoding layers, and in the final layer of the autoencoder, $\hat{\mathbf{X}}$ is constructed to have the same dimensionality as the input $\mathbf{X}$. The parameters of the network are then trained to minimize the same reconstruction error MSE (Eq. 1) as PCA.

The convolutional autoencoder (CAE) is an extension to the autoencoder which faciltates the analysis of data on regular grids, such as images. Here, each data point can be indexed by two pixel positions. In the convolutional and deconvolutional layers, the weights consist of small image filter kernels, and the product of the weights and the data consist of spatial convolutions. Denoting the feature k of the pixel indexed by i and j by $\mathbf{x}^k_{i,j}$, the preactivations $\mathbf{W}\mathbf{x}$ are computed using

$$(\mathbf{W}\mathbf{x})^l_{i,j} = \sum_{k,l,m} \mathbf{W}^{lk}_{i-l,j-m} \mathbf{x}^k_{i,j}\,. \tag{3}$$

Convolutional layers operate as a set of local image filters with the capacity to extract patterns that increase in complexity with depth [8]. They have the benefit of greatly reducing the number of learnable parameters per layer, and faciliate training by seeking out features which encode the structure of local image patches [9].

Finally, CAEs have pooling layers that coarse-grain the image plane after each convolution. We use 2×2 max pooling, with unpooling layers which use piecewise constant 2×2 upsampling. We use two regularization strategies to improve training regularity and smoothness of learned filters. In the first, we use image flipping and weight decay with strength β. In the second, we use noise injection (denoising autoencoder), applying pixel-wise Gaussian noise with mean 0 and standard deviation γ to the images.

Autoencoders presented here are trained using stochastic gradient descent for 1000 epochs with a learning rate of 0.01, which is updated using a Nesterov scheme with a momentum of 0.975, and a batch size of 128. Convolutional layers use linear activation functions (we have not been able to successfully train non-linear activations). Fully connected layers use rectified linear (ReLU) activation functions. Decoding layers have the reverse structure of the encoding layers, but we do not tie the weights between the encoding and decoding layers. Convolutional boundary conditions use valid convolutions (no padding), with decoding convolutional layer dimensions computed to produce the correct size output. As preprocessing, the temperature fields are normalized to mean 0 and standard deviation 1 using the global mean and standard deviation.

TABLE I
ARCHITECTURES OF CONVOLUATIONAL AUTOENCODERS

label	encoding architecture	β	γ
A1	CL7,32-PL-FC40	0.00025	N/A
A2	CL5,32-PL-CL5,64-PL-FC40	0.00025	N/A
B1	CL5,32-PL-CL5,64-PL-FC40	N/A	N/A
B2	CL5,32-PL-CL5,64-PL-FC40	N/A	0.1
B3	CL5,32-PL-CL5,64-PL-FC40	N/A	0.5

CLn,m is a convolutional layer with an $n \times n$ receptive field and m features, PL is a 2×2 pooling layer, and FCm is a fully connected layer with m neurons.

III. RESULTS

A. CCSM4

We train encoders with data from the Community Climate System Model - version 4 (CCSM4) [10]. We employ 150 years of surface temperature (T_s) monthly climatology data from the first ensemble, pre-industrial control run of the 5th version of the Climate Model Intercomparison Project (CMIP5). The data was regridded to a 3.75×3.75 degree grid (T31 grid). The resulting dataset has 1800 samples of 48×96 T_s fields.

The trained architectures (A1 and A2, table I) do a good job at recovering the global structure of the temperature field T_s. There are differences in local features in some regions (not shown). In table II we compare the MSE (Eq. 1) of reconstructions using CAEs A1 and A2 and PCA. Architecture A1 performs better than A2, but PCA performed better than both autoencoders. The weights of the first convolutional layer indicated that this type of regularization was not effective (Fig. 1, upper left). There is some repetitive structure between weights, and they are noisy, indicating that the neural networks are not well trained. One possible reason for this is that there are not enough samples (N=1800). To investigate this, we now use a dataset with a much larger number of samples.

B. IPSL-CM5A-LR

Here, we use near-surface air temperature T_{as} from the low resolution pre-industrial control run produced

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

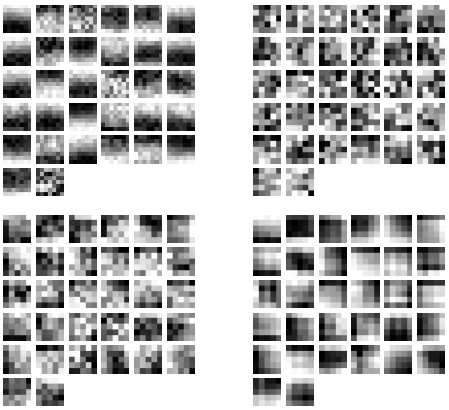


Fig. 1. Weights from the first convolutional layer ($\mathbf{W}^1$) in the trained architectures A1 (top left) and B1 (top right), B2 (bottom left) and B3 (bottom right). Brightness scale is arbitrary, only the patterns visible in the plots are meaningful here.

TABLE II
MEAN SQUARED ERROR OF RECONSTRUCTIONS [$^\circ$K^2].

CCSM4-T31	PCA	A1	A2	
	0.8022	1.5495	2.1608	
IPSL-CM5A-LR	PCA	B1	B2	B3
	4.9014	4.2415	4.3806	4.5243

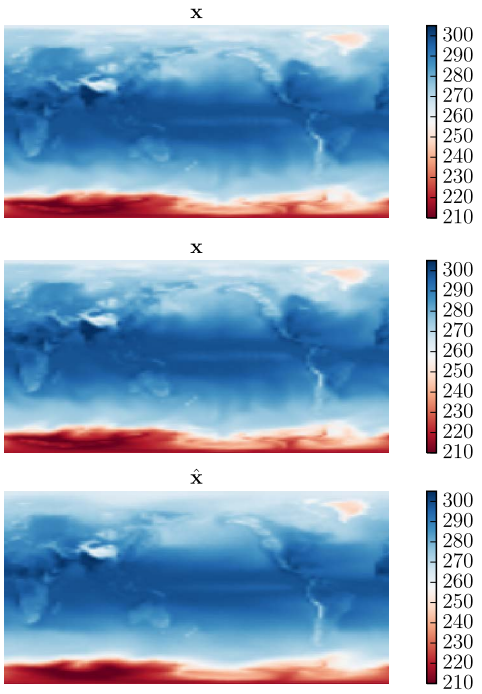


Fig. 2. Comparison of T_{as} [$^\circ$K] from the original dataset (top) and the reconstruction using architectures B1 (middle) and B2 (bottom), sample 28618 from the IPSL-CM5A-LR dataset.

by the IPSL-CM5A-LR model, part of the CMIP5 first ensemble [11]. The data is on a 1.9×3.75 degree grid, 96×96, and we use daily output for 100 years between 1800 and 1900 (N=36500 samples), although 600 years of data are available (219000 samples). As a result, the dataset has 36500 samples of 96×96 T_{as} fields, resulting in $\mathbf{X}$ with shape $(N = 36500) \times (M = 9216)$.

We implement an autoencoder architecutre, B1, which is similar to A2, but without any type of regularization (table I). The MSE using B1, shown in table II, is smaller than the error obtained with PCA. However, the weights are still noisy (Figure 1, upper right). To remedy this, we explored architectures B2 and B3 (table I) which are regularized using injected noise. The weights become smoother with more noise (Fig 1, lower left and right), but the errors are larger, as shown in table II.

In figure 2 we show the reconstructed temperature fields, which are very similar to the temperature in the original dataset. Large scale features of the global temperature patterns are preserved. Smaller scale features in regions such as over the Antarctic peninsula, the North Atlantic and the South Pacific are filtered out. Some small scale features above high elevation topography, such as the Andes and the Himalayas, appear to be well preserved.

IV. DISCUSSION

The results of this work in progress indicate that there is potential to devise deep autoencoders for dimensionality reduction of climate data. Noise in the filters of the trained networks indicate that finding effective representations is dependent on regularization and the availability of data.

Future developments will focus on testing networks with data not used for training/validating, using larger datasets, using nonlinear convolutional activation functions, implementing other regularization methods (e.g. dropout), and using deeper networks, with the aim of improving the reconstruction of small scale features. Future analysis will include the investigation of patterns extracted by the convolutions.

ACKNOWLEDGMENTS

This work was supported by the Office of Science (BER), U. S. Department of Energy. Autoencoders are built using Python libraries Theano [12], Lasagne [13], and nolearn [14]. We thank A. Jonko for providing the CCSM4-T31 data. We acknowledge the WCRP-WGCM, which is responsible for CMIP. We thank the DOE and NCAR for developing the CCSM4, producing and making available their model output.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] C. M. Little, M. Oppenheimer, and N. M. Urban, "Upper bounds on twenty-first-century antarctic ice loss assessed using a probabilistic framework," *Nature Clim. Change*, vol. 3, pp. 654–659, 07 2013.

[2] C. M. Little, N. M. Urban, and M. Oppenheimer, "Probabilistic framework for assessing the ice sheet contribution to sea level change," *Proceedings of the National Academy of Sciences*, vol. 110, no. 9, pp. 3264–3269, 2013.

[3] I. Ross, *Nonlinear Dimensionality Reduction Methods in Climate Data Analysis*. PhD thesis, University of Bristol, Jan. 2009.

[4] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks". *Adv. Neural Inf. Process. Syst.*, pages 10971105, 2012.

[6] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint* arXiv:1409.1556 (2014).

[7] Masci, Jonathan, et al. "Stacked convolutional auto-encoders for hierarchical feature extraction." *International Conference on Artificial Neural Networks.* Springer Berlin Heidelberg, 2011.

[8] Lee, Honglak, et al. "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations." *Proceedings of the 26th annual international conference on machine learning.* ACM, 2009.

[9] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.

[10] P.R. Gent, G. Danabasoglu, L.J. Donner, M.M. Holland, E.C. Hunke, S.R. Jayne, D.M. Lawrence, R.B. Neale, P.J. Rasch, M. Vertenstein, P.H. Worley, Z. Yang, and M. Zhang, "The Community Climate System Model Version 4". *J. Climate*, vol. 24, pp. 49734991.

[11] J.-L. Dufresne, M.-A. Foujols, S. Denvil, A. Caubel, O. Marti, O. Aumont, Y. Balkanski, S. Bekki, H. Bellenger, R. Benshila, S. Bony, L. Bopp, P. Braconnot, P. Brockmann, P. Cadule, F. Cheruy, F. Codron, A. Cozic, D. Cugnet, N. de Noblet, J.-P. Duvel, C. Ethé, L. Fairhead, T. Fichefet, S. Flavoni, P. Friedlingstein, J.-Y. Grandpeix, L. Guez, E. Guilyardi, D. Hauglustaine, F. Hourdin, A. Idelkadi, J. Ghattas, S. Joussaume, M. Kageyama, G. Krinner, S. Labetoulle, A. Lahellec, M.-P. Lefebvre, F. Lefevre, C. Levy, Z. X. Li, J. Lloyd, F. Lott, G. Madec, M. Mancip, M. Marchand, S. Masson, Y. Meurdesoif, J. Mignot, I. Musat, S. Parouty, J. Polcher, C. Rio, M. Schulz, D. Swingedouw, S. Szopa, C. Talandier, P. Terray, N. Viovy, and N. Vuichard, "Climate change projections using the ipsl-cm5 earth system model: from cmip3 to cmip5," *Climate Dynamics*, vol. 40, no. 9, pp. 2123–2165, 2013.

[12] R. Al-Rfou, G. Alain, A. Almahairi et al. 2016 "Theano: A Python framework for fast computation of mathematical expressions" *CoRR*, abs/1605.02688.

[13] S. Dieleman, J. Schlter, C. Raffel, E. Olson, S.K. Snderby, D. Nouri, D. Maturana, M. Thoma, E. Battenberg, J. Kelly, J. De Fauw, M. Heilman, dogo149, B. McFee, H. Weideman, takacsg84, peterderivaz, Jon, instagibbs, K. Rasul, CongLiu, Britefury, J. Degrave. 2015 "Lasagne: First release" doi: 10.5281/zenodo.27878

[14] D. Nouri. 2014. "nolearn: scikit-learn compatible neural network library". https://github.com/dnouri/nolearn

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

MAPPING PLANTATION IN INDONESIA

Xiaowei Jia[1], Ankush Khandelwal[1], James Gerber[2], Kimberly Carlson[3], Paul West[2], Vipin Kumar[1]

***Abstract*—Plantation is a key driver of deforestation in Southeast Asia. Many governments and companies aim to ensure that the plantation growth meets rigorous sustainability standards. Such regulations depend on the capacity of monitoring plantations in large regions. In this paper we propose a two-stage automatic method to map plantations on a yearly scale. The results on Kalimantan region of Indonesia well demonstrate the effectiveness of the proposed method.**

I. INTRODUCTION

Tropical forests are important as they store ∼50% of all carbon stored in terrestrial vegetation. Tropical forests of Southeast Asia are unique because of very high diversity of wildlife. Nowadays Southeast Asia is the largest producer of palm oil plantation, which has become one of the biggest drivers of deforestation in this region. The loss of tropical forests in Southeast Asia has led to a huge impact on climate as conversions of these forests to other land cover types lead to massive release of CO_2. Various efforts have been put in place to reduce carbon emissions and to achieve sustainable ways of doing commercial activities. To these reasons, it is crucial to develop high quality monitoring systems on the plantation.

In particular, remote sensing data acquired through various earth observation satellites provide immense opportunity to monitor land use/land cover (LULC) changes. For instance, recent studies have shown advances in identifying soybean and sugarcane in Brazil using remote sensing data [1], [2]. Besides, a wide variety of methods have been proposed that use remote sensing at different spatial and temporal scale for monitoring changes in land cover. However, a vast majority of these methods focus on detecting deforestation activity only [3], [4], [5]. For instance, the widely used global deforestation product [6] does not differentiate between forest and plantations.

[1]Department of Computer Science and Engineering, University of Minnesota, {jiaxx221,khand035}@umn.edu, kumar@cs.umn.edu
[2]Institute on the Environment, University of Minnesota, {jsgerber,pcwest}@umn.edu
[3] Department of Natural Resources and Environmental Management, University of Hawai'i Mānoa, kimberly.carlson@hawaii.edu

The main challenges of mapping plantation lie in several aspects. First, we are provided with limited ground-truth information. Most available plantation datasets are created via visual interpretation of satellite imagery, and consequently have either low precision or low recall. Moreover, there exists strong data heterogeneity in several aspects. To identify plantations, we need to differentiate between plantations and a variety of land covers, e.g. agriculture, forest, etc. Besides, the same land cover may look different over space and over time.

To tackle these challenges we propose our two-stage framework to map plantations over different years. With the available imperfect ground-truth datasets, we first propose an effective strategy to sample training data and to learn a multi-class classification model. Then we implement Hidden Markov Model (HMM) to discover the latent transition relationship and mitigate the classification errors at each time step.

II. DATASETS

In this work we utilize MODIS data which are available for every day at 500 meters spatial resolution on a global scale. We create 8-day composite images by taking per-pixel seven-band spectral values with least noise from the corresponding 8-day interval. Based on MODIS data, we wish to apply the classification technique to map plantations for each year. We will test the proposed method in MODIS tile h29v09, which covers most Kalimantan region of Indonesia.

The first step for classification is to collect training samples using the available ground-truth datasets. Even though yearly plantation maps are not available, there exist a few datasets prepared by different organizations that provide plantations maps for a single year or a few years.

A. Tree Plantation Dataset

Tree Plantation dataset (TP) [7] is created by Transparent World and is available on Global Forest Watch. In this dataset, the plantation locations are manually labeled based on Landsat images on 2013 and 2014. In total this data this dataset covers 260483 locations

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

in our region of interest. Based on the assessment on random stratified sampling [7], it is reported that precision of the dataset is around 78.87% and the recall is around 93.86%.

B. RSPO Dataset

RSPO dataset is provided by Roundtable on Sustainable Palm Oil (RSPO) [8], and covers all the locations in the region of study. In this region, each location is categorized into one of 19 land cover types on 2000, 2005 and 2009 by RSPO dataset. While it does not provide a detailed assessment of accuracy, the comparison with the high-resolution images from Digital Globe shows that RSPO dataset is accurate (high precision), but misses many real plantation areas (low recall).

III. METHOD

Here we present our method in two stages, as shown in Fig. 1. First, we introduce an ensemble classification method to map plantations at each time step. Then we will investigate the latent land cover transition and post-process the results.

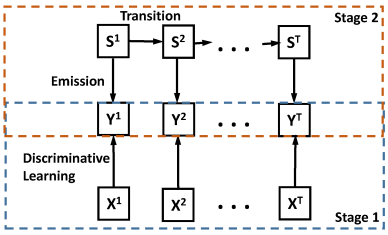


Fig. 1. The flow chart in two stages.

A. Ensemble Learning Model

The mapping of plantation is difficult from machine learning perspective since it requires the differentiation between plantation and a variety of other land cover types. If we directly merge all the non-plantation classes as the negative class and conduct binary classification, the heterogeneity within the negative class will greatly hamper the classification performance.

To this end we propose a three-class ensemble learning method to identify plantations. Specifically, we aggregate undisturbed forests and the disturbed forests with crossing roads but not yet logged into "forest", and aggregate the remaining land covers as "other". We keep "forest" separate since tropical forests look similar with tree plantations. Then we train three binary classifiers between each pair of classes from {"plantation", "forest", "other"}. The final predicted label is obtained by majority voting.

To train each classifier, we take equal amount of samples for each sub-class (defined by RSPO) within "forest" and "other" in case the training process is dominated by the land cover type with large population, e.g., forest and grassland. Besides, we combine the spectral features from different years so that the learned model can be applied on different years. In addition, since TP has high recall, we will use TP dataset to further filter the selected training samples.

B. Temporal Transition

The latent transition relationship among different land cover types offers opportunity to mitigate the individual classification errors. For instance, given a yearly sequence of {forest, forest, plantation, plantation, forest, plantation}, the third "forest" is highly likely to be a classification error since usually plantation would not be converted back to forest.

To capture the transition relationship, we utilize Hidden Markov Model (HMM), which models the transitions probability among latent states by the transition matrix T and the mapping relationship between the latent state and the observed class by the emission matrix E, as shown in Fig. 1. In particular T_{ij} represents the transition probability from state i and state j, and E_{ik} denotes the emission probability from state i to the observed class k. The joint probability of the latent states and the observed classes can be described as:

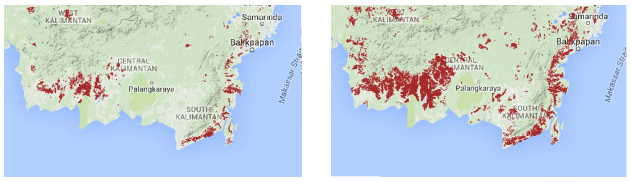
$$P(y^{1:T}, s^{1:T}) = P(s^1)E_{s^1 y^1} \prod_{t=2}^{T} T_{s^{t-1}s^t} E_{s^t y^t}. \quad (1)$$

In our problem each latent state represents a real land cover type and each observed class $k \in \{\text{plantation}, \text{forest}, \text{other}\}$. Specifically, T_{ij} is initialized as the proportion of locations in land cover i at any time step from $\{t_1, t_2, ..., t_{T-1}\}$ to be converted to land cover j at next time step. On the other hand, each entry in emission matrix E_{ik} represents the probability for a real land cover class i to be classified as class $k \in \{\text{plantation}, \text{forest}, \text{other}\}$. In this way the emission matrix E can capture the confusion between land cover classes. With the obtained transition matrix and emission matrix, we can fix the yearly prediction on each location via Viterbi algorithm [9].

IV. RESULTS AND DISCUSSION

Based the proposed method we can generate yearly plantation maps. For instance, we show our generated plantation maps in 2002, and 2014 in Fig. 2. Then

we will evaluate the effectiveness of our method by comparing to multiple baselines.



(a) (b)

Fig. 2. The generated plantation maps in 2002 (a), and 2014 (c). The plantation locations are marked in brown color.

TABLE I
COMPARISON TO DIFFERENT LEARNING STRATEGIES - YEARLY RECALL FROM 2005 TO 2009, OVERALL PRECISION AND OVERALL RECALL.

	yearly					*overall*	
Method	2005	2006	2007	2008	2009	prec	rec
$PALM_u$	0.849	0.799	0.814	0.832	0.847	0.783	0.862
$PALM_{np}$	0.828	0.740	0.743	0.761	0.815	0.750	0.833
$PALM_s$	0.780	0.736	0.749	0.756	0.760	0.736	0.643
PALM	0.867	0.810	0.823	0.837	0.858	0.846	0.868

We conduct the ensemble learning with Deep Belief Network (DBN) for each binary classifier. To show the effectiveness of our proposed ensemble learning model and the sampling strategy (termed PALM for **P**lantation **A**nalysis by **L**earning from **M**ultiple land covers), we compare to the following baselines:

$PALM_u$: Here we uniformly sample from the entire "other" class rather than take equal amount of samples from each land cover type.

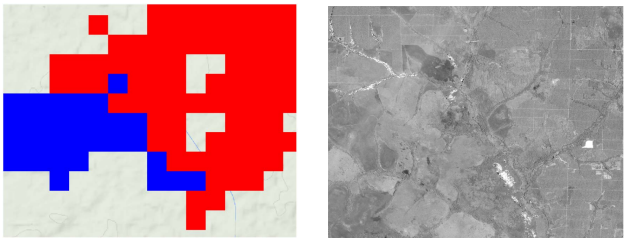
$PALM_{np}$: We implement the proposed learning method without using post-processing process.

$PALM_s$: We implement our ensemble learning strategy using Support Vector Machine (SVM) with RBF kernel.

Then we introduce the involved metrics in measuring the performance. Based on the generated yearly map, we measure the yearly recall from 2005 to 2009. The yearly recall is computed based on the interpolation of the provided maps in 2000, 2005 and 2009 by RSPO dataset. Since RSPO dataset has low recall, we cannot well estimate the precision on each year. Instead, we measure the overall precision using the Tree Plantation dataset and the overall recall using the RSPO dataset (on 2009) based on all the detected plantation locations through 2001 to 2014.

From the results shown in Table I, we can observe that the performance of $PALM_u$ is not as good as our approach since the training is dominated by the land cover types with large population. In this way the trained classifier is highly likely to misclassify the small classes, e.g., urban area, as plantation, and consequently leads to low precision. Furthermore, we can observe that PALM outperforms $PALM_s$ by a considerable margin due to the effectiveness of DBN in learning from complex feature space. Moreover, the comparison between $PALM_{np}$ and PALM demonstrates the effectiveness of post-processing.

In addition we compare our generated map to TP and RSPO datasets. We show a case study in Fig. 3. Here the red color denotes the detected plantation locations outside RSPO and blue color denotes the detected locations by TP but missed by our method. From the high-resolution image in Fig. 3 (b) we can verify that the red colored region is real plantation while the blue colored region is not. Therefore we can conclude that our method can detect more real plantations outside RSPO while avoiding the false positives in TP.



(a) (b)

Fig. 3. The comparison with TP and RSPO. (a) The detected regions. (b) The corresponding high-resolution image from Digital Globe.

V. ACKNOWLEDGEMENT

This work was funded by the NSF Awards 1029711 and 0905581, and the NASA Award NNX12AP37G. Access to computing facilities was provided by NASA Earth Exchange (NEX) resources and Minnesota Supercomputing Institute.

VI. CONCLUSION

In this work we propose an automatic method to map plantations on an annual scale. The proposed method is implemented in two stages. First, we conduct ensemble learning to distinguish plantations from other land covers. Then we post-process the predicted results on different years using HMM. The results on Kalimantan region of Indonesia demonstrates that the effectiveness of the method in generating yearly plantation maps. Besides we show through a case study that our method can achieve a better balance of precision and recall than existing datasets.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] B. F. T. Rudorff, M. Adami, D. A. Aguiar, M. A. Moreira, M. P. Mello, L. Fabiani, D. F. Amaral, and B. M. Pires, "The soy moratorium in the amazon biome monitored by remote sensing images," *Remote Sensing*, vol. 3, no. 1, pp. 185–202, 2011.

[2] B. F. T. Rudorff, D. A. Aguiar, W. F. Silva, L. M. Sugawara, M. Adami, and M. A. Moreira, "Studies on the rapid expansion of sugarcane for ethanol production in são paulo state (brazil) using landsat data," *Remote sensing*, vol. 2, no. 4, pp. 1057–1076, 2010.

[3] A. Hoscilo, S. E. Page, K. J. Tansey, and J. O. Rieley, "Effect of repeated fires on land-cover change on peatland in southern central kalimantan, indonesia, from 1973 to 2005," *International Journal of Wildland Fire*, vol. 20, no. 4, pp. 578–588, 2011.

[4] M. C. Hansen, S. V. Stehman, P. V. Potapov, T. R. Loveland, J. R. Townshend, R. S. DeFries, K. W. Pittman, B. Arunarwati, F. Stolle, M. K. Steininger, *et al.*, "Humid tropical forest clearing from 2000 to 2005 quantified by using multitemporal and multiresolution remotely sensed data," *Proceedings of the National Academy of Sciences*, vol. 105, no. 27, pp. 9439–9444, 2008.

[5] B. A. Margono, S. Turubanova, I. Zhuravleva, P. Potapov, A. Tyukavina, A. Baccini, S. Goetz, and M. C. Hansen, "Mapping and monitoring deforestation and forest degradation in sumatra (indonesia) using landsat time series data sets from 1990 to 2010," *Environmental Research Letters*, vol. 7, no. 3, p. 034010, 2012.

[6] M. C. Hansen, P. V. Potapov, R. Moore, M. Hancher, S. Turubanova, A. Tyukavina, D. Thau, S. Stehman, S. Goetz, T. Loveland, *et al.*, "High-resolution global maps of 21st-century forest cover change," *science*, vol. 342, no. 6160, pp. 850–853, 2013.

[7] R. Petersen, E. Goldman, N. Harris, S. Sargent, D. Aksenov, A. Manisha, E. Esipova, V. Shevade, T. Loboda, N. Kuksina, *et al.*, "Mapping tree plantations with multispectral imagery: preliminary results for seven tropical countries," *World Resources Institute, Washington, DC*, 2016.

[8] P. Gunarso, M. E. Hartoyo, F. Agus, Killeen, T. J., and J. Goon, "Roundtable on sustainable palm oil, kuala lumpur, malaysia," *Reports from the technical panels of the 2nd greenhouse gas working group of the Roundtable on sustainable palm oil*, 2013.

[9] G. D. Forney Jr, "The viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

FROM CLIMATE DATA TO A WEIGHTED NETWORK BETWEEN FUNCTIONAL DOMAINS

Ilias Fountalis[1], Annalisa Bracco[2], Bistra Dilkina[3], Constantine Dovrolis[1]

***Abstract*—We propose δ-MAPS, a method that analyzes spatio-temporal data first to identify the distinct spatial components of the underlying system, referred to as "domains", and second to infer the connections between them. A domain is a spatially contiguous region of highly correlated temporal activity. Domains may be spatially overlapping. Different domains may have correlated activity, potentially at a lag, because of direct or indirect interactions. The proposed edge inference method examines the statistical significance of each lagged cross-correlation between two domains, infers a range of lag values for each edge, and assigns a weight to each edge based on the covariance of the two domains. We illustrate the application of δ-MAPS on data from climate science.**

I. INTRODUCTION

Spatio-temporal data become increasingly prevalent and important for both science (e.g., climate, systems neuroscience, seismology) and enterprises (e.g., the analysis of geotagged social media activity). The spatial scale of the available data is often determined by an arbitrary grid, which is typically larger than the true dimensionality of the underlying system. One major task is to identify the distinct semi-autonomous components of this system and to infer their (potentially lagged) interconnections.

A common approach to reduce the dimensionality of spatio-temporal data is to apply EOF (standard or rotated) analysis. In climate science, EOF analysis has been used to identify teleconnections between distinct climate regions [1], [2]. However, the orthogonality between EOF components complicates the interpretation of the results making it difficult to identify the distinct underlying modes of variability and to separate their effects, as clearly discussed in [3].

Another broad family of spatio-temporal dimensionality reduction methods is based on unsupervised clustering [4]. These groups of algorithms are quite different but they share some common characteristics: the resulting clusters may not be spatially contiguous [5], [6], every grid cell needs to belong to a cluster (potentially excluding only outliers) [7], [8], and the number of clusters is often required as an input parameter [9] - none of these algorithms account for the fact that clusters may overlap. In particular, the lack of spatial contiguity makes it hard to distinguish between correlations due to spatial diffusion (or dispersion) phenomena from correlations that are due to remote interactions between distinct effects.

An approach of increasing popularity is to first construct a correlation-based network between individual grid cells, after pruning cross-correlations that are not statistically significant – see [10]. Then, some of these methods analyze the (binary or weighted) cell-level network directly based on various centrality metrics, k-core decomposition, spectral analysis, etc. (e.g., [11], [12], [13]) or they first apply a community detection algorithm (potentially able to detect overlapping communities, e.g., [14], [15]) on the cell-level network and then analyze the resulting communities in terms of size, density, location, overlap, etc. (e.g., [16], [17], [18], [19]). A community however may group together two regions that are, first, not spatially contiguous, and second, different in terms of how they are connected to other regions.

To overcome these limitations we propose δ-MAPS, an inference method that first identifies domains - spatially contiguous regions, homogeneous to the underlying variable. Domains might be overlapping and not all grid cells need to belong to a domain. At a second step, δ-MAPS identifies connections between the domains constructing a *domain-level* network. The network is modeled as a directed and weighted graph. The weight of a network edge captures the magnitude of the interaction between domains while the direction of the edge (and the lag associated to it) captures temporal ordering of events.

Corresponding author: I. Fountalis, fountalis@gatech.edu [1]School of Computer Science, Georgia Tech [2]School of Earth and Atmospheric Sciences, Georgia Tech [3] School of Computational Science and Engr, Georgia Tech

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

II. δ-MAPS

Informally a domain is a spatially contiguous region that participates in the same function. The functional relation between the grid cells of a domain results in highly correlated temporal activity. If we accept this premise it follows that each domain will have an epicenter of action. We identify such epicenters as local maxima in the correlation field between each grid cell and its K nearest neighbors. Formally, a domain is a set of grid cells that includes the epicenter, is spatially contiguous, and the average correlation between its grid cells is higher than a threshold δ. A domain may not have sharp spatial boundaries. Instead of searching for the discrete boundary of a domain, it is more reasonable to compute a domain as the *largest possible set of cells* that satisfies the previous three constraints. The threshold δ determines the minimum degree of homogeneity that a set of grid cells should have to form a domain. Domains might be overlapping.

To construct the functional domain network, we associate to each domain a signal, defined as the cumulative anomaly of the time series of its grid cells. Connections between the domains are not assumed to be instantaneous, thus we search for connections in a lag range $\{-\tau_{max} \dots \tau_{max}\}$. Using the appropriate statistics to control for the presence of autocorrelations [20] and the FDR procedure to account for the multiple testing problem [21] we identify significant correlations for a given false discovery rate q. The edges of the network are weighted in terms of the covariance between the domain signals, the covariance is calculated in respect to the maximum significant correlation in absolute sense. To each edge we assign a lag or a range of lags (if multiple significant correlations exist) capturing the time points at which the domains are connected. The latter also determines the direction of an edge.

III. APPLICATION IN CLIMATE SCIENCE

Here we apply δ-MAPS in the context of climate science. Climate scientists are interested in *teleconnections* between different regions, and they often rely on EOF analysis to uncover them [1]. Here, we analyze the monthly *Sea-Surface Temperature* (SST) field from the HadISST dataset [22], covering 50 years (1956-2005) at a spatial resolution of $2.0^o \times 2.5^o$, and we focus on the latitudinal range of $[60^oS; 60^oN]$ to avoid sea-ice covered regions. Following standard practice, we preprocess the time series to form *anomalies*, i.e., remove the seasonal cycle and remove any long-term trend at each grid-point (using the Theil-Sen estimator).

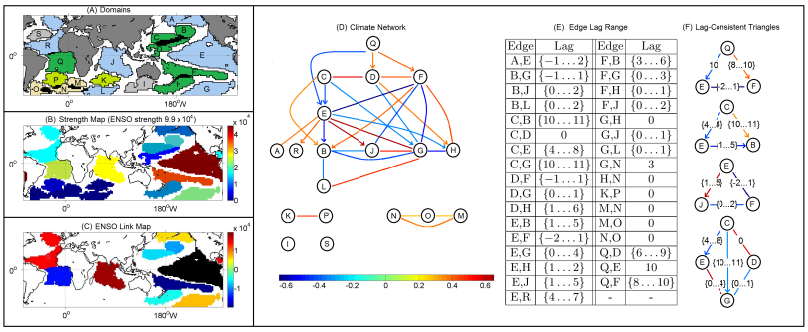
δ-MAPS is applied as follows. We set the local neighborhood to the K=4 nearest cells and tbe homogeneity threshold δ to 0.37. In the edge inference stage, the lag range is $\tau_{max} = 12$ months (a reasonable value for large-scale changes in atmospheric wave patterns), and q is set to 3% (we identify about 30 edges and so we expect no more than one false positive).

Fig. 1-A shows the identified domains. The spatial dimensionality has been reduced from about 6000 grid cells to 18 domains. 65% of the sea-covered cells belong to at least one domain; the overlapping regions are shown in black and they cover 2% of the grid cells that belong to a domain. The largest domain (domain E) corresponds to the El Niño Southern Oscillation (ENSO), which is also the most important in terms of node strength (see Fig. 1-B). Other strong nodes are domain F (part of the "horseshoe-pattern" surrounding ENSO), domain J (Indian ocean) and domain Q (subtropical Atlantic). The strength of the edges associated with ENSO are shown in Fig. 1-C. These findings are consistent with known facts in climate science regarding ENSO and its positive correlation with the Indian ocean and north tropical Atlantic, and negative correlations with the regions that surround it in the Pacific (horseshoe-pattern) [23].

Fig. 1-D shows the inferred domain-level network. The color code represents the (signed) cross-correlation for each edge. The lag range of each edge is shown in Fig. 1-E; some edges are not directed because their lag range includes τ=0. The network consists of five weakly-connected components. If we analyze the largest component (which includes ENSO) as a signed network (i.e., some edges are positive and some negative) we see that it is *structurally balanced* [24]. A graph is structurally balanced if it does not contain cycles with an odd number of negative edges. A structurally balanced network can be partitioned in a "dipole", so that positive edges only appear within each pole and negative edges appear only between the two poles. In Fig. 1-A, the nodes of these two poles are colored as blue and green (the smaller disconnected components are shown in other colors).

Focusing on the lag range of each edge, domain Q seems to play a unique role, as it temporally precedes all other domains in the inferred network. Specifically, its activity precedes that of domains D, E and F by about 5-10 months. The lead of south tropical Atlantic SSTs (domain Q) on ENSO has recently received significant attention in climate science [25]. Our results suggest that SST anomalies in domain Q may impact a large portion of the climate system.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO



Edge	Lag	Edge	Lag
A,E	{−1...2}	F,B	{3...6}
B,G	{−1...1}	F,G	{0...3}
B,J	{0...2}	F,H	{0...1}
B,L	{0...2}	F,J	{0...2}
C,B	{10...11}	G,H	0
C,D	0	G,J	{0...1}
C,E	{4...8}	G,L	{0...1}
C,G	{10...11}	G,N	3
D,F	{−1...1}	H,N	0
D,G	{0...1}	K,P	0
D,H	{1...6}	M,N	0
E,B	{1...5}	M,O	0
E,F	{−2...1}	N,O	0
E,G	{0...4}	Q,D	{6...9}
E,H	{1...2}	Q,E	10
E,J	{1...5}	Q,F	{8...10}
E,R	{4...7}	-	-

Fig. 1. (A) The identified domains. The color of each domain corresponds to the connected component it belongs to (the blue and green nodes belong to two different poles of the same component). (B) Color map for domain strength. The strength of ENSO (domain E) is shown at the top. (C) Edges to and from ENSO (shown in black). (D) The climate network. The color of each edge represents the corresponding cross-correlation. (E) The lag range associated with each edge. (F) Examples of lag-constistent triangles.

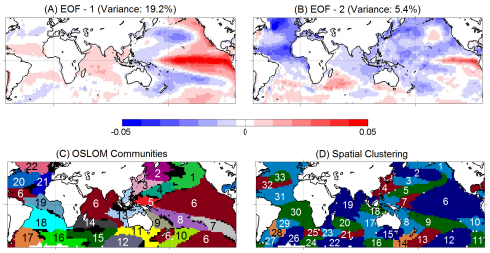


Fig. 2. (A),(B) The first two components of EOF analysis. (C) Communities identified by OSLOM. Each community has a unique number and color. (D) Areas identified by spatial clustering.

Switching to lag inference, we say that a triangle is *lag-consistent* if there is at least one value in the lag range associated with each edge that would place the three nodes in a consistent temporal distance with respect to each other. For instance, in the case of the first triangle of Fig. 1-F, the triangle is lag-consistent if the edge from Q to F has a lag of 8 months and the edge between E and F has lag -2 months (meaning that the direction would be from F to E); several other values would make this triangle lag-consistent. We have verified the lag-consistency of every triangle in the climate network. One exception is the triangle between domains (C, D, G), shown at the bottom of Fig. 1-F. However, the large lag in the edge from C to G can be explained with the triangle between domains (C, E, G), which is lag-consistent. We emphasize that the temporal ordering that results from these lag relations should not be misinterpreted as causality; we expect that several of the edges we identify are only due to indirect correlations, not associated with a causal interaction between the corresponding two nodes.

For comparison purposes, Fig. 2 shows the results of EOF analysis, community detection, and spatial clustering on the same dataset. The first EOF explains only about 19% of the variance, implying that the SST field is too complex to be understood with only one spatial component. On the other hand, the joint interpretation of multiple EOF components is problematic due to their orthogonal relation [3]. The anti-correlation between ENSO and the horseshoe-pattern regions is well captured in the first component but several other important connections, such as the negative and lagged relation between the south subtropical Atlantic and ENSO (domains Q and E, respectively), are missed.

Fig. 2-C shows the results of the overlapping community detection method OSLOM. Following [17], the input to OSLOM is a correlation-based cell-level network. Correlations less than 30% are ignored. The weight of each edge is set to the maximum absolute correlation between the corresponding two cells, across all considered lags. OSLOM identifies 22 communities. Community 6 is not spatially contiguous; it covers ENSO, the Indian ocean, a region in the north tropical Atlantic, and a region in south Pacific. This is a general problem with community detection methods: they cannot distinguish high correlations due to a remote connection from correlations due to spatial proximity. In the context of climate, the former may be due to atmospheric waves or large-scale ocean currents while the latter may be due to local circulations.

Finally, Fig. 2-D shows the results of a spatial clustering method [26], with the same homogeneity threshold δ we use in δ-MAPS. That method ensures that every cluster (referred to as "area") is spatially contiguous but it also requires that there is no overlap between areas and it attempts to assign each grid cell to an area. Consequently, it results in more areas (compared to the number of domains), some of which are just artifacts of the spatial parcellation process. Further, the spatial expanse of an area constrains the computation of subsequent areas because no overlaps are allowed.

Acknowledgements This work was funded from the Department of Energy, Climate and Environmental Sciences Division, SciDAC: Earth System Model Development under Grant DE-SC0007143, and from the National Science Foundation under Grant DMS 1049095.

REFERENCES

[1] H. Von Storch and F. W. Zwiers, *Statistical analysis in climate research*. Cambridge university press, 2001.

[2] M. Vejmelka, L. Pokorná, J. Hlinka, D. Hartman, N. Jajcay, and M. Paluš, "Non-random correlation structures and dimensionality reduction in multivariate climate data," *Climate Dynamics*, vol. 44, no. 9-10, pp. 2663–2682, 2015.

[3] D. Dommenget and M. Latif, "A cautionary note on the interpretation of eofs," *Journal of Climate*, vol. 15, no. 2, pp. 216–225, 2002.

[4] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[5] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter, "Discovery of climate indices using clustering," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 446–455, ACM, 2003.

[6] J. Kawale, S. Liess, A. Kumar, M. Steinbach, P. Snyder, V. Kumar, A. R. Ganguly, N. F. Samatova, and F. Semazzi, "A graph-based approach to find teleconnections in climate data," *Statistical Analysis and Data Mining*, vol. 6, no. 3, pp. 158–179, 2013.

[7] T. Blumensath, T. E. Behrens, and S. M. Smith, "Resting-state fmri single subject cortical parcellation based on region growing," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 188–195, Springer, 2012.

[8] D. Birant and A. Kut, "St-dbscan: An algorithm for clustering spatial–temporal data," *Data & Knowledge Engineering*, vol. 60, no. 1, pp. 208–221, 2007.

[9] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, "A whole brain fmri atlas generated via spatially constrained spectral clustering," *Human brain mapping*, vol. 33, no. 8, pp. 1914–1928, 2012.

[10] M. A. Kramer, U. T. Eden, S. S. Cash, and E. D. Kolaczyk, "Network inference with confidence from multivariate time series," *Physical Review E*, vol. 79, no. 6, p. 061916, 2009.

[11] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, "The backbone of the climate network," *EPL (Europhysics Letters)*, vol. 87, no. 4, p. 48007, 2009.

[12] J. Ludescher, A. Gozolchiani, M. I. Bogachev, A. Bunde, S. Havlin, and H. J. Schellnhuber, "Improved el niño forecasting by cooperativity detection," *Proceedings of the National Academy of Sciences*, vol. 110, no. 29, pp. 11742–11745, 2013.

[13] A. A. Tsonis and P. J. Roebber, "The architecture of the climate network," *Physica A: Statistical Mechanics and its Applications*, vol. 333, pp. 497–504, 2004.

[14] A. Lancichinetti, F. Radicchi, J. J. Ramasco, and S. Fortunato, "Finding statistically significant communities in networks," *PloS one*, vol. 6, no. 4, p. e18961, 2011.

[15] Y. Zhang and D.-Y. Yeung, "Overlapping community detection via bounded nonnegative matrix tri-factorization," in *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 606–614, ACM, 2012.

[16] M. P. McGuire and N. P. Nguyen, "Community structure analysis in big climate data," in *Big Data (Big Data), 2014 IEEE International Conference on*, pp. 38–46, IEEE, 2014.

[17] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, "An exploration of climate data using complex networks," *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 1, pp. 25–32, 2010.

[18] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, "Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science," *Statistical Analysis and Data Mining*, vol. 4, no. 5, pp. 497–511, 2011.

[19] K. Steinhaeuser and A. A. Tsonis, "A climate model intercomparison at the dynamics level," *Climate dynamics*, vol. 42, no. 5-6, pp. 1665–1670, 2014.

[20] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[21] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[22] N. Rayner, D. E. Parker, E. Horton, C. Folland, L. Alexander, D. Rowell, E. Kent, and A. Kaplan, "Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century," *Journal of Geophysical Research: Atmospheres*, vol. 108, no. D14, 2003.

[23] S. A. Klein, B. J. Soden, and N.-C. Lau, "Remote sea surface temperature variations during enso: Evidence for a tropical atmospheric bridge," *Journal of Climate*, vol. 12, no. 4, pp. 917–932, 1999.

[24] D. Easley and J. Kleinberg, *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.

[25] B. Rodríguez-Fonseca, I. Polo, J. García-Serrano, T. Losada, E. Mohino, C. R. Mechoso, and F. Kucharski, "Are atlantic niños enhancing pacific enso events in recent decades?," *Geophysical Research Letters*, vol. 36, no. 20, 2009.

[26] I. Fountalis, A. Bracco, and C. Dovrolis, "Enso in cmip5 simulations: network connectivity from the recent past to the twenty-third century," *Climate Dynamics*, vol. 45, no. 1-2, pp. 511–538, 2015.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

EMPLOYING SOFTWARE ENGINEERING PRINCIPLES TO ENHANCE MANAGEMENT OF CLIMATOLOGICAL DATASETS FOR CORAL REEF ANALYSIS

Mark Jenne[1], Alex Zimmerman[2], Hasan Kurban[1], Claudia Johnson[2], M.M. Dalkilic[1]

***Abstract*—The challenges presented by data to scientific inquiry and hypothesis testing in an oceanographic setting are not new problems. Indeed, the challenges are at least a century old. The problems are not with the data itself, but rather with the attention to the management of the "data ecology" in the information systems. Data needs to be accessible as an input to scientific inquiry—a requirement that goes far beyond simply centralizing the available data. Our research focuses on the development of a proof-of-concept system that properly handles an information ecology. The power of such a strong foundation is then demonstrated in two ways: (1) through our data driven hypothesis generation system, built and employed for analysis of the relationship between coral disease and temperature in the Caribbean; (2) through programmatic search for patterns and anti-patterns that verify, falsify, or demonstrate no discernible relationship for a set of variants on a particular temperature–disease hypothesis.**

I. MOTIVATION

Many oceanographic data repositories have come online in the last few decades. Some repositories are large oceanographic datasets (World Ocean Database (WOD) [1] and World Ocean Atlas (WOA) [2]), while others have more specific content (ReefBase Coral Bleaching GIS [3] and Global Coral Disease Database [4]). Data stored in these repositories, particularly the WOD and WOA, are vast and invaluable. Making the data accessible as a proprietary product, however, is not sufficient for driving large-scale analyses needed to understand the effects of climate change on coral reefs.

The focus of our research is to make climate and coral data available for scientific inquiry in a data-driven approach. We recognize that significant effort and resources are required to ensure longevity of data in an information system. Our proof-of-concept software for data governance and robust management of the data ecology is developed with this in mind. We demonstrate the power of such an approach by building an algorithm for a high-level analysis of coral disease in relation to ocean temperature in the Caribbean on top of our information ecology framework that uses not thousands, nor hundreds of thousands, but millions of data points. Together, the components of our system allow for programmatic search of the coral disease and temperature space to form testable hypotheses—a methodology referred to as data-driven hypothesis generation. It is these data-driven approaches that allows us to perform large-scale analyses needed to address the questions looming large for coral reefs under climate change.

The robustness of our information ecology management system is further demonstrated through a reciprocal approach where the complete data space is searched for patterns and anti-patterns that verify, falsify, or demonstrate no discernible relationship for a temperature–disease hypothesis formulated as an conditional rule-set and fed back into the system. Thus our system acts as a framework for both hypothesis generation and testing.

II. METHOD

The software systems behind the two primary components in this research are: (1) the information ecology framework; (2) the algorithms for the analysis of the relationship between coral disease and temperature. We include only an overview of these components here, but provide formal notation for the algorithms behind the data-driven hypothesis generation and hypothesis pattern search. The configuration for the system, or experimental parameterization, behind the data-driven hypothesis extraction and testing is then described.

Corresponding author: M Jenne, mjenne@indiana.edu [1]School of Informatics and Computing, Indiana University [2]Department of Geological Sciences, Indiana University

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

Rather than following the traditional computational science approach of ushering all of the data to the algorithm and building out the algorithm to incorporate the elements of transformation, management, and processing of the data, we isolate the non-research related procedures and push them to the data. Together, this set of procedures and the controlling software around them, form the backbone of what we are calling our information ecology framework. This system accomplishes two major goals: (1) provisioning a robust data manager with all necessary extraction, transformation, and loading procedures; (2) isolating the processes involved in the scientific inquiry of the algorithm development.

With the data ecology framework in place, the algorithmic components focus on the search of the data space and extraction of isolated trends in the data for hypothesis testing (Alg. 1) and assessment of those trends against a particular hypothesis or hypotheses formulated as a logical rule-set (Alg. 2). As a first pass, our approach leverages a naive quasi-clustering technique for establishing spatial bounds for the geographic extent of coral disease outbreaks and is followed by a search of the large temperature data space for local representative ocean temperature data. The new spatially and temporally associated data are then used to produce visual analytic tools for expert analysis from which individual, testable hypotheses are extracted for further consideration. Hypotheses formed from these trends can then be plugged back in as rule-sets guiding pattern search through the data. This methodology employs data-driven hypothesis generation by trading in the necessity for explicit, testable claims to drive experimental setup in favor of general pattern search within the data pertaining to the relationship between coral disease and temperature.

Reviewing individual geographic locations reveals a potential causal relationship between thermal stress anomalies followed by disease outbreaks. Or stated succinctly: for a particular geographic location, where the annual average sea surface temperature exceeds the regional average during the time period of 1970 to 2009 by some threshold (antecedent), we expect to see an increase in coral disease at that location in the following year (consequent). Forming a similar metric to those presented by Selig, et al [5] for testing temperature–coral disease trends, we translate this hypothesis into a logical rule-set and programmatically search all geographic locations for all instances that support and refute the hypothesis. The consequent of this rule is whether there is an increase [verifies], a reduction [falsifies], or no change [inconclusive] in

Algorithm 1 Coral Disease Temperature Analysis

INPUT data $\{\Delta_1, \Delta_2\}$, config Φ
OUTPUT Temperature–Disease Timelines $A_1, \ldots, A_n \in \mathsf{A}$
%% assume that each A_i is a tuple $(lat, lon, \mathsf{D} \in \Delta_1, \mathsf{T} \in \Delta_2, \mathsf{Y})$
%% where each $Y_i \in \mathsf{Y}$ is $(y, \mathsf{D}_i \subset \mathsf{D}, \mathsf{T}_i \subset \mathsf{T}, \mathsf{C})$
%% y is the year, D_i is the subset of coral diseases at this location for year y, T_i is the subset of temperatures at this location for year y
%% C is the list of corals affected by disease at this location
%% cluster disease instances
for $\mathbf{x} \in \Delta_1$ **do**
  flag $\leftarrow$ false
  for $A_i \in \mathsf{A}$ **do**
    if $\mathbf{x}$.distance$(A_i.lat, A_i.lon) \leq \Phi.radius$ **then**
      $A_i.\mathsf{D} \leftarrow A_i.\mathsf{D} \cup \mathbf{x}$
      $flag \leftarrow true$
    end if
  end for
  if !flag **then**
    $\mathsf{A} \leftarrow \mathsf{A} \cup A(\mathbf{x})$
  end if
end for
%% associate temperature data with disease clusters
for $A_i \in \mathsf{A}$ **do**
  for $Y_j \in A_i.\mathsf{Y}$ **do**
    $itr \leftarrow 0$
    while $\|resultSet\| = 0 \wedge itr < \Phi.maxItr$ **do**
      $results \leftarrow Query(\Delta_2, A_i, Y_j.y, \Phi.rad + (\Phi.rad * (itr/2)))$
    end while
    if $\|resultSet\| = 0$ **then**
      $A_i.\mathsf{T} \leftarrow Query(\Delta_2, Y_j.y)$
    else
      $A_i.\mathsf{T} \leftarrow resultSet$
    end if
  end for
end for
%% process temperature-disease timelines
for $A_i \in \mathsf{A}$ **do**
  for $Y_j \in A_i.\mathsf{Y}$ **do**
    $Y_j.\mathsf{D}_j \leftarrow A_i.\mathsf{D}.Where(x => x.year = Y_j.y)$
    $Y_j.\mathsf{T}_j \leftarrow A_i.\mathsf{T}.Where(x => x.year = Y_j.y)$
    $Y_j.\mathsf{C} \leftarrow Y_j.\mathsf{D}_j.Select(x.genusSpecies => x)$
  end for
end for

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

coral disease instances following a true evaluation of the rule antecedent. Four variants of the rule with the threshold ranging from 1°C to 4°C are tested. Results are presented in Table 1.

Algorithm 2 Temperature–Disease Hypothesis Search

INPUT Temperature–disease timelines $A_1, \ldots, A_n \in \mathsf{A}$, Hypothesis Rule R
OUTPUT Pattern sets that verify V, falsify F, and demonstrate no discernible relationship I
%% assume that each A_i and all of the constituent variables have the same meaning as those presented in Algorithm 1.
for $A_i \in \mathsf{A}$ **do**
  for $Y_j \in A_i.\mathsf{Y}$ **do**
    if $AntecedentMatch(R.Ant, Y_j.\mathsf{T}_j)$ **then**
      if $ConsequentVerify(R.Con, Y_j.\mathsf{D}_j)$ **then**
        $\mathsf{V} \leftarrow \mathsf{V} \cup Y_j$
      end if
      if $ConsequentFalsify(R.Con, Y_j.\mathsf{D}_j)$ **then**
        $\mathsf{F} \leftarrow \mathsf{F} \cup Y_j$
      end if
      if $ConsequentInconclusive(R.Con, Y_j.\mathsf{D}_j)$ **then**
        $\mathsf{I} \leftarrow \mathsf{I} \cup Y_j$
      end if
    end if
  end for
end for

III. EVALUATION

The results presented here are in the context of both a big data problem and large-scale analyses. Making use of our data ecology framework, our algorithm for data-driven hypothesis generation regarding the temperature-coral disease relationship in the Caribbean was able to integrate and process the complete coral disease catalog presented by ReefBase and the complete ocean temperature data set hosted in the WOD. Grouping the 5,038 coral disease records into spatial clusters yielded 293 distinct geographic locations for analysis. At each location, respective temperature data subsets were selected from the more than 62 million data points available. The resulting coral disease and temperature sets were grouped together and visualized for extraction of testable hypotheses. We now address components of a compound hypothesis stating that a 2°C temperature rise and pH reduction of about 0.1 are more than sufficient to cause extensive stress and mortality to corals [6]. Our hypothesis regarding regional thermal stress anomalies preceding coral disease outbreaks was formed as a logical rule-set and fed into the system to see if the trends in the data verify or falsify the hypothesis. The results are found in TABLE I. It is informative that in the Caribbean, the data show verification/falsification counts are similar for ≤ 2 °C, but > 2 °C the hypothesis appears valid. As importantly, we observe that the data show relatively few instances of coral diseases at these temperatures, likely because these temperatures are rarely observed in the Caribbean, especially when all depths of the ocean are pooled together. This data-driven hypothesis generation technique suggests, for the high temperatures, an analysis of the shallow water temperatures separate from the deeper, cooler waters would be warranted, and would further test the hypothesis relating high temperatures to coral diseases.

TABLE I: Compound Hypothesis Analysis

TSA	Verify	Falsify	Inconclusive
≥ 1°C	68	67	1625
≥ 2°C	29	28	539
≥ 3°C	11	3	126
≥ 4°C	4	0	23

Cases that verify, falsify, or demonstrate no discernible relationship for the temperature–coral disease hypothesis for *thermal stress anomalies (TSA)* ranging from 1°C to 4°C.

This data-driven hypothesis generation and testing approach explores a proof-of-concept through a hypothesis rule-set and, as such, suffers from an incomplete picture of the biotic data. Incorporation of additional data *e.g.*, reef coverage, coral counts, coral mortality, would help the construction of a more complete model.

Here we have demonstrated the benefit that a robust information ecology management system lends to hypothesis generation and testing. Use of our system made the individual data sets involved easily accessible as input to our scientific inquiry, which allowed us to perform Caribbean-wide analyses in exploring the relationship between ocean temperature and coral disease. Proper data management in concert with these data-driven approaches will further allow us to perform large-scale analyses needed to address the questions looming large for coral reefs under climate change.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] T. Boyer, J. A. O. Baranova, C. Coleman, H. Garcia, A. Grodsky, D. Johnson, R. Locarnini, A. Mishonov, T. O'Brien, C. Paver, J. Reagan, D. Seidov, I. Smolyar, and M. Zweng, "World ocean database," *NOAA Atlas NESDIS*, vol. 72, p. 209, 2013.

[2] R. A. Locarnini, A. V. Mishonov, J. I. Antono, T. P. Boyer, H. E. Garcia, O. K. Baranova, M. M. Zweng, C. R. Paver, J. R. Reagan, D. R. Johnson, M. Hamilton, and D. Seidov, "World ocean atlas 2013, volume 1: Temperature," *NOAA Atlas NESDIS*, vol. 73, p. 40, 2013.

[3] "Reefbase coral diseases." Accessed: 2016-06-15.

[4] "Global coral disease database." Accessed: 2016-06-15.

[5] E. R. Selig, C. Drew Harvell, J. F. Bruno, B. L. Willis, C. A. Page, K. S. Casey, and H. Sweatman, "Analyzing the relationship between ocean temperature anomalies and coral disease outbreaks at broad spatial scales," *Coral reefs and climate change: science and management*, pp. 111–128, 2006.

[6] O. Hoegh-Guldberg, P. J. Mumby, A. J. Hooten, R. S. Steneck, P. Greenfield, E. Gomez, C. D. Harvell, P. F. Sale, A. J. Edwards, K. Caldeira, *et al.*, "Coral reefs under rapid climate change and ocean acidification," *science*, vol. 318, no. 5857, pp. 1737–1742, 2007.

[7] S. Levitus, "Climatological atlas of the world ocean," *Eos, Transactions American Geophysical Union*, vol. 64, no. 49, pp. 962–963, 1983.

[8] T. H. Davenport and L. Prusak, *Information Ecology: Mastering the Information and Knowledge Environment*. Oxford University Press, 1st ed., 1997.

[9] A. de Geus, *The Living Company: Habits for Survival in a Turbulent Business Environment*. Harvard Business School Press, 1st ed., 1997.

[10] T. R. McClanahan, M. Ateweberhan, C. A. Muhando, J. Maina, and M. S. Mohammed, "Effects of climate and seawater temperature variation on coral bleaching and mortality," *Ecological Monographs*, vol. 77, no. 4, pp. 503–525, 2007.

[11] R. J. Jones, J. Bowyer, O. Hoegh-Guldberg, and L. L. Blackall, "Dynamics of a temperature-related coral disease outbreak," *Marine Ecology Progress Series*, vol. 281, pp. 63–77, 2004.

[12] J. W. Porter, P. Dustan, W. C. Jaap, K. L. Patterson, V. Kosmynin, O. W. Meier, M. E. Patterson, and M. Parsons, "Patterns of spread of coral disease in the florida keys," in *The Ecology and Etiology of Newly Emerging Marine Diseases*, pp. 1–24, Springer, 2001.

[13] W. L. Hürsch and C. V. Lopes, "Separation of concerns," 1995.

[14] E. Gamma, R. Helm, R. Johnson, and J. Vlissides, *Design patterns: elements of reusable object-oriented software*. Pearson Education India, 1995.

6th International Workshop on Climate Informatics
September 22-23, 2016
CI 2016 Hosted by the National Center for Atmospheric Research in Boulder, CO

Profiler Guided Manual Optimization for Accelerating Cholesky Decomposition on R Environment

V.B. Ramakrishnaiah[1], R.P. Kumar[1, 3], J. Paige[2], D. Hammerling[3], D. Nychka[3], R. Loft[3]

Abstract— **fields is a spatial statistics package in R that is supported by the National Center for Atmospheric Research (NCAR) and is widely used to analyze spatial data. We made use of the Matrix Algebra on GPU and Multicore Architectures (MAGMA) [1][2][3] library to accelerate the Cholesky decomposition in this package. The acceleration of the Cholesky decomposition was motivated by a) its role as a core computation in spatial statistics, b) its relative simplicity in implementing on GPU architectures, and c) its suitability for parallelization. A key benefit of this project is linking a high level and flexible data language with a lower level and parallel support for computationally intensive steps. In this project, we observed some unexpected behaviors such as multiple GPUs being slower than a single GPU, and in-place decomposition being faster than deep copy. CPU and GPU profiling helped to explain the unconventional behavior observed in the multi-GPU executions and to develop strategies to accelerate the Cholesky decomposition. These strategies include accelerating the underlying C function, reducing the function call overheads in R and optimizing the R environment. We were able to optimize the code and the environment to get a speedup greater than 75x for large matrices. We also integrated our accelerated C functions with Julia and drew a performance comparison between R and Julia. Julia was found to significantly reduce the function call overheads when compared to R. We also uncovered a way to improve the MAGMA functions themselves by replacing the intra-node, inter-GPU communications with direct device-to-device calls.**

I. INTRODUCTION

We depend on physical models all the time, and climate models play an important role in predicting a plausible future geospatial environment. One of the biggest computational challenges in the analysis of spatial data is determining the parameters that control how the spatial field is correlated as a function of distance. Maximum Likelihood Estimation is an accurate way to estimate these covariance parameters but is computationally intensive as the number of spatial locations grows. This is due to the fact that the core computation is the Cholesky decomposition of a positive definite matrix, the covariance matrix of the spatial observations. Thus, for n spatial locations the computational complexity grows as $O(n^3)$ [4]. The other parts of the spatial analysis are not as demanding and so it is appropriate to implement the bulk of the computations in a higher-level language such as R, which most data analysts are familiar with. R provides simple interfaces to C, so we incorporated a parallel linear algebra library into R to take advantage of multiple cores and GPUs for the Cholesky decomposition. In particular, the Matrix Algebra on GPU and Multicore Architectures (MAGMA) [1][2][3] library has good performance and is open source. One advantage of R is that this decomposition can be overloaded for the R Cholesky method and so, beyond adding the specialized MAGMA functions that support the overloading, few additional changes were required to accelerate the `fields` package.

The accelerated code demonstrated some surprising behavior that highlights the practical issues of harnessing coprocessors for scientific computation. The multi-GPU implementation of the Cholesky decomposition was slower than the single GPU implementation and the deep copy version of the code was slower than the shallow copy version.

II. Initial Assessment

There are R packages such as **HiPLAR** [5] that make use of PLASMA [1][6] and MAGMA libraries to accelerate linear algebra computations, but invoking them from the `fields` package will cause additional overheads of loading the corresponding R environments (objects) and will slow down the computations. So we decided to directly invoke the MAGMA library from the `fields` package. MAGMA provides two functions that compute the Cholesky decomposition: 'dpotrf_m' is a CPU function making GPU calls, and 'dpotrf_mgpu' is a GPU function. We used the NVIDIA visual profiler to profile the code and found that the 'dpotrf_mgpu' performed better than 'dpotrf_m'. But when we integrated it with the R environment the performance was worse. Since the profiling results demonstrated the working of the underlying C code, we wanted to get an idea about the overheads in R calls. So, we timed different sections of the code separately and found that R calls incur significant overheads while calling C functions. In addition, given R's legacy as a serial code with modest memory demands, we noticed that the R environment could be optimized for using MAGMA. All testing was performed on the Caldera nodes of the Yellowstone [7] supercomputing environment.

III. Optimization Approaches

With insight from the profiler and timing results we used the following approaches. First, we accelerated the underlying C functions. We found that

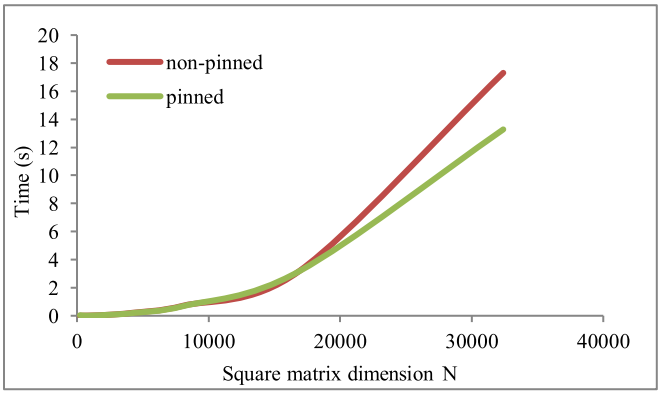


Figure 1: Performance improvement by allocating single precision matrices on pinned memory.

the inability to allocate pinned memory in R had a significant impact on the performance of 'dpotrf_mgpu' [8]. We had multiple versions of the code including a single precision version, which was targeted towards users willing to trade off accuracy for speed. (A separate study demonstrated that single precision arithmetic is adequate for the likelihood computations [9].) Since R always uses double precision for calculations, we forced a manual copy to a single precision variable in C. This variable was allocated on pinned memory, which improved the performance of the single precision code. Figure 1 shows the improvement in performance of the code by allocating the single precision variable on pinned memory. We can see that the improvement in performance is significant for larger data sizes as the page-able memory starts swapping data with the virtual memory (secondary storage).

To include compiler level optimizations, we rebuilt MAGMA with the latest Intel 2015 compiler [10] and incorporated it with R to compile shared objects. This provided us some additional performance benefits, which can be seen in Figure 2.

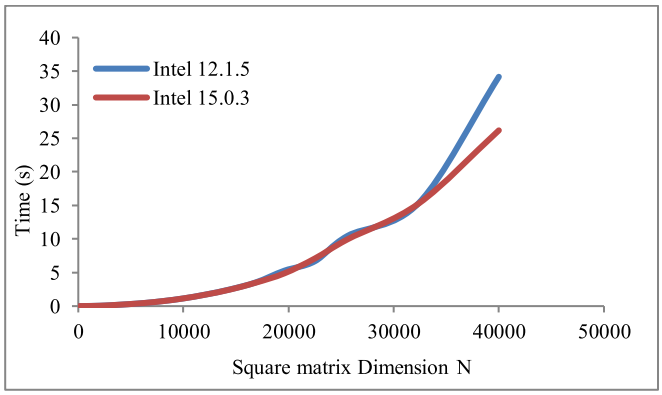


Figure 2: Improvement with Intel 2015 compiler.

Finally, to reduce the function call overheads in R we used the following techniques:

- Preloaded the Intel MKL libraries, so that they are readily available during the time of execution.
- Created a single shared object for dynamic loading in R.
- Moved dynamic loads to the highest level of program flow to avoid R environment overheads.

With these strategies, we gained some additional performance benefits compared to a naïve implementation of the MAGMA library. This can be seen in Figure 3, and we can see that the performance

benefits are modest but more pronounced for large data sizes.

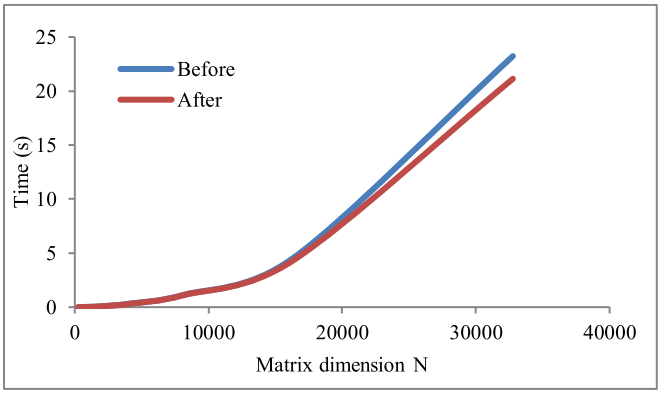


Figure 3: Speedup from reduced function call overheads in R.

IV. Performance Improvement with Julia

The overheads in R led us to investigate calling these C functions from a more recently developed environment called 'Julia'. Julia is a high level language similar to R but is well suited for high performance computing applications. Because it is a recent development, it has the disadvantage of a smaller user base and much fewer contributed packages for data analysis and statistics. Julia, however, has performance close to the C programming language and new function libraries are being developed on a regular basis. Using Julia to call the underlying C functions in our code reduced the time and memory overheads of calling C functions. For example, the execution time for computing the Cholesky Decomposition for a matrix of size 32,400 was about 16.21s when called from Julia, whereas, R was taking about 28.97s. This is because the C functions can be called without any additional 'glue' code, which improves performance.

V. Conclusion and Future Work

We achieved up to 75x (single precision) and 65x (double precision) speedup compared to the default, single thread Cholesky decomposition on a single CPU (Figure 4). This significantly improved the Kriging performance (Figure 5), which is an important technique used in analysis of climate models. In Figure 4, we see that the multiple-GPU version performs better than the single GPU version, which indicates good program scalability. The amount of memory used in the case of deep copy is larger, resulting in more data swaps to the virtual memory. That is the reason why the deep copy version is slower than the in-place version. This was verified by using the 'perf' tool in Linux to monitor the last-level cache and translation look-aside buffer misses. Comparison of Julia with R showed that overheads in Julia were significantly lower compared to R. We observed that the accelerated C functions computing the Cholesky Decomposition performed better using Julia (greater than 80% improvement) compared to the R environment. For future work, we plan to replace the MAGMA code with direct device-to-device communication to avoid involving the CPU.

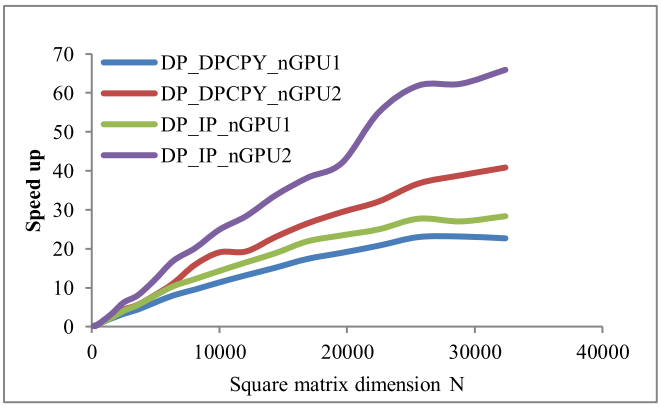


Figure 4: Improvement in Cholesky Decomposition performance. Abbreviations: DP=Double Precision, DPCPY=Deep copy, IP=In place, nGPU=number of GPUs.

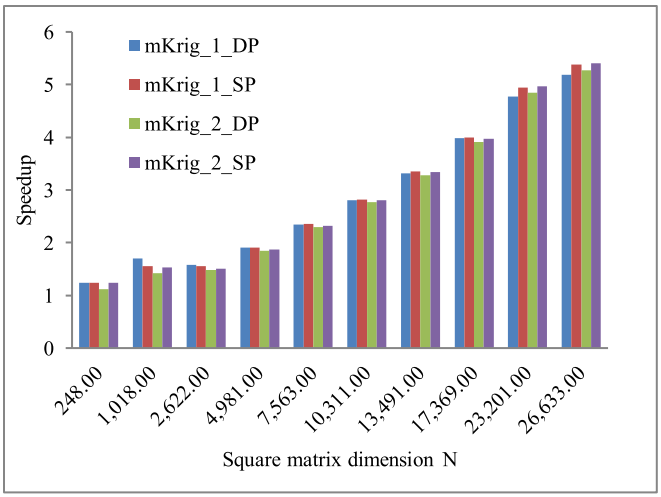


Figure 5: Improvement in Kriging performance. Abbreviations: DP=Double Precision, SP=Single Precision, mKrig_x=number of GPUs used for Kriging.

REFERENCES

[1] Agullo, E., Demmel, J., Dongarra, J., Hadri, B., Kurzak, J., Langou, J., et al. Numerical Linear Algebra on Emerging Architectures: The PLASMA and MAGMA projects. *Journel of Physics: Conference series , 180* (012037).

[2] Tomov, S., Dongarra, J., Volkov, V., & Demmel, J. *MAGMA library.* University of Tenessee and University of California, Knoxville, TN, and Berkeley, CA.

[3] http://icl.cs.utk.edu/magma/.

[4] Katzfuss, M., & Cressie, N. (2012). Bayesian hierarchical spatio-temporal smoothing for very large datasets (Vol. 23). Environmetrics.

[5] https://developer.nvidia.com/hiplar.

[6] http://icl.cs.utk.edu/plasma/.

[7] *https://www2.cisl.ucar.edu/resources/yellowstone*.

[8] *http://devblogs.nvidia.com/parallelforall/how-optimize-data-transfers-cuda-cc/*.

[9] Paige, J., Lyngaas, I., Ramakrishnaiah, V., Hammerling , D., Kumar, R., & Nychka, D. (2015). *Incorporating MAGMA into the 'fields' spatial statistics package.* Boulder: UCAR/NCAR.

[10] *https://software.intel.com/en-us/articles/optimizing-without-breaking-a-sweat*.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

GLOBAL MONITORING OF SURFACE WATER EXTENT DYNAMICS USING SATELLITE DATA

Anuj Karpatne[1], Ankush Khandelwal[1] and Vipin Kumar [1]

***Abstract*—Freshwater, which is only available in inland water bodies such as lakes, reservoirs, and rivers, is increasingly becoming scarce across the world and this scarcity is posing a global threat to human sustainability. A global monitoring of surface water bodies is necessary for policy-makers and the scientific community to address this problem. The promise of data-driven approaches coupled with the availability of remote sensing data presents opportunities as well as challenges for global monitoring. One of the major challenges in monitoring surface water is the presence of a rich variety of land and water bodies across the world, that show varying and sometimes overlapping characteristics in remote sensing signals. This heterogeneity within the land and water classes makes it difficult to use traditional classification approaches for differentiating between all types of land and water bodies at a global scale. Our research aims at developing predictive models that address this challenge for creating the first global monitoring system of surface water dynamics. This system can greatly enhance our understanding of the interplay between climate change, human actions, and surface water dynamics.**

I. MOTIVATION AND BACKGROUND

Inland water bodies, which include all water sources contained within landmasses, such as lakes, reservoirs, and rivers, are important natural resources as they sustain every form of terrestrial life on Earth [1]. Increased incidence of adverse events such as dwindling ground water, shrinking freshwater bodies, rapidly degrading water quality, severe droughts, and devastating floods not only pose a significant threat (see Figure 1) to the sustainability of humans but also to the Earth's ecosystem. As a result, managing inland water has become one of the major 21st century challenges for the world [2]. A global water monitoring system that can provide timely and accurate information about the available water stocks across the world is critical for managing water resources.

Corresponding author: Anuj Karpatne, anuj@cs.umn.edu [1] University of Minnesota

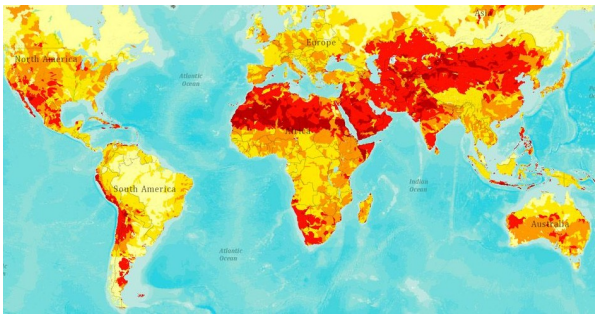
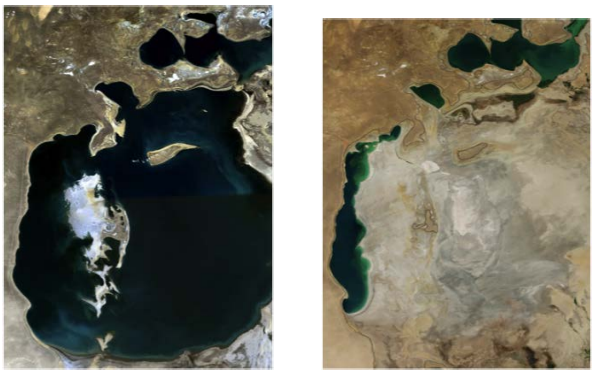


Figure 1. Global map showing regions facing water threats, indicated in red [3].

A global monitoring system will enable a number of advances in understanding the dynamics of water resources and its management. First, monitoring water dynamics can help in assessing the impact of human actions on the state of inland water bodies. As an example, the Aral sea has been shrinking since the 1960s (shown in Figure 2) due to the undertaking of several irrigation projects by the Soviet Union, which has brought the lake to the verge of extinction in 2014. Second, information pertaining to the dynamics of inland water bodies will aid in discovering relationships between changes occurring in different water bodies



(a) Image in 1989 (b) Image in 2014

Figure 2. Satellite images from NASA Earth Observatory showing the shrinking of Aral Sea starting from 1960's to present day.

and their interactions with other climatic processes, such as heat waves and precipitation extremes. Third, a global monitoring system would also facilitate the forecasting of water stocks and risks in the future, which when coupled with information about the projected water demands can help in devising policies for managing water in a timely and effective manner.

The potential in creating a global water monitoring system is enabled by the availability of remote sensing data, acquired via satellites orbiting the Earth. Remote sensing data has made it possible to obtain observations even for the most inaccessible regions of the Earth that could not have been otherwise obtained using ground-based sensors [4]. Remote sensing data provides global coverage of a variety of physical attributes about the Earths surface at fine spatial resolutions and frequent time intervals. This information can be appropriately leveraged for distinguishing water bodies from land bodies [5], [6]. Data-driven approaches that mine useful information from large data sources and have found success in various applications (e.g. finance, advertising, and social network analysis) offer a promise in constructing a global water monitoring system using remote sensing data. However, traditional data-driven approaches are not suitable for addressing the unique challenges faced in analyzing remote sensing data at a global scale, as discussed in detail in Section 2. This has restricted the application of existing water monitoring approaches to local and regional scales as highlighted in [7], while no effort has yet been made to monitor the dynamics of inland water bodies at a global scale. This motivates the need for novel research in data-driven approaches for global monitoring of inland water dynamics.

II. Goals and Challenges

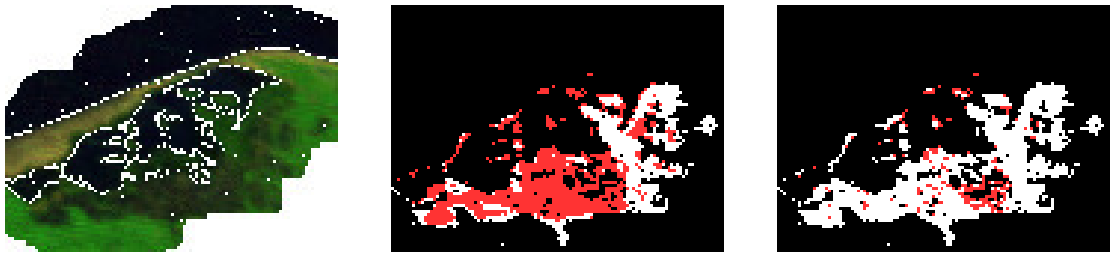
The ultimate objective of this project is to develop predictive models that are able to identify whether a particular location on the Earth at a given time is water or land (binary classes) using remote sensing data. However, a major challenge in learning predictive models for global water monitoring is the fact that land and water bodies appear very different in remote sensing signals in different regions of the Earth, due to the presence of varying geographies, topographies, and climatic conditions across the world. Furthermore, the same land or water body can show different characteristics at different times, due to the presence of Earths seasonal cycles and inter-annual changes. This heterogeneity within land and water bodies results in a multi-distribution of both classes, where different pairs of land and water modes show different degrees of separability in the feature space of remote sensing variables. In such scenarios, the traditional approach of learning predictive models that differentiate between all varieties of land and water modes would suffer from a number of limitations. First, the performance of such an approach may be reasonable for certain pairs of modes that are easily separable from each other, but may be poor for pairs of modes that are highly overlapping in the feature space, termed as pairs of confusing modes. For example, a pair of land and water modes, being observed in different regions and times, may show similar remote sensing characteristics, making them difficult to be differentiated from each other. Second, the presence of such pairs of confusing modes can also impact the performance over other modes in the data, that are reasonably separable in the feature space. Third, the learning of a traditional predictive model can be biased towards certain modes in the data that have been favorably represented in the training set, resulting in improper learning of the classifier over modes that have been under-represented during training.

III. Methods

In our research, we have explored ensemble learning methods that can address the address the aforementioned challenge of heterogeneity within the water (positive) and land (negative) classes for global surface water monitoring. These methods can be briefly described as follows.

In the presence of heterogeneity within the two classes, every pair of positive and negative modes requires the learning of a different classifier that is designed to differentiate between instances belonging to the given pair of modes. This can be achieved by learning an ensemble of classifiers, where every classifier differentiates between a different pair of positive and negative modes. Such an approach would help in ensuring adequate representation of every mode in the learning of the classifier ensemble, along with maintaining diversity among the classifiers. In contrast to traditional ensemble learning approaches (e.g. bagging) that use random partitions of the input space, constructing ensemble classifiers in accordance with the multi-modal structure of every class results in improved classification performance across all pairs of positive and negative modes. In our previous work [8], we have explored various strategies for constructing ensembles of classifiers that take into account the heterogeneity within the two classes.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO



(a) Remote sensing image (b) Errors of baseline approach (in red)) (c) Errors of proposed approach (in red)

Figure 3. Classification results on a land category near Burullus lake, Egypt, shown by red and white pixels in Figures 3(b) and 3(c). Red pixels represent classification errors while white pixels represent correctly classified instances.

It should be noted that even though land and water bodies exist as multiple modes in the feature space, only a small subset of land and water modes appear in the vicinity of a given water body at a certain time, which may be easily separable from each other. To make use of this property, we consider binary classification problems where the test data arrives as groups of test instances, termed as test scenarios, and every test scenario involves only a subset of all the positive and negative modes in the data. Note that the concept of a test scenario depends on the grouping structure present in the test data, which is common in a number of real-world classification problems. For example, a test scenario could comprise of instances observed in the vicinity of the same water body at the same time-step, which are contextually similar to each other and thus require the same learning of a classifier. In such a setting, different pairs of positive and negative modes may emerge or disappear in different test scenarios, and even though some modes may be participating in class confusion at a global scale, the subset of modes appearing in a given test scenario can be considered to be locally separable among each other. This shows a promise in using information about the context of a test scenario for overcoming class confusion.

In our previous work [9], we have developed ensemble learning methods that are able to make use of the distribution of unlabeled instances in a test scenario for assigning local weights to ensemble classifiers. This helps in selecting classifiers that are locally relevant in the context of a given test scenario, and discarding classifiers that are irrelevant or show poor prediction performance. Thus, by using locally adaptive weights on ensemble classifiers in accordance with the context of a test scenario, we are able to provide significantly better classification performance even in the presence of class confusion. Figure 3 shows differences in the classification results of the proposed approach and a baseline approach (that does not make use of the context of test instances) over a certain land mode in Burullus lake, Egypt in Feb 2000, that participates in class confusion at a global scale (shown in red and white pixels in Figures 3(b) and 3(c)). It can be seen that the errors of the proposed approach are significantly fewer than that of the baseline approach in the local context of this lake at this time, which shows the power in using information about the local context of test instances.

IV. BROADER VISION AND IMPACTS:

As a first step towards creating a global surface water monitoring system using data-driven approaches, we have created a preliminary version of a web-viewer for visualizing changes occurring in water bodies: http://z.umn.edu/monitoringwater. This viewer is able to capture a variety of dynamics occurring in surface water, e.g. melting of glacial lakes in Tibet, shrinking water bodies in Brazil and California due to droughts, construction of dams and reservoirs, and changes in river morphology such as river migration and delta erosion (see http://z.umn.edu/waterslides for further details). We envision a global water monitoring system to be a key enabler in identifying changes in surface water and studying their interactions with climate change and human actions. Furthermore, the predictive learning approaches developed as part of this project will have wide applicability in several real-world applications that involve heterogeneity in data populations, e.g. detecting ecosystem disturbances using remote sensing data and predicting disease risks using health-care data.

V. ACKNOWLEDGMENTS

This work was supported by the National Science Foundation Award 1029711 and the NASA Award NNX12AP37G. Access to data storage and computing facilities was provided by the University of Minnesota Supercomputing Institute and the NASA Earth Exchange.

REFERENCES

[1] C. J. Vörösmarty, P. Green, J. Salisbury, and R. B. Lammers, "Global water resources: vulnerability from climate change and population growth," *Science*, vol. 289, no. 5477, pp. 284–288, 2000.

[2] J. Hall, D. Grey, D. Garrick, F. Fung, C. Brown, S. Dadson, and C. Sadoff, "Coping with the curse of freshwater variability," *Science*, vol. 346, no. 6208, pp. 429–430, 2014.

[3] F. Gassert, M. Landis, M. Luck, P. Reig, and T. Shiao, "Aqueduct global maps 2.0," *Water Resources Institute (WRI): Washington, DC*, 2013.

[4] A. Karpatne, J. Faghmous, J. Kawale, L. Styles, M. Blank, V. Mithal, X. Chen, A. Khandelwal, S. Boriah, K. Steinhaeuser, *et al.*, "Earth science applications of sensor data," in *Managing and Mining Sensor Data*, pp. 505–530, Springer, 2013.

[5] E. P. Crist and R. C. Cicone, "A physically-based transformation of thematic mapper data— the TM tasseled cap," *IEEE Transactions On Geoscience and Remote Sensing*, vol. 22, no. 3, pp. 256–263, 1984.

[6] H. Gao, C. Birkett, and D. P. Lettenmaier, "Global monitoring of large reservoir storage from satellite remote sensing," *Water Resources Research*, vol. 48, no. 9, 2012.

[7] A. Karpatne, A. Khandelwal, X. Chen, V. Mithal, J. Faghmous, and V. Kumar, "Global monitoring of inland water dynamics: State-of-the-art, challenges, and opportunities," in *Computational Sustainability* (K. Morik, K. Kersting, and J. Laessig, eds.), Springer, 2015 (in production).

[8] A. Karpatne, A. Khandelwal, and V. Kumar, "Ensemble learning methods for binary classification with multi-modality within the classes," *SIAM International Conference on Data Mining (SDM)*, 2015 (accepted).

[9] A. Karpatne, A. Khandelwal, and V. Kumar, "A local learning algorithm for classification with multi-modal data: An application in global lake monitoring," *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015 (in review).

[10] C. J. Gleason and L. C. Smith, "Toward global mapping of river discharge using satellite images and at-many-stations hydraulic geometry," *Proceedings of the National Academy of Sciences*, vol. 111, no. 13, pp. 4788–4791, 2014.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

QUANTIFYING TROPICAL CYCLONE RISK USING POISSON MODELING

Erica M. Staehling[1*] and Ryan E. Truchelut[2,1*]

Abstract— **Hurricane landfall risk has substantial social and economic implications, yet extant methods of diagnosing annual Atlantic tropical cyclone (TC) activity demonstrate no skill in diagnosing U.S. hurricane landfalls. Using extended landfall activity and reanalysis datasets, we produced a novel Landfall Diagnostic Index (LDI), which captures more of the interannual variability in U.S. hurricane landfalls than current genesis indices or even Atlantic TC count itself. LDI also offers physical insight into why indices that successfully capture TC activity fail to diagnose landfalls: there is an inherent tension between conditions likely to steer hurricanes toward the U.S and conditions favorable for TC development.**

I. INTRODUCTION

The first empirical index to condense multiple oceanic and atmospheric fields into a single metric that expresses the relative permissibility of cyclogenesis was introduced in 1979 [1] and there has been renewed interest in the development of such indices in recent years with the expanded availability of high-quality atmospheric re-analysis datasets as in [2], [3], [4], [5], and [6]. Genesis indices (GIs) have a variety of contemporary uses, having garnered attention as potential proxies for changes in the number of TCs in climate change simulations, especially in situations where only large-scale fields are available due to computational limitations, as shown by [7], [8], [9], and [10].

We compared and evaluated the performance of extant GIs in diagnosing aggregate TC activity as well as U.S. hurricane landfall activity, focusing on U.S. hurricane landfalls because the vast majority of TC economic damage in the Atlantic is due to high-intensity landfalling events and the historical record of U.S. hurricane landfalls is accurate earlier than the full Atlantic TC record [11], [12]. We obtained hurricane landfall records from the International Best Track Archive for Climate Stewardship (IBTrACS [13]) and atmospheric fields from the NOAA/CIRES 20th Century Reanalysis version 2 (20CRv2 [14]), which overlap and are reliable from 1900 through 2012. These choices afford us a temporal domain more than twice the length of previous GI studies, with n = 113 years. We used 20CRv2 to reproduce extant GIs for the historical record, yielding comparable or better performance relative to previously published results, despite the fact that these GIs were trained using different reanalyses. This builds confidence that 20CRv2 is a sufficiently accurate basis on which to build a diagnostic model.

Since hurricane landfall count has a significant relationship with overall TC activity (R^2 = 0.166, p = 0.0027, n = 50 for 1966-2015), it is tempting to assume GIs that show skill in diagnosing TC activity are also useful proxies for landfall incidence, but this is not the case. We calculated and regressed the seasonal mean value of each extant GI averaged over both the entire Atlantic basin and the extended main development region (EMDR), onto annual Atlantic TC count and U.S. hurricane landfall count, and found that most GIs averaged over the EMDR explain about half of the interannual variance in Atlantic TC activity. In contrast, there is no significant correlation evident between any of the GIs and the seasonal hurricane landfall count.

TC count and intensity have energetic and moisture implications for the atmospheric system at large, but it is landfall activity that has direct impacts on human populations and the coastal environment. While several predictive studies of U.S. hurricane landfall activity exist in the literature in [15], [16], [17], analogous approaches to diagnose landfalls are rare [18]. In this study, we investigate the utility of adapting and expanding the GI methodology to historical records of U.S. hurricane landfall to create a Landfall Diagnostic Index (LDI).

Corresponding author: E. Staehling, erica@weathertiger.com
[1]Research and Development Division, WeatherTiger LLC, [2]Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL. *Authors contributed equally to this research.

II. NEW INDEX DEVELOPMENT

The lack of landfall diagnostic skill in GIs points to the importance of factors other than overall activity for assessing landfall risks to coastal regions, including genesis location [19], intensification processes [20], and large-scale steering patterns [15]. To address this utility gap, we adapted the index generation methodology of [5] to produce LDI, namely employing empirical Poisson linear regression modeling to build a diagnostic index from a database of physically explicable predictors. This study differs from [5] in that we substituted hurricane landfall count for TC count as the diagnosed quantity, used time- and area-averaged fields, and considered additional predictors related to large-scale steering and TC intensity regulating processes.

Using stepwise forward Poisson linear regression modeling, we tested potential fields drawn from the 20CRv2 for inclusion in LDI. Our final version of LDI includes upper-tropospheric horizontal divergence ($\nabla_h \cdot \boldsymbol{v}$) between 250-100 hPa (the difference in vertical motion between these levels), meridional wind (v) averaged over 500-650 hPa, zonal shear vorticity ($\partial u/\partial y$) at 1000 hPa, and relative sea surface temperature (rSST) [21]. To determine the relative sensitivity of hurricane landfall activity to the four terms, we normalized the time series of each component, producing normalized LDI:

$$\mathrm{LDI}_\sigma = \exp(0.45+0.47(\nabla_h \cdot \boldsymbol{v})-0.22(v)-0.10(\partial u/\partial y)+0.06(\mathrm{rSST})),$$

so that the magnitudes of term coefficients reveal the sensitivity to proportional changes in each component.

As a check against overfitting, we performed cross-validation tests [22], comparing both in-sample and out-of-sample LDI performance against several baseline diagnostic methodologies, including the constant mean count of annual hurricane landfalls, a linear fit regressing year onto landfall count, a ten-year trailing average of landfalls, the best-performing extant GIs, and a linear regression model incorporating seasonally-averaged Atlantic Multi-decadal Oscillation, extended multivariate ENSO index, and North Atlantic Oscillation values. Both in-sample and out-of-sample LDI show significant diagnostic skill beyond all baseline methods with greater than 95% confidence.

In-sample and out-of-sample LDI explain 31.4% and 26.6% of hurricane landfall interannual variability, respectively, with $p = 0.014$ and $p = 0.068$ relative to TC count per Wilcoxon signed rank tests [23]. As Atlantic TCs are a necessary but not sufficient condition for U.S. hurricane landfall, to outperform TC count, LDI must draw skill from genesis location, intensification, or steering patterns, offering insight into processes that influence landfall risk. Comparing the annual expected value of U.S. hurricane landfalls indicated by LDI with the observational record, the index generally correctly sorts seasons into less active than normal, near normal, and more active than normal, with the mean value of both LDI and observed hurricane landfalls in the top decile of LDI approximately twice the historical mean of 1.73 U.S. hurricane landfalls per year, and the mean value in the bottom decile less than half the mean.

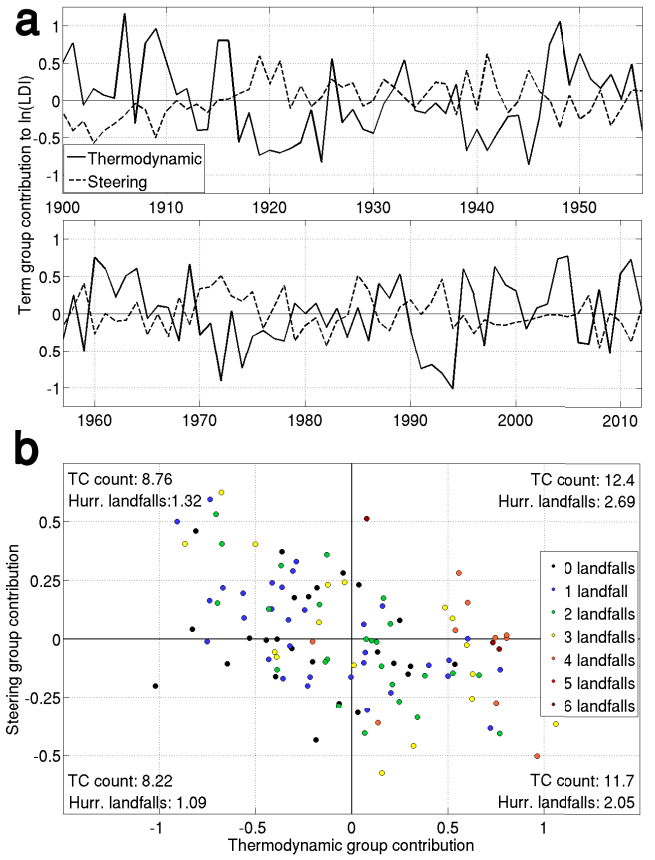


Figure 1. (**a**) The interannual variation of the thermodynamic (solid; divergence and rSST) and steering (dashed; meridional wind and zonal shear vorticity) term group contributions to ln(LDI). (**b**) Scatter plot of the contributions of the term groups, with aggregate TC count and landfall outcomes by quadrant.

III. PHYSICAL INTERPRETATION

Each of LDI's four components is closely tied to processes that develop, steer, and regulate the intensity of TCs in the Atlantic basin, offering physical insight into sources of interannual variance. Since LDI is an exponential sum, we examined the anomaly of each component to quantify its individual contribution [5]. We call the divergence and rSST components the "thermodynamic terms," as they are physically linked to storm-scale convective, energetic [24], and circulation processes [25], and call the meridional wind and shear vorticity components the "steering terms," linked to large-scale TC track patterns [15].

The contributions of the thermodynamic and steering groups to ln(LDI) are shown in Fig. 1 to be negatively correlated ($R = -0.495$, $p = 2.6 \times 10^{-8}$), suggesting a tension between environmental conditions conducive for development and landfall-favoring steering patterns. The scatterplot of steering versus thermodynamic group annual ln(LDI) contributions shown in Fig. 1b includes TC and hurricane landfall counts averaged within the quadrants dividing positive and negative thermodynamic and steering term contributions. Moving from negative to positive steering term contribution while holding the sign of thermodynamic term contribution fixed, there is a marginal increase in quadrant-averaged TC count (less than 7%) and a much larger increase in hurricane landfall count (roughly 20-30%). Correlating each term group with the 500 hPa geopotential height anomalies from the re-analysis data shows the patterns associated with the thermodynamic and steering morphologies most favorable for U.S. landfalls are in near-diametric opposition over much of the Northern Hemisphere, further evidence of an intrinsic tension between the two term groups of LDI.

IV. CONCLUSIONS

LDI successfully diagnoses a significant portion of U.S. hurricane landfall variance in the twentieth century and reveals potential reasons for the historical difficulty in understanding interannual variability in U.S. landfall activity, elucidating the physical relationship between TC count and hurricane landfall count. As GIs have been invoked in the context of climate models unable to resolve TC-scale structures [27], LDI could serve as a proxy for changes in the number of landfalling hurricanes in climate change simulations, especially since even models and downscaling efforts able to simulate TCs directly are presently unable to reliably reproduce realistic landfall statistics [28]. We have shown that it is important to take the demonstrated tension between conditions favorable for genesis and conditions favorable for landfall into account when designing studies that address direct impacts on human populations, and using diagnostic indices to quantify tropical cyclone risk is a useful step toward that end [29].

ACKNOWLEDGMENTS

Partial financial support provided by a Research Assistantship from Florida State University. Support for the Twentieth Century Reanalysis Project is provided by the U.S. Department of Energy, Office of Science Innovative and Novel Computational Impact on Theory and Experiment (DOE INCITE) program, and Office of Biological and Environmental Research (BER), and by the National Oceanic and Atmospheric Administration Climate Program Office (NOAA CPO).

REFERENCES

[1] W. Gray, "Hurricanes: their formation, structure and likely role in the tropical circulation," in *Meteorology over Tropical Oceans*, D. B. Shaw, Ed., Royal Meteorological Society, Berkshire, 1979, pp. 155-218.

[2] S. Camargo, A. Sobel, A. Barnston, and K. Emanuel, "Tropical cyclone genesis potential index in climate models," *Tellus* 59A, 428-443, 2007.

[3] K. Emanuel, "Tropical cyclone activity downscaled from NOAA-CIRES reanalysis," *J. Adv. Model. Earth Syst.* 2, 1908-1958, 2010.

[4] M. McGauley and D. Nolan, "Measuring environmental favorability for tropical cyclogenesis by statistical analysis of threshold parameters,". *J. Clim.* 24, 5968-5997, 2011.

[5] M. Tippett, S. Camargo, and A. Sobel, "A Poisson regression index for tropical cyclone genesis and the role of large-scale vorticity in genesis," *J. Clim.* 24, 2335-2357, 2011.

[6] C. Bruyère, G. Holland, and E. Towler, "Investigating the use of a genesis potential index for tropical cyclones in the North Atlantic basin," *J. Clim.* 25, 8611-8626, 2012.

[7] B. Ryan, I. Watterson, and J. Evans, "Tropical cyclone frequencies inferred from Gray's yearly genesis parameter: validation of GCM tropical climates," *Geophys. Res. Lett.* 19, 1831-1834, 1992.

[8] J. Royer, F. Chauvin, B. Timbal, P. Araspin, and D. Grimal, "A GCM study of the impact of greenhouse gas increase on the frequency of occurrence of tropical cyclones," *Clim. Change* 28, 307-353, 1998.

[9] S. Camargo, K. Emanuel, and A. Sobel, "Use of a genesis potential index to diagnose ENSO effects on tropical cyclone genesis," *J. Clim.* 20, 4819-4834, 2007.

[10] S. Camargo, M. Tippett, A. Sobel, G. Vecchi, and M. Zhao, "Testing the performance of tropical cyclone genesis indices in future climates using the HIRAM model," *J. Clim.* 27, 9171-9196, 2014.

[11] K. Emanuel, "Global warming effects on U.S. hurricane damage," *Wea. Clim. Soc.* 3, 261-268, 2011.

[12] C. Landsea, R. Pielke, A. Mestas-Nuñez, and J. Knaff, "Atlantic basin hurricanes: indices of climatic changes," *Clim. Change* 42, 89-129, 1999.

[13] K. Knapp, M. Kruk, D. Levinson, H. Diamond, and C. Neumann, "The international best track archive for climate stewardship, IBTrACS," *Bull. Amer. Meteor. Soc.* 91, 363-376, 2010.
[14] G. Compo et al., "The twentieth century reanalysis project," *Q. J. R. Meteorol. Soc.* 137, 1-28, 2011.
[15] M. Saunders and A. Lea, "Seasonal prediction of hurricane activity reaching the coast of the United States," *Nature* 434, 1005-1008, 2005.
[16] G Lehmiller, T. Kimberlain, and J. Elsner, "Seasonal prediction models for North Atlantic basin hurricane location," Mon. Weather Rev., 125, 1780 – 1791, 1997.
[17] J, Elsner and T. Jagger, "Prediction models for annual U.S. hurricane counts," J. Clim. 19, 2935-2952, 2006.
[18] S. Tolwinski-Ward, "Uncertainty quantification for a climatology of the frequency and spatial distribution of North Atlantic tropical cyclone landfalls," J. Adv. Model. Earth Syst., 7, 305–319, 2015.
[19] J. Kossin, S. Camargo, and M. Sitkowski, "Climate modulation of North Atlantic hurricane track," *J. Clim.* 23, 3057-3076, 2010.
[20] K. Emanuel, "Increasing destructiveness of tropical cyclones over the past 30 years," *Nature* 436, 686-688, 2005.
[21] G. Vecchi and B. Soden, "Effect of remote sea surface temperature change on tropical cyclone potential intensity," *Nature* 450, 1066-1070, 2007.
[22] J. Elsner and C. Schmertmann, "Assessing forecast skill through cross validation," *Wea. Forecasting* 9, 619–624, 1994.
[23] F. Wilcoxon, "Individual comparisons by ranking method," *Biometrics Bulletin* 1, 80-83, 1945.
[24] S. Garner, I. Held, T. Knutson, and J. Sirutis, "The role of wind shear and thermal stratification in past and projected changes of Atlantic tropical cyclone activity," *J. Clim.* 22, 4723-4734, 2009.
[25] M. Zhao and I. Held, "TC-permitting GCM simulations of hurricane frequency response to sea surface temperature anomalies projected for the late-twenty-first century," *J. Clim.* 25, 2995-3009, 2012.
[26] C. Landsea, R. Pielke, A. Mestas-Nuñez, and J. Knaff, "Atlantic basin hurricanes: indices of climatic changes," *Clim. Change* 42, 89-129, 1999.
[27] J. Royer, F. Chauvin, B. Timbal, P. Araspin, and D. Grimal, "A GCM study of the impact of greenhouse gas increase on the frequency of occurrence of tropical cyclones," *Clim. Change* 28, 307-353, 1998.
[28] T. Knutson et al., "Dynamical downscaling projections of 21st century Atlantic hurricane activity: CMIP3 and CMIP5 Twentieth Century simulations," J. Clim. 26, 6591-6617, 2013.
[29] E. Staehling and R. Truchelut, "Diagnosing United States hurricane landfall risk: An alternative to count-based methodologies," *Geophys. Res. Lett.* 43, doi:10.1002/2016GL070117, 2016.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

OPTIMAL TROPICAL CYCLONE INTENSITY ESTIMATES WITH UNCERTAINTY FROM BEST TRACK DATA

Suz Tolwinski-Ward[1]

***Abstract*—The destructive wind power of a tropical cyclone TC is best characterized by its maximum wind speed, but this quantity is notoriously hard to measure. Central pressure has historically been a much more accurately-measured proxy quantity for the intensity of a TC, especially before the advent of aircraft reconnaissance missions and modern satellite observing technologies. Physically-derived relationships between central pressure and maximum wind speed (or WPRs, wind-pressure relationships) have thus often been used to translate the more certain pressure observations into estimates of wind speed. Typically, however, estimates have assimilated both sets of observations only informally, and have neglected the uncertainty in the measurements themselves, as well as in their relationship. A Bayesian hierarchical modeling strategy is employed to create joint optimal estimates of wind speed and central pressure, as well as to quantify the uncertainty in the estimates. Drawing on both maximum windspeed and pressure observations across a large number of storms, the model is also used to derive estimates of bias and uncertainty in the pressure-wind relationship.Here we present a proof-of-concept using Atlantic Basin tropical cyclones from 2004 through the present. Ultimately, the formulation can be used to create a homogeneous set of TC intensities across basins and time, which can enable better regional comparisons and trend analysis.**

I. MOTIVATION

Maximum tropical cyclone (TC) windspeed is the main quantity of interest for assessing the risks of hurricane winds to life and property. However, it is a difficult variable to measure for a combination of reasons. Traditional instruments that measure local windspeed directly can break and fail in extreme winds, and networks of these instruments are sparse in space so that the probability of having measured the global maximum windspeed over a storm's domain is extremely unlikely. Modern remote sensing instruments, meanwhile, may be able to make observations over the full domain of the storm, but the measurements are indirect and tend to reflect conditions at higher atmospheric levels, rather than at the ground where conditions are most relevant to risk. Changing measurement instrumentation over time likely introduces temporal biases in TC windspeed estimates [1] (hereafter referred to as TS12), while differing instrumentation, observing conventions, and possibly environments also makes it difficult to compare estimates for TCs occurring in different oceanic basins [2].

The central pressure of a tropical cyclone is an alternative proxy measure for windspeed hazard intensity, because the lower the central pressure of a cyclone tends to be, the faster its winds circulate. Estimates of the relationship between windspeed and pressure can be derived from the first-order physical approximation of hurricane winds as resulting from gradient wind balance. The relationship derived by [3] (henceforth referred to as KZ07) and refined by [4] (hereafter CK09) is well-known and used throughout the scientific literature and in the insurance industry to convert between maximum windspeeds and central pressures, with dependencies in the relationship on storm latitude, size, and environmental pressure. The uncertainty and potential bias in this relationship is typically neglected in conversions, however. Several different agencies collect and maintain so-called "best tracks" datasets– databases of historical TC dates, locations and intensities throughout the observational period. A comparison of Atlantic Basin best tracks windspeed and windspeed estimated via the Knaff-Zehr wind-pressure relationship (KZ WPR), both plotted versus best tracks pressure values, demonstrates both a need to account for both bias and uncertainty (Fig 1). Some tracks in these best tracks datasets contain estimates of windspeed, others contain only estimates of central pressure, and some contain estimates of both variables. The present research strives to combine all sources of information and uncertainty objectively to infer optimal maximum windspeed estimates and accompanying

STolwinski-Ward@air-worldwide.com [1]AIR Worldwide Corporation, Boston, MA

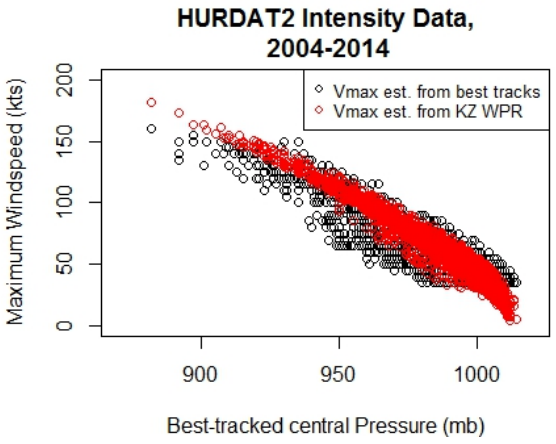


Fig. 1. Maximum windspeed from the best tracks dataset (black) and predicted (red) by the KZ WPR as a function of best-tracked central pressure for historical tropical cyclones in the North Atlantic Basin.

space-and-time-dependent estimate uncertainties for all best-tracks data. The resulting consistent set of historical maximum windspeed estimates should thus be consistent across time and space, and can therefore facilitate studies of the effects of climate variability and change on the destructive hazard to society posed by tropical cyclone activity.

II. METHOD

We develop a Bayesian hierarchical model (see [5] for an introduction) for the present problem with the following structure:

$$\begin{aligned} \textit{(posterior)} \quad & [V, \Delta P | \hat{V}, \hat{\Delta P}, \hat{R}_{34}, \hat{c}, \hat{\phi}, \vec{\theta}] \propto \\ \textit{(data-level)} \quad & [\hat{V}|V, \theta_V][\hat{\Delta P}|\Delta P, \theta_{\Delta P}] \\ & \times [\hat{R}_{34}|R_{34}, \theta_{R34}][\hat{c}|c, \theta_c][\hat{\phi}|\phi, \theta_\phi] \\ \textit{(process level)} \quad & \times [V, \Delta P | R_{34}, c, \phi, \theta_{KZ}] \\ \textit{(parameter level)} \quad & \times [\theta_V][\theta_{\Delta P}][\theta_{R34}][\theta_c][\theta_\phi][\theta_{KZ}] \end{aligned} \tag{1}$$

where the variables used are defined in the following table:

Variable	Units	Summary
V, $\hat{V}$	kts	The underlying "true" and observed/best-tracked values of the maximum windspeed, resp.
ΔP, $\Delta\hat{P}$	hPa	Underlying/best-tracked values of the pressure deficit (diff. in surface pressure from TC periphery to central minimum).
R_{34}, $\hat{R}_{34}$	nmi	Underlying/best-tracked values of the azimuthal average of the radius of gale-force winds.
c, $\hat{c}$	kts	Underlying/best-tracked values of TC forward speed.
ϕ, $\hat{\phi}$	°N	Underlying/best-tracked TC latitude.
θ_{KZ}	-	Parameters of the KZ07 wind-pressure relationship.
θ_X	-	Parameters quantifying measurement uncertainty or bias of process X.

The inference for the proof-of-concept results presented here is performed using tracks from the HURDAT2 database from 2004 - 2014 (2004 is the earliest year of the database in which values for R34, the radius of gale-force winds, is included). In the following subsections, we describe the choices for each model component in the general structure specified above.

A. *Data-level models.*

All data-level models are assumed to be independent of one another, and to be described using mean-zero Gaussian model errors. (Note that this does not imply that the process X itself is Gaussian, only that the best-track data are normally distributed about the underlying true process due to measurement noise and the process of best-tracking.) The modeling choices for each data model case are then reduced to how to specify the error covariances.

a) $[\hat{V}|V, \theta_V]$ *and* $[\hat{\Delta P}|\Delta P, \theta_{\Delta P}]$*:* This model describes how the best-tracked estimate of maximum windspeed $\hat{V}$ and pressure deficit ΔP for a given cyclone and time relate to the "true" unobserved values V and ΔV. For present purposes, we assume simple Gaussian models to account for measurement error:

$$\begin{aligned} \hat{V} &\sim N(V, \sigma_V^2) \\ \Delta\hat{P} &\sim N(\Delta P, \sigma_{\Delta P}^2) \end{aligned} \tag{2}$$

b) $[\hat{R}_{34}|R_{34}, \theta_{R34}]$, $[\hat{c}|c, \sigma_x^2]$,$[\hat{\phi}_t|\phi, \sigma_x^2]$*:* In principal, the uncertainty in the radius of gale-force winds R_{34}, the forward speed c, and the latitude ϕ along the

tracks can be explicitly modeled at the data-level as well. The model for $\hat{R}_{34}$ should treat storm asymmetry, and positional uncertainties σ_x^2 (documented in TS12 for modern observing platforms) can be used to derive the models for $\hat{c}$ and $\hat{\phi}$. For the current proof-of-concept results, however, we neglect the uncertainty in the HURDAT values of these variables, using the observations as perfect reflections of the truth. Their uncertainty will instead get accounted for implicitly in estimates of the uncertainty of the KZ WPR relationship.

B. Process-level model.

c) $[V, \Delta P | R_{34}, c, \phi, \theta_{KZ}]$*:* The strategy for describing prior understanding of the maximum windspeed and pressure deficit processes themselves is conservative: it is only assumed that V and ΔP are related to one another by the WPR within the confines of an error structure to be specified. No assumptions are made about the temporal evolution or other features of the processes themselves are made, and in that sense the prior may be viewed as rather noninformative, and providing only a vague constraint. Thus the process-level structure may be re-written as

$$\begin{aligned}[V, \Delta P | R_{34}, c, \phi, \theta_{KZ}] &= [V | \Delta P, R_{34}, c, \phi, \theta_{KZ}] \\ &\quad \times [\Delta P, R_{34}, c, \phi] \\ &= [V | \Delta P, R_{34}, c, \phi, \theta_{KZ}]\end{aligned}$$

where the last equality follows from the first given the assumption of a flat joint prior on variables ΔP, R_{34}, c, and ϕ.

We assume that the true values of time series V and ΔP along a given track are related through the KZ07 relationship, plus a WPR discrepancy δ_{WPR}, which accounts for systematic errors in the WPR, plus an error term ϵ:

$$V \sim f_{KZ}(\Delta P, V, c, \phi, R_{34}) + \delta_{WPR} + \epsilon \qquad (4)$$

The function f_KZ is taken from KZ07 and CK09, and has the form

$$\begin{aligned}f_{KZ} &= 1.5c^{0.63} + a_0 - a_1 S(V, R_{34}, \phi) - a_2\phi - \\ &\quad a_3 \Delta P + a_4 \sqrt{|\Delta P|} \qquad (5)\end{aligned}$$

The KZ model discrepancy δ_k is assumed to be linear in pressure deficit:

$$\delta_k = \alpha_k + \beta_k \Delta P \qquad (6)$$

C. Parameter-level models.

1) Data-level model parameters: The parameters θ_V and $\theta_{\Delta P}$ of the simple proof-of-concept data-level models for the windspeed and pressure deficit data are just the variances of the measurement noise, σ_v^2 and $\sigma_{\Delta P}^2$. In principal, informative priors can be put on these parameters, and their posterior values derived in light of the information contained in the data and the KZ wind-pressure relationship. For present purposes, where we use data derived from modern observing platforms for the 2004-2014 HURDAT2 data, we fix these uncertainties at the constant, but intensity-dependent values given in TS12. There are no parameters θ_{R34}, θ_c, and θ_ϕ in the present version of the model, given that we currently neglect explicit uncertainty modeling for each of the associated variables.

2) Process-level model parameters: The linear coefficients of the KZ WPR discrepancy function, α_0 and α_1, are both given flat priors on the set of all reals. The error variance for the KZ WPR is given a flat prior on the positive reals (and so its full conditional has a conjugate inverse gamma form, which is easy to sample.

III. PRELIMINARY RESULTS

Results derived from even the simplified model assumptions here shed light on the uncertainty and bias of the KZ WPR. At least for Atlantic Basin data, the relationship estimating tropical cyclone wind from pressure is generally negatively biased, with a greater magnitude of bias for more intense storms. The linear discrepancy function δ corrects for this bias. The median and 95% credible intervals for the posterior distribution of the discrepancy as a function of pressure deficit is shown in the upper panel of Figure 2. The uncertainty in the bias-corrected KZ WPR has a standard deviation σ_{KZ} with a symmetric posterior distribution, with a median value around 22 knots (lower panel of 2).

Estimates of maximum windspeed, the most relevant quantity for hazard risk analysis, can be improved by combining the information in both best-tracked measures of intensity, windspeed and pressure, along with the physically-based information in the KZ WPR. According to the preliminary analysis performed here, combining the three information sources also reduces the uncertainty in the maximum windspeed estimates in most cases, particularly for the most intense tropical cyclones. One such example is shown in Fig 3, where the bias correction is especially evident at the highest intensities along the track of hurricane Wilma from 2005, and the credible intervals containing the same probability mass around the central estimates

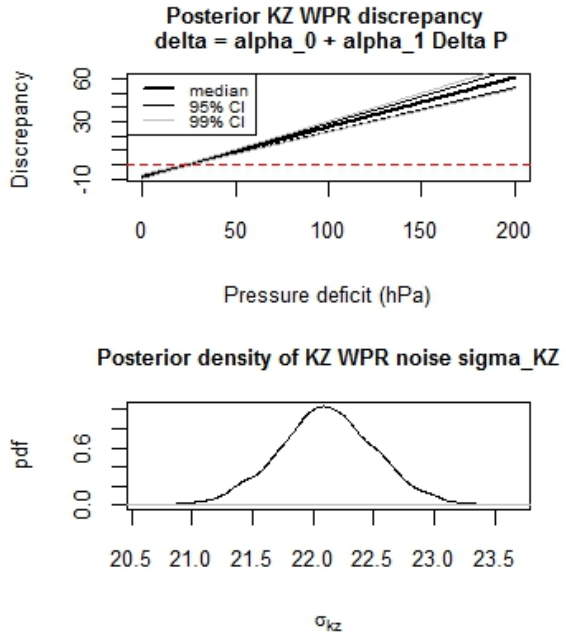


Fig. 2. Top panel: inferred distributions of the discrepancy function for the KZ WPR for the Atlantic Basin. Bottom panel: inferred distribution of the error variance for the KZ WPR.

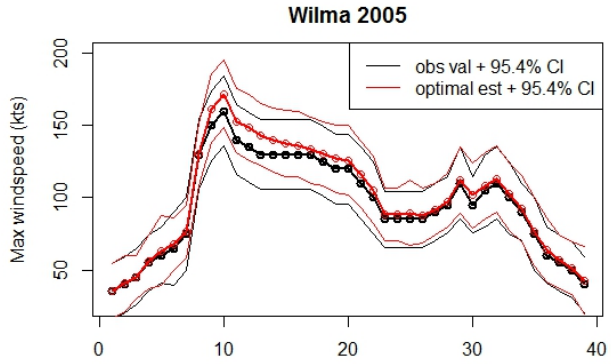


Fig. 3. Best-tracked maximum windspeed estimates for Wilma, 2005, and estimate derived by combining windspeed, pressure, and KZ WPR relationship information.

are also narrowed by the inclusion of two additional sources of information (the pressure and physics-based information in the KZ WPR) in addition to the Vmax observational data alone.

IV. Future Directions

In future development of this work, considerable effort will be spent gathering estimates of time-dependent observational uncertainties from the literature to reflect the decreasing accuracy of windspeed and pressure observations with time earlier in the historical datasets. This development will enable the same type of uncertainty quantification analysis presented here to be applied throughout the observational record for more rigorous quantification of uncertainties during analysis of long-term trends in the data. The author also hopes to apply the current methodology to other basins, where different measurement techniques by agencies outside the Atlantic may result in differences in the KZ WPR bias and uncertainty results. The effect of uncertainty in the radius of gale-force winds should also be explicitly modeled and quantified, so that the methodology may be applied additionally when these size data are missing (which is true prior to 2004 in the Atlantic basin). Finally, the realism of the data-level models $[\hat{V}|V, \theta_V]$ and $[\Delta\hat{P}|\Delta P, \theta_{\Delta P}]$ can be enhanced by more formally modeling the best-tracking process. In particular, scientists who perform the best-tracked estimates from raw observations likely assume some smoothness in time in the time series of V and ΔP; this smoothing can be modeled and its degree inferred within the context of the Bayesian hierarchical model presented here.

References

[1] R. Torn and C. Snyder, "Uncertainty of tropical cyclone best-track information," *Weather and Forecasting*, vol. 27, pp. 715–729, 2012.

[2] K. Knapp and M. Kruk, "Quantifying interagency differences in tropical cyclone best-track wind speed estimates," *Monthly Weather Review*, vol. 138, pp. 1459–1473, 2010.

[3] J. Knaff and R. Zehr, "Reexamination of tropical cyclone wind-pressure relationships," *Weather and Forecasting*, pp. 71–88, 2007.

[4] J. Courtney and J. Knaff, "Adapting the knaff and zehr wind-pressure relationship for operational use in tropical cyclone warning centres," *Australian Meteorological and Oceanographic Journal*, vol. 58, pp. 167–179, 2009.

[5] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*. Chapman and Hall/CRC, 2 ed., 2003.

6th International Workshop on Climate Informatics
September 22-23, 2016
CI 2016 Hosted by the National Center for Atmospheric Research in Boulder, CO

EXTREME WEATHER PATTERN DETECTION USING DEEP CONVOLUTIONAL NEURAL NETWORK

Yunjie Liu[1], Evan Racah[1], Prabhat [1], Joaquin Correa [1], Amir Khosrowshahi[2], David Lavers[3], Kenneth Kunkel[4], Michael Wehner[1], William Collins[1]

***Abstract*—Characterizing frequency and intensity of extreme weather events in future climate projections remains one of the grant challenges in climate change research. Various algorithms have been developed for detecting extreme weather events in climate simulations; however, these schemes are largely build upon human expertise in specifying multi-variate thresholds of relevant climate properties in defining event. In other words, these schemes require practitioner hand craft features that best characterize event. In doing so, competing algorithms often produce very different results even on same data set. Alternative schemes for extreme weather event detection that avoid subjective manual feature engineering and thresholds specifying, eventually produce consistent and trustworthy results are in urgent demand. In this work, we employ deep convolutional neural network for classifying weather events, with the ultimate goal to develop automatic systems capable of detecting and tracking multiple extreme events in climate simulations simultaneously. We demonstrate success on classifying tropical cyclones, atmospheric rivers and weather fronts. The promising results point to the applicability of convolutional neural network for learning a broad class of extreme weather events, potentially transforming the quantitative evaluation of climate change in the future.**

I. Motivation

An array of methods have been developed for detecting and tracking extreme weather events in large climate simulations. Most widely used algorithms are built on human expertise on providing multi-variate thresholds for relevant climate variables to define event. For instance, tropical cyclones are strong rotating weather systems that are characterized by low pressure and warm temperature core with high circulating wind. However, there is no universally agreement of a set of criteria that best characterize tropical cyclone ([1]). The "Low" pressure and "Warm" temperature are interpreted differently among researchers based on their understanding of event characteristics, therefore, different thresholds are proposed. Very different, sometimes contradictory, results are reported by employing these algorithm on same data set and difficulties of assessing results between approaches remains high ([2], [3]).

Recent advances in deep learning have demonstrated exciting and promising success on pattern recognition in natural image ([4], [5], [6]) and speech ([7], [8]). These success have demonstrated the potential transformative power of deep learning to replace hand crafting feature engineering with direct feature learning from training data. Most of the state-of-art deep learning architectures for visual pattern recognition are build on the hierarchical feature learning convolutional neural network. Modern convolutional neural network tend to be deep and large with many hidden layers and millions of hidden units, making them very flexible in learning a broad class of patterns simultaneously from training data. In this paper, we formulate the problem of detecting extreme weather events as classic supervised visual pattern recognition problem that consists two components: classification and localization. We develop deep convolutional architecture and mainly address the classification task in this write up.

II. Method

A classic deep convolutional neural network comprised of several convolutional layers followed by a small amount of fully connected layers. Inbetween two successive convolutional layers, sub-sampling operation (e.g. max pooling) is performed typically. The inputs of a convolutional neural network are presumably *(m,n,p)* images, where *m* and *n* are the width and height of an image in pixel, *p* is the number of color channel of image. The output of a convolutional neural network is a vector of *q* probability units, corresponding to the number of categories to be classified (e.g. for binary classifier *q*=2) in the classification task, and is a vector of coordinates that describe the location of object in input image in the localization task. The most

Corresponding author: Yunjie Liu, yunjieliu@lbl.gov, [1]Lawrence Berkeley National Lab, [2]Nervana Systems, [3]Scripps Institution of Oceanography, [4]National Oceanic and Atmospheric Administration

important layer in a convolutional neural network is the convolutional layer, which performs convolution operation between kernels and the input image (or feature maps). A convolutional layer contains k kernels with the size (i,j,p), where i,j are the width and height of the kernel, typically much smaller than the height and width of input image. p always equal to the number of color channel of input image (e.g. a color image has red, green, and blue channels, thus p=3). Each of the kernels independently convolves cross the input images (or feature maps) and dot product is computed between the entry of kernel and the local region it covers in the input image (or feature maps). Dot product at each location is further passed to the non-linear activation function of hidden units, generating k feature maps, each corresponding to a particular kernel. The entries of kernels (a.k.a the parameters of convolutional layer) have to be learned in the process of training. Convolve kernels across input image will produce larger outputs for certain sub-regions than others, which allows features to be extracted from input and preserved in the feature maps regardless of where the feature locates in the input image.

The pooling layer sub-samples the feature maps from convolutional layer over a (s,t) contiguous region, where s,t are the width and height of the sub-sampling window. For example, in max pooling operation, only the maximum value within the sub-sample window is retained. This operation reduces the resolution of feature maps with the depth of layers. Pooling layer does not have any parameter to learn.

Fully connected layer, unlike the convolutional layer, has full connections to all hidden units in previous layer. The last fully connected layer also does the high level reasoning based on the feature vectors from previous layer and produce final class scores for objects in input image or predict coordinates of objects in input image.

Training deep architecture is known to be difficult ([9], [10]). It requires carefully tuning model parameters and learning parameters. The parameter tuning process, however, can be tedious and non-intuitive. In this study, we employ a Bayes' frame work of hyper-parameter optimization technique to facilitate parameter selecting.

Referring to AlexNet ([4]), we build classification system with two convolutional layers followed by two fully connected layers. Inbetween two consecutive convolutional layer, max sub-sampling is performed. Details of the architecture and layer parameters can be found in Table I.

III. EVALUATION

A. Data

We collected ground truth labeling of tropical cyclone, atmospheric river and weather front obtained via multivariate threshold based criteria implemented in TECA ([11]), and manual labeling by experts ([12], [13]). Training data comprise of image patches, in which relevant variables are extracted from global climate simulation or reanalysis over a prescribed box that bounds an event, then stacked together. The dimension of the box is based on domain knowledge of events spatial extent in real word. For instance, tropical cyclone radius are typically with in range of 100 kilometers to 500 kilometers, thus a box of size 500 kilometers by 500 kilometers is likely to capture most of tropical cyclones. The chosen climate variables are also based on domain expertise. To facilitate model training, the prescribed box is placed over an event and location is slightly adjusted such that event is approximately at the center. Image patches are cropped and centered correspondingly. Because the spatial extent of climate events varies and the spatial resolution of simulation and reanalysis data is non-uniform, final training images prepared differ in their size among the three types of events. This is one of the most important limitations that prevent us developing one signal convolutional neural network to classify all three types of event simultaneously. The class labels of images are "containing events" and "not containing events". In other words, we formulate the problem as binary classification task. A summary of the attributes of training images is listed in Table III, attributes of original reanalysis and model simulation data are documented in Table II. Training data for localization task is prepared in a similar way, but differ in the prescribed box size and events location within the box. Specifically, the weather events are randomly located within a much larger box. The ground truth event location in the box is computed and saved. These prepared data set are split into "training" and "testing" subsets.

B. Classification

To illustrate performance of deep convolutional neural network comparing to other methods, we also trained four widely used algorithms on the same classification task. Table IV summarizes the performance of all five algorithms on classifying tropical cyclone, atmospheric river and weather front. Deep convolutional neural networks perform best regarding classification accuracy. In addition, the systems are well trained and

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

TABLE I: Classification CNN architecture and layer parameters. The convolutional layer parameters are denoted as <kernel size>-<number of feature maps> (e.g. 5x5-8). The pooling layer parameters are denoted as <pooling window> (e.g. 2x2). The fully connected layer parameter are denoted as <number of units> (e.g. 2). Non-linear activation function of hidden unit is shown in parentheses.

	Conv1 (ReLu)	Pool1	Conv2 (ReLu)	Pool2	Fully (ReLu)	Fully (Sigmoid)
Tropical Cyclone	5x5-8	2x2	5x5-16	2x2	50	2
Weather Fronts	5x5-16	2x2	5x5-16	2x2	400	2
Atmospheric River	12x12-8	3x3	12x12-16	2x2	200	2

TABLE II: Original Data Sources

Climate Dataset	Time Frame	Temporal Resolution	Spatial Resolution (lat x lon degree)
CAM5.1 historical run	1979-2005	3 hourly	0.23x0.31
ERA-Interim reanalysis	1979-2011	3 hourly	0.25x0.25
20 century reanalysis	1908-1948	Daily	1x1
NCEP-NCAR reanalysis	1949-2009	Daily	1x1

TABLE III: Dimension of image, diagnostic variables (channels) and labeled data set size for classification task (PSL: sea surface pressure, U: zonal wind, V: meridional wind, T: temperature, TMQ: vertical integrated water vapor, Pr: precipitation)

Events	Image Dimension	Variables	Total Examples
Tropical Cyclone	32x32	PSL,V-BOT,U-BOT, T-200,T-500,TMQ, V-850,U-850	10,000 +ve 10,000 -ve
Atmospheric River	148 x 224	TMQ, Land Sea Mask	6,500 +ve 6,800 -ve
Weather Front	27 x 60	T-2m, Pr, PSL	5,600 +ve 6,500 -ve

do not suffer from over-fitting [1]. We believe this is mostly because of the shallow and small size of the architecture (4 learnable layers) and the weight decay regularization. Deeper and larger architecture would be inappropriate for this study due to limited amount of training data. Fairly good train and test classification results also suggest that the deep convolutional neural network are able to efficiently learn representations of climate pattern from labeled data and make predictions based on learned representations. Traditional threshold based detection method requires human experts to carefully examine extreme event before determining a set of thresholds for relevant variables that best characterize events. In contrast, as shown in this study, deep convolutional neural network are able to learn climate pattern from labeled data directly, thus potentially overcome subjective judgment in traditional detecting scheme.

However, the superior performance of deep convolutional neural network comes at the cost of expensive to train comparing to other competing methods. As many researcher has pointed out that deep neural networks are hard to train ([9], [10]). Large architecture sometimes need to be trained for weeks. Despite the high computation expense in training, deep convolutional neural network is still promising as a generic algorithm for event detecting by integrating classification and localization, since one time evaluation is needed at testing after the architecture is trained. Other methods would require sliding window cross the entire image to perform detecting. This is very computation expensive, because classifier needs to perform classification at every window it encountered.

Further analyzing the classification results, we observed that random forest performs best among four methods we chose, probably because random forest ensembles a sets of classifier to reach a conclusion. Tropical cyclone classification is a relatively easy task. Simple logistic regression can achieve over 95% of accuracy. It is likely due to the fact that the characteristics of tropical cyclone are easy and well defined. Further investigating the results, we see logistic regression and support vector machines achieve comparable testing accuracy for weather fronts but the training accuracy is below that of deep convolutional neural network. The difference between training and testing accuracy are fairly large (5%), indicating the models are not well trained given training data or the model is not suitable for this task.

[1] over fitting is a phenomenon that an algorithm, due to its complexity, adapts to the noise in training data rather than learning fundamental relations. Such that it is not able to generalize well to unseen test data. Often, an over-fitted algorithm performs comparatively well on training data, but not as well on unseen testing data. The training and testing performance of a well trained learning algorithm should closely follow each other.

TABLE IV: Classification cccuracy of five classification methods

ConvNet			Logistic		KNN		SVM		RandForest	
Event Type	Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)	Train (%)	Test (%)
Tropical Cyclone	99.3	99.1	96.8	95.9	98.1	97.9	97.0	95.9	99.2	99.4
Atmospheric River	90.5	90.0	81.9	82.7	79.7	81.7	81.6	83.0	87.9	88.4
Weather Front	88.7	89.4	84.7	89.8	72.5	76.5	84.4	90.2	81.0	87.5

IV. CONCLUSION

We present deep convolutional neural network as a novel method for classifying extreme weather events. The classification system achieves fairly high classification accuracy (89%-99%), outperforming K-nearest neighbor, logistic regression, support vector machine and random forest algorithms. Deep convolutional neural network is powerful because it can learn high-level representation of pattern from data directly replacing hand feature engineering that is often subjective. Integrating object classification and localization will be a novel approach for extreme event detecting that does not rely on cherry picking features and thresholds. In this manual script we emphasize on the classification. However, the early result of localization is also promising. To the best of our knowledge, this is the first attempt to develop new detecting algorithm that does not require hand feature engineering, thus can overcome drawbacks of traditional detecting scheme. The successful application could be a precursor for tackling a broader class of pattern detection problem in climate science. These detecting results are also critical information for characterizing extreme events behavior from past to future.

ACKNOWLEDGMENTS

This research used resources of the National Energy Research Scientific Computing Center. This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, Applied Mathematics program of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Partial support was provided by the National Oceanic and Atmospheric Administration (NOAA) Climate Program Office Applied Research Center, by NOAA through the Cooperative Institute for Climate and Satellites North Carolina under Cooperative Agreement NA14NES432003, and by Strategic Environmental Research and Development Program Contract W912HQ-15-C-0010.

REFERENCES

[1] D. S. Nolan and M. G. McGauley, "Tropical cyclogenesis in wind shear: Climatological relationships and physical processes," in *Cyclones: Formation, Triggers, and Control*, pp. 1–36, 2012.

[2] U. Neu, M. G. Akperov, N. Bellenbaum, R. Benestad, R. Blender, R. Caballero, A. Cocozza, H. F. Dacre, Y. Feng, K. Fraedrich, *et al.*, "Imilast: a community effort to intercompare extratropical cyclone detection and tracking algorithms," *Bulletin of the American Meteorological Society*, vol. 94, no. 4, pp. 529–547, 2013.

[3] M. Horn, K. Walsh, M. Zhao, S. J. Camargo, E. Scoccimarro, H. Murakami, H. Wang, A. Ballinger, A. Kumar, D. A. Shaevitz, *et al.*, "Tracking scheme dependence of simulated tropical cyclone response to idealized climate simulations," *Journal of Climate*, vol. 27, no. 24, pp. 9197–9213, 2014.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.

[5] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Internaltional Conference on Learning Representation (ICLR)*, 2015.

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[7] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 6645–6649, IEEE, 2013.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, pp. 3104–3112, 2014.

[9] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *The Journal of Machine Learning Research*, vol. 10, pp. 1–40, 2009.

[10] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[11] Prabhat, S. Byna, V. Vishwanath, E. Dart, M. Wehner, W. D. Collins, *et al.*, "Teca: Petascale pattern recognition for climate science," in *Computer Analysis of Images and Patterns*, pp. 426–436, Springer, 2015.

[12] K. E. Kunkel, D. R. Easterling, D. A. Kristovich, B. Gleason, L. Stoecker, and R. Smith, "Meteorological causes of the secular variations in observed extreme precipitation events for the conterminous united states," *Journal of Hydrometeorology*, vol. 13, no. 3, pp. 1131–1141, 2012.

[13] D. A. Lavers, G. Villarini, R. P. Allan, E. F. Wood, and A. J. Wade, "The detection of atmospheric rivers in atmospheric reanalyses and their links to british winter floods and the large-scale climatic circulation," *Journal of Geophysical Research: Atmospheres*, vol. 117, no. D20, 2012.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

INFORMATION TRANSFER ACROSS TEMPORAL SCALES IN ATMOSPHERIC DYNAMICS

Nikola Jajcay[1,2] and Milan Paluš[1]

***Abstract*—Earth climate, in general, varies on many temporal and spatial scales. In particular, air temperature exhibits recurring patterns and quasi-oscillatory phenomena with different periods. Although these oscillations are usually weak in amplitude, they might have non-negligible influence on temperature variability on shorter timescales due to cross-scale interactions, recently observed by Paluš [1]. In this letter, we show how to discern possible cross-scale interactions and, if present, how to quantify their effect on air temperature in Europe.**

I. MOTIVATION

Better understanding of the complex dynamics of the Earth atmosphere and climate is one of the challenges for contemporary science and society. Large amount of experimental data requires new mathematical and computational approaches. Considering the climate system as a complex network of interacting subsystems [2] is a new paradigm bringing new data analysis methods helping to detect, describe and predict atmospheric phenomena [3]. A crucial step in constructing climate networks is inference of network links between climate subsystems [4]. Directed links determine which subsystems influence other subsystems, i.e. uncover the drivers of atmospheric phenomena. Inference of causal relationships from climate data is an intensively developing research field [5], [6], [7], [8], [9], [10], [11]. Typically, a causal relation is sought between different variables or modes of atmospheric variability. Paluš [1] has open another view at the complexity of atmospheric dynamics by uncovering causal relations or information flow between dynamics on different time scales in the same variable. Here we further develop research in this direction.

Air temperature dynamics vary on a large range of spatial and temporal scales. In this letter, we will focus on relatively small temporal range from the annual scale to near-decadal time scales. Long instrumental temperature records available from various European stations allow us to study long term temperature variability and to identify recurring patterns, e.g. climate oscillations on interannual scale. Plaut [12] found climate oscillation with the period about 7-8 years in central England temperature record, Paluš and Novotná identified oscillatory phenomenon with similar, 7-8 years period in various central Europe surface air temperature (SAT, hereafter) records (Prague, Berlin, De Bilt, ...) [13]. Grieser et al. reported 7-8 year oscillations in SAT records from western and northern Europe [14], while Pišoft found spatial patterns of occurrence of the 8 year cycle at various geopotential heights in the NCEP reanalyzed temperature series [15]. These cycles have been usually observed and identified using subtle detection techniques as singular spectrum analysis [16], Monte Carlo SSA [17] and others, since their amplitude is usually low and they are hidden in overall temperature variability.

Once the oscillations (or quasi-oscillations) are detected, we can employ causality measures to detect possible cross-scale information transfer. Following Paluš [1], we expect that the phase of slow oscillatory phenomena interacts with and influences the amplitude of faster frequencies, thus we try to estimate the causality measures between the time series of phase of 7-8 year cycle and the amplitude of faster cycles.

II. METHOD

The information transfer between various scales (in particular, from above mentioned 7-8 year cycle to faster time scales) could be studied in the framework of phase dynamic approach [18], where we compute the instantaneous phase and amplitude of a quasi-oscillatory phenomenon using either Hilbert transform [18] to obtain imaginary part of time series, or complex continuous wavelet transform [19], which provides both the bandpass filtering of the signal (for

Corresponding author: N. Jajcay, jajcay@cs.cas.cz [1]Dept. of Nonlinear Dynamics and Complex Systems, Institute of Computer Science, Academy of Sciences of the Czech Republic, Prague, Czech Republic[2]Dept. of Atmospheric Physics, Charles University in Prague, Czech Republic

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

selected central period) and the estimation of the instantaneous phase and amplitude. Mathematically,

$$\psi(t) = s(t) + i\hat{s}(t) = A(t)e^{i\phi(t)} \quad (1)$$

$$\phi(t) = \arctan\frac{\hat{s}(t)}{s(t)} \quad (2)$$

$$A(t) = \sqrt{s^2(t) + \hat{s}^2(t)} \quad (3)$$

where $s(t)$ is the original (SAT) time series, $\hat{s}(t)$ is its imaginary counterpart, $\phi(t)$ is its instantaneous phase and $A(t)$ is its instantaneous amplitude.

Among a number of notions and estimators of causality (for a review see e.g. [20]) we utilize the conditional mutual information [21] which can be used to infer the causal influence (in the Granger sense) between two two time series. Namely, we estimate the functional

$$I(\phi_1(t);\, A_2(t+\tau) \,|\, A_2(t), A_2(t-\eta), \ldots, A_2(t-m\eta)), \quad (4)$$

where τ is forward-time lag and η is backward-time lag in the $(m+1)$-dimensional condition. Using this approach, Paluš [1] found the above mentioned slow 7-8 year phase to influence the amplitude on shorter time scales. Furthermore, we tried to estimate the effect using the conditional means technique on the amplitude of the annual cycle of SAT and also overall variability of SAT anomalies (SATA hereafter, from the SAT time series the yearly climatology is subtracted).

III. Amplitude of the Annual Cycle

The strongest mode of variability in the European temperatures is without a doubt the annual cycle. Still, its amplitude varies in time and space (e.g. [22]). The period of various oscillatory phenomena found in Prague SAT data fluctuates in a wide range, however the most frequent period is close to 8 years (see [1] Fig. 3a and references therein). The cycles, obtained from time series of SAT recorded in Prague-Klementinum [23], we are interested in are plotted in Fig. 1. The apparent relationship between the amplitude of the annual cycle (AAC hereafter, yellow in Fig. 1, corresponds well with the "climatological amplitude", defined as the difference between the means of daily temperature above the upper and below the lower quartile in each year, yellow circles in Fig. 1) and the 8-year cycle (red in Fig. 1, Pearson correlation coefficient -0.86) can be now studied using the conditional means technique.

The conditional means utilize simple binning approach, where the phase interval $(-\pi, \pi)$, representing the full cycle, is divided equidistantly into 8 bins. Note, that this means, that one bin equal approximately one year of the cycle. For each bin we then evaluate

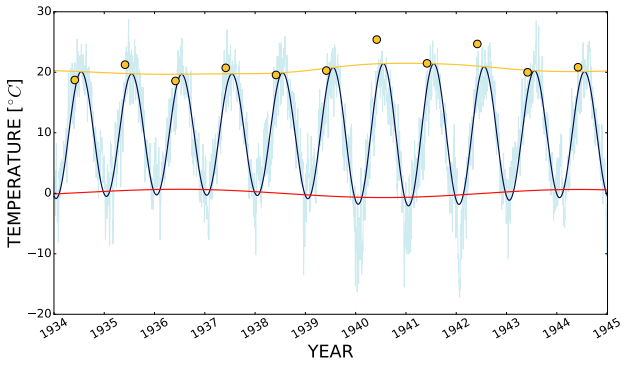


Fig. 1. Cycles in the SAT data in the period 1. January 1934 to 31. December 1944. Shown are SAT daily station data from Prague-Klementinum (light blue), the wavelet reconstruction $A_{8y}(t) \cdot \cos\phi_{8y}(t)$ of the 8 year cycle (red), the wavelet reconstruction of the annual cycle $A_{1y}(t) \cdot \cos\phi_{1y}(t)$ (dark blue), the wavelet amplitude $A_{1y}(t)$ (yellow) and the climatological amplitude (yellow dots) as the difference between warmest and coldest 25% of the year.

the mean (or other statistical measure) to obtain the discretized estimate of the conditional mean of the studied variable. If the 8 year cycle has no influence on the studied variable, the conditional means in all bins would be the same (within statistical fluctuations), equal to the unconditional, global mean.

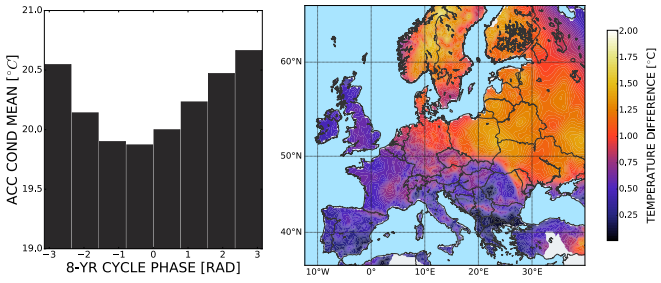


Fig. 2. (left) Conditional means for the amplitude of the annual cycle, $A_{1y}(t)$, for the Prague-Klementinum SAT data with the period 1 January 1950 to 31 December 2013, conditioned on the phase of the 8 year cycle, $\phi_{8y}(t)$. (right) Spatial variability of the effect of the 8 year cycle on the amplitude of SAT annual cycle in Europe. Differences of the maximum and minimum conditional means of the ECA&D reanalysis SAT annual cycle amplitude, $A_{1y}(\mathbf{x}, t)$, conditioned on the phase of the 8 year cycle, $\phi_{8y}(\mathbf{x}, t)$

The histogram of conditional means of AAC is presented in Fig. 2 (left). The maximum mean, conditioned on the phase of the 8 year cycle, is located in the eight bin (equivalent to last year of the 8 year cycle) with value of 20.66°C, while the minimum, in the fourth bin, is at 19.87°C. This implies that through the 8 year cycle the AAC changes, on average, within the range of 0.79°C. This is the average change for the eight cycles in the period 1 January 1950 to 31 December

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

2013. When using different segments of data, the results differ, suggesting that this effect is nonstationary. This is probably due to the nonstationarity of the temperature data and cross-scale interactions themselves. The evolution of conditional means in different temporal windows could be found in Jajcay et al. [24], along with statistical testing against synthetic time series generated under the null hypotheses of linear autoregressive process (autoregressive surrogates of order 1 [17]) and linear process with same spectrum (Fourier transform surrogates [25]).

Since the effect is nonstationary, we presume that the effect is also variable in space. Thus, the gridded temperature reanalysis dataset (E-OBS [26], 1 January 1950 to 31 December 2013) underwent the same AAC conditional means analysis using the phase of CCWT-extracted 8 year cycle. The marked influence of the phase of the 8 year oscillatory mode on the amplitude of the annual cycle can be seen over central, northern and eastern Europe.

IV. Overall SATA variability

Since the 8 year cycle is supposed to have an effect on various temporal scales, not just the annual cycle [1], now we explore its effect on the overall variability represented by the surface air temperature anomalies (SATA). As before, we utilize the conditional means technique, but this time, we are computing the mean of SATA directly, conditioned on the phase of the 8 year cycle. The results, presented in Fig. 3 (left) show "cold" bins in the beginning and the end of the cycle (with minimum of $0.16°$C in the eight bin) and the "warm" bins in the middle of the cycle (with maximum of $1.3°$C located in the fourth bin). This gives us overall effect on the temperature anomalies of $1.2°$C in magnitude.

This effect, similarly as with the AAC effect, differs in time and space. Our analysis showed, that it also differs with seasons, with strongest effect visible in the winter months (December - February, DJF), when the differences between the maximum and the minimum mean winter temperature, conditioned on the phase of the 8 year cycle, could reach approximately 4-5°C in the station SATA data from central Europe. During the summer season, the effect is not statistically significant, i.e. it is not distinguishable from random temperature variability (caused by the intrinsic chaotic nature of the temperature dynamics, see Supporting information in [24]).

The spatial variability of the effect on the winter temperatures in Europe is presented in Fig. 3 (right). The differences range from about $1°$C in Spain to the maximum of $6.5°$C in Finland and adjacent areas of Russia. The pattern is similar to that in Fig. 2 (right) and in central and eastern Europe they both resemble (the inverse of) mountain topography: the effect of the 8 year cycle is strong in the lowlands from the North and Baltic Seas southward and weakens at the mountain ranges of the Alps and Carpathians. The interaction of variable jet stream with the mountain topography apparently has a modulating effect on the influence of the 8 year cycle on the temperature variability in Europe.

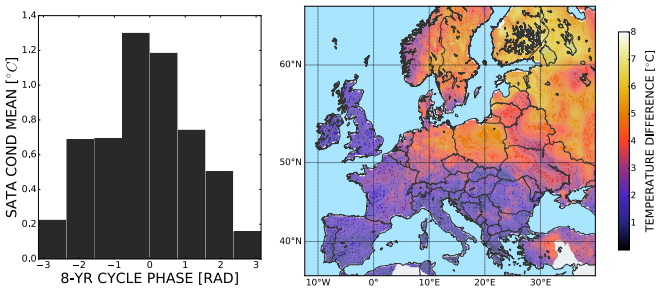


Fig. 3. (left) Conditional means for the SAT anomalies for the Prague-Klementinum SAT data with the period 1 January 1950 to 31 December 2013, conditioned on the phase of the 8 year cycle, $\phi_{8y}(t)$. (right) Spatial variability of the effect of the 8 year cycle on the amplitude of SATA winter season DJF in Europe. Differences of the maximum and minimum conditional means of the ECA&D reanalysis SATA DJF time series, conditioned on the phase of the 8 year cycle, $\phi_{8y}(\mathbf{x}, t)$

V. Conclusions

Considering air temperature variability in a range of time scales, Paluš [1] presented a statistical evidence for a cross-scale-directed information flow from slower to faster scales. We applied a simple, conditional means approach to quantify the effect of this information transfer, in particular, the effect of the phase of the 8 year cycle on the amplitude of the annual cycle and also overall SATA variability. We showed that this effect is strongest in the winter, DJF, months in which the change in temperature could be up to 4-5°C. These results suggest that the weak in amplitude 7-8 year cycle plays an important role in the temperature variability on interannual and shorter time scales. Therefore, this phenomenon deserves a further study to understand its mechanisms. Detailed discussion could be found in [24].

Acknowledgments

The data used are listed in the references. This study was supported by the Ministry of Education, Youth

and Sports of the Czech Republic within the Program KONTAKT II, project LH14001.

REFERENCES

[1] M. Paluš, "Multiscale atmospheric dynamics: Cross-frequency phase-amplitude coupling in the air temperature," *Physical Review Letters*, vol. 112, no. 7, pp. 1–5, 2014.

[2] A. A. Tsonis and P. J. Roebber, "The architecture of the climate network," *Physica A: Statistical Mechanics and its Applications*, vol. 333, pp. 497–504, 2004.

[3] S. Havlin, D. Y. Kenett, E. Ben-Jacob, A. Bunde, R. Cohen, H. Hermann, J. Kantelhardt, J. Kertész, S. Kirkpatrick, J. Kurths, *et al.*, "Challenges in network science: Applications to infrastructures, climate, social systems and economics," *The European Physical Journal Special Topics*, vol. 214, no. 1, pp. 273–293, 2012.

[4] M. Paluš, D. Hartman, J. Hlinka, and M. Vejmelka, "Discerning connectivity from dynamics in climate networks," *Nonlinear Processes in Geophysics*, vol. 18, no. 5, pp. 751–763, 2011.

[5] I. Ebert-Uphoff and Y. Deng, "Causal discovery for climate research using graphical models," *Journal of Climate*, vol. 25, no. 17, pp. 5648–5665, 2012.

[6] Y. Deng and I. Ebert-Uphoff, "Weakening of atmospheric information flow in a warming climate in the community climate system model," *Geophysical Research Letters*, vol. 41, no. 1, pp. 193–200, 2014.

[7] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths, "Identifying causal gateways and mediators in complex spatio-temporal systems," *Nature communications*, vol. 6, 2015.

[8] E. H. van Nes, M. Scheffer, V. Brovkin, T. M. Lenton, H. Ye, E. Deyle, and G. Sugihara, "Causal feedbacks in climate change," *Nature Climate Change*, vol. 5, no. 5, pp. 445–448, 2015.

[9] D. Smirnov and I. Mokhov, "Estimation of interaction between climatic processes: Effect of sparse sample of analyzed data series," *Izvestiya, Atmospheric and Oceanic Physics*, vol. 49, no. 5, pp. 485–493, 2013.

[10] G. Wang, P. Yang, X. Zhou, K. L. Swanson, and A. A. Tsonis, "Directional influences on global temperature prediction," *Geophysical Research Letters*, vol. 39, no. 13, 2012.

[11] A. Hannart, J. Pearl, F. Otto, P. Naveau, and M. Ghil, "Causal counterfactual theory for the attribution of weather and climate-related events," *Bulletin of the American Meteorological Society*, vol. 97, no. 1, pp. 99–110, 2016.

[12] G. Plaut, M. Ghil, and R. Vautard, "Interannual and interdecadal variability in 335 years of central england temperatures," *Science*, vol. 268, no. 5211, p. 710, 1995.

[13] M. Paluš and D. Novotná, "Enhanced Monte Carlo Singular System Analysis and detection of period 7.8 years oscillatory modes in the monthly NAO index and temperature records," *Nonlinear Processes in Geophysics*, vol. 11, no. 5/6, pp. 721–729, 2004.

[14] J. Grieser, S. Trömel, and C.-D. Schönwiese, "Statistical time series decomposition into significant components and application to European temperature," *Theoretical and applied climatology*, vol. 71, no. 3-4, pp. 171–183, 2002.

[15] P. Pišoft, J. Mikšovský, and M. Žák, "An analysis of the spatial distribution of approximate 8 years periodicity in NCEP/NCAR and ERA-40 temperature fields," *The European Physical Journal Special Topics*, vol. 174, no. 1, pp. 147–155, 2009.

[16] R. Vautard, P. Yiou, and M. Ghil, "Singular-spectrum analysis: A toolkit for short, noisy chaotic signals," *Physica D: Nonlinear Phenomena*, vol. 58, no. 1, pp. 95–126, 1992.

[17] M. R. Allen and L. A. Smith, "Monte carlo SSA: Detecting irregular oscillations in the presence of colored noise," *Journal of Climate*, vol. 9, no. 12, pp. 3373–3404, 1996.

[18] A. Pikovsky, M. Rosenblum, and J. Kurths, *Synchronization: a universal concept in nonlinear sciences*, vol. 12. Cambridge university press, 2003.

[19] C. Torrence and G. P. Compo, "A practical guide to wavelet analysis," *Bulletin of the American Meteorological society*, vol. 79, no. 1, pp. 61–78, 1998.

[20] K. Hlaváčková-Schindler, M. Paluš, M. Vejmelka, and J. Bhattacharya, "Causality detection based on information-theoretic approaches in time series analysis," *Physics Reports*, vol. 441, no. 1, pp. 1–46, 2007.

[21] M. Paluš and M. Vejmelka, "Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections," *Phys. Rev. E*, vol. 75, p. 056211, May 2007.

[22] I. Zveryaev, "Climatology and long-term variability of the annual cycle of air temperature over Europe," *Russian Meteorology and Hydrology*, vol. 32, no. 7, pp. 426–430, 2007.

[23] A. Klein Tank, J. Wijngaard, G. Können, R. Böhm, G. Demarée, A. Gocheva, M. Mileta, S. Pashiardis, L. Hejkrlik, C. Kern-Hansen, *et al.*, "Daily dataset of 20th-century surface air temperature and precipitation series for the European Climate Assessment," *International Journal of Climatology*, vol. 22, no. 12, pp. 1441–1453, 2002.

[24] N. Jajcay, J. Hlinka, S. Kravtsov, A. A. Tsonis, and M. Paluš, "Time scales of the European surface air temperature variability : The role of the 7 8 year cycle," *Geophysical Research Letters*, vol. 43, pp. 1–8, 2016.

[25] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, "Testing for nonlinearity in time series: the method of surrogate data," *Physica D: Nonlinear Phenomena*, vol. 58, no. 1-4, pp. 77–94, 1992.

[26] M. Haylock, N. Hofstra, A. Klein Tank, E. Klok, P. Jones, and M. New, "A european daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006," *Journal of Geophysical Research: Atmospheres*, vol. 113, no. D20, 2008.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

Identifying precipitation regimes in China using model-based clustering of spatial functional data

Haozhe Zhang[1], Zhengyuan Zhu[1], Shuiqing Yin[2,3]

***Abstract*—The identification of precipitation regimes is important for many purposes such as agricultural planning, water resource management, and return period estimation. Since precipitation and other related meteorological data typically exhibit spatial dependency and different characteristics at different time scales, clustering such data presents unique challenges. In this short paper, we develop a flexible model-based approach to identify precipitation regimes in China by clustering spatial functional data. Though the focus of this study is on precipitation data, this methodology is generally applicable to other environmental data with similar structure.**

I. INTRODUCTION

The study of precipitation in meteorology and climatology has a significant society impact. For example, drought and flood are two of the most serious meteorological disasters in China, with a direct economic loss of 177 billion Chinese Yuan and annual average of 1256 deaths each year during the period 2001-2014 [1]. Obvious seasonal and interannual variations of precipitation in China affected by Asian monsoon and complex terrain are the main reasons for the frequent drought and flood disasters. Dividing a large geographical area into more homogeneous precipitation regimes [2] has been shown to be useful for precipitation prediction, flood zone management, and regional extreme analysis [3]. Precipitation data has complex characteristics on multiple scales and typically has spatial and temporal dependence, which makes delineating precipitation regimes a non-trivial task. Motivated by this critical need, in this paper we develop a clustering approach for spatial functional data, and apply it to the precipitation data in China.

Regionalization problem has been studied extensively in the meteorological literature. The empirical orthogonal function (EOF) analysis has been widely used for regionalization problems in environmental science [4], [5], which is equivalent to principal component analysis in statistics. The EOF is used in [2] to analyze the normalized monthly mean precipitation data from 1961 to 2006 at 400 stations and obtained a precipitation regionalization focusing on seasonal and interannual variations. However, the seasonal advance and retreat of the summer monsoon rain belt in East Asia behave in a manner with a step of 10-15 days [6], which can not be accurately described using monthly data, and daily rainfall data may be more useful to describe this summer monsoon effect accrurately. Due to the limitation of EOF method, unevenly distributed stations in space can significantly affect the loading patterns. For example, the station density in the western and eastern parts of China is very different, therefore, some stations in the eastern part of China were ignored in the EOF analysis, which led to loss of information.

The motivation of this research is to identify precipitation regimes in China using precipitation data. In this article, we propose a model-based approach to clustering spatial functional data by incorporating both spatial and geographic information in the procedure. In section III, we introduce the functional linear model for observed data and Markov model for cluster memberships with geographic covariates. In section IV, we apply the proposed method to precipitation data.

II. DATA

The data we analyze in this study is the daily precipitation data of 824 meteorological stations in the mainland China from 1951 through 2012. They were provided by the National Meteorological Information Center, China Meteorological Administration. The proportion of the missing days was 0.04%. Only those stations with more than 50 years' complete data are included in the analysis, so there are 722 stations in total used in the analysis. The locations of these meteorological stations are shown in Fig 1.

Corresponding author: Zhengyuan Zhu, zhuz@iastate.edu
[1]Department of Statistics, Iowa State University, Ames, IA [2]State Key Laboratory of Earth Surface Processes and Resource Ecology [3]School of Geography, Beijing Normal University, Beijing, China

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

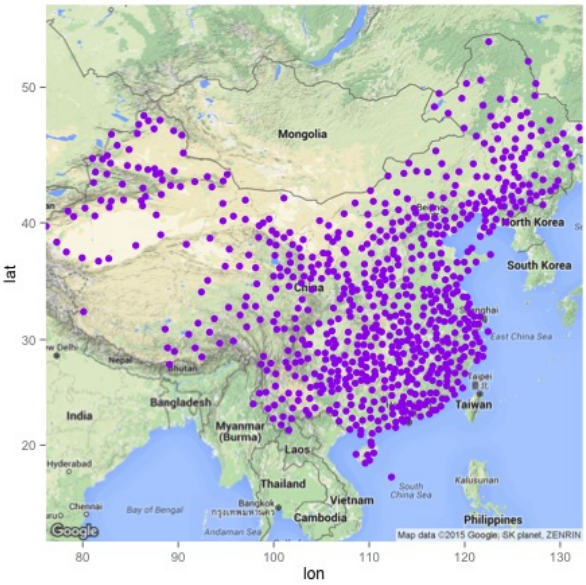


Fig. 1. Spatial distribution of meteorological stations

III. MODEL

Assume Y_{ij} is the precipitation data observed in station i at time point t_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, n_i$. Denote $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{i,n_i})^T$. Let Z_i be the cluster membership, called latent variable, following a multinomial distribution with support $\{1, \ldots C\}$. Here, C is the number of clusters and is a tuning parameter. $Z_i = k$ if $\boldsymbol{Y}_i$ belongs to kth cluster. We call $\{(\boldsymbol{Y}_i, Z_i) : i = 1, \ldots, n\}$ the complete dataset.

A. Functional linear model for observed data

Given the cluster membership Z_i, we assume $\boldsymbol{Y}_i|(Z_i = k)$ follows a multivariate normal distribution with a functional representation:

$$\boldsymbol{Y}_i|(Z_i = k) = \boldsymbol{S}_i(\boldsymbol{\alpha}_k + \boldsymbol{\gamma}_i) + \boldsymbol{\epsilon}_i, \quad (1)$$

$$\boldsymbol{\gamma}_i \sim N(0, \boldsymbol{\Gamma}), \quad \boldsymbol{\epsilon}_i \sim N(\boldsymbol{0}, \sigma^2 I), \quad (2)$$

where $i = 1, \ldots, n, k = 1, \ldots, C$. In the functional linear model, $\boldsymbol{S}_i = (\boldsymbol{s}(t_{i1})^T, \ldots, \boldsymbol{s}(t_{in_i})^T)^T$ is the basis matrix for ith curve. $\boldsymbol{s}(\cdot)$ is a vector of basis functions, which can be B-spline, Fourier or functional principal component. But the row number of basis matrix can vary across different curves to allow irregularly spaced time points and slight missing of data. $\boldsymbol{\alpha}_k$ is the coefficient and needs to be estimated. The data in the same cluster share the same coefficient $\boldsymbol{\alpha}_k$. The difference of $\{\boldsymbol{\alpha}_k\}$ reflects the heterogeneity across clusters. We assume the independence between distinct curves given cluster memberships. However, the within-curve dependence is accounted by the random effect $\boldsymbol{\gamma}_i$, since $\text{cov}(Y_{ij}, Y_{ij'}) =$the (j, j') element of $\boldsymbol{S}_i\boldsymbol{\Gamma}\boldsymbol{S}_i^T$. $\boldsymbol{\epsilon}_i$ can be regarded as the measurement error or stochastic error. Note that $\boldsymbol{\gamma}_i$ and $\boldsymbol{\epsilon}_i$ are confounded. Therefore, some constraint should be imposed for identifiability [7]. We require that

$$\boldsymbol{S}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{S} = \boldsymbol{I}, \quad (3)$$

where $\boldsymbol{S}$ is the basis matrix evaluated over a fine lattice of time points that covers the full range of the data and $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I} + \boldsymbol{S}\boldsymbol{\Gamma}\boldsymbol{S}^T$.

B. Markov model for cluster membership

To fully address the joint distribution of complete data $(\boldsymbol{Y}_i, Z_i)$, we need to specify the distribution of Z_i. Here, we assume the cluster membership follows a Markov model in space. We assume the following probability mass function of cluster memberships in the Markov model

$$P(Z_i = k|\boldsymbol{Z}_{\partial i}) = \frac{exp\{U_{ik}(\theta)\}}{N_i(\theta)}, \quad (4)$$

where $U_{ik}(\theta) = \theta\sum_{j\in\partial i} I(Z_j = k)$ is called the energy function and $N_i(\theta) = \sum_{k=1}^{C} exp\{U_{ik}(\theta)\}$ is the normalizing constant. θ is the interaction parameter that reflects the degree of interaction among nearby sites in Markov random field. The above distribution is called the Gibbs distribution [8], which originates from statistical physics but is widely used in spatial statistics.

There are several ways to incorporate geographic covariates in the Markov model. One way is to generalize the definition of distance from Euclidean distance to "geographic distance" by spatial deformation. For instance, if there is a high mountain between two sites, then the distance between them can be set to be much larger than their euclidean distance on the earth but the geometric properties of Euclidean distance are still kept. The change of the definition of distance may lead to the respective change of neighbors. This method has been introduced in many papers in spatial statistics, to name a few, [9], [10], etc. The second way is to extend the energy distribution by imposing a function $f_{i,j}(\cdot)$ on $I(Z_j = k)$, i.e. $\tilde{U}_{ik}(\theta) = \theta\sum_{j\in\partial i} f_{i,j}\{I(Z_j = k)\}$ and $\tilde{N}_i(\theta) = \sum_{k=1}^{C} exp\{\tilde{U}_{ik}(\theta)\}$, where $f_{i,j}\{Z_{ik}\}$ is a function affected by the geographical covariates between site i and one of its neighbors, i.e. site j.

IV. RESULTS

We applied this method to identify the precipitation regimes in China. Here, we focus on the interseasonal patterns of precipitation. The extension of this method

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

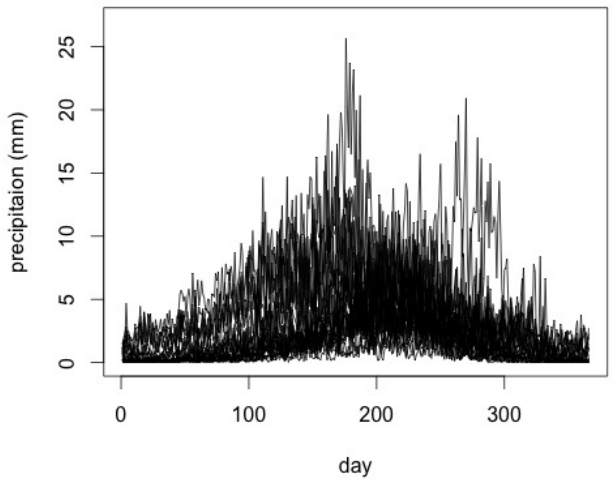


Fig. 2. The averaged daily precipitation in some stations

Fig. 3. Regionalization of precipitation regimes in China

to multi-scale functional and scalar data will be addressed in the following full paper. As a consequence, the averaged daily precipitation records within a year are used in the clustering. The detailed procedure of summarizing data is that, first we get the daily precipitation in each year from 1963 to 2012, then calculate the mean of these 50 curves. Some curves are illustrated in Fig 2. We used the second approach introduced in Section III-B to incorporate geographical covariate. If the elevation difference between two stations is larger than 1000m [11], we no longer consider them to be the "neighbors" in the Markov random field even if they are closest in terms of distance.

The final cluster assignments are shown in Fig 3. The results of clustering are consistent with the stepwise manner of East Asian monsoon. The seasonal advance and retreat of the summer monsoonal airflow and monsoon rain belt in East Asia behave in a stepwise manner (Ding, 2004). When the East Asian summer monsoon advances northward, it undergoes three standing stages (South China and northern South China Sea from mid-May to early June; 25−30°N from mid-June to mid-July; and 40−45°N during the last 10 days of July to mid-August), and two stages of abrupt northward shifts (the first 10 days of June and around mid-July). In early or mid-August the rainy season of North China comes to end, with the major monsoon rain belt disappearing. From the end of August to early September the monsoon rain belt moves back to South China again.

V. DISCUSSION

In this short paper, we develop a flexible model-based approach to cluster precipitation data which utilizes the spatial and geographical information. There are still several important aspects of this method needed to be addressed, such as the selection of cluster numbers, how to evaluate the uncertainty of clustering assignments, etc. The parameter estimation, simulation study, model selection, extension to multi-scale data and uncertainty assessment will be introduced and addressed in the following full paper.

REFERENCES

[1] *China Meteorological Administration (CMA): Statistical Yearbooks of Meteorological Disasters in China*. Meteorology Publishing House, 2015.

[2] C. Li-Juan, C. De-Liang, W. Hui-Jun, and Y. Jing-Hui, "Regionalization of precipitation regimes in China," *Atmospheric and Oceanic Science Letters*, vol. 2, no. 5, pp. 301–307, 2009.

[3] J. R. M. Hosking and J. R. Wallis, *Regional frequency analysis: an approach based on L-moments*. Cambridge University Press, 2005.

[4] D. White, M. Richman, and B. Yarnal, "Climate regionalization and rotation of principal components," *International Journal of Climatology*, vol. 11, no. 1, pp. 1–25, 1991.

[5] C. B. Uvo, "Analysis and regionalization of northern european winter precipitation based on its relationship with the north atlantic oscillation," *International Journal of Climatology*, vol. 23, no. 10, pp. 1185–1194, 2003.

[6] Y. Ding, "Seasonal march of the east-asian summer monsoon," *East Asian Monsoon*, vol. 2, no. 30, p. e53, 2004.

[7] G. M. James and C. A. Sugar, "Clustering for sparsely sampled functional data," *Journal of the American Statistical Association*, vol. 98, no. 462, pp. 397–408, 2003.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

[8] H. Jiang and N. Serban, "Clustering random curves under spatial interdependence with application to service accessibility," *Technometrics*, vol. 54, no. 2, pp. 108–119, 2012.

[9] P. D. Sampson and P. Guttorp, "Nonparametric estimation of nonstationary spatial covariance structure," *Journal of the American Statistical Association*, vol. 87, no. 417, pp. 108–119, 1992.

[10] E. B. Anderes and M. L. Stein, "Estimating deformations of isotropic gaussian random fields on the plane," *The Annals of Statistics*, pp. 719–741, 2008.

[11] L. Gerlitz, O. Conrad, and J. Böhner, "Large-scale atmospheric forcing and topographic modification of precipitation rates over high asia-a neural-network-based approach," *Earth System Dynamics*, vol. 6, no. 1, p. 61, 2015.

6th International Workshop on Climate Informatics
September 22-23, 2016
CI 2016 Hosted by the National Center for Atmospheric Research in Boulder, CO

RELATIONAL RECURRENT NEURAL NETWORKS FOR SPATIO-TEMPORAL INTERPOLATION FROM MULTI-RESOLUTION CLIMATE DATA

Guangyu Li[1], Yan Liu[1]

***Abstract*—Spatio-temporal interpolation is a fundamental task in climate modeling. A series of models have been developed and achieved successes to a certain extent. However, in many real world applications, we are confronted with spatiotemporal climate data in *multiple resolutions*. This raises significant challenges to existing solutions. To address this problem, we generalized recent advance in variational autoencoder for modeling sequential data to modeling relational (e.g. spatially related) multi-resolution time series. Specifically, we proposed a generative model, namely Relational Recurrent Neural Network, to jointly model temporal dynamics and spatial structures within multi-resolution climate data as an effective solution to spato-temporal interpolation. Experiments on a wind speed dataset show that Relational RNN can capture spatio-temporal relationships and achieve considerable improvement over state-of-art methods.**

I. MOTIVATION

Spatio-temporal interpolation, i.e., estimating unobserved values in locations or time of interest, is a fundamental task in climate modeling [1] A series of models have been developed for this task by modeling spatial and temporal dependencies simultaneously, such as deterministic interpolation methods (e.g., Inverse Distance Weighting (IDW) based and regression-based methods), and stochastic models (e.g., spatio-temporal Kriging and spatio-temporal Gaussian process) [2].

Even though the idea of jointly modeling spatial and temporal relationship is intriguing, existing methods are mainly designed to model time series with the same temporal resolution (i.e. sampled at the same frequency). It is well known that in climate domain, many time series observations come from different sources and may have various temporal resolutions. To apply existing models to multiresolution data, researchers typically adopt the preprocessing step, i.e., aggregating high resolution observations to lowest available resolutions or interpolating low resolution data into highest resolutions. In this pre-filtering process, potentially useful information may get lost and mis-specification may be induced. Hence, effectively handling multi-resolution data becomes an essential task in spatio-temporal interpolation.

Very recently, variational-autoencoder(VAE)-based deep generative models have recently shown to be an effective framework to extract higher representation and model flexible density over data space [3]. Given a high-dimensional data, $\boldsymbol{x} = \{\boldsymbol{x}_n\}_{n=1}^N$, the VAE introduce a set of latent random variables $\boldsymbol{z}_n$ to capture higher representation in observation $\boldsymbol{x}_n$ through conditional probability $p(\boldsymbol{z}_n \mid \boldsymbol{x}_n)$ (stochastic encoder), and further model observation density through conditional probability $p(\boldsymbol{x}_n \mid \boldsymbol{z}_n)$ (stochastic decoder). Note that the VAE usually parameterizes these probabilities with highly flexible mapping such as L layers neural networks, and then make efficient inference with Variational Bayes methods. By leveraging variation induced by latent random variables and flexibility of neural network mapping, VAE has already shown promise as a generative model in modeling many kinds of complicated data density. [3]

In this paper, we generalize the representation capability of VAE from solely modeling sequential data to modeling spatio-temporal data. Specifically, we propose a generative model based on a recurrent version of VAE [4], namely Relational Recurrent Neural Networks (Relational RNN), to handle multi-resolution time series in spatio-temporal interpolation. The intuition is that we use variational recurrent autoencoder to map each time series into a fixed-length random vectors as higher-level representations in each time window (e.g., 1 hour), and then model the spatial structures between these random vectors via multivariate Gaussian random fields. Since the model input is a set of fixed-time window frames, all multi-resolution time series would

Corresponding author: Guangyu Li, guangyul@usc.edu
[1]Department of Computer Science, University of Southern California

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

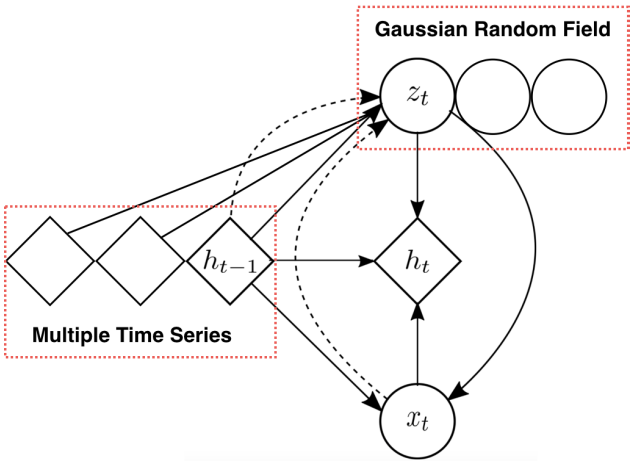


Fig. 1. Graphical illustration of Relational RNN.

be synchronized automatically without information loss or mis-specification. In addition, we train the model with autoencoder reconstructure likelihood and random field likelihood jointly to make an unified generative model.

II. METHOD

Suppose we aim to model N climate time series $\mathbf{X}_{1:N}$ with different temporal resolutions $r_{1:N}$. Firstly, time series are represented as a sequence of time window frames. Note that the number of samples within each frame may vary among time series due to their different temporal resolutions. For each time series $\{\mathbf{x}_{n,t}\}_{t=1}^{T}$, we model it through a variational recurrent autoencoder with a series of latent variables $\{\mathbf{z}_{n,t}\}_{t=1}^{T}$ and hidden states $\{\mathbf{h}_{n,t}\}_{t=1}^{T}$. At the same time, we apply a multivariate Gaussian random field (GRF) to model the joint distribution of all latent variable for each time step $\{\mathbf{z}_{n,t}\}_{n=1}^{N}$.

For time series n at each time step t, the prior of latent variable $\boldsymbol{z}_{n,t}$ is conditioned on latent variables of all time series at previous time step.

$$\boldsymbol{z}_{n,t} \sim \mathcal{N}(\boldsymbol{\mu}_{0,t}, diag(\boldsymbol{\sigma}_{0,t}^2)), \tag{1}$$

where $[\boldsymbol{\mu}_{0,t}, \boldsymbol{\sigma}_{0,t})] = \varphi^{prior}(\boldsymbol{h}_{1:N,t-1})$. The posterior (encoder) and generate distribution (decoder) are similar with regular variational autoencoder, defined as follows:

$$\begin{aligned}\boldsymbol{z}_{n,t} \mid \boldsymbol{x}_{n,t} &\sim \mathcal{N}(\boldsymbol{\mu}_{z,t}, diag(\boldsymbol{\sigma}_{z,t}^2)),\\ \text{where } [\boldsymbol{\mu}_{z,t}, \boldsymbol{\sigma}_{z,t})] &= \varphi^{enc}(\varphi^{\boldsymbol{x}}(\boldsymbol{x}_{n,t}), \boldsymbol{h}_{n,t-1})\end{aligned} \tag{2}$$

$$\begin{aligned}\boldsymbol{x}_{n,t} \mid \boldsymbol{z}_{n,t} &\sim \mathcal{N}(\boldsymbol{\mu}_{x,t}, diag(\boldsymbol{\sigma}_{x,t}^2)),\\ \text{where } [\boldsymbol{\mu}_{x,t}, \boldsymbol{\sigma}_{x,t})] &= \varphi^{dec}(\varphi^{\boldsymbol{z}}(\boldsymbol{z}_{n,t}), \boldsymbol{h}_{n,t-1})\end{aligned} \tag{3}$$

While the hidden state $\boldsymbol{h}_{n,t}$ is updated with recurrent equation.

$$\boldsymbol{h}_{n,t} = f_\theta(\varphi^{\boldsymbol{x}}(\boldsymbol{x}_t), \varphi^{\boldsymbol{z}}(\boldsymbol{z}_t), \boldsymbol{h}_{n,t-1}) \tag{4}$$

Here $\varphi^{prior}, \varphi^{enc}, \varphi^{dec}$ generate distribution parameters and $\varphi^{\boldsymbol{x}}, \varphi^{\boldsymbol{x}}$ extract features from $\boldsymbol{x}_{n,t}$ and $\boldsymbol{z}_{n,t}$. Generally, they can be any valid function such as neural network in our case.

To model relationship among time series, we apply a multivariate GRF to all latent variable at each time step $\{\mathbf{z}_{n,t}\}_{n=1}^{N}$.

$$(\boldsymbol{z}_{1,t}, \boldsymbol{z}_{2,t}, ...\boldsymbol{z}_{N,t})^T \sim \mathcal{N}(\boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z) \tag{5}$$

Note that if there is prior knowledge of dependenceamong time series i.e. spatial structure, we could add constraints on covariance matrix i.e. kronecker product structure, or define GRF based on precision matrix to better illustrate the conditional independence. For large scale time series modeling, we could even switch to Gaussian Markov random field instead for higher efficiency. [5]

The parameters are estimated by maximizing the likelihood of the data. Specifically, the objective function involve two parts, i.e. timestep-wise variational lower bound and multivariate GRF likelihood.

$$\begin{aligned}&\sum_{t=1}^{T} p(\boldsymbol{z}_{1:N,t} \mid \boldsymbol{\mu}_Z, \boldsymbol{\Sigma}_Z) + \sum_{n=1}^{N} \Bigg(\mathbb{E}_{q(\boldsymbol{z}\le T_n \mid \boldsymbol{x}<T_n)} \Big(\sum_{t=1}^{T_n} \\ &(-\mathrm{KL}(q(\boldsymbol{z}_t \mid \boldsymbol{x} \le T_n, \boldsymbol{z} < T_n) \parallel p(\boldsymbol{z}_t \mid \boldsymbol{x} < T_n, \boldsymbol{z} < T_n)) \\ &+ \log p(\boldsymbol{x}_t \mid \boldsymbol{z} \le T_n, \boldsymbol{x} < T_n)) \Big) \Bigg)\end{aligned} \tag{6}$$

where $p(\boldsymbol{z}_t \mid \boldsymbol{x} < T_n, \boldsymbol{z} < T_n)$, $q(\boldsymbol{z}_t \mid \boldsymbol{x} \le T_n, \boldsymbol{z} < T_n)$, $p(\boldsymbol{x}_t \mid \boldsymbol{z} \le T_n, \boldsymbol{x} < T_n)$ correspnd to equation (1) (2) (3) repesctively.

After we train the generative model, interpolation is carried out in a streaming fashion. That is, the interpolation result of current time step is provided before next time window arrives [6]. Suppose we are interested in predicting $\boldsymbol{x}_{n,t}$ based on $\boldsymbol{h}_{t-1}$, $\boldsymbol{x}_{-n,t}$ and $\boldsymbol{z}_{-n,t}$. First, we predict $\boldsymbol{z}_{-n,t}$ from $\boldsymbol{h}_{t-1}$ and $\boldsymbol{x}_{-n,t}$ with (2), then compute condition distribution of $\boldsymbol{z}_{n,t} \mid \boldsymbol{z}_{-n,t}$ through multivariate GRF. Finally, we obtain the distribution of $\boldsymbol{x}_{n,t}$ from $\boldsymbol{h}_{n,t-1}$ and expectation of $\boldsymbol{z}_{n,t}$ with (3).

III. EXPERIMENTS

We evaluate the proposed Relational RNN model on the task of wind speed interpolation. The data comes from NREL Western Wind Resource Dataset which

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

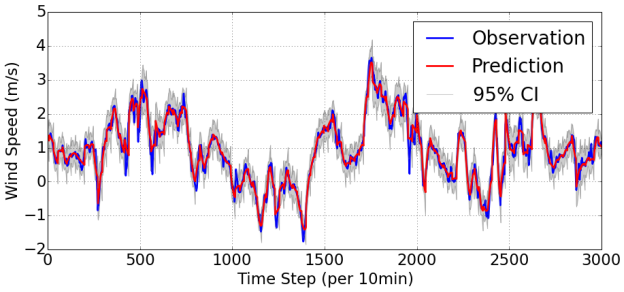


Fig. 2. Snapshot of Relational RNN Interpolation Results

TABLE I
PREDICTION PERFORMANCE COMPARISION BETWEEN RELATIONAL RNN AND BASELINE METHODS

	Relational RNN	Regression	STGP
MAE	0.1023	0.1237	0.1344

have wind speed data sampled every 10 minutes over 32043 locations in western United States for 3 years, 2004, 2005, 2006. We choose a subset of 7 locations and down-sample time series into three frequencies i.e. sampled per 20min, 30min, 60min, to generate multiple temporal resolutions settings. We set time window as 4 hours and interpolate 1 locations based on others in the streaming scenario. The proposed model is compared with one deterministic baseline, regression-based model, and one stochastic baseline, Spatio-temporal Gaussian Process.

Figure 2 shows a snapshot of the interpolation result and corresponding confidence interval. The overall results are summarized into Table 1. As we can see, Relational RNN achieves significantly better performance compared with the two baseline models.

IV. CONCLUSION

We have introduced a generative model, Relational Recurrent Neural Network (Relational RNN), for spatio-temporal modeling. The proposed model can handle multi-resolution time serise data directly without information loss. The experiment shows that the Relational RNN model is capable to capture complex spatio-temporal structure in an interpolation task. In addition, with the flexible fixed-length random vector representations, Relational RNN can reveal important information regarding spatial dependencies, such as correlation and spatial clustering within spatio-temporal data. The possible extensions above will be pursued in our future work.

REFERENCES

[1] A. Appice, A. Ciampi, D. Malerba, and P. Guccione, "Using trend clusters for spatiotemporal interpolation of missing data in a sensor network," *Journal of Spatial Information Science*, vol. 2013, no. 6, pp. 119–153, 2013.

[2] J. Lindström, A. Szpiro, P. D. Sampson, S. Bergen, and L. Sheppard, "Spatiotemporal: An r package for spatio-temporal modelling of air-pollution," *J stat softw (in press)(http://cran. rproject. org/web/packages/SpatioTemporal/index. html)*, 2013.

[3] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[4] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Advances in neural information processing systems*, pp. 2980–2988, 2015.

[5] H. Rue and L. Held, *Gaussian Markov random fields: theory and applications*. CRC Press, 2005.

[6] X. Zhong, A. Kealy, and M. Duckham, "Stream kriging: Incremental and recursive ordinary kriging over spatiotemporal data streams," *Computers & Geosciences*, vol. 90, pp. 134–143, 2016.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

OBJECTIVE SELECTION OF ENSEMBLE BOUNDARY CONDITIONS FOR CLIMATE DOWNSCALING

Andrew Rhines[1], Naomi Goldenson[1]

***Abstract*—Computational constraints often lead to downscaling experiments being limited to a small ensemble of simulations. For small downscaling ensembles, random sampling of boundary condition datasets may not adequately cover the range of future climate scenarios. We describe a procedure for efficiently and objectively choosing boundary conditions for a prescribed number of ensemble members. The cost function used to assess distance between boundary condition datasets can be specified to maximize ensemble spread in several ways depending on the needs of the application. We provide an example case study using tropical sea surface temperatures from the CESM Large Ensemble.**

I. MOTIVATION

Downscaling is used to provide higher resolution estimates of the impacts of large-scale climate change. Many different downscaling methods have been proposed and tested, and can generally be categorized as either statistical or dynamical techniques [1], though in some cases the two classes of methods are combined in order to include empirical bias corrections [2]. Statistical downscaling can increase the resolution and accuracy of global climate model (GCM) output through the inclusion of additional covariates known to affect local climate conditions. Dynamical downscaling uses a nested high resolution physical model that is driven by the coarse resolution output of a GCM through specification of atmospheric boundary conditions at the edges of the domain, and typically also of surface boundary conditions over the ocean. The high resolution numerical grid permits for simulation of smaller scale physical processes and topography not resolved by the GCM. By improving the representation of fine-scale features such as topography, land use, and hydrology, GCM biases can be reduced and more accurate projections of climate change can be provided.

Corresponding author: A Rhines, arhines@atmos.uw.edu
[1]Department of Atmospheric Sciences, University of Washington

A challenge associated with dynamical downscaling is that increased resolution comes at a high computational cost that often limits our ability to simulate a complete range of climate scenarios. For example, large ensemble simulations from GCMs have been used to assess uncertainty due to physical parameterizations (e.g., [3]) and the role of internal variability in multiple climate change scenarios using the Community Earth System Model Large Ensemble (CESM-LE, [4]). However, it is computationally prohibitive to dynamically downscale each ensemble member, raising the question of how to objectively select a subset of the ensemble members. Constructing boundary conditions from scratch is undesirable as it leads to simulations that are not dynamically self-consistent. To our knowledge, only one attempt has been made to provide an alternative to simple random sampling; [5] presented a method of selecting m members from an ensemble of size $N > m$, using hierarchical clustering of Empirical Orthogonal Function (EOF) expansion coefficients (referred to as principal components, or PCs, in climate literature).

We propose a method of selecting m members from $N >> m$ potential boundary condition datasets by clustering, using a flexible similarity measure. We describe the clustering method in Section II, discuss several useful similarity measures in Section III, and discuss an application of the method to the Community Earth System Model (CESM) Large Ensemble.

II. CLUSTERING METHOD

[5] used hierarchical agglomerative clustering, setting the number of clusters on the basis of their computational constraint, m. Hierarchical clustering classifies members among the m clusters, but does not directly address the question of which member from each cluster to use for downscaling. We address this with exemplar clustering, a class of methods that both clusters the data and selects a representative member from each. We use k-medoids clustering with the Partitioning Around

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

Medoids (PAM) algorithm [6], as it is in common use. PAM is a greedy algorithm that may result in a local rather than global optimum; we note that other methods of exemplar clustering could be used in its place, including affinity propagation [7] and other variants of the k-medoids algorithm (e.g., [8]).

The clustering method is as follows:

1) Randomly initialize by selecting, without replacement, $k < N$ members as cluster medoids.
2) For each member, find the nearest medoid.
3) Until the cost — the sum of the intra-cluster distances, $1 - S$ — no longer increases:
   a) Compute the cost of switching each non-medoid member with each medoid.
   b) Perform the switch for each sequential comparison only if the cost decreases.

III. Similarity Measures for GCM Ensembles

Defining a similarity measure, S, between the output of different simulations requires collapsing the temporal and spatial dimensions of the data. We describe two complementary approaches that address temporal and spatial similarity in different ways. Method I uses EOFs to measure similarity in terms of the temporal development of the leading modes of variability in the simulations. Method II permits for shuffling of the data in time, comparing the range of synoptic conditions experienced throughout the simulations. The first method is useful when temporal trends are of greatest interest, e.g., when considering how internal variability can affect short-term trends. The second method would be employed when attempting to capture the full range of conditions irrespective of their relative timing.

A. EOF-Based Measures

Whereas [5] performed principal components analysis (PCA) on average temperatures within specific geographic regions, we retain the full spatial field. Given some variable of interest, Y, we construct maps for each ensemble member, $\mathbf{Y}_{(i)}$, where i is the ensemble number, rows map to a spatial index over latitude and longitude, and columns map to a temporal index. We note that while emergent modes of variability such as the Pacific Decadal Oscillation (PDO) and the El Niño Southern Oscillation (ENSO) are often defined by EOFs [9], their spatial projections can differ substantially between different models, observations, and even different simulations using the same model. Thus an important step is to first compute combined EOFs, wherein each ensemble member is treated as an independent realization of the same dynamical system. Combined EOFs are computed by first constructing the combined observation matrix,

$$\mathbf{Y} = \left[\mathbf{Y}_{(1)}|\mathbf{Y}_{(2)}|\dots|\mathbf{Y}_{(N)}\right].$$

If monthly or daily data are used, anomalies $\mathbf{Y}'$ are generally used in place of $\mathbf{Y}$ by removing the seasonal cycle across all ensemble members and years. Removing the seasonal cycle at this stage is more precise than doing so in the individual $\mathbf{Y}_i$; using multiple ensemble members reduces the standard error of the estimated seasonal cycle by a factor of $\sqrt{N}$ provided that each ensemble member is longer than the dominant timescales of variability in the model. The leading modes of variability can then be identified via standard EOF analysis on $\mathbf{Y}'$. Given that the matrix may be unusually large due to the number of ensemble members, it may be necessary to employ a sequential or truncated variant of the singular value decomposition (e.g., [10]). The PC timeseries, $\mathbf{x}_{(i)}(t)$, for each ensemble member are then obtained by projecting the combined EOF spatial patterns of each mode back onto the original de-seasonalized data. Similarity between ensemble members is then computed using some norm over $\mathbf{x}(t)$,

$$S_{ij} = \|\mathbf{x}_{(i)} - \mathbf{x}_{(j)}\|,$$

which is then supplied directly to the PAM algorithm.

B. Time-Agnostic Measures

The need also arises to compare simulations in terms of the uniqueness of spatial patterns represented across all times. For example, one might wish to ensure that different modes of variability are represented in each of their possible combined phases irrespective of the time at which this occurs, capturing situations such as those where strong El Niño events occur simultaneously with the negative, positive, and neutral phases of the PDO. To achieve this, we wish to compute the similarity matrix,

$$S_{ij} = \|\mathbf{Y}_{(i)} - \mathbf{Y}_{(j)}\mathbf{P}_{(i,j)}\|,$$

where $\mathbf{P}_{(i,j)}$ is the permutation matrix minimizing

$$\arg\min_{\mathbf{P}} \|\mathbf{Y}_{(i)} - \mathbf{Y}_{(j)}\mathbf{P}\|.$$

As with the EOF-based approach we use the L_1 norm, though other cost functions could be trivially substituted instead. Finding each $\mathbf{P}$ using a naive auction algorithm given N_t temporal samples would require $\mathcal{O}(N_t!)$ time, which is clearly impractical for this problem given that N_t will typically be at least 30 for annual means or 360

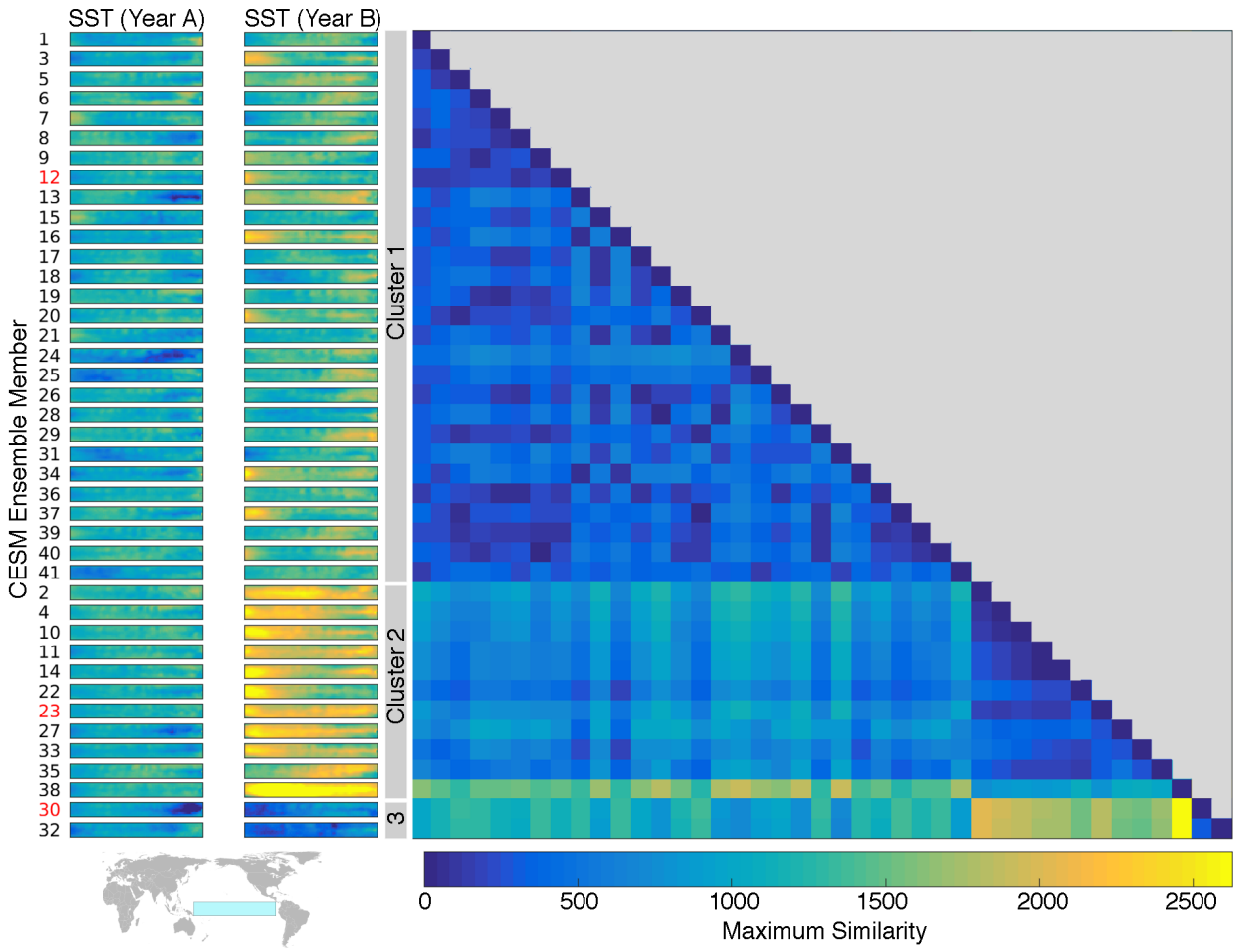


Fig. 1. Example of time-agnostic clustering using sea surface temperatures from the tropical Pacific in 41 members of the CESM Large Ensemble. A two-year block is randomly chosen for each ensemble member, and the Hungarian algorithm is used to compute the similarity between all pairs of ensemble members while ignoring the temporal ordering of the maps. Clusters are identified using k-medoids with $m = 3$, and the medoids for each cluster are highlighted in red. Years (labeled 'A' and 'B') are aligned relative to the medoid for each cluster. The similarity matrix $\mathbf{S}$ is shown at right, and is sorted to match the maps to the left.

for monthly data over a thirty year simulation. However, it is possible to use the Hungarian algorithm [11] to solve the assignment problem in $\mathcal{O}(N_t^3)$. By symmetry, computation of $\mathbf{S}$ requires doing this $M(M+1)/2$ times, yielding a complexity of $\mathcal{O}(N^2 N_t^3)$. In practice we are able to compute $\mathbf{S}$ within 100 hours on a single core, even for the largest ensemble problems involving several centuries of monthly data from 40 ensemble members. Furthermore, the problem can be easily parallelized at two different stages.

We now provide a brief example application of the time-agnostic clustering using the CESM Large Ensemble. For ease of display and because it contains the ENSO region that is the dominant source of interannual variability through teleconnections, we focus on sea surface temperatures in the tropical Pacific. We draw only two years ($N_t = 2$) from each of the 41 ensemble members, and use the Hungarian algorithm to find the best possible match between the maps in each pair of ensemble members. We then perform k-medoids clustering using $m = 3$ clusters. The results (Fig. 1) show that the algorithm is identifies three distinct sets of conditions. Cluster 1 is the largest, and represents ENSO-neutral conditions. Cluster 2 is second largest, and contains 11 El Niño events of varying strength. Cluster 3 is the smallest, containing two members with La Niña events. Despite each of these events occurring with vastly different frequencies in this particular subset of the ensemble, the algorithm ensures that one representative member of each cluster is identified for use in a 3-simulation downscaling experiment.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

ACKNOWLEDGMENTS

Funding was provided by the James S. McDonnell Foundation (award number 220020421).

REFERENCES

[1] R. L. Wilby and T. Wigley, "Downscaling general circulation model output: a review of methods and limitations," *Progress in Physical Geography*, vol. 21, no. 4, pp. 530–548, 1997.

[2] E. Zorita and H. Von Storch, "The analog method as a simple statistical downscaling technique: comparison with more complicated methods," *Journal of climate*, vol. 12, no. 8, pp. 2474–2489, 1999.

[3] N. Massey, T. Aina, M. Allen, C. Christensen, D. Frame, D. Goodman, J. Kettleborough, A. Martin, S. Pascoe, and D. Stainforth, "Data access and analysis with distributed federated data servers in climateprediction. net," *Advances in Geosciences*, vol. 8, pp. 49–56, 2006.

[4] J. Kay, C. Deser, A. Phillips, A. Mai, C. Hannay, G. Strand, J. Arblaster, S. Bates, G. Danabasoglu, J. Edwards, *et al.*, "The community earth system model (cesm) large ensemble project: A community resource for studying climate change in the presence of internal climate variability," *Bulletin of the American Meteorological Society*, vol. 96, no. 8, pp. 1333–1349, 2015.

[5] T. Mendlik and A. Gobiet, "Selecting climate simulations for impact studies based on multivariate patterns of climate change," *Climatic Change*, vol. 135, no. 3, pp. 381–393, 2016.

[6] L. Kaufman and P. J. Rousseeuw, "Partitioning around medoids (program pam)," *Finding groups in data: an introduction to cluster analysis*, pp. 68–125, 1990.

[7] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *science*, vol. 315, no. 5814, pp. 972–976, 2007.

[8] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[9] C. Deser, M. A. Alexander, S.-P. Xie, and A. S. Phillips, "Sea surface temperature variability: Patterns and mechanisms," *Annual Review of Marine Science*, vol. 2, pp. 115–143, 2010.

[10] I. Mogotsi, "Christopher d. manning, prabhakar raghavan, and hinrich schütze: Introduction to information retrieval," *Information Retrieval*, vol. 13, no. 2, pp. 192–195, 2010.

[11] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

LONG-LEAD PREDICTION OF EXTREME PRECIPITATION CLUSTER VIA A SPATIO-TEMPORAL CONVOLUTIONAL NEURAL NETWORK

Yong Zhuang[1], Wei Ding[1]

***Abstract*—A reliable long-lead (5-15 days ahead) prediction of extreme precipitation cluster is vitally important for regional flooding forecasting. A significant research effort is to develop methods for making long-lead flood forecasts using machine learning techniques,as current physics-based numerical simulation models can be extremely complex to account for compounding uncertainty in measurements and modeling. Accurate precipitation forecasts by numerical weather prediction models are limited to a few days lead-time, because non-linearity in the governing equations of the atmosphere creates a sensitive dependence on initial conditions. We design a novel Spatio-Temporal Convolutional Neural Network (ST-CNN) to fully utilize the spatial and temporal information and automatically learn underlying patterns of precipitation precursors from data for extreme precipitation cluster prediction. We validate the ST-CNN model using 62 years historical precipitation data collected in the State of Iowa, USA, from 1948-2010.**

I. INTRODUCTION

According to the U.S. Geological Survey [1] , floods were the number-one natural disaster in the United States in terms of number of lives lost and property damage during the 20th century. Regional flooding is often produced by long sequences of slowly moving, low-pressure or frontal storm systems including decaying hurricanes or tropical storms accruing over periods of several days to several weeks. A reliable long-lead (5-15 days ahead) prediction of extreme precipitation event is vitally important for mitigating flood damage. Accurate precipitation forecasts by numerical weather prediction models are limited to a few days lead-time because the non-linearity in the governing equations of the atmosphere creates a sensitive dependence on initial conditions that causes an effort in the initial conditions to double after just a few days, thus making long-range forecasts (longer than 7 days) practically impossible. Understanding the future trend of climate requires accurately identifying the precipitation precursors. Long-lead predictions have to consider variables in a long time period and large spatial neighborhoods, which involves an enormous amount of potentially influencing variables.

The goal of this study is to integrate machine learning and data mining methods with hydrological science and atmospheric science to detect interesting spatio-temporal patterns from this huge feature space to improve long-lead forecasting of extreme precipitation events. We design and implement a new Spatio-Temporal Convolutional Neural Network (ST-CNN) model to automatically learn the dependency of meteorological variables on spatio-temporal neighborhoods and summarize the patterns of local neighboring groups of neurons, to predict heavy precipitation cluster in 10 days ahead. We evaluate ST-CNN using 62 years historical meteorological data collected in the State of Iowa, USA.

II. RELATED WORK

Over the last three decades, a great deal of attention in statistics and machine learning has been directed toward extreme weather prediction[2] [3]. Most of them rely on meteorological inputs that usually come from observation networks and radar [4], and require a complex and meticulous simulation of the physical equations in the atmosphere model.

In recent years, machine learning feature selection methods, which aim to select a subset of relevant features from an original feature set, often at a scale of millions of features, have become popular in climate research for constructing forecasting models. For instance, Wu et al. used Online Streaming Feature

{yong.zhuang001, wei.ding}@umb.edu [1]Department of Computer Science, University of Massachusetts Boston, Boston, MA

Selection (OSFS) for heavy precipitation prediction [5][6][7]; Wang et al. applied the fast-OSFS algorithm for extreme flood forecasting [8][9]. Although feature selection methods can simplify forecasting models for easier interpretation and time efficiency, the approach usually requires domain scientists to provide initial feature sets that are closely related with the problem domain.

Neural-network-based machine learning approaches have been very successfully used for detecting high level patterns from raw low-level features without much intervention with prior domain knowledge [10]. Anctil et al. used artificial neural network (ANN) technique to forecast rainfall [11]. The results show that ANN forecasting models can get superior results to those obtained by linear regression models. More recently, Shi et al. implemented a new convolutional long short term memory (LSTM) deep neural network for precipitation nowcasting [12]. This model is trained on two dimensional radar map time series data. Their study showed that deep networks reveal a great potential on various climate problems. In our study, we explore a new Convolutional Neural Network architecture to learn patterns from meteorological variables in spatio-temporal grids for the long-lead prediction of extreme precipitation clusters.

III. METHOD

We formulate the long-lead prediction of extreme precipitation cluster as a classification problem with multiple spatio-temporal tensor data as inputs.

A. Spatio-Temporal Tensor Features

If we use the positions of cells in a data matrix to represent spatio grids, then one observed variable over a spatial region of m by n can be listed in a $m \times n$ matrix, which consists m rows and n columns. Then the matrices of k variables, which are collected at the same time t, can be stacked as a $m \times n \times k$ cuboid Q_t. If the observations are recorded periodically, we get a sequence of cuboids $Q_{t_1}, Q_{t_2}, ..., Q_{t_q}$ (in this study, we use $q = 10$ days records of 9 meteorological variables over a 32 by 32 region, and the size of cuboid is $32 \times 32 \times 9$). Thus, the multiple spatio-temporal sequences can be represented by a tensor $\chi \in \mathbb{R}^{m \times n \times k \times q}$. Then the long-lead ($x$ time-stamps ahead) prediction of extreme precipitation cluster problem can be formulated as follows:

$$P(C_{t_{q+x}}) = \operatorname{argmax} P(C_{t_{q+x}} | Q_{t_1}, Q_{t_2}, ..., Q_{t_q}) \quad (1)$$

Here $C_{t_{q+x}}$ denotes the class label of the $(t_{q+x})^{th}$ time-stamp, and the objective function selects the mostlikely outcome class (extreme precipitation cluster vs. non-extreme precipitation cluster) in a x time-stamps lead time given previously known spatio-temporal tensor data sequences.

B. Spatio-temporal Data Analysis

Our goal is let a CNN model automatically identify interesting and physically meaningful spatio-temporal patterns from data for precipitation cluster precursor identification. Certain pre-cursor patterns in the synoptic domain may indicate the development and movement of strong storms, including the location of fronts or strong horizontal temperature gradients, the presence of an upstream trough / ridge axis or a strong jet streak or a change in low-level winds.

C. The Spatio-Temporal Convolutional Neural Network

A Convolutional Neural Network (CNN) architecture is usually formed by a stack of distinct layers that transform the input volume into an output volume through a differentiable function. In this study, we build our ST-CNN architecture with six layers, including two convolutional layers, two max polling layers, and two fully connected layers. The ST-CNN is designed to use those layers of neurons (learning units) to automatically detect very local and detailed representations of a broad class of patterns from tensor data at the convolutional layers, and then summarize those local patterns to build high level features in the max pooling layers. The configuration of our architecture is depicted in Figure 1. The output volume is a vector that includes the class scores of the binary class labels of the extreme precipitation clusters after the calculation of the fully connected layers. We use back propagation algorithm to search for minimum of loss function in weight space and apply L_2 regularization to prevent overfitting.

Convolutional Layer: In ST-CNN architecture, there are two convolutional layers, which consist of multiple 3D filters(kernels) with the size of 5×5 (in this project, we choose to detect local patterns in a 5 by 5 neighborhood in the convolutionary layer). The first convolutional layer has 10 filters and the second one has 15 filters. Each filter takes inputs from a cuboid section of the previous layer. The weights for this cuboid section are the same for each filter in the convolutional layer to reduce the number of neural network parameters to be learned. Thus, the convolutional layer is just a feature map convolution of the previous layer. The task of the convolutional layer is to automatically

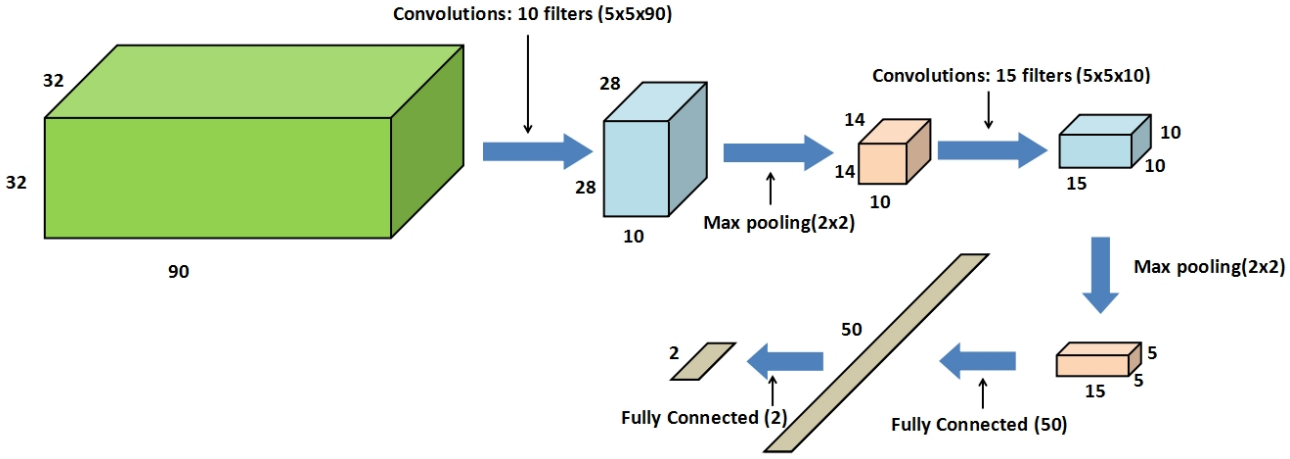


Fig. 1. The ST-CNN Architecture and its layer parameters. The ST-CNN contains two convolutional layers with filter size (5×5), two pooling layer with filter size (2×2), and two fully connected layer with 50 neurons and 2 neurons respectively. The input is a $32 \times 32 \times 90$ tensor which is associated with nine meteorological variables of 10 days over a 32 by 32 region. Nine meteorological variables are PW, T850, U300, U850, V300, V850, Z300, Z500, and Z1000 [13]. The final output is a vector that includes the class scores of the binary class labels (extreme precipitation cluster vs. non-extreme precipitation cluster).

learn local meaningful patterns that are associated with class labels.

Pooling Layer: Each convolutional layer is followed by a pooling layer which takes small cuboid blocks from the convolutional layer and sub-samples it to produce a single output from that block. In other words, each pooling layer is the summary of the patterns learned by convolutional layer. Here, pooling layers are max-pooling layers with 2×2 filters, which let the patterns of previous convolutional layers be reduced at half size and the outputs of adjacent pooling units do not overlap. The function of the pooling layer is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network, and also to control over-fitting.

Fully Connected Layer: ST-CNN has two fully connected layers with 50 neurons and 2 neurons respectively. Each one takes all neurons in the previous layer and connects it to every single neuron it has. The final layer outputs the class score vector of the binary class label (extreme precipitation cluster vs. non-extreme precipitation cluster).

IV. EXPERIMENTS

A. Data

The dataset we used for experiments is the historical meteorological data collected in the State of Iowa, USA from January 1st, 1948 to December 31^{st}, 2010 [13]. A total of 23,011 samples over 63 years. We chose nine meteorological variables from the dataset (Figure 1, Table I), which are collected at different pressure surfaces and typically used by meteorologists for making forecasts, as meteorological predictor variables. To enhance efficacy, we only chose the samples collected during the rainy season (April to October) every year, which might have correlation with extreme precipitation events. The samples in (1948-2000) are used as training set to learn the prediction model, and the remaining 10 years data are used as test set to evaluate the prediction model.

TABLE I
METEOROLOGICAL VARIABLES.

PW	Precipitable Water
T850	850hPa Temperature
U300	300hPa Zonal Wind
U850	850hPa Zonal Wind
V300	300hPa Meridional Wind
V850	850hPa Meridional Wind
Z300	300hPa Geopotential Height
Z850	850hPa Geopotential Height
Z1000	1000hPa Geopotential Height

B. Spatio-temporal Feature Space Construction

In order to build a feature space with the spatial and temporal information of the meteorological variables,

we create the raw spatio-temporal feature input space for ST-CNN as following steps:

Step 1. Choose 1,024 locations, which are uniformly distributed wrapped around the State Iowa (32 latitudes and 32 longitudes).

Step 2. Sample 9 meteorological variables from 1,024 locations in the same day as the feature cuboid (32 rows and 32 columns) of one day.

Step 3. Repeat Step 2 until the feature cuboids of 10 continuous days are accumulated, then stack these cuboids as a tensor (32 x 32 x 90) which is the experimental sample of the last day in the 10 continuous days.

Step 4. Repeat Step 1-3 until all experimental samples are created.

C. Class Label Creation

Here, we use historical spatial average precipitation data (the mean of daily precipitation totals from 22 observation stations divided by the standard deviation) of the State Iowa from the same time period to create the class labels. We define any 14 days periods as extreme precipitation clusters and label them as a positive sample if the total amount of precipitations of the 14 days reaches a historical high level (i.e., above the 95% percentile of the historical records). Otherwise, we label it as a negative sample.

V. Preliminary Results and Conclusion

Here we compare our model with the streaming feature selection method OSFS [5], and use Accuracy and F-measure for evaluation. Particular, Accuracy ($\frac{TP+FP}{TP+TN+FP+FN}$, where TP is true positive, TN is true negative, FP is false positive, FN is false negative) refers to the closeness of a predicted class label to a known class label. And F-measure ($\frac{2*TP}{2*TP+FP+FN}$) conveys the balance between the exactness and the completeness.

TABLE II
EXPERIMENTS RESULTS.

Event	Metrics	OSFS	CNN
Extreme precipitation	Accuracy	0.712	0.708
Extreme precipitation	F-measure	0.748	0.743

Table II summarizes the performance of our CNN model predicting extreme precipitation clusters. Four learnable layers (two convolutional layers and two pooling layers) were able to produce comparable results. Our next research work will focus on improving the architecture of our model in convolutional and pooling layers and understand how to interpret the features learned by the ST-CNN with physically meaningful patterns among these meteorological variables.

Acknowledgments

We thank the team of Dr. Shafiqul Islam and Dr. David Small from Tufts University for their help on historical meteorological data collection.

References

[1] H. Jr and et al., "US geological survey natural hazards science strategy: Promoting the safety, security, and economic well-being of the nation," *US Geological Survey*, 2013.

[2] Johnson and et al., "Development of a $European$ flood forecasting system," *International Journal of River Basin Management*, pp. 49–59, 2003.

[3] H. Kaixun and et al., "Long-lead term precipitation forecasting by hierarchical clustering-based bayesian structural vector autoregression.," *IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC)*, 2016.

[4] Reyniers and et al., "Quantitative precipitation forecasts based on radar observations: Principles, algorithms and operational systems," *Institut Royal Mtorologique de Belgique*, 2008.

[5] X. Wu and et al., "Online feature selection with streaming features," *IEEE transactions on pattern analysis and machine intelligence*, pp. 1178–1192., 2013.

[6] Y. Di and et al., "Developing machine learning tools for long-lead heavy precipitation prediction with multi-sensor data," *Networking, Sensing and Control*, pp. 63–68, 2015.

[7] Z. Yong and et al., "An evaluation of big data analytics in feature selection for long-lead extreme floods forecasting.," *IEEE 13th International Conference on Networking, Sensing, and Control (ICNSC)*, 2016.

[8] D. Wang and et al., "Towards long-lead forecasting of extreme flood events: a data mining framework for precipitation cluster precursors identification," *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1285–1293, 2013.

[9] K. Yu and et al., "Classification with streaming features: An emerging-pattern mining approach," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2015.

[10] W. Tong and et al., "Text simplification using neural machine translation," *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[11] Anctil and et al., "Impact of the length of observed records on the performance of ANN and of conceptual parsimonious rainfall-runoff forecasting models," *Environmental Modelling Software*, pp. 357–368, 2004.

[12] Xingjian and et al., "Convolutional $LSTM$ network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, pp. 802–810, 2015.

[13] Kalnay and et al., "The ncep/ncar 40-year reanalysis project," *Bulletin of the American meteorological Society*, pp. 437–471, 1996.

6th International Workshop on Climate Informatics
September 22-23, 2016
CI 2016 Hosted by the National Center for Atmospheric Research in Boulder, CO

MULTIPLE INSTANCE LEARNING FOR BURNED AREA MAPPING USING MULTI -TEMPORAL REFLECTANCE DATA

Guruprasad Nayak, Varun Mithal, Vipin Kumar
University of Minnesota

***Abstract*—Mapping burned area on a global scale typically requires the use of a weak signal like Active Fire for training the burned scar classification model. Since these weak signals typically are inaccurate with respect to temporal and spatial pinpointing of the event occurrence, the use of Multiple instance learning paradigm to model the occurrence of the event in a wider spatio-temporal window is demonstrably beneficial than using the exact date of the weak signal. In this work, we demonstrate the use of MIL algorithm to model the temporal uncertainty of the weak signal. We further propose an noise-robust extension to the MIL paradigm for learning on sequence data.**

I. INTRODUCTION

Mapping forest fires is essential for efficient and sustainable land management and for maintaining a healthy forest population on our planet. Remote sensing enables us to use machine learning algorithms to achieve this in an efficient and automated fashion. Given training data for burned and unburned locations, a classifier can be trained that identifies burned regions. However, due to the presence of large heterogeneity in landcovers and seasons across fires in different regions and times, it is prohibitively expensive to gather training data for every kind of fire occurrence. Also, it is much harder to pin point the exact span of time when the fire happened at a location in comparison to determining if the location experienced a fire activity at some point in a wider time window. For instance, while the presence of Active Fire signal at a location can be used as a proxy to assume that the location experienced a fire sometime in that year, the date at which the active fire signal was observed at the location seldom shows presence of a *burn scar* on the ground. The scar typically has a delay of a few time steps after the AF date, the duration of which may vary depending on the region.

Corresponding author: G Nayak, nayak@cs.umn.edu

In the traditional setting of classification, one is provided with features and labels for all training instances and task at hand is to learn a function that maps a given instance to a class. However, in a lot of practical settings, labels are hard to acquire for sufficient number of training instances for the training set to be representative enough of the whole population. One of the ways to handle label scarcity is through the paradigm of Multiple Instance Learning (MIL)[1]. In a MIL setting, the training set includes features for all instances but labels are available only for groups of instances, called *bags*. Typically, the task then is to learn a classifier that can learn to classify a new *bag* into either one of the predefined classes.

In this paper, we propose a method that adapts MIL for sequence data. Given a set of sequences with binary labels for each, we treat each sequence as a *bag*. The task then is to a learn a classifier that distinguishes positive bags from the negative ones. This scenario is different from the traditional Multiple Instance scenario where the instances in the bag are assumed to be independently and identically distributed. In sequence data, there exists a temporal auto-correlation which implies that the occurrence of a meaningful event will manifest itself in contiguous time steps in the sequence. This is the notion that we use to define positiveness of a bag in our MIL algorithm. In other words, a bag is considered positive only if there exists a contiguous span of time steps of at least a certain length that appear positive. If the features spuriously appear positive for a time step but the positiveness doesn't seem to persist, the sequence is not considered positive. The model is cognizant of these random perturbations in the features that cause some time steps to erroneously look positive and these error rates are learned for the data set during the training process. The figure 1 shows examples of such positive and negative sequences.

By casting the fire mapping problem in the MIL

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

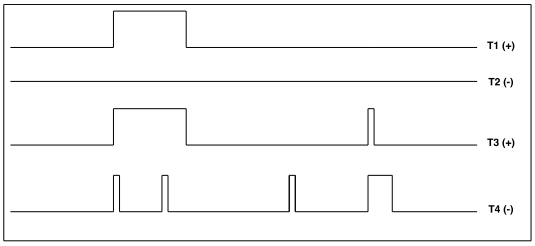


Fig. 1: Caricature explaining the different kinds of positive and negative time series. The first and the third time series are positive while the second and the fourth time series correspond to negative locations.

setting, we need less supervision, since we only need to know about the occurrence of fire in the year and not the exact time when it happened. The other trouble with using multispectral data for classification is that it is plagued with data noise. Especially in the tropics, cloud cover and smoke corrupt a lot of data in fire prone regions. Since the proposed method accounts for data noise, it is expected to help circumvent this problem.

Multiple Instance learning is the task of learning a classifier that can distinguish positive bags of instances from negative ones. Different MIL algorithms differ in the assumptions they make about the relationship between the class label of a bag and the class label of the instances within it[2], [1]. The first work on MIL by Dietterich et. al [3] assumed that a bag is labelled positive if at least one of it's member instances is labeled positive. A bag is labeled negative if all of it's member instances are labeled negative. Algorithms have been proposed that generalize this notion to call a bag positive if at least k of it's member instances are positive [4], [5], [6]. MIL algorithms have also been proposed that treat instances to have some structure instead of assuming them to be i.i.d.[7], [8]. Most popular time series classification techniques like Nearest neighbor classifier using dynamic time warping and shapelets make use of distance metric of some kind defined on sequences.[9], [10].While these techniques perform very well on univariate time series with no noise, but since they use a distance measure (computed from the nearest neighbor or from the shapelet subsequence) to decide the classification, learning complex decision boundaries in multivariate feature space is harder without having lots of training data. The problem is further complicated if the input time series are noisy.

II. METHOD

A. Proposed model

Figure 2 shows a graphical model for the proposed solution. Each *bag*(sequence) I is a sequence of observations $\{x_1, x_2, \ldots, x_T\}$ over T time steps. $E \in \{0, 1\}$ is the class label assigned to the bag. For each time step i, $y_i \in \{0, 1\}$ denotes the instance level class label according to the classifier β and feature x_i. To account for the presence of randomly occurring (not temporally auto-correlated) noise in the features, a parameter α is learned that captures the probability of the class label at a time step truly being what the features at that time step say it should. If $e_i \in \{0, 1\}$ denotes the membership of the time step i in an event i.e it's *true* class label, then $\alpha =< \alpha_1, \alpha_2, \alpha_3, \alpha_4 >$ is defined as,

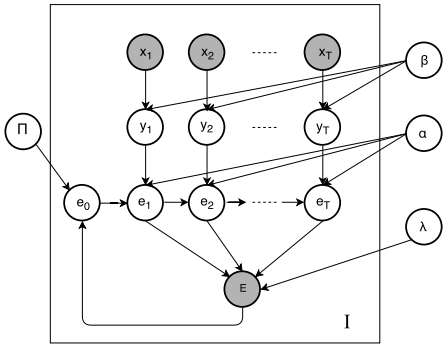


Fig. 2: Proposed graphical model for the noise robust version of MIL for multivariate timeseries data

$$\begin{aligned}\alpha_1 &= \Pr(e_i = 1 | e_{i-1} = 1, y_i = 1)\\ \alpha_2 &= \Pr(e_i = 1 | e_{i-1} = 0, y_i = 1)\\ \alpha_3 &= \Pr(e_i = 1 | e_{i-1} = 1, y_i = 0)\\ \alpha_4 &= \Pr(e_i = 1 | e_{i-1} = 0, y_i = 0)\end{aligned}$$

The parameter λ encodes the minimum number of positive instances in the bag required to label the bag positive. e_i's are assumed to follow a first order Markov dependence and the parameter π encodes the probability of starting state. In this work, $\Pr(y_i | x_i; \beta)$ is assumed to be logistic. Also, to make an exact inference possible, $\Pr(E_I | \mathbf{e}; \lambda)$ is assumed to be decomposable over instances in the bag I. Specifically, in this work, we assume,

$$\begin{aligned}\Pr(E_I = 1 | \mathbf{e}; \lambda) &= e^{c_i - \lambda} && c_i < \lambda,\\ &= 1 && otherwise\end{aligned}$$

where $c_i = \sum_i e_i$ is the count of positive instances in the bag.

B. Training the model

We propose an Expectation-Maximization (EM) solution for learning the parameters $\theta = \{\beta, \alpha, \lambda, \pi\}$ in the above graphical model. The update for the mth iteration of EM will be done as

$$\arg\max_{\theta} \sum_{I=1}^{N} \sum_{\mathbf{e},\mathbf{y}} Q(\mathbf{e}, \mathbf{y}) \log\left[\Pr(\mathbf{x_I}, E_I, \mathbf{e}, \mathbf{y}; \theta)\right]$$

where $Q(\mathbf{e}, \mathbf{y})$ depends on the parameter values in the $m-1$th iteration and is defined as,

$$\Pr(\mathbf{e}, \mathbf{y} | \mathbf{x_I}, E_I; \theta_{\mathbf{m-1}})$$

Using the independence assumptions defined in the graphical model, the joint probability $\Pr(\mathbf{x_I}, E_I, \mathbf{e}, \mathbf{y}; \alpha, \beta, \lambda)$ can be factorized as

$$\Pr(E_I|\mathbf{e};\lambda)\Pr(e_0|E_I;ps)\Pr(\mathbf{x_I}) \times \prod_{i=1}^{T}\Pr(y_i|x_i;\beta)\prod_{i=1}^{T}\Pr(e_i|e_{i-1},y_i;\alpha)$$

Thus, the optimal parameters at the mth iteration are chosen as,

$$\alpha_m = \arg\max_{\alpha} \sum_{I=1}^{N}\sum_{\mathbf{e},\mathbf{y}} Q(\mathbf{e},\mathbf{y}) \sum_{i=1}^{T} \log \Pr(e_i|e_{i-1},y_i;\alpha)$$

$$\beta_m = \arg\max_{\beta} \sum_{I=1}^{N}\sum_{\mathbf{e},\mathbf{y}} Q(\mathbf{e},\mathbf{y}) \sum_{i=1}^{T} \log \Pr(y_i|x_i;\beta)$$

$$\lambda_m = \arg\max_{\lambda} \sum_{I=1}^{N}\sum_{\mathbf{e},\mathbf{y}} Q(\mathbf{e},\mathbf{y}) \log \Pr(E_I|\mathbf{e};\lambda)$$

$$\pi_m = \arg\max_{\pi} \sum_{I=1}^{N}\sum_{\mathbf{e},\mathbf{y}} Q(\mathbf{e},\mathbf{y}) \log \Pr(e_0|E_I;\pi)$$

It is possible to solve exactly for each one of the above optimizations using dynamic programming algorithms. Each update iteration in the EM algorithm takes $O(NT^2)$ time, where N is the number of training sequences and T is the length of each.

III. Experiments

Active fire is known to have a lot of errors, with recall dropping below acceptable levels in a lot of regions, especially the tropics. However, the presence of Active fire at a location at some time typically indicates the occurrence of a burn activity in a wide spatial and temporal window around it. This aspect of supervision via Active Fire is very naturally captured in the MIL paradigm. In [11], we used a 3-stage classification procedure[12] to generate historical burned scar maps for the tropical regions of Amazon and Indonesia. The produced maps have more than 3 times the amount of burned area detection with equally good precision as the state-of-the-art NASA product. The procedure uses Active Fire and multispectral MODIS data to train the classifier. We used the original formulation of MIL with a logistic learning as base classifier to learn burn scar signatures in the first stage. The scars were then refined with and spatially enhanced with active fire proximity in subsequent stages to produce the final maps. The final classification maps can be viewed online at http://z.umn.edu/fireviewersept.

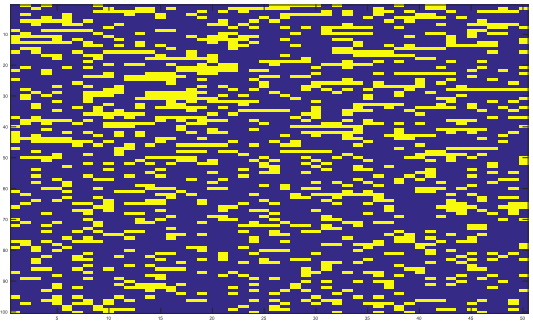


Fig. 3: Instance level labels for the synthetic dataset. Each row represents a time series i.e a bag. Yellow columns in a row represent positive instances in the bag and blue columns represent negative instances.

We demonstrate the proposed extension to the MIL paradigm for sequence data on a synthetic data set. Every sequence in the data, both positive and negative has a 20% of the time steps randomly marked positive, i.e the feature values at these time steps will be taken from the positive distribution. However, each of the positive sequences has a continuous stretch of 5 time steps with a positive signature. The rest of the time steps in both kinds of sequences are taken from the negative distribution. Figure 3 below shows the kind of instances at each time step of every sequence. The two kinds of instances are drawn from two different Gaussian distributions. Each row is a sequence and the top half of the sequences are positive. The positive time steps in each sequence are marked yellow while negative ones are marked blue. A MIL algorithm was trained using the original formulation of Deittrich (MILR) and with the proposed algorithm (MIS) that incorporates event continuity. Logistic regression was used as the base classifier for each algorithm. The algorithms were trained on 200 sequences and were tested separately on 1000 sequences. MILR failed to learn the signature of the positive event and marked every sequence as positive. MIS, on the other hand was able to distinguish the two kinds of sequences with an accuracy of 87.3%.

Acknowledgments

This research was supported in part by the National Science Foundation under Grants IIS-1029711 and IIS-0905581, and NASA grant NNX12AP37G. Access to computing facilities was provided by the University of Minnesota Supercomputing Institute.

6th International Workshop on Climate Informatics
September 22-23, 2016
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] J. Amores, "Multiple instance classification: Review, taxonomy and comparative study," *Artificial Intelligence*, vol. 201, pp. 81–105, 2013.

[2] J. Foulds and E. Frank, "A review of multi-instance learning assumptions," *The Knowledge Engineering Review*, vol. 25, no. 01, pp. 1–25, 2010.

[3] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[4] H. Hajimirsadeghi, J. Li, G. Mori, M. Zaki, and T. Sayed, "Multiple instance learning by discriminative training of markov networks," *arXiv preprint arXiv:1309.6833*, 2013.

[5] L. Duan, W. Li, I. W.-H. Tsang, and D. Xu, "Improving web image search by bag-based reranking," *IEEE Transactions on Image Processing*, vol. 20, no. 11, pp. 3280–3290, 2011.

[6] W. Li, L. Duan, D. Xu, and I. W.-H. Tsang, "Text-based image retrieval using progressive multi-instance learning," in *2011 International Conference on Computer Vision*, pp. 2049–2055, IEEE, 2011.

[7] D. Kotzias, M. Denil, N. De Freitas, and P. Smyth, "From group to individual labels using deep features," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 597–606, ACM, 2015.

[8] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-iid samples," in *Proceedings of the 26th annual international conference on machine learning*, pp. 1249–1256, ACM, 2009.

[9] L. Ye and E. Keogh, "Time series shapelets: a new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 947–956, ACM, 2009.

[10] X. Xi, E. Keogh, C. Shelton, L. Wei, and C. A. Ratanamahatana, "Fast time series classification using numerosity reduction," in *Proceedings of the 23rd international conference on Machine learning*, pp. 1033–1040, ACM, 2006.

[11] V. Mithal, G. Nayak, A. Khandelwal, V. Kumar, N. Oza, and R. Nemani, "Mapping burned areas in tropical forests using modis data," under submission, University of Minnesota, 2016.

[12] V. Mithal, G. Nayak, A. Khandelwal, V. Kumar, N. Oza, and R. Nemani, "RAPT: Rare class prediction in absence of true labels," under submission, University of Minnesota, 2016.