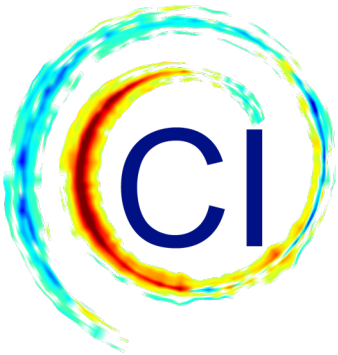


Proceedings of the 7th International Workshop on **Climate Informatics: CI 2017**

Volume editors
Vyacheslav Lyubchich
Nikunj C. Oza
Andrew Rhines
Eniko Szekely

Series editors
Imme Ebert-Uphoff
Claire Monteleoni
Doug Nychka

NCAR Technical Notes
NCAR/TN536+PROC

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

NCAR IS SPONSORED BY THE NSF

How to Cite this Document

V. Lyubchich, N.C. Oza, A. Rhines, E. Szekely (Eds.), I. Ebert-Uphoff, C. Monteleoni, D. Nychka (Series Eds.), Proceedings of the 7th International Workshop on Climate Informatics: CI 2017. NCAR Technical Note NCAR/TN-536+PROC, Sept 2017, doi: 10.5065/D6222SH7.

The ISBN number (optional use) for this document is 978-0-9973548-2-9.

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.

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

To be added to the workshop mailing list, please send an email to climate.informatics.workshop@gmail.com.

7th International Workshop on Climate Informatics
CI 2017

Table of Contents

Peer-Reviewed Papers

Foreword by the CI 2017 Workshop Chairs

Climate informatics continues to develop as an area of collaboration for statisticians, climate and computer scientists. With trendy names, such as “data science”, “deep learning”, and “big data analytics”, coming and going, the climate informatics scientists do their best in unraveling the history of the Earth’s climate and making meaningful predictions about the future. Moreover, we start seeing a number of research projects where local climate features, such as Indian monsoons or California rains, are linked to global climate, and where implications of climate change for local agriculture, insurance, and critical infrastructure are assessed in great details.

It has been the seventh International Workshop on Climate Informatics and we believe it had much success in accelerating discovery at the intersection of these disciplines. For the 2017 workshop, participants convened at the National Center for Atmospheric Research (NCAR) in Boulder, Colorado between September 20-22.

The first day of the workshop was an optional Hackathon led by David John Gagne II. The Hackathon used the Rapid Analytics and Model Prototyping (RAMP) platform created by our continued collaborator Balázs Kégl and colleagues at the Paris-Saclay Center of Data Science. The RAMP platform was used to host a team-based prediction challenge, wherein groups of attendees were tasked with predicting extreme precipitation events in California several months in advance. The Hackathon also served to give attendees more opportunities to interact and discuss their research, methods, and potential collaborations. A summary of the hackathon is included in these proceedings.

The main workshop (September 21-22) featured five invited speakers, a poster session, short talks by several early career scientists, and two panel discussions. Invited speakers covered many topics from across the spectrum of climate informatics: Alexis Hannart (Ouranos) explained how causal counterfactual theory can be used to attribute events in the climate system; Robert Lund (Clemson University) spoke about statistical challenges posed by analyzing trends in real-world observations; Elisabeth Moyer (University of Chicago) explained how climate models can be most efficiently used to augment limited observations; Prabhat (Lawrence Berkeley National Laboratory) presented recent advances in using deep learning at scale to identify extreme events such as hurricanes and atmospheric rivers; and Sai Ravela (MIT) described how physical constraints can be used to improve machine learning in the context of data assimilation systems.

We would like to thank all participants for submitting papers on a wide range of interesting topics. Following reviews by the program committee, 27 short papers were selected for presentation at the workshop and are published in these proceedings. Of these, authors of six outstanding papers were also invited to give short oral presentations of their research: Savini Samarasinghe and Marie McGraw, Robert Wills, Seongchan Kim, Matthias Demuzere, Zachary Butler, and Eniko Szekely. Thanks to generous funding from the National Science Foundation, NCAR, and the Elsevier journal Artificial Intelligence, 14 early career authors were provided with travel grants to attend the workshop.

We would also like to thank the many people on the organizing committee and at NCAR, whose hard work was so crucial for the workshop's success. First and foremost, we would like to thank the steering committee for years of leadership and funding efforts that have ensured continuity

of the workshop: Doug Nychka, Imme Ebert-Uphoff, and Claire Monteleoni. The chairs of the program committee, Nikunj Oza and Eniko Szekely, did a fantastic job of not only orchestrating the paper reviews, but also countless other tasks that helped to make the workshop run smoothly. The members of the program committee (Wei Ding, Yulia Gel, Mohammad Gorji Sefidmazgi, Sara Graves, Vipin Kumar, Stefan Liess, Nikunj Oza, Brian Smoliak, Eniko Szekely, Pierre Tandeo, and Martin Tingley) provided thorough and timely reviews of the submitted papers, and we thank them for volunteering so much of their time to do so. We are indebted to David John Gagne II and Balázs Kégl for designing and running the Hackathon, which we think can serve as a model for future events. We are also grateful for the work of Mohammad Gorji Sefidmazgi, the Budget and Travel Chair, and Erich Seamon, the Publicity Chair. Last (but certainly not least), Michelle Patton and Cecilia Banner from NCAR provided an enormous amount of support with planning, facilities, logistics, transportation, website updates, and coffee. Their efforts, along with long-term support and guidance from Doug Nychka, have been in large part responsible for the success of the workshop.

Finally, we would like to thank our sponsors: The National Science Foundation of the United States (NSF), the National Center for Atmospheric Research (NCAR), the Artificial Intelligence Journal, the Research Network for Statistical Methods for Atmospheric and Oceanic Sciences (STATMOS), and the NVIDIA corporation, for their financial support — not only this year, but also at previous workshops. We hope the workshop will continue to serve as the leading venue for interdisciplinary research in climate informatics and we look forward to your future participation and support.

Andy Rhines
Vyacheslav Lyubchich
CI2017 Workshop Co-Chairs

Organizing Committee

Workshop Co-Chairs:
Andrew Rhines, University of Washington
Slava Lyubchich, University of Maryland Center for Environmental Science

Program Committee Co-Chairs:
Nikunj C. Oza, NASA
Eniko Szekely, New York University

Hackathon Chair
David John Gagne II, NCAR

Publicity and Publications Chair:
Erich Seamon, University of Idaho

Travel and Budget Chair:
Mohammad Gorji, Syntelli Solutions Inc.

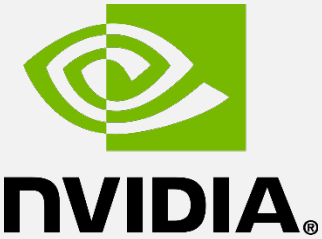
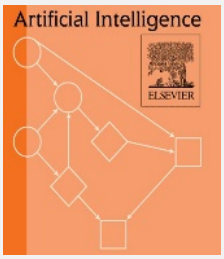
Steering Committee:
Imme Ebert-Uphoff, Colorado State University
Claire Monteleoni, George Washington University
Doug Nychka, National Center for Atmospheric Research

Program Committee Members:
Wei Ding, University of Massachusetts Boston
Yulia Gel, University of Texas Dallas
Mohammad Gorji, Syntelli Solutions Inc.
Sara Graves, University of Alabama
Vipin Kumar, University of Minnesota
Stefan Liess, University of Minnesota
Nikunj Oza, NASA
Brian Smoliak, WindLogics
Eniko Szekely, New York University
Pierre Tandeo, IMT-Atlantique
Martin Tingley, Netflix

Local Administrative Support:
Michelle Patton, NCAR
Cecilia Banner, NCAR

Acknowledgements of Sponsors

We gratefully acknowledge the generous contributions of the following sponsors who have helped make CI 2017 possible:

The Research Network for Statistical Methods for Atmospheric and Oceanic Sciences	
The NVIDIA Corporation	
Artificial Intelligence Journal - A division of IJCAI	
The National Science Foundation	
The National Center for Atmospheric Research	

Workshop Location

The workshop was held again at the Mesa Laboratory of 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 every year since 2012.

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.

NCAR Mesa Laboratory in Boulder, CO

Workshop Participants

Many of the Workshop Participants (for a complete list, see next page)

Photo credit: Brian Bevirt.

List of Registered Workshop Participants

Attendee	Affiliation
Ayet, Alex	Ecole Normale Supérieure
Barnes, Elizabeth	Colorado State University
Berdahl, Mira	LANL
Butler, Zachary	UCI
Chen, Chen	University of Chicago
Dash, Yajnaseni	Indian Institute of Technology Delhi
Demuzere, Matthias	Ghent University
Ebert-Uphoff, Imme	Colorado State University
Fernandez, Arturo	UC Berkeley
Fitchett, Stephanie	Open to Opportunities
Ford, Bruce	Clear Science, Inc.
Gagne, David	National Center for Atmospheric Research
Gel, Yulia	University of Texas, Dallas
Hall, David	Univ of CO Boulder, Dept. Computer Science
Hammerling, Dorit	NCAR
Hannart, Alexis	Ouranos
Hickey, Jason	Google
Hong, Seungkyun	KISTI
Hubey, Mark	Montclair State University
Kashinath, Karthik	Lawrence Berkeley National Lab
Kim, Jae Youp	KISTI
Kim, Seongchan	KISTI
Kim, Sookyung	Lawrence Livermore National Laboratory
Kumar, Vipin	University of Minnesota
Li, Yi	Northeastern University
Liu, Ke	JPL, California Institute of Technology
Loft, Rich	NCAR
Lund, Robert	Clemson University
Lyubchich, Vyacheslav	UMCES
Matsuoka, Daisuke	JAMSTEC
McGinnis, Seth	NCAR

Attendee	Affiliation
McGraw, Marie	Atmospheric Science, Colorado State University
Mitra, Adway	ICTS-TIFR
Monteleoni, Claire	George Washington University
Mouatadid, Soukayna	University of Toronto
Moyer, Elizabeth	University of Chicago
Nadiga, Balu	Los Alamos National Lab
Nychka, Doug	NCAR
OConnor, Alison	Charles River Analytics
Oza, Nikunj	NASA Ames Research Center
Posey, Stan	NVIDIA
Prabhat	Lawrence Berkeley National Laboratory
Ravela, Sai	MIT
Ramea, Kalai	Palo Alto Research Center, Inc.
Rhines, Andy	University of Washington
Rupe, Adam	UC Davis
Sain, Steve	Jupiter
Samarasinghe, Savini	Colorado State University
Schmude, Johannes	IBM
Seo, Sungyong	University of Southern California
Song, Sa-Kwang	KISTI
Staib, Matthew	MIT
Szekely, Eniko	New York University (now EPFL Switzerland)
Tierney, Craig	NVIDIA
Tingley, Martin	Netflix
Trailovic, Lidia	CIRES
Urban, Nathan	Los Alamos National Laboratory
Vandal, Thomas	Northeastern University
Weatherhead, Elizabeth	U. Colorado at Boulder
Weitzel, Nils	University of Bonn
Wills, Robert	University of Washington
Wu, Wei	NCAR / University of Wyoming
Yun, Sooin	Univ of IL at Urbana-Champaign, Statistics Dept
Zhang, Yawen	University of Colorado Boulder

HACKATHON INFORMATION

Hackathon Agenda, Wednesday, September 20th, 2017

9:30	Bus pickup at Marriott Courtyard
10:00 – 10:30	Arrival at NCAR Mesa Lab: registration, coffee and introduction
10:30 – 12:30	Session 1
12:30 – 1:30	Lunch (provided) 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	Group presentations and closing
6:00	Bus departs NCAR for Marriott Courtyard

Shown here: Majority of Hackathon Participants

Photo credit: Brian Bevirt.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

THE 2017 CLIMATE INFORMATICS HACKATHON

David John Gagne II[1]

***Abstract*—The 2017 Climate Informatics Hackathon focused on predicting extreme winter precipitation in northern California at seasonal time scales. The contest dataset, methods, participants, successes, and challenges are discussed.**

I. MOTIVATION

California receives most of its annual rainfall during the winter when storms fueled by moisture from the tropical Pacific impact the state. This past winter was the wettest on record for northern California, resulting in massive floods and over 1 billion dollars in damage. Storm runoff into Lake Oroville led to extensive releases of water along the Oroville Dam spillway, which was damaged in the process. Some of the flooding associated with dams in northern California could be managed better with more accurate seasonal and subseasonal forecasts of rainfall. If water managers had a skilled forecast of expected rainfall, then they could change the distribution of water in northern California to be more resilient to large rainfall events. The mitigation process can take weeks to complete, so seasonal forecasts are needed for effective mitigation. Current operational seasonal precipitation guidance from the NOAA Climate Prediction Center has no skill above climatology for northern California and is not presented in a way that is useful for water managers at the California Department of Water Resources. Current seasonal precipitation forecasting relies primarily on teleconnection indices, such as ENSO. However, these indices individually are poorly correlated with northern California winter rainfall. Other teleconnections should also have some correlation with California rainfall, but finding the most important connections and how they interact is not a task that can be easily done manually.

The goal for the 2017 Climate Informatics Hackathon is to use the November-averaged atmospheric fields to predict the probability of at least 750 mm of rain in northern California between December and February. The observational record for northern California rainfall only goes back to the early 1920s, which would provide a very limited sample size for machine learning or statistical models. Therefore, we are going to use climate model output from simulations run over the last 1000 years. By using climate model output, we hope to sample better the range of possible combinations of weather patterns and rainfall and fit more complex ML and statistical models.

Corresponding author: David John Gagne II, dgagne@ucar.edu
[1]National Center for Atmospheric Research, Boulder, Colorado

II. DATA AND METHODS

The Hackathon focused on climate model output from the NCAR Community Earth System Model (CESM) Last Millennium Ensemble (LME) [1]. Each LME member was run from 850 AD to 2005, and the atmospheric model used 2° grid spacing. The LME consists of 12 members with full climate forcing while other members only include forcing from one major climate source, such as volcanic eruptions, solar variability, land use, greenhouse gases, orbital changes, and ozone-aerosols. The full forcing members all used the same physics and forcing but had their initial atmospheric temperature fields randomly perturbed by 10^{-14} K to examine the natural variability within the model. Since each full forcing member represents an independent path through the same climate, the full forcing members were grouped into training (4 members), public testing (4 members), and private testing (3 members) sets. Participants were given the following November-averaged global spatial fields as inputs for their machine learning models: mean sea level pressure (PSL), surface temperature (TS), precipitable water (TMQ), 500 mb zonal wind (U_500), and 500 mb meridional wind (V_500).

The Hackathon data, starting kit, and leaderboard were hosted through the Rapid Analytics and Model Prototyping (RAMP) website, developed by the Paris-Saclay Center of Data Science. Unlike the machine learning contest site Kaggle, in which contest participants submit predictions from locally trained models, the RAMP site requires participants to submit Python code describing their feature extraction and machine learning model processes. The submitted code is then run on the RAMP web server to train and evaluate

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

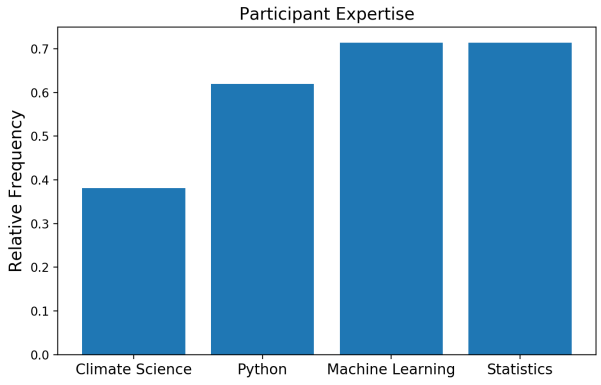


Fig. 1. Relative frequency of participant expertise in areas relevant to the Hackathon.

each participant's machine learning model. After a specified closed period, the source code for each team's submissions are then made visible to all participants for copying and modification by other teams. This system encourages participants to build models that can run in a reasonable amount of time, discourages cheating, and increases the amount of cooperation among different teams.

The submissions from each team are ranked based on the Brier Skill Score, which is the ratio of the submission Brier Score to the Brier Score of predicting the test set mean probability (climatology) for each event. Participants were provided code for a baseline method that performed spatial principal component analysis (PCA) on the PSL, TS, and TMQ anomaly fields. Then the top 5 principal components were input into a logistic regression with a LASSO penalty to encourage sparse coefficients.

III. Participants

The Hackathon attracted 21 participants from a diverse array of backgrounds. Out of all the participants, 12 came from universities, 5 from government or national labs, and 3 from the private sector. A survey of the participants revealed that 70% had expertise in machine learning or statistics, over 60% had expertise in Python, and less than 40% were experts in Python. All but one of the climate science participants had experience with Python or with machine learning and statistics. Based on their self-reported expertise, the participants were divided into 6 teams of 3 to 4 people so that each area of expertise would be covered by at least one team member.

The scores of the top submissions from each team are shown in Table I. The top team improved on the baseline Brier Skill Score by 15% but also had a lower AUC than the baseline. The top team used LASSO logistic regression to select a sparse set of spatial points and then fed them into a gradient boosted classifier. Other successful methods included removing the southern hemisphere mid and upper latitudes from the PCA procedure and adding wind variables. Other teams submitted more complex methods, including neural networks and decision tree ensembles, but most of these methods overfit to noise in the data.

TABLE I
THE TOP SUBMISSION FROM EACH TEAM RANKED BY BRIER SKILL SCORE.

Team	Brier Skill Score	AUC	Contributivity
4	0.150	0.788	40
6	0.146	0.798	17
2	0.139	0.801	7
5	0.137	0.798	11
7	0.133	0.794	0
3	0.133	0.794	0
Base	0.130	0.792	0

IV. Discussion

The Hackathon featured both many areas of success and areas for improvement. The combination of Jupyter Notebooks, Anaconda Python environments, and the RAMP system helped participants quickly start working on the problem with less time spent troubleshooting technical issues. Randomly assigned teams with diverse skill sets led to more collaboration and a sustained enthusiasm for the problem throughout the day. Group presentations at the end helped to showcase how everyone contributed to the task and the wide array of approaches that were tried. There was some confusion among participants about the Brier Skill Score and not enough explanation in the starting kit about what it measures. While some effort was made by the organizer to encourage the interpretation of the machine learning models, none of the teams discovered any new physical insights from their machine learning models. In order to encourage more model interpretation in future hackathons, more interpretation code examples may be needed, and more participants with a background in climate science should be encouraged to attend. The Hackathon site will continue to remain open and accept submissions. Please visit https://ramp.studio/problems/california_rainfall to download the starting kit, data, and create submissions.

Acknowledgments

Special thanks go to Balazs Kegl and the RAMP team for hosting the Hackathon RAMP site and provid-

ing support and feedback during the development and Hackathon process. Jeanine Jones inspired the contest problem with a talk at the Big Data and the Earth Sciences: Grand Challenges Workshop, and Sloan Coats assisted with suggesting and acquiring the CESM LME data. Andy Rhines helped with reviewing and testing the hackathon code.

REFERENCES

[1] B. L. Otto-Bliesner, E. C. Brady, J. Fasullo, A. Jahn, L. Landrum, S. Stevenson, N. Rosenbloom, A. Mai, and G. Strand, "Climate variability and change since 850 ce: An ensemble approach with the community earth system model," *Bulletin of the American Meteorological Society*, vol. 97, no. 5, pp. 735–754, 2016. [Online]. Available: https://doi.org/10.1175/BAMS-D-14-00233.1

WORKSHOP AGENDA

CI Workshop, Thursday, September 21, 2017

7:45 GreenRide shuttle 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. Alexis Hannart:** *Methods for Attributing Climate Trends and Events: Overview and Challenges*

10:00 – 10:30 Coffee break

10:30 – 11:30 **Spotlight presentations** (15 minutes each)

1) Savini Samarasinghe and Marie McGraw: *A study of causal links between the Arctic and the midlatitude jet-streams*
2) Robert Wills: *Extracting Modes of Variability and Change from Climate Model Ensembles*
3) Seongchan Kim: *DeepRain: ConvLSTM Network for Precipitation Prediction using Multichannel Radar Data*
4) Matthias Demuzere: *Sensitivity of global ecosystems to climate anomalies in observations and earth system models*

11:30 – 11:40 Group photo

11:40 – 1:00 Lunch (cafeteria serves food from 11:30 a.m. - 1:30 p.m., cash only)

1:00 – 2:00 **Invited talk. Sai Ravela:** *What can Systems Science teach Machine Learning?*

2:00 – 3:00 **Poster highlights** (2 minutes each)

3:00 – 3:30 Coffee break

3:30 – 4:30 **Panel discussion**

4:30 – 6:30 **Poster session and reception**, Mesa Lab Cafeteria

6:40 GreenRide shuttle departs NCAR, returns to Marriott Courtyard

CI Workshop, Friday, September 22, 2017

7:45 GreenRide pickup at Marriott Courtyard to NCAR

8:15 – 8:30 Continental breakfast

8:30 – 9:30 **Invited talk. Robert Lund**: *Climate Data Homogenization: Informatics and the Changepoint Problem*

9:30 – 10:00 Coffee break

10:00 – 11:00 **Invited talk. Mr. Prabhat**: *Deep Learning for Extreme Weather Detection*

11:00 – 11:30 **Spotlight presentations** (15 minutes each)

1) Zachary Butler: *Fire Event Prediction for Improved Regional Smoke Forecasting*
2) Eniko Szekely: *Pattern extraction in dynamical systems using information geometry: application to tropical intraseasonal oscillations*

11:30 – 11:45 **Sponsors' session. Craig Tierney**: *Applications of Deep Learning in Climate Science and Data Analysis*

11:45 – 1:00 Lunch and posters (cafeteria serves food from 11:30 a.m. - 1:30 p.m., cash only)

1:00 – 2:00 **Invited talk. Liz Moyer:** *Insights from Model Emulation for Climate Research*

2:00 – 3:00 **Panel discussion**

3:00 – 3:15 **Concluding remarks**

3:15 – 3:30 Coffee break

3:30 – 3:55 **Hackathon Presentation**

3:55 – 5:15 **Community-building via hiking around NCAR**

5:30 GreenRide departs NCAR, returns to Marriott Courtyard

Invited Talks

Alexis Hannart, Ouranos, Montreal, CA.

Title: Methods for Attributing Climate Trends and Events: Overview and Challenges

Abstract: Investigating causal links between climate forcings, whether natural or anthropogenic, and observed responses – ranging from the large scale evolution of climate over the instrumental era, to occurrences of local extreme weather events, as well as the socioeconomic impacts of both – represent a significant research effort in the climate sciences. Studies addressing these questions combine observations, physical insights, and climate model simulations in order to evidence such causal links, under a probabilistic setting able to handle the many uncertainties at play. This talk will attempt to give an overview of the various statistical methods that were designed to perform this task, as well as their recent and ongoing evolution, with an emphasis on the potentialities of data science in this field.

Sai Ravela, Center for Global Change Science, MIT, Cambridge, MA.

Title: What can Systems Science teach Machine Learning?

Abstract: Practitioners in the earth sciences will easily agree that it takes data and physics (models) to deliver predictive skill. The use of physical constraints within statistical procedures and the augmentation of physical models with statistical ones are, for example, well established practices. Data Science /ML has made rapid advances in many fields and garnered enormous interest. It offers new pathways in data model symbiosis for prediction and uncertainty quantification.

In this talk, I will discuss some learning related research and teaching efforts in EAPS at MIT, and then present two applications of the symbiosis. In one direction, I will show using typical Geophysical inference problems that the rush to nonparametric statistics must be tempered in the presence of Model error. I then argue that there is hope that this situation can be handled through learning, presenting a particular case using ensemble learning and manifold learning. In the other direction, I will use an example of neural reduced models for uncertainty quantification to argue a dynamic data driven application systems paradigm to learning is essential. I will argue that the current variational approaches are ineffective for learning to predict uncertainty, generative/sampling approaches also being of relatively poor efficacy. I will argue that even from a data assimilation perspective, neural learning seems dated.

I then propose a tractable variational information theoretic learning approach to train for uncertainty that enables quantification of posterior parameter uncertainties, demands minimal sampling, allows training under non Gaussian error distributions, facilitates adaptive sampling for learning and facilitates exploration of sparsity. I will show that the same approach is also useful for the more general data assimilation problem.

Robert Lund, Mathematical Sciences, Clemson University, Clemson, SC.

Title: Climate Data Homogenization: Informatics and the Changepoint Problem

Abstract: This talk overviews climate homogenization issues, presenting recent advances in statistical homogenization techniques. First, the need to homogenize climate data is justified. Attention is thereafter focused on how to estimate the number of changepoints and their locations in time-ordered data sequences. A penalized likelihood objective function is developed for the task from minimum description length (MDL) information theory principles. Optimizing the objective function yields estimates of the changepoint number(s) and location time(s). The MDL penalty depends on where the changepoint(s) lie, but not solely on the total number of changepoints (such as classical AIC and BIC penalties). Specifically, configurations with changepoints that occur relatively closely to one and other are penalized more heavily than sparsely arranged changepoints. The techniques allow for autocorrelation in the observations and mean shifts at each changepoint time. The fundamental methods are modified to handle series with trends, seasonal features, and scenarios where a ``metadata" record exists documenting some, but not necessarily all, of station move times and instrumentation changes. Applications to climate time series are presented throughout and computational and informational issues are espoused upon.

Mr. Prabhat, Data & Analytics Services, Lawrence Berkeley National Laboratory, Berkeley, CA.

Title: Deep Learning for Extreme Weather Detection

Abstract: Deep Learning has revolutionized the fields of computer vision, speech recognition, robotics and control systems. Can Deep Learning be applied to solve pattern detection problems in climate science?

This talk will present our efforts in applying Deep Learning for detecting and localizing extreme weather events (e.g. tropical cyclones, extra-tropical cyclones, atmospheric rivers, fronts) in simulation and observational datasets. We have successfully developed supervised convolutional architectures for the binary classification tasks of detecting weather patterns in centered, cropped patches. We have subsequently extended our architecture to a semi-supervised formulation, which is capable of learning a unified representation of multiple weather patterns, predicting bounding boxes and object categories, and has the capability to detect novel patterns (w/ few, or no labels). We will briefly present our efforts in scaling the semi-supervised architecture to 9600 nodes of the Cori supercomputer, obtaining 15 PF performance. The talk will conclude with a list of open challenges in Deep Learning, and speculations about the role of AI in climate science.

Liz Moyer, Atmospheric Science, University of Chicago, Chicago, IL.

Title: Insights from Model Emulation for Climate Research

Abstract: not available

Sponsors' session.

Craig Tierney, NVIDIA

Title: Applications of Deep Learning in Climate Science and Data Analysis

Abstract: Data science has been evolving quickly with the application of Deep Learning to the big data challenges we see today. By using deep learning with NVIDIA GPUs, new insights in climate informatics can be realized. In this talk we will provide a overview of NVIDIA technology and tools how they have been successfully applied to challenges in analyses of climate data.

For questions, regarding the above topic, contact Craig Tierney (ctierney@nvidia.com).

For questions regarding free access for faculty to NVIDIA's GPU clusters, contact Stan Posey (sposey@nvidia.com) or go to http://www.nvidia.com/object/io-128392.html.

7th International Workshop on Climate Informatics
CI 2017

Peer-Reviewed Papers

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

SPATIO-TEMPORALLY CONSISTENT SIMULATION OF DAILY RAINFALL OVER INDIA

Adway Mitra[1]

***Abstract*—Simulation of rainfall over a region for long time-sequences should be able to preserve the known spatial and temporal characterestics to be of practical use. Rainfall over India is very heterogeneous, making its simulation a big challenge. General Circulation Models (GCMs) are unable to do so and various rainfall generators using stochastic processes are also difficult to apply. In this work, we consider two Bayesian models based on conditional distributions of latent variables that describe weather conditions at specific locations and over the whole country. During model parameter estimation from observed data, we use spatio-temporal smoothing using Markov Random Field. Also, we use a nonparametric spatial clustering based on Chinese Resturant Process to identify homogeneous regions, which are utilized by one of our proposed models. We compare the simulations by these models with daily rainfall over India in 2000-2014, and evaluate their spatio-temporal properties.**

I. INTRODUCTION

The impact of rainfall is enormous in certain parts of the world such as India. Future rainfall projections are needed for impact assessment and feasibility studies of any projects. Process models, like biophysical crop models and hydrological models for reservoirs require future rainfall data as input, which can be provided by rainfall simulations. However, such simulations must be accurate, and preserve as many of the characteristics of the real rainfall data as possible. Climate models of varying levels of complexity have been developed to simulate meteorological variables worldwide. A class of such models called General Circulation Models (GCMs) are quite popular, and they provide simulations of rainfall over India, conditioned on simulated climatic conditions all over the world. Some of them have been found to be reasonably accurate in preserving certain properties of Indian Monsoon rainfall, such as inter-annual and intra-seasonal variability [3]. However, most of the models are unable to capture spatio-temporal properties of Indian rainfall.

Corresponding author: Adway Mitra, adway.mitra@icts.res.in
[1]ICTS-TIFR, Bangalore, India

Another approach is Stochastic Rainfall Generators. Introduced by C.W. Richardson [6], they model rainfall occurrence, rainfall volume and sometimes other climatic variables like temperature using conditional probability distributions (as in a Bayesian Network), conditioned on rainfall occurrence. Most of these stochastic simulators use a training dataset to fit various parameters of these distributions, and then long temporal sequences of meteorological variables are simulated by sampling repeatedly from these distributions. Various statistics of interest are computed from this simulated data, and they are compared with the corresponding statistics from the observed data. This is the general approach prescribed by the Intergovernmental Panel on Climate Change (IPCC) [16].

Most of the stochastic rainfall generators simulate daily rainfall occurrence (binary) and rainfall volume (real-valued) separately using a latent variable, such as in [8]. Temporal coherence of the rainfall occurrence variable is maintained using Markovian or Semi-Markovian([17]) approach. Originally location-specific point processes were studied [7], then multi-site processes were introduced [9] to capture spatial correlations. Most recent stochastic simulators like [5], [18], [19] achieve spatial correlations by using Gaussian Processes. But they need to choose suitable covariance functions, which is often difficult. A concise but comprehensive survey on stochastic daily rainfall generators is available in [1]. These stochastic rainfall simulations have been used in Argentina [18], Sweden [19], USA [17], and various countries in Africa [11], [12]. However, not too much work has been done for India, except some attempts like [10]. One reason for that is Indian rainfall is spatio-temporally very heterogeneous. A detailed study of these variabilities is presented in [2].

In this work, we aim to build stochastic simulation models for daily Indian monsoon rainfall, based on latent variables, conditional distributions, and coherent zones within the landmass. We propose two stochastic rainfall generators and study how accurately they can reproduce spatio-temporal properties of Indian rainfall.

II. VARIABLES AND PARAMETERS

Suppose there are S locations, and the total number of days is T. Any location s has a set of neighboring locations $NB(s)$, according to the grid coordinates. Only locations lying on Indian geo-political landmass are considered. At each location s and day t, $X(s,t)$ denotes the volume of rainfall received, while $Y(t)$ denotes the aggregate rainfall received by the entire country on that day. When these variables are measured from the dataset, we denote them as $X^{DATA}(s,t)$ and $Y^{DATA}(t)$. When we consider simulation outputs by a model M, they are denoted as $X^M(s,t)$ and $Y^M(t)$.

Next, we introduce two latent variables that indicate the rainfall conditions. Each state of binary variable $Z(s,t)$ represents a distribution over the rainfall volume at location s and day t, one state $(Z = 1)$ peaked at higher value and the other $(Z = 2)$ close to 0. In other words, $X(s,t) \sim Gamma(\alpha_{skt}, \beta_{skt})$ *where* $k = Z(s,t)$ where (α, β) are the parameters of a Gamma distribution dependent on Z, and potentially varying across locations and time. We also consider $U(t)$ that takes 3 values and indicates the rainfall conditions over the entire country. $U = 1$ is associated with active spells [2] when most of the S locations are in state $Z = 1$. But $U = 2$ is associated with the pre-onset and break spells [2], when most of the S locations are in state $Z = 2$. $U = 3$ signifies normal conditions.

Our stochastic simulators use conditional distributions of local climate conditions on each day $Z(s,t)$, based on all-India climatic conditions $U(t)$ and local conditions on the previous day $Z(s,t-1)$. To estimate these conditional distributions from the training data, we first need to infer the Z and U variables during this period. To maintain spatio-temporal coherence of these estimates Z^{DATA} and U^{DATA}, we use a Markov Random Field, with vertices corresponding to each of these variables, and also vertices corresponding to observed variables $X(s,t)$ and $Y(t)$. Edges are put between spatio-temporally adjacent Z-variables (temporal edges between $Z(s,t)$-$Z(s,t+1)$, spatial edges between $Z(s,t)$-$Z(s',t)$ where $s' \in NB(s)$), $Z(s,t)$-$X(s,t)$ variables for each spatio-temporal location, $Z(s,t)$-$U(t)$ and $U(t)$-$Y(t)$ variables for each day. Edge potentials functions are defined on edges to enforce spatio-temporal coherence, which take high values when the edge's end-vertices are equal. The Z and U variables are inferred conditioned on the X and Y variables, using Gibbs Sampling. Parameters like α, β are also estimated simultaneously. For simplicity, α, β are made independent of time. The inferred Z^{DATA} and U^{DATA} are used to estimate conditional distributions.

III. COHERENT ZONE DETECTION

To improve spatial coherence, we now attempt to partition the landmass into coherent zones, so that Z-varaiables can be made specific to zones rather than locations. In the literature, various attempts at regionalization of the Indian landmass has been made based on rainfall characteristics [20] mostly with respect to annual statistics. We attempt to identify sets of locations where each of them can be assigned the same value of Z every day. For this purpose we use the Z^{DATA} assignments into the framework of spatial clustering. Since we do not now the number of clusters, i.e. coherent zones to be formed, we make use of Nonparametric approaches based on Chinese Restaurant Process, like [14], [15].

Consider each locations s is assigned to a coherent zone $H(s)$, and a set V of canonical binary vectors $\{V_1, V_2, \dots\}$ of dimension T (number of days), each of which corresponds to the Z-vectors for a coherent zone. The Z-vector of each location is a somewhat corrupted version of $V_{H(s)}$, where an expected fraction p of the binary entries are flipped, i.e. on an expected number Tp of all the T days, the local weather state at any location is different from the weather state of its corresponding zone. The number of zones to be created clearly depends on p, let this number be K_p.

Now, we introduce the generative model based on Spatially Coherent Chinese Restaurant Process (SCCRP). For each location s, we assign to it a zone id $H(s)$, which can be among the zones assigned to the neighboring locations, or a separate zone. This ensures that all the zones are spatially coherent; no location is assigned to a zone unless at least one of its neighboring locations is also assigned to that zone. As with normal Chinese Restaurant Process, if we consider the assignment process sequentially, the probability of assigning any location s to a zone k is proportional to the number of locations n_k already assigned to it, and that of assigning s to a new coherent zone is proportional to a constant α. Once this has been done, the binary Z-vector for that location s is generated by flipping each of the elements of V_k with a probability p. We use Gibbs Sampling to perform the inference on H, with V re-estimated with each iteration. Finally we get K_p coherent zones, which depends on p.

IV. SIMULATION MODELS

In the first model $M1$, we simulate the all-India conditions U from a conditional distribution $prob(U(t) = l|U(t-1) = m) = \lambda_{lm}$, and then the local conditions Z for each location from conditional distribution

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

$prob(Z^{M4}(s,t) = n | Z^{M4}(s,t-1) = l, U^{M4}(t) = m) = \pi_{slmn}$. λ and π are estimated from Z^{DATA} and U^{DATA}. The model is as follows:

$$U^{M1}(1) \sim \hat{\lambda}; U^{M1}(t) \sim \lambda_n \textbf{ where } n = U^{M1}(t-1)$$
$$Z^{M1}(s,t) \sim \pi_{slm}; X^{M1}(s,t) \sim Gamma(\alpha_{sk}, \beta_{sk})$$
$$\textbf{where } m = U^{M1}(t); l = Z^{M1}(s,t-1); k = Z^{M1}(s,t);$$
$$\forall s \in \{1,S\}, t \in \{1,T\}$$

This model ensures temporal coherence of both Z and U, and it also captures the relation between local and all-India conditions each day, but not spatial coherence of Z. So we propose Model $M2$, but using an additional variable C for weather state at zone z. The π distributions are now defined over these zones instead of locations. Once the zonal weather states C have been simulated according to π, the local weather states $Z(s,t)$ are selected by setting them equal to the corresponding zonal state $C(H(s),t)$ with probability p, and the reverse of the zonal state with probability $(1-p)$. The model is as follows:

$$U^{M2}(1) \sim \hat{\lambda}; U^{M2}(t) \sim \lambda_n \textbf{ where } n = U^{M2}(t-1)$$
$$C^{M2}(z,t) \sim \pi_{zlm}; \forall z \in \{1,K_p\}, t \in \{1,T\}$$
$$\textbf{where } m = U^{M2}(t); l = C^{M2}(z,t-1);$$
$$Z^{M2}(s,t) \sim Ber(c,p); X^{M2}(s,t) \sim Gamma(\alpha_{sk}, \beta_{sk})$$
$$c = C^{M2}(H(s),t); k = Z^{M2}(s,t); \forall s \in \{1,S\}, t \in \{1,T\}$$

V. EVALUATION OF SPATIO-TEMPORAL PROPERTIES

We use two gridded datasets of daily rainfall over India - with resolutions of $100Km - 100Km$ and $25Km - 25Km$. We use every even year in the period 2000-2014 for training, from which we estimate the model parameters and coherent regions. The simulation is then done for 15 years, and the results are compared against the data for the period 2000-2014. We also compare the results by 16 General Circulation Models which were identified by [3] to be somewhat suitable for Indian monsoon simulation.

The number of locations S is 357 for low-resolution and 4964 for the high-resolution dataset, while $T = 122 * 15$, where 122 is the number of days in monsoon per year. $NB(s)$ for location s is taken as the set of locations surrounding it in the rectangular grid system, i.e. each location (except those on border or sea shore) has 8 neighbors. For coherent zone identification we use $p = 0.9$, which forms $K_{0.9} = 129$ for the low-resolution dataset and $K_{0.9} = 248$ for the high-resolution one.

For each simulated dataset, we compute the mean and standard deviation of X at all locations, and also Y. We compute dMX and dSX: the mean relative error in these quantities, i.e. $dMX = mean_s \frac{|mn_s^{MODEL}(X) - mn_s^{DATA}(X)|}{mn_s^{DATA}(X)|}$ where $mn_s(X)$ is the mean of X at location s across all the days, and $dSX = mean_s \frac{|sd_s^{MODEL}(X) - sd_s^{DATA}(X)|}{sd_s^{DATA}(X)|}$ where $sd_s(X)$ is the standard deviation of X at location s across all the days. We also measure SY - the standard deviation of spatial aggregate rainfall. Also, to see how well *local extreme rainfall* are simulated, we measure $X100$: the total number of daily rainfall events of over 100mm. Next, we come to mean lengths of wet spells $wetln$, the mean number of successive days that a location receives over 10mm of rainfall. Next, we evaluate scr *-mean spatial correlation* -of each location with its neighboring locations on same day. The mean is computed across locations and days. We also have spatial patterns - S-dimensional vector of rainfall volume at each location. We compute this pattern each day and compute its correlation with the pattern of the previous day, and the mean correlation across all days is evaluated as *mean temporal correlation* tcr. Again, we compute the mean spatial pattern across all the days for both the data and the simulation, and evaluate their correlation as $spatcr$.

Model	dMX	dSX	SY	X100	wetln	scr	tcr	spatcr
LDATA	0	0	1212	2542	1.9	0.58	0.37	1
GCM	0.46	0.45	1261	748	3.1	0.71	0.66	0.58
LModel1	0.11	0.11	729	2472	1.8	0.15	0.28	0.92
LModel2	0.16	0.13	819	2690	1.8	0.23	0.27	0.9
HDATA	0	0	1212	2542	1.9	0.58	0.37	1
HModel1	0.17	0.13	936	33272	1.7	0.08	0.29	0.9
HModel2	0.18	0.14	1036	32158	1.7	0.24	0.28	0.88

TABLE I
PERFORMANCE EVALUATION OF THE PROPOSED MODELS AND GCMS AGAINST LOW-AND-HIGH RESOLUTION DATA FOR 2000-2014.

The above quantities are computed for the datasets, proposed models, and the selected GCMs. In Table 1 we show these results. For brevity, we show the mean of each quantity measured from all the selected GCMs. In case of the proposed models, the results are provided for both high-resolution and low-resolution data. For each model, the reported numbers are the mean over several simulations. It is clear that the location-wise mean and variance are estimated better by the proposed models than GCMs. They also simulate the temporal coherence properties much better compared to the overestimation by GCMs. They are also able to produce the spatial patterns accurately unlike GCMs. Finally, Model2 improves upon Model1 for spatial correlation by using coherent zones. The overestimation of this quantity by GCMs is because they use coarser grids,

which are downscaled by us for comparison. Although we do not show the performance of individual GCMs, none of them outperform any of the proposed models. However, some GCMs are better than the rest in certain respects. The details are available in the full version of this work [21].

VI. Further Work and Conclusion

To understand the complexities of the process, and strengths or weaknesses of different approaches, we need to study a more exhaustive list of models. Such a study is available in the full version of this work [21]. Here, we explore 6 models, including the ones discussed here, and compare their merits and demerits. Another aspect that needs to be evaluated is *conditional simulation* in which the simulation by the models is carried out conditioned on some available information [18], such as the total all-India rainfall each day, or rainfall at a small number of locations each day. Such conditional simulations are very useful and are also easier to evaluate since they are linked to real data. In [21] we present detailed evaluations of the models for both types of conditional simulation.

This work shows the potential of stochastic simulation of Indian rainfall by Bayesian generative models. The coherent zones provide a way to reduce model parameters and improve spatial correlations, without using Gaussian Processes for which design of covariance function is very difficult for the diverse landmass. We aim to produce more efficient and accurate generative models for Indian rainfall.

Acknowledgement This work was partially funded by Airbus India.

References

[1] Ailliot, Pierre and Allard, Denis and Monbet, Valérie and Naveau, Philippe, *Stochastic weather generators: an overview of weather type models*, Journal de la Société Française de Statistique, 2015, Vol 156(1), pp 101–113

[2] S. Gadgil, *The Indian monsoon and its variability*, Annual Review of Earth and Planetary Sciences, 2003, Vol. 31(1), pp 429–467

[3] Jayasankar, CB and Surendran, Sajani and Rajendran, Kavirajan, *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, 2015, Vol. 42(9), pp 3513–3520

[4] Mendoza, Pablo A and Clark, Martyn P and Barlage, Michael and Rajagopalan, Balaji and Samaniego, Luis and Abramowitz, Gab and Gupta, Hoshin, *Are we unnecessarily constraining the agility of complex process-based models?*, Water Resources Research, 2015, Vol. 51(1), pp 716–728

[5] Kleiber, William and Katz, Richard W and Rajagopalan, Balaji, *Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes*, Water Resources Research, 2012, Vol. 48(1)

[6] C.W. Richardson, *Stochastic simulation of daily precipitation, temperature, and solar radiation*, Water Resour. Res., 1981, 17 (1), pp. 182190

[7] Rodriguez-Iturbe, Ignacio and Cox, DR and Isham, Valerie, *A point process model for rainfall: further developments*, Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences, 1988, Vol. 417(1853), pp 283–298

[8] Glasbey, CA and Nevison, IM, *Rainfall modelling using a latent Gaussian variable*, Modelling Longitudinal and Spatially Correlated Data (1997), pp 233–242

[9] Wilks, DS, *Multisite generalization of a daily stochastic precipitation generation model*, Journal of Hyrology (1998), Vol 210(1), pp 178–191

[10] Dabral, PP and Pandey, Ashish and Baithuri, N and Mal, BC, *Stochastic modelling of rainfall in humid region of NorthEast India*, Water Resources Management, 2008, Vol. 22(10), pp 1395–1407

[11] Cowden, Joshua R and Watkins, David W and Mihelcic, James R, *Stochastic rainfall modeling in West Africa: parsimonious approaches for domestic rainwater harvesting assessment*, Journal of Hydrology (2008), Vol. 361(1), pp 64–77

[12] Munang Tingem, Mike Rivington, Sayed Azam-Ali and Jeremy Colls, *Assessment of the ClimGen stochastic weather generator at Cameroon sites*, African Journal of Environmental Science and Technology, 2007, Vol. 1 (4), pp. 86-92

[13] Q.Fu, A.Banerjee, S.Liess, and P.K.Snyder, *Drought detection of the last century: An MRF-based approach*, SIAM International Conference on Data Mining (SDM), 2012, pp 24–34

[14] Ghosh, Soumya and Ungureanu, Andrei B and Sudderth, Erik B and Blei, David M, *Spatial distance dependent Chinese restaurant processes for image segmentation*, Advances in Neural Information Processing Systems, 2011, pp 1476-1484

[15] Socher, Richard and Manning, Christopher D, *A Gibbs Sampler for Spatial Clustering with the Distance-dependent Chinese Restaurant Process*, researchgate.net

[16] http://www.ipcc-data.org/guidelines/pages/weather_generators.html

[17] Kim, Dongkyun and Kim, Jongho and Cho, Yong-Sik, *A poisson cluster stochastic rainfall generator that accounts for the interannual variability of rainfall statistics: validation at various geographic locations across the united states*, Journal of Applied Mathematics, 2014

[18] Verdin, Andrew and Rajagopalan, Balaji and Kleiber, William and Podestá, Guillermo and Bert, Federico, *A conditional stochastic weather generator for seasonal to multi-decadal simulations*, Journal of Hydrology, 2015

[19] Baxevani, Anastassia and Lennartsson, Jan, *A spatiotemporal precipitation generator based on a censored latent Gaussian field*, Water Resources Research, 2015, Vol 51(6), pp 4338–4358

[20] Srinivas, VV, *Regionalization of Precipitation in India–A Review*, Journal of the Indian Institute of Science, 2013, Vol 93(2), pp 153–162

[21] Mitra, A. *Bayesian approach to Spatio-temporally Consistent Simulation of Daily Monsoon Rainfall over India* , 2017, arxiv preprints.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

FINDING ACTIVE AND BREAK SPELLS OF INDIAN MONSOON BY MARKOV RANDOM FIELDS

Adway Mitra[1], Amit Apte[1], Rama Govindarajan[1], Vishal Vasan[1], Sreekar Vadlamani[2]

***Abstract*—The Indian summer Monsoon brings rainfall to most parts of India, during mid-May to mid-October each year. It exhibits considerable spatial and temporal variations across years and within each year. Meteorologists have been studying the variations of various properties of the monsoon, including short phases of unusually high (active) or low (break) rainfall using threshold-based approaches. In this work, we propose a framework based on Markov Random Field (MRF) to analyse grid-level daily data of precipitation and cloud cover. The MRF assigns state variables to (gridpoint,day) pairs indicating the climatic condition. Spatial and temporal coherence of such states is ensured by the edge potential functions of the MRF. We use these variables to identify active/break phases of the monsoon, taking into consideration the climate states at all locations instead of only the all-India spatial-mean as done by existing methods. We also identify common spatial patterns of rainfall and clouds associated with these phases.**

I. INTRODUCTION

India receives considerable rainfall every year during the months June to September (JJAS), from the Monsoon. However, the amount of rainfall varies greatly across years, across days within each year, and also across locations [1], [2]. Low-rainfall years are known to have significant negative effects on food-grain production, and hence the lives of a billion people. Indian Meteorological Department (IMD) and climate scientists have, for more than a century, tried to identify spatially coherent regions which have uniform rainfall patterns, days and years in which rainfall is excess or deficient, how the spatial distribution of rainfall changes across days in a year and across years, and so on. Recently, the widespread availability of data and the huge progress in Data Science has resulted in advanced statistical methods being used in multiple domains of study, including climate. So far, not much effort has been made in studying the Indian monsoon through the lens of Data Science/Machine Learning. The aim of this work is to make a data-driven study of the Indian monsoon climate, with special emphasis on discovering small-scale or local properties, and how they are related to large-scale or all-India patterns.

Corresponding author: Adway Mitra, adway.mitra@icts.res.in
[1]ICTS-TIFR, Bangalore, India [2]TIFR-CAM, Bangalore, India

II. ACTIVE AND BREAK SPELLS

In most years, the monsoon season has one or more periods of 3 or more continuous days when the mean rainfall volume over the Monsoon Zone [2] (a large rectangular region over Central and Northern India, representative of All-India Rainfall) is exceptionally high/low. These periods are called "active spells" and "break spells" respectively. Break spells are caused when the cloud cover vanishes over Central and Northern India as characterised by [3], [4], while [5] defines these phases based on the strength of winds over the Bay of Bengal. On the other hand, [6],[7] and [8] define active and break spells directly in terms of the aggregate rainfall over the Monsoon Zone or the entire country[9]. The mean rainfall for each day is compared against the climatogical mean for that date, and accordingly each day is marked as "active" or "break", and 3 or more consecutive days marked this way are identified as spells. [10] uses smaller subdivisions of India and identifies regional dry and wet spells. It is recognized by most of the above studies that active and break spells are associated with characteristic spatial patterns of rainfall.

In contrast to the existing approaches we define "active" or "break" spells based on the climatic conditions of individual locations, encoded by discrete state variables, instead of spatial aggregate. We aim to replace hard thresholds by a probabilistic approach where *spatio-temporal coherence* plays an important role in determining these state variables.

III. SPATIO-TEMPORAL MARKOV RANDOM FIELD

Markov Random Fields have been used in various applications of Computer Science for the past 20 years, but more recently it has emerged as a powerful approach for modeling spatio-temporal data [11]. In the geospatial domain, Markov and Gaussian Random Fields have

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

been used to model ocean temperature [12], detection of droughts [13], downscaling/disaggregation of a process observed at low resolution into high resolutions [14], or for spatial interpolation. Sometimes, they have been used to model a latent process which drives an observed process. Most of these works consider continuous latent variables, but in this work we will use discrete latent variables to answer specific questions.

Consider S grid-locations and D days in the daily rainfall dataset for a year. For each location s, on day d, we denote by Z_{sd} the *state of the climate*, and by X_{sd} the observed measurement of a climatic variable, such as rainfall. The Z-variables are unobserved or latent, and must be inferred. However, each of them can take values in $\{1, 2 \ldots, K\}$. In our model, we will consider 3 types of edges: 1) **temporal edges** between Z_{sd} and $Z_{s,d+1}$ i.e., between state variables of a particular location on successive days; 2) **spatial edges** between Z_{sd} and $Z_{s'd}$ i.e., between state variables of neighboring locations on a particular day; 3) **data edges** between Z_{sd} and X_{sd}, i.e., between a state variable and the corresponding observed measurement on a particular location and day. The model is shown in Figure 1.

We set the potential function associated with the data edges as the conditional PDF of the observed value over the continuous space of observations, conditioned on the discrete state variable. In this work, the observed variable X_{sd} is the rainfall or OLR recorded at location s on day d. For rainfall, we consider a state space of size $K = 2$, signifying wet days ($Z_{sd} = 1$) and dry days ($Z_{sd} = 2$). For both states, we consider a Gamma distribution over the observed quantity of rainfall, with parameters specific to the state, as $\psi_D(Z_{sd}, X_{sd}) = Gamma(X_{sd}; \alpha_{sk}, \beta_{sk})$ where $Z_{sd} = k$

Obviously, state-specific parameters ensure that heavier rainfall is likely on wet days ($Z_{sd} = 1$) and low or no rainfall is likely on dry days ($Z_{sd} = 2$). The use of Gamma distribution to model local daily rainfall has been used in the Climate Science community particularly in rainy seasons [15], [16].

We define the potential functions of temporal and spatial edges between the state variables to promote *spatial and temporal coherence*. These functions take high values if the Z-variables connected by the edge take equal value, and low values if they are different, as $\psi_T(Z_{sd}, Z_{s,d+1}) = exp(a\mathcal{I}(Z_{sd} = Z_{s,d+1}))$ (temporal coherence) and $\psi_S(Z_{sd}, Z_{s'd}) = exp(c(s, s')\mathcal{I}(Z_{sd} = Z_{s'd}))$ (spatial coherence). Here $\mathcal{I}$ is the indicator function and s' is any neighboring grid location of s. a is temporal-coherence parameter (assumed constant), and spatial coherence parameter c is specific to the locations (s, s') because not every pair of neighboring locations have the same degree of correlation.

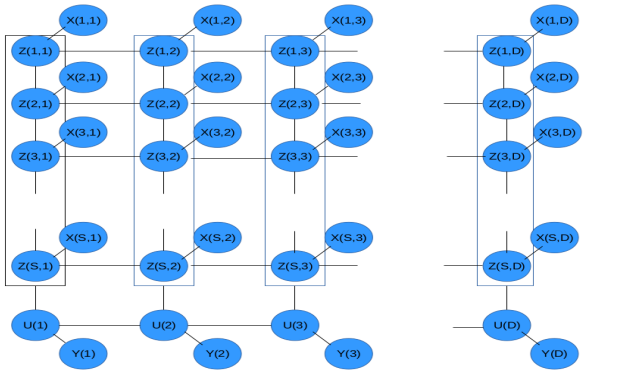


Fig. 1. Spatio-Temporal Markov Random Field for Indian Rainfall. Each column represents one day, each row represents a location. Z: latent state variable, X: observed rainfall volume. Horizontal edges are "emporal", vertical edges "spatial", "data edges" connect Z, X nodes. The U-nodes below represent daily All-India state variables, connected to all local state variables on same day

Since active and break spells of the Indian monsoon are defined not for individual locations but for the country as a whole, we incorporate into our model all-India state variables U_d which indicate whether day d is an "active day" (1), "break day" (2) or "normal day" (3). However, as already mentioned, we wish to define the all-India state based on the individual locations, rather than on the all-India spatial mean rainfall. We let the locational state variables $\{Z_{sd}\}_s$ "vote" to decide the all-India state variable U_d. Hence, we set edges from all the locational variables Z_{sd} to U_d on which edge potentials are defined, allowing U_d to take a "vote" on the locational states, and each value of U_d is characterized by the number of locations that are in state 1 (wet), i.e. $\psi_{SS}(U_d, Z_{sd}) = exp(\mathcal{I}(U_d = Z_{sd})/S)$. At the same time, we associate U_d with Y_d, the spatial mean rainfall on day d as done in literature, but all-India spatial mean instead of the "monsoon zone", as $\psi_{SD}(U_d, Y_d) = \mathcal{N}(Y_d; \mu_{U_d}, \sigma_{U_d})$. Additionally, since active and break phases are defined as intervals of at least 3 consecutive days, we enforce temporal coherence on the Z_d-variables through edge potentials, as $\psi_{ST}(U_d, U_{d+1}) = b\mathcal{I}(U_d = U_{d+1})$. Note that here we have replaced the Gamma distribution by Gaussian distribution for the observed rainfall quantity, since the distribution of spatial-mean daily rainfall volume is more symmetric, unlike location-specific daily rainfall.

The most important task now is to infer the latent variables Z, while estimating the parameters like α, β. We first make an initial estimate of the Z and U-variables by ignoring the edges and putting thresholds

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

on the observations X,Y. The coherence parameters a, b, c are estimated based on these initial values of Z, U. For example, the spatial coherence paramaters for an edge connecting two locations are set proportional to the number of days that the initial estimates of Z at the two locations are equal. After that, we infer Z and U-variables using Gibbs Sampling, by sampling each Z_{sd} or U_d by turn, conditioned on the rest.

This same model can also be used for discrete representation of cloud cover, measured by the proxy variable of Outgoing Longwave Radiation (OLR). In this case we use three states (1-very cloudy, 2-not cloudy, 3-moderate cloudy) for the Z-value at each location and also for the all-India variable U. We use Gaussian distribution (instead of Gamma) for the data edge potentials, i.e. $\psi_D(Z_{sd}, X_{sd}) = \mathcal{N}(X_{sd}; \mu_{sk}, \sigma_{sk})$ where $Z_{sd} = k$. These differences are due to the natures of the distribution of daily rainfall and daily OLR data, at grid-scale and all-India scale. Grid-wise daily OLR data is symmetric, while grid-wise daily rainfall data has a long right tail. To distinguish between the variables associated with rainfall and OLR, we use the notations $\{XC, YC, ZC, UC\}$ for OLR and $\{XR, YR, ZR, UR\}$.

Two main properties of the proposed model are interpretability and non-separability. The latent variables Z at any spatio-temporal is easily interpretable- it indicates whether or not the location is under significant rainfall or cloud cover. Same holds for U also. Some models for spatio-temporal data model the spatial and temporal coherence/covariance properties separately, such models are called separable. However, this is not a realistic assumption, but non-separable models [17] often suffer from computational inefficiency. Our model defines spatio-temporal covariance implicitly through local interactions between spatio-temporally neighboring variables, making it non-separable. Since the latent variables are discrete, we also avoid computational challenges, and the algorithm is very efficient.

IV. EXPERIMENTAL EVALUATION

We use gridded daily data for precipitation and cloud cover for experiments. The **precipitation** dataset that we use was published by Rajeevan *et al.* [7]. It records daily rainfall at 357 grid-points all over Indian landmass, each $100KM \times 100KM$ in size, for the period 1901-2011. **Cloud cover** is quantified by Outgoing Longwave Radiation (OLR) for which re-analyzed data is available at (https://www.esrl.noaa.gov/psd/data/gridded/data.interp_OLR.html). This data is on a $250KM \times 250KM$ worldwide grid system, available daily since 1975. For our analysis we interpolated to the same grid system as the Indian rainfall dataset.

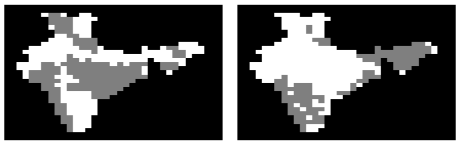


Fig. 2. Locations that are frequently in "wet" state during active spells (left) and break spells (right), shown in grey

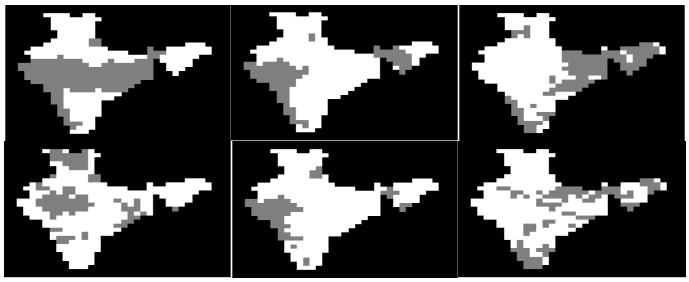


Fig. 3. Column 1: 1 August 2006, "active" by both MRF and [8], Column 2: 6 July 2006, "active" by [8] but normal by MRF, Column 3: 18 July 2007, "break" by [8] but normal by MRF. Top row shows locations assigned $Z = 1$ by MRF in grey, bottom row shows locations receiving more rainfall than their daily mean.

Active/break spells are intervals of at least 3 days of unusually high/low rainfall. In our model, the all-India state variables $\{U_d\}$ are used to determine these spells, based on both the spatial mean rainfall, and number of locations in wet state. We compare the spells identified by our method against those identified by [8], with respect to the all-India mean rainfall on those days, and the mean number of locations in "wet" state as identified by our method. The results, tabulated in Table 1, show the merits of our approach, especially for the "break" days. Clearly, the "break spells" identified by the proposed method have much less mean all-India rainfall, and fewer locations in the "wet" state. In some years, [8] doesn't identify any active or break spells,

	ACTIVE SPELLS				BREAK SPELLS			
	Mean AIR		Mean #(Z=1)		Mean AIR		Mean #(Z=1)	
Year	MRF	[8]	MRF	[8]	MRF	[8]	MRF	[8]
2000	11.34	**11.72**	182.44	**193.13**	**3.08**	5.4	**48.2**	76.33
2001	11.15	**12.45**	170.68	**187**	4.45	4.48	**87.56**	89.25
2002	10.79	–	154.6	–	**3.49**	4.39	**73.9**	79.8
2003	11.88	**13.02**	170.63	**174**	6.53	–	117.29	–
2004	**13.5**	13.2	182.5	**234.33**	**4.5**	9.5	**82.03**	113.17
2005	**12.03**	11.59	**172.68**	171	**3.16**	4.53	**46.4**	63.94
2006	**10.68**	10.34	**169.66**	148.22	4.84	–	107.75	–
2007	11.12	**11.86**	**147.6**	137.5	**6.11**	7.33	**83.76**	101.75

TABLE I
MEAN ALL-INDIA RAINFALL, AND MEAN NUMBER OF "WET" LOCATIONS IN ACTIVE(LEFT) AND BREAK(RIGHT) SPELLS IDENTIFIED BY THE PROPOSED METHOD, AND BY [8]

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

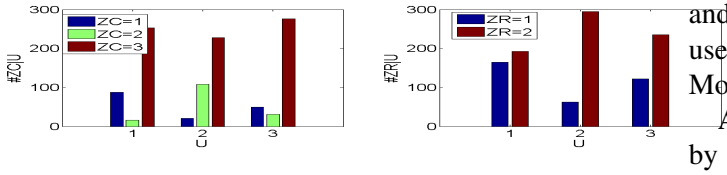


Fig. 4. Mean number of locations in different states w.r.t. rainfall and cloud-cover during active, break, normal spells

and the spells identified by our method in those years have unusually high or low values.

It is well known that during active spells, the western ghats and the monsoon zone of north/central India receive heavy rainfall, while in break spells rainfall is restricted to the north-eastern provinces, northern foothills of Himalayas and parts of Southern peninsula. These are corroborated by the state assignment to locations by our method, as shown in Figure 1, for the period 2000-2007.

In the period 2000-07, 88% of the days marked as "active" by [8] are also marked "active" by the proposed MRF, while 57% of days marked as "break" by [8] are also marked so by MRF. The differences between active and break days identified by ours and threshold-based approaches are indicated in Figure 2. Days like 6 July 2006 have heavy rainfall in some parts of India, for which they are classified as "active" by [8], even though large parts of the country remain dry. Again, days like 18 July 2007 are "break" by [8] though several locations receive above-mean rainfall. The proposed method takes local effects into consideration, and identify these days as "normal". Clearly on all these days, the local state assignments by MRF is far more coherent than using local thresholds, which allows us to define "zones" of rainfall.

We also consider the cloud-cover state variables (3 states) during these periods. We denote by ZR and ZC the latent variables corresponding to states of rainfall and cloud cover respectively. In Figure 3, we show the mean number of locations have high (ZC=1), low (ZC=2) or medium (ZC=3) cloud cover and high or low rainfall during active (U=1), break (U=2) and normal (U=3) days. Clearly, the number of locations having high cloud cover and rainfall are directly related to these spells.

V. CONCLUSION

We propose a new approach to represent Indian monsoon climate, using spatio-temporally coherent discrete state variables. We showed here how this representation can be used to make a new characterization of active and break spells, and their spatial patterns. We plan to use this representation to define local onset dates and Monsoon Intra-Seasonal Oscillation Index (MISO).

Acknowledgement This work was partially funded by Airbus India.

REFERENCES

[1] D. Sikka and S. Gadgil, "On the maximum cloud zone and the itcz over indian, longitudes during the southwest monsoon," *Monthly Weather Review*, vol. 108, no. 11, pp. 1840–1853, 1980.

[2] S. Gadgil, "The indian monsoon and its variability," *Annual Review of Earth and Planetary Sciences*, vol. 31, no. 1, pp. 429–467, 2003.

[3] K. Ramamurthy, "Monsoon of india: some aspects of the breakin the indian southwest monsoon during july and august," *Forecasting manual*, vol. 1, no. 57, pp. 1–57, 1969.

[4] R. Krishnan, C. Zhang, and M. Sugi, "Dynamics of breaks in the indian summer monsoon," *Journal of the atmospheric sciences*, vol. 57, no. 9, pp. 1354–1372, 2000.

[5] B. Goswami and R. A. Mohan, "Intraseasonal oscillations and interannual variability of the indian summer monsoon," *Journal of Climate*, vol. 14, no. 6, pp. 1180–1198, 2001.

[6] S. Gadgil and P. Joseph, "On breaks of the indian monsoon," *Journal of Earth System Science*, vol. 112, no. 4, pp. 529–558, 2003.

[7] M. Rajeevan, J. Bhate, J. Kale, and B. Lal, "High resolution daily gridded rainfall data for the indian region: Analysis of break and active," *Current Science*, vol. 91, no. 3, 2006.

[8] M. Rajeevan, S. Gadgil, and J. Bhate, "Active and break spells of the indian summer monsoon," *Journal of earth system science*, vol. 119, no. 3, pp. 229–247, 2010.

[9] V. Krishnamurthy and J. Shukla, "Intraseasonal and seasonally persisting patterns of indian monsoon rainfall," *Journal of climate*, vol. 20, no. 1, pp. 3–20, 2007.

[10] N. Singh and A. Ranade, "The wet and dry spells across india during 1951–2007," *Journal of Hydrometeorology*, vol. 11, no. 1, pp. 26–45, 2010.

[11] M. Haran, "Gaussian random field models for spatial data," *Handbook of Markov Chain Monte Carlo*, pp. 449–478, 2011.

[12] M. Lavine and S. Lozier, "A markov random field spatio-temporal analysis of ocean temperature," *Environmental and Ecological Statistics*, vol. 6, no. 3, pp. 249–273, 1999.

[13] Q. Fu, A. Banerjee, S. Liess, and P. K. Snyder, "Drought detection of the last century: An mrf-based approach," in *Proceedings of the 2012 SIAM International Conference on Data Mining*, pp. 24–34, SIAM, 2012.

[14] D. J. Allcroft and C. A. Glasbey, "A latent gaussian markov random-field model for spatiotemporal rainfall disaggregation," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 52, no. 4, pp. 487–498, 2003.

[15] G. J. Husak, J. Michaelsen, and C. Funk, "Use of the gamma distribution to represent monthly rainfall in africa for drought monitoring applications," *International Journal of Climatology*, vol. 27, no. 7, pp. 935–944, 2007.

[16] O. Vlček and R. Huth, "Is daily precipitation gamma-distributed?: Adverse effects of an incorrect use of the kolmogorov–smirnov test," *Atmospheric Research*, vol. 93, no. 4, pp. 759–766, 2009.

[17] N. Cressie and H.-C. Huang, "Classes of nonseparable, spatio-temporal stationary covariance functions," *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1330–1339, 1999.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

TRACKING THE PROPAGATION OF PLANETARY SCALE CLOUD ZONES OVER INDIAN OCEAN AND SOUTH ASIA WITH MARKOV RANDOM FIELDS

Adway Mitra[1], Amit Apte[1], Rama Govindarajan[1], Vishal Vasan[1], Sreekar Vadlamani[2]

***Abstract*—Much of South and South-East Asia receives its rainfall from planetary-scale cloud zones that originate over Indian Ocean south of the Equator, and propagate northwards in Summer. Across the year, different locations in this region are frequently under cloud cover and rainfall. Daily grid-scale data of Outgoing Longwave Radiation (OLR) and precipitation contains a lot of spatial and temporal variation, and it is impossible to visualize and identify large-scale movement patterns of cloud zones from it. We propose an approach where discrete latent variables are defined to indicate whether a location is under an active cloud zone or not on each day. A spatio-temporal Markov Random Field is constructed with the latent variables, and they are inferred based on daily local observations of rainfall and OLR. Their spatio-temporal coherence is ensured by edge potentials. Location and movement patterns of large and coherent cloud zones are clearly visible from these discrete variables.**

I. INTRODUCTION

Many countries in South Asia receive most of their annual rainfall from a monsoon season [1], roughly from May to September. It is known that this season is caused by the northward migration of a huge, planetary-scale cloud-band called Inter-Tropical Convergance Zone (ITCZ) [2] from the Equatorial region over Indian Ocean to the landmasses of South Asia. As a result, much of the South Asian landmass is under medium or high cloud cover and rainfall on several days of the monsoon season, unlike in other parts of the year. However, over most of the Indian Ocean and in various landmasses of South-East Asia this seasonality of climate is less pronounced and more evenly distributed through the year. In this work, we are interested in tracking the daily movements of planetary-scale cloud zones (including but not limited to ITCZ). The influence of an active cloud zone over a location is manifested through daily cloud cover (measureable by OLR) and rainfall there. Using observations of these two quantities, we attempt to identify the "active zones" each day, i.e. locations over which such cloud zones are active. Since our aim is to identify, track and study large-scale systems, we need the representation to be concise. Such conciseness can be obtained from a discrete representation that has spatial and temporal coherence. In doing so, we may need to make a trade-off with the magnitude of local rainfall/cloud cover, since we are not interested in strong local rainfall events unless they are related to a larger system. To make the trade-off process flexible and to avoid thresholds, we turn to a probabilistic approach: Markov Random Field.

Corresponding author: Adway Mitra, adway.mitra@icts.res.in
[1]ICTS-TIFR, Bangalore, India [2]TIFR-CAM, Bangalore, India

II. THRESHOLD-BASED APPROACH

We used gridded datasets at spatial resolution of $25Km - 25Km$ for daily rainfall and Outgoing Longwave Radiation (a proxy for cloud cover) over the South Asian region, including both land and sea, $50°E - 120°E$ and $30°S - 30°N$. The results presented here is for a particular year: 2010, though similar results were obtained in other years also (not shown here for brevity). We represent by $X^R(s,d)$ and $X^C(s,d)$, the rainfall and OLR measurements at location s, date d. Binary variable $U(s,d)$ indicates whether locations s on day d is part of an "active zone" ($U = 2$) or not ($U = 1$).

An obvious solution to this problem is to put thresholds on the rainfall and OLR values. On each day, if the OLR at a particular grid-location is below a threshold (say $220W/m^2$), and the precipitation is above a threshold (say $5mm$) we may say that the location is under a active zone on that day, and set U-variables accordingly. However, such assignment is not at all spatially coherent, and hence the daily maps are spatially oversegmented with no clear patterns. Temporal coherence is also missing due to lots of changes between any two successive days. Moreover, it is difficult to choose the thresholds reasonably. The

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

maps for two days in the year 2010 (20 Jan and 20 July) are shown in Figure 1. A quantitative measure of the lack of spatio-temporal coherence is provided later.

To improve this, we considered intervals of days around each day to get a more stable picture of whether or not a location is under influence of cloud zone. Similarly, we considered the spatial neighborhood of each spatial location (grid). We attempted spatio-temporal smoothing of the variables and $X^R(s,d)$, $X^C(s,d)$ are replaced by the mean over $\{X(s,d-t),\ldots,X(s,d+t)\}$ for a period t and $\{X(s',d)\}$ where s' is a spatial neighbor of location s. The slightly improved results are shown in last row of Figure 1.

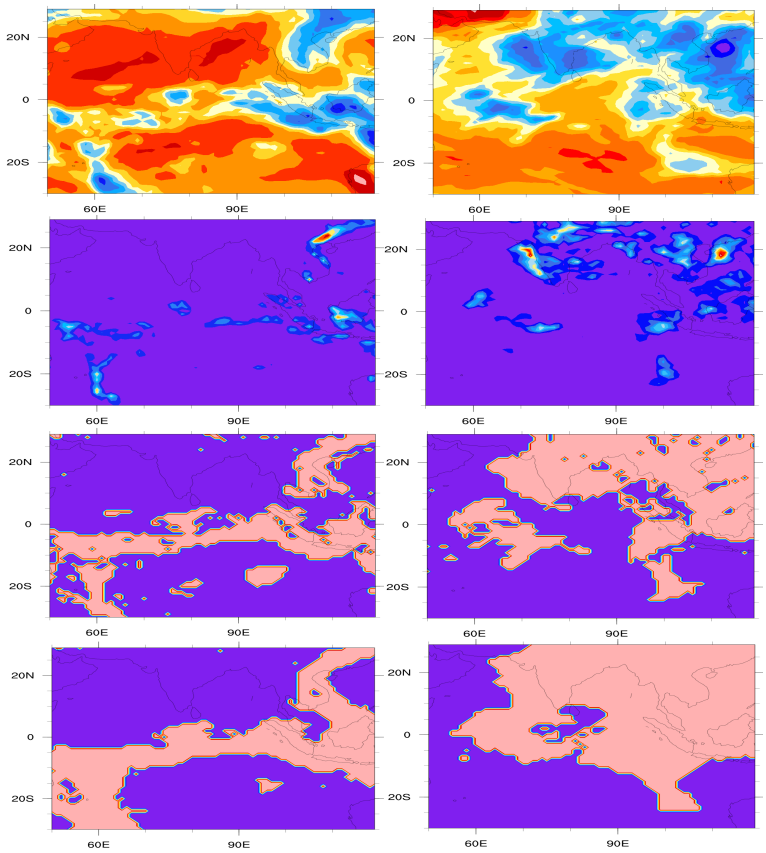


Fig. 1. Map of OLR (row 1), rainfall (row 2) and U-variables (row 3) by putting thresholds on raw data ($250W/m^2$ for OLR, 2mm for rain) on 20 Jan and 20 July, 2010. Row 4: U-variables (bottom) after spatio-temporal smoothing and putting thresholds on raw data ($220W/m^2$ for OLR, 5mm for rain) In row 1, red: weak cloud, blue: strong cloud. In row 2, purple: less rain, red: heavy rain. In rows 3,4: Pink: $U=2$, i.e. under cloud cover

III. MARKOV RANDOM FIELD

A major drawback of this approach is the use of strict thresholds, and there is no explicit attempt to identifying spatio-temporally coherent structures. We now attempt a Latent Variable approach based on Graphical Models to address these issues. A similar approach was made in [4] to locate the ITCZ over Pacific Ocean, but the model did not include rainfall and it was supervised.

For each spatio-temporal location (s,d), consider discrete variables $Z^C(s,d), Z^R(s,d), U(s,d)$ which indicate whether or not it is under strong cloud cover, heavy rainfall and a combination of both. In the previous analysis, these were assigned on the basis of thresholds on (X^R, X^C). Now, we define a Markov Random Field (MRF) [3] on these Z and U-variables, along with edge potential functions that enforce spatio-temporal coherence, and infer their values through probabilistic inference. In this model, we define Z^R and Z^C-variables to have 4 discrete states, corresponding to very low (1), low(2), high (3) and very high (4) amounts of rainfall and cloud cover. The values associated with these states are not separated by hard thresholds but allowed to overlap, so that spatial and temporal coherence of these state-assignments is maintained. The U-variables are binary, as earlier.

We consider a graph where each node corresponds to one of these variables like $Z^C(s,d), Z^R(s,d), U(s,d)$, for a spatio-temporal location (s,d). We also have nodes corresponding to $X^C(s,d), X^R(s,d)$. *Spatial edges* are added between pairs of nodes associated with spatially adjacent locations ((s,d) and (s',d) where s' and s are spatial neighbors), corresponding to each type of latent variable. Similarly, *temporal edges* are added between pairs of nodes associated with successive days ((s,d) and $(s,d+1)$), corresponding to each type of latent variable. Also, nodes corresponding to each observed variable $X^R(s,d)$ (or $X^C(s,d)$) are connected by *data edges* to discrete variables $Z^R(s,d)$ (or $Z^C(s,d)$). Finally $U(s,d)$ is linked to the Z^R, Z^C-variables for a period $\{d-t,\ldots,d+t\}$ of days by *state edges* for every spatio-temporal location (s,d). The intermediate discrete variable Z is important to design and compute potentials for U, as each U-variable must connect to an interval of $2t+1$ days.

We define *edge potential* functions Ψ on each of the edges that enforce spatial and temporal coherence of these discrete variables. Specifically, for spatial edges the potential is defined as $\Psi^R((s,d),(s',d)) = exp(-|Z^R(s,d) - Z^R(s',d)|)$, $\Psi^C((s,d),(s',d)) = exp(-|Z^C(s,d) - Z^C(s',d)|)$, for temporal edges also the potentials are defined accordingly. Similarly, for state edges between $U(s,d)$ and $Z(s,d')$ (where $d' \in \{d-t,\ldots,d+t\}$), the potential functions are defined as $\Psi^R_U((s,d),(s,d')) = exp(-|Z^R(s,d) - 2U(s,d')|)$, $\Psi^C_U((s,d),(s,d')) = exp(-|Z^C(s,d) - 2U(s,d')|)$. Clearly, each of these functions takes high value when

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

the discrete variables connected by that edge are equal or close.

Each discrete state represents a distribution over the real space of the corresponding observed variables. For example, the distribution for $Z^C = 1$ is peaked at low values, while that of $Z^C = 4$ is peaked at high values. We use the Gamma distribution for rainfall and Gaussian distribution for OLR values, i.e. $X^R(s,d) \sim Gamma(\alpha_{sk}, \beta_{sk})$ where $k = Z^R(s,d)$, $X^C(s,d) \sim \mathcal{N}(\mu_{sk}, \sigma_{sk})$ where $k = Z^C(s,d)$. These PDFs are used as the edge potentials for each data edge. The distribution parameters are dependent on locations, to account for the natural spatial variability of rainfall.

The total likelihood of the model is the product of all the edge potential functions discussed above. The discrete state variables Z^R, Z^C, U and the distribution paramaters are unknown. So we first initialize them at each location individually, by setting thresholds on the observations based on the mean and variance at each location for rainfall and OLR. After that probabilistic inference is carried out using Gibbs Sampling [5], where each of these latent variables are sampled, conditioned on all of the rest. The conditional distribution of each variable is proportional to the product of the potential functions on all edges attached to its node in the MRF. By the iterative process we generate enough samples for all the variables, from which we estimate their optimal values.

Two main properties of the proposed model are interpretability and non-separability. The latent variables Z^R, Z^C, U at any spatio-temporal is easily interpretable- it indicates whether or not the location is under significant rainfall or cloud cover or both. Some models for spatio-temporal data model the spatial and temporal coherence/covariance properties separately, such models are called separable. However, this is not a realistic assumption, but non-separable models [6] often suffer from computational inefficiency. Our model defines spatio-temporal covariance implicitly through local interactions between spatio-temporally neighboring variables, making it non-separable. Since the latent variables are discrete, we also avoid computational challenges, and the algorithm is very efficient.

IV. RESULTS

In Figure 2, we show maps corresponding to assignments of Z^R,Z^C and U on two different days. It is obvious that the maps of Z^R and Z^C are more spatially coherent than those of X^R and X^C in Figure 1. The binary map of U is also clearly far more coherent in Figure 2.

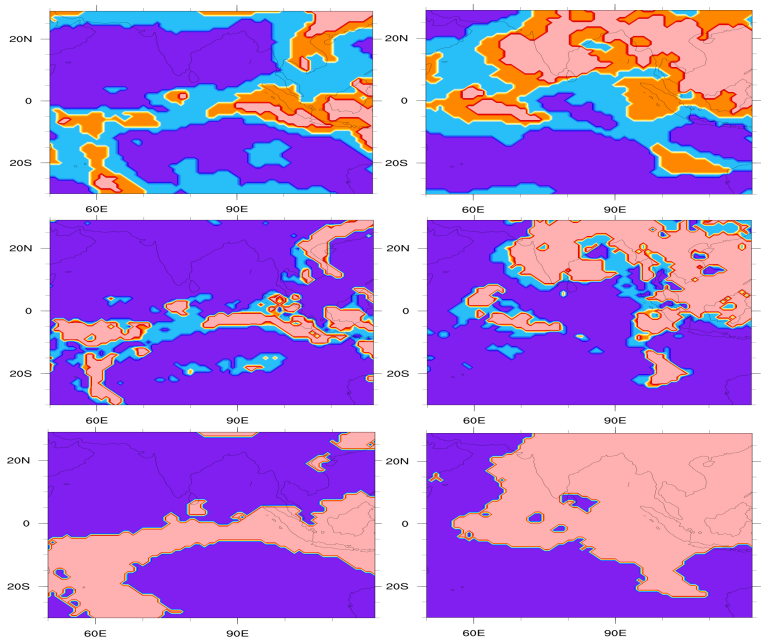


Fig. 2. Map of discrete states for OLR (Z^C,top), rainfall (Z^R,middle) and U-variables (bottom) by proposed model on 20 Jan and 20 July, 2010. Pink: $U = 2$, i.e. under cloud cover

We now attempt to express these results quantitatively. For comparison, we choose such thresholds so that the number of spatio-temporal locations assigned to $U = 2$ are roughly comparable $(33\% - 40\%)$ in all methods. To quantify *temporal coherence*, we compute on how many days the U-variable at each location was equal to that of the previous day, and compute its mean over all locations. In case of the threshold-based approach, this turns out to be 257 out of the 365 days in the year, while after smoothing it improves to 342 days. In the MRF-based approach this is 358 days.

We quantify *spatial coherence* in two ways - locally and globally. For *local spatial coherence*, we compute how many spatially neighboring grids of each grid-point has the same value of U on each day, and compute its mean over all locations and days. Using threshold-based approach on the raw data, this value is 6.9 (where each location has 8 neighbors in the grid system). In both the threshold-with smoothing and the MRF approaches, this value comes to around 7.4. But to measure global spatial coherence, we compute the number of *connected components* in the maps of U on each day, where we consider each grid-location as a node in a graph, and neighboring locations having same value of U are connected by edges. The number of connected components in such graphs is measured for each day. Smaller the number of connected components, more coherent is the graph. In the MRF-based approach, the mean number

of connected components is 10, while in the threshold-based approach it is 25, and for thresholding after smoothing the value is 15. Clearly, the MRF approach creates coherent maps of U.

In Figure 3 we show the tracking of cloud zone marked by U-variables on maps from the period 5 May to 28 June 2010, when the band expands northwards from the equatorial region to cover much of South Asia including India, Sri Lanka and Burma.

V. CONCLUSION

Tracking of large-scale cloud zones is important to understand the physics of the climate systems. However, detecting or tracking them is not an easy task because they are not always manifested in the local daily readings of climatic variables, and using thresholds on these readings fails to capture any coherent structure. In this work we showed how this shortcoming can be overcome using graphical models that derive statistical strength from spatio-temporal neighborhooods. While temporal effects were taken into account by considering a window of 15 days, it should be noted that the window length is a tuneable parameter with bearings on the results. In a more detailed study, we plan to explore this issue. It is also interesting that though the Markov Random Field imposed local spatial coherence through interactions between neighboring spatial locations, yet global spatial structures emerged which are visible as large coherent zones in the maps. The temporal stability of these structures allow us to track them. A follow-up on this work would be to quantify the movement patterns of these zones, and build a dynamical model.

Acknowledgement This work was partially funded by Airbus India.

REFERENCES

[1] S. Gadgil, *The Indian monsoon and its variability*, Annual Review of Earth and Planetary Sciences, 2003, Vol. 31(1), pp 429–467

[2] D.R.Sikka, S.Gadgil, *On the Maximum Cloud Zone and the ITCZ over Indian Longitudes during the South-West Monsoon*, Monthly Weather Review, 1980, Vol 108, pp 1840–1853

[3] R.Kindermann and J.L. Snell, *Markov Random Fields and their Applications*, 1980, American Mathematical Society

[4] C.L.Bain, J. De Paz, J. Kramer, G. Magnusdottir, P.Smyth, H. Stern, and C Wang, *Detecting the ITCZ in Instantaneous Satellite Data using Spatiotemporal Statistical Modeling: ITCZ Climatology in the East Pacific*, Journal of Climate, 2011, Vol 24(1), pp 216-230

[5] R.M. Neal, *Probabilistic Inference using Markov Chain Monte Carlo Methods*, 1993

[6] Cressie N, Huang HC; *Classes of nonseparable, spatio-temporal stationary covariance functions*, Journal of American Statistical Association, 1994, pp 1330–1340

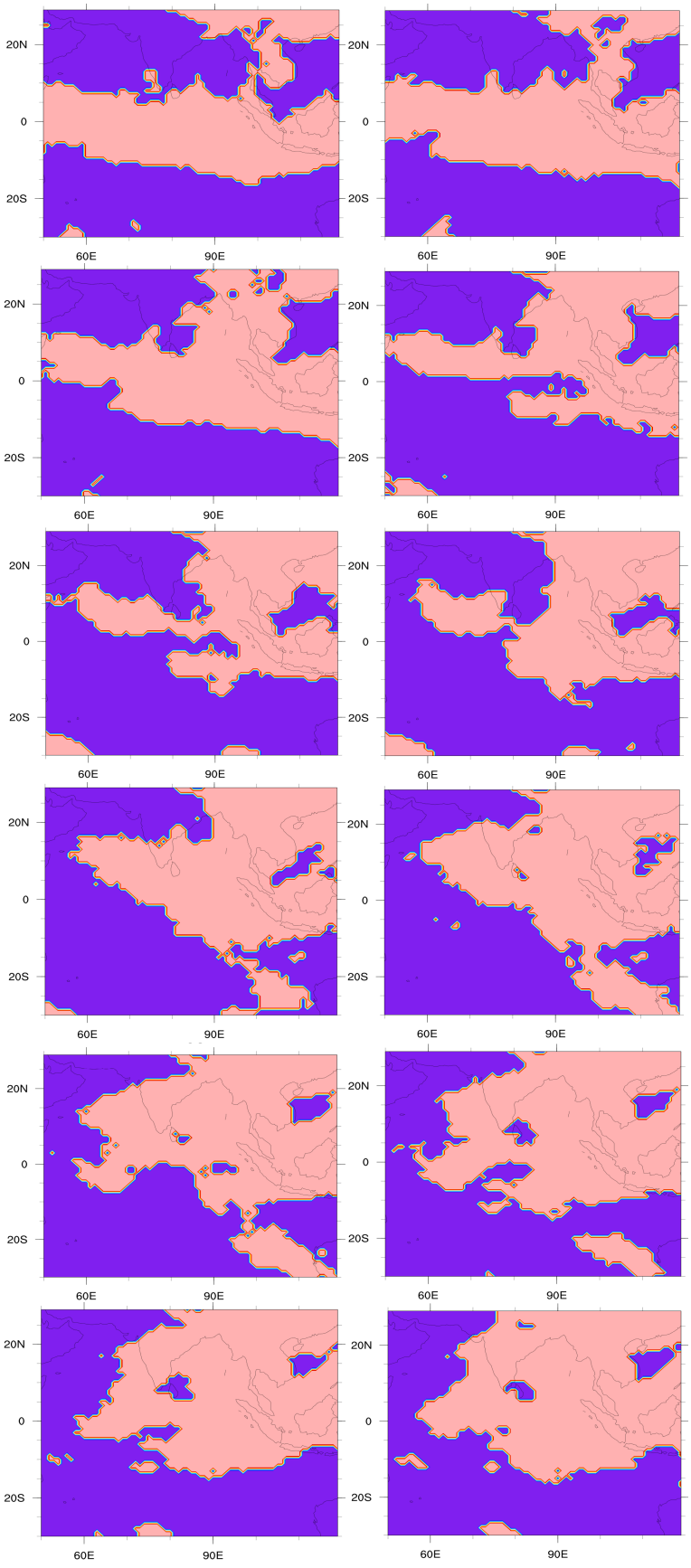


Fig. 3. Maps of U-variables by proposed model from 5 May and 28 June, 2010 at 5-day intervals, showing northward propagation of the pink band to cover South Asia. Pink: $U = 2$, i.e. under cloud cover

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

PREDICTABILITY OF ATTRIBUTES OF ANNUAL AND MONTHLY RAINFALL OVER INDIA

Adway Mitra[1], Ashwin Seshadri[2]

***Abstract*—We evaluate predictability of discrete attributes of monthly and annual rainfall over India using conditional statistics. Three attributes are studied: whether rainfall in a given year is above or below its long-term mean ("state"), whether it differs from the long-term mean by at least one standard deviation ("extreme"), and whether it is above or below the previous year's rainfall ("phase"). For monthly scale we consider rainfall volumes scaled by the corresponding calendar-month's mean to account for seasonal effects. We consider an attribute to be more predictable if its probability distribution conditioned on attributes in the previous month or year is sharper than its unconditional distribution. Conditional probability calculations reveal that phase is more predictable than state and extremes because it is likely to reverse sign between successive years and months. Also, dry years are more likely than wet years to be followed by extreme years. We also examine conditional probability relations between attributes at grid-level and all-India scales, revealing that all-India extremes entail widespread grid-level phase and state of the corresponding sign.**

I. INTRODUCTION

A large section of India's population depends on rainfall for agriculture [2]. Prediction of rainfall in India is a major scientific challenge [1], which meteorologists and others have attempted for decades using various approaches. Most approaches seek to predict All-India spatial aggregate rainfall (and sometimes for smaller "homogeneous zones"), either for entire years or for the monsoon season (June-September) that accounts for 75% of India's annual rainfall [7], [5]. Such prediction for any year includes annual rainfall in the past years, to take into account variability on different timescales, as well as climatic variables in certain parts of the world (e.g. sea surface temperature over Pacific and Indian Ocean) in the same year [9], [8]. This is a difficult task and the efforts are yet to yield adequate results [6]. Monthly rainfall prediction with long lead-times, such as in the previous year, is rarely attempted [11]. Moreover, prediction of all-India rainfall is not sufficient as Indian rainfall has a lot of spatial variation [10], and local predictions are necessary for all stakeholders [3], [4]. In this work, we consider the simpler task of prediction of certain attributes of annual and monthly rainfall, and the relationship between these attributes at all-India and local scales. We study how well these attributes can be predicted based only on their past values, without invoking external conditions. Since these are discrete attributes, we cannot use regression-based techniques which are used for prediction of all-India rainfall or ENSO indices. So we make use of discrete conditional distributions. This approach allows us to make probabilistic forecasts for the attributes, something which regression-based techniques do not allow.

Corresponding author: Adway Mitra, adway.cse@gmail.com[1]ICTS-TIFR, Bangalore, India [2]Divecha Center for Climate Change, IISc, Bangalore, India

II. PHASE, STATE, EXTREME

Let us denote by $Y(t)$ the aggregate annual rainfall over India in year t. Also, let μ and σ be the mean and standard deviation of Y, measured over a century. In years that $Y(t) > \mu$ we say Y is in *positive state*, $SY(t) = 1$, otherwise $SY(t) = -1$. Similarly, if $Y(t) > \mu + \sigma$ we call it a *positive extreme* with $ZY(t) = 1$, and if $Y(t) < \mu - \sigma$ it is a *negative extreme* with $ZY(t) = -1$, and otherwise it is a normal year with $ZY(t) = 0$. These two quantities have been used quite frequently in climate sciences [13], including for All-India annual rainfall [12]. Additionally, we also define *positive phase*, $PY(t) = 1$ if $Y(t) > Y(t-1)$ and negative phase $PY(t) = -1$ otherwise, i.e. phase is the direction of change in the time-series. These attributes can be defined for each location also, i.e. we can have annual time-series of local rainfall $X(s,t)$ based on which we can have local annual state SX, phase PX and extreme ZX.

We also consider corresponding quantities for monthly time-series $MY(t)$ and $MX(s,t)$ at all-India aggregate and local scales respectively. These time-series consider months across years in chronological

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

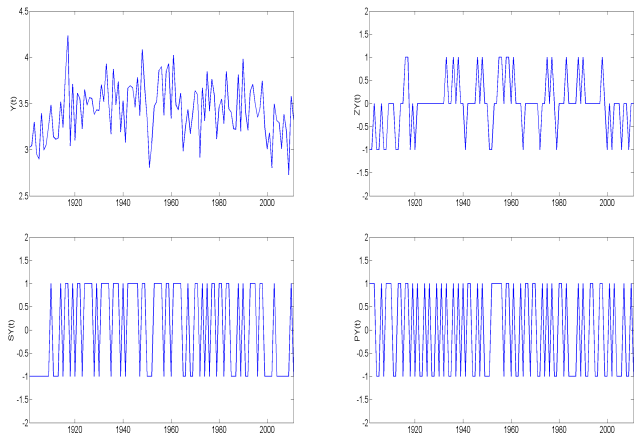


Fig. 1. Time-series of annual All-India Rainfall (AIR) and its attributes over 1901-2011: AIR $Y(t)$; Extremity $ZY(t)$; Phase $PY(t)$; State $SY(t)$ in clockwise order

order. However due to seasonal effects in Indian rainfall, different months have very different long-term means. So we consider mean and standard deviation μ_m, σ_m separately for each of the 12 calendar-months, and divide $MY(t)$ by μ_m where m is the calendar-month $MM(t)$ corresponding to month t. This is done specific to locations for MX. The phase, state, extreme (MPY, MSY, MZY) and (MPX, MSX, MZX) are computed accordingly.

III. Conditional Distributions of Attributes

Now, we consider the time-series of phase, state and extremes respectively. Figure 1 shows the time-series of Y, PY,SY and ZY for the period 1901-2011. These figures, in case of Y and PY, indicate a *mean-reverting behavior* - a high year is likely to be followed by a low year, and vice versa. This raises the possibility of prediction.

We aim to see if $PY(t)$, $SY(t)$, $ZY(t)$ can be predicted from $PY(t-1)$, $SY(t-1)$, $ZY(t-1)$. Since these are all binary variables we cannot use regression, so instead we estimate the conditional distributions of each attribute, based on the attributes in the previous year, i.e. $pr(PY(t)|PY(t-1)), pr(PY(t)|SY(t-1)), pr(SY(t)|ZY(t-1))$ and so on. They are computed as relative frequencies. These conditional distributions are shown Table 1, where each column stands for the condition, i.e. an attribute in the previous year $t-1$. Each row stands for the attribute's value in the current year t. We also evaluate the unconditional distribution of each attribute.

First of all, Table 1 shows that the unconditional distributions are not informative for making predictions, as the phase and state in any year can take any of the two values with almost equal frequency. But the distributions $pr(PY(t)|PY(t-1))$ are clearly "sharper", suggesting that the phase will reverse with around 66% probability. In other words, previous year's phase is a reasonable predictor for current year's phase. But previous year's state is an even better predictor, as $pr(PY(t)|SY(t-1))$ distribution is even sharper. Current year's phase will be the opposite of previous year's state with about 75% probability. Also, an extreme event of either type will almost surely cause the rainfall volume in the next year to change in the reverse direction, as indicated by as $prob(PY(t)|ZY(t-1))$. This tendency to reverse is called *Mean-reverting behavior*. These results show that phase for a given year can be predicted quite well from previous year's attributes, but this is not true for state, as indicated by conditional distributions $pr(SY(t)|SY(t-1))$, $pr(SY(t)|PY(t-1))$ and $pr(SY(t)|ZY(t-1))$. The conditional distributions are quite close to the unconditional distribution of $SY(t)$. Finally, both types of extremities are rare, but years of negative phase and negative state are more likely to be followed by a year of extremity of either type, while years of positive state/phase are more likely to be followed by a normal year. These general patterns hold for individual grid-locations also.

We carried out a statistical testing for these attributes, under the null hypothesis that the PY,SY,ZY time-series were drawn independently at each time-step according to the unconditional (marginal) distribution of the attributes. We simulated sample time-series according to this hypothesis, and the test statistic was whether or not the state transition probabilities were around the values measured in Table 1. The null hypothesis was rejected at 1% significance level.

We also explore distributions of these attributes, conditioned on multiple attributes of the previous year. For state and extreme this did not make much difference, but it did for phase when conditioned on both phase and state of the previous year. In particular, when both phase and state in previous year are same, the probability of current year's phase being the reverse is increased further to nearly 80%. But when $SY(t-1)$ and $PY(t-1)$ have opposite signs, $PY(t)$ is more likely to take the sign opposite to $SY(t-1)$.

These analyses of attributes have consequences for the annual aggregate rainfall $Y(t)$ also. In Figure 1 we plot the probability density function of $Y(t)$ conditioned on $PY(t-1)$,$SY(t-1)$ and $ZY(t-1)$. The results are clearly consistent with the attribute conditional distributions studied above, as the previous

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

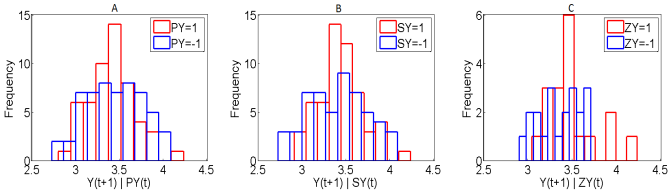


Fig. 2. Distribution of Y conditioned on attributes P, S, Z in previous year

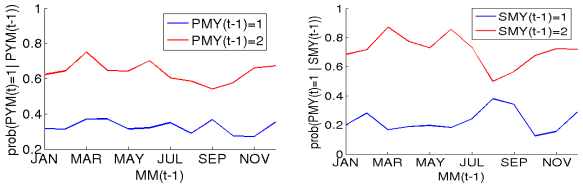


Fig. 3. Distribution of phase and state attributes for individual calendar-months conditioned on previous month

	. .	$PY=1$	$PY=-1$	$SY=1$	$SY=-1$	$ZY=1$	$ZY=-1$	$ZY=0$
$PY=1$	0.50	0.32	0.64	0.25	0.74	0.17	0.95	0.44
$PY=-1$	0.50	0.68	0.36	0.75	0.26	0.83	0.05	0.56
$SY=1$	0.52	0.51	0.54	0.52	0.53	0.50	0.45	0.55
$SY=-1$	0.48	0.49	0.46	0.48	0.47	0.50	0.55	0.45
$ZY=1$	0.16	0.09	0.23	0.14	0.19	0.17	0	0.21
$ZY=-1$	0.18	0.15	0.20	0.11	0.25	0.06	0.25	0.18
$ZY=0$	0.66	0.76	0.57	0.75	0.56	0.77	0.75	0.61

TABLE I
DISTRIBUTION OF CURRENT YEAR'S ATTRIBUTES (ALONG EACH ROW), CONDITIONED ON PREVIOUS YEAR'S ATTRIBUTES (ALONG EACH COLUMN)

	. .	$PMY=1$	$PMY=-1$	$SMY=1$	$SMY=-1$	$ZMY=1$	$ZMY=-1$	$ZMY=0$
$PMY=1$	0.49	0.33	0.64	0.23	0.71	0.11	0.86	0.49
$PMY=-1$	0.51	0.67	0.36	0.77	0.29	0.89	0.14	0.51
$SMY=1$	0.47	0.48	0.46	0.48	0.46	0.53	0.4	0.47
$SMY=-1$	0.53	0.52	0.54	0.52	0.54	0.47	0.6	0.53
$ZMY=1$	0.16	0.15	0.17	0.16	0.15	0.17	0.13	0.16
$ZMY=-1$	0.15	0.14	0.16	0.12	0.18	0.06	0.24	0.15
$ZMY=0$	0.69	0.71	0.67	0.72	0.67	0.77	0.63	0.69

TABLE II
DISTRIBUTION OF CURRENT MONTH'S ATTRIBUTES (ALONG EACH ROW), CONDITIONED ON PREVIOUS MONTH'S ATTRIBUTES (ALONG EACH COLUMN)

year's state or phase have a reverting effect on current year's rainfall.

Similarly, this is repeated for the monthly-scale attributes, where we compute conditional distributions like $pr(PMY(t)|SMY(t-1))$. These results are shown in Table 2. Broadly the same patterns of predictability hold in this case also, where the state SMY is not predictable because its unconditional and conditional distributions are all close to uniform. The phase variable PMY is usually the reverse of the previous month's phase and state, and an extremity in any month makes the phase reversal in the next month almost certain. Unlike the annual scale, the distribution of $ZMY(t)$ is not strongly affected by $PMY(t-1)$ and $SMY(t-1)$. But a positive extreme month has an increased probability of being followed by a normal month, while a negative extreme month is more likely to be followed by another extreme month.

The aforementioned results are aggregates across calendar months. In addition we examine for each calendar-month (e.g. July) how its phase and state can be predicted based on the attributes of the previous month (in case of July, based on June). This is important for making operational forecasts involving the progression of monsoon. The results are shown in Figure 2, where we plot the probability of positive phase in the next month, conditioned on the phase (left panel) and state (right panel) in the current calendar-month.

The figure shows that predictability of a month's phase based on the previous month's phase does not vary much across different calendar months, although negative phase in March, June, November or December indicate a higher probability (over 0.7) of positive phase in the following months, and positive phase in March, April and September indicate a reasonable probability (nearly 0.4) of positive phase in the following months also. Again, a negative state in March and June strongly indicate that the following month will have higher relative rainfall (positive phase), but negative state in August means September is unlikely to be any better. A positive state in August improves the chances of having a better September.

IV. ATTRIBUTES ACROSS SPATIAL SCALES

We now study another aspect - how attributes of all-India aggregate rainfall in a year affect those at grid-scale in the same year. In other words, we now study the conditional distribution of $PX(s,t), SX(s,t), ZX(s,t)$, based on $PY(t), SY(t), ZY(t)$. This is important because local information is very important for activities like agriculture, and yet most predictions are made at an aggregate scale. It is important to relate local attributes to all-India aggregate attributes.

We consider 357 grid-locations over India, each of size $100Km - 100Km$. For every location s, we count the fraction of years in which its phase agrees with the all-India phase, i.e. $|t : PX(s,t) = PY(t)|$. We repeat the same for state and extremities. Since we cannot show the results for all locations, we show in Figure 3 the distribution of these fractions. It turns out that most locations follow the all-India attributes in at least 50% of all the years, and some locations do so in even 70% of the years.

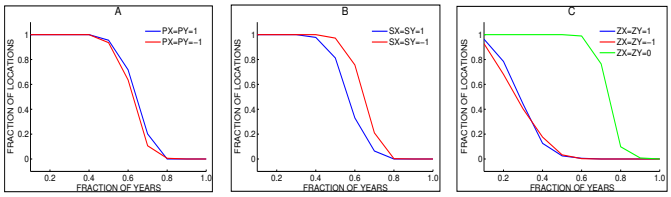


Fig. 4. The fractions of locations (Y-axis) which have their local attributes equal to attributes of AIR in at least a certain fraction of the years (X-axis).

	. .	$PY = 1$	$PY = -1$	$SY = 1$	$SY = -1$	$ZY = 1$	$ZY = -1$	$ZY = 0$
$E(PX = 1)$	0.51	0.64	0.38	0.58	0.43	**0.67**	0.41	0.49
$E(PX = -1)$	0.49	0.36	0.62	0.42	0.57	0.33	0.59	0.51
$E(SX = 1)$	0.47	0.55	0.39	0.57	0.36	**0.67**	0.27	0.47
$E(SX = -1)$	0.53	0.45	0.61	0.43	0.64	0.33	**0.73**	0.53
$E(ZX = 1)$	0.15	0.19	0.11	0.20	0.10	**0.29**	0.07	0.14
$E(ZX = -1)$	0.15	0.09	0.20	0.08	0.22	0.05	**0.30**	0.13
$E(ZX = 0)$	0.70	0.72	0.69	0.72	0.68	0.66	0.63	0.73

TABLE III
SPATIAL DISTRIBUTION OF LOCAL ANNUAL ATTRIBUTES CONDITIONED ON ALL-INDIA ANNUAL ATTRIBUTES IN SAME YEAR

Next, we compute the expected number of locations (spatial distribution) having different values of local attributes in each year, conditioned on the all-India attributes. The results are shown in Table 3, where each column stands for a value of an all-India attribute, and each row for a local attribute value. The unconditional spatial distribution of each attribute is also shown. Clearly, there is a general positive correlation between local and all-India attributes of the same type. But all-India extremes have the strongest association with local attributes, as they are associated with significantly more local extremes of the same type. Years of all-India extreme also entail widespread local phases and states of the corresponding sign. Similar patterns hold in the relations between monthly-scale attributes.

V. CONCLUSIONS

We made a study of simple attributes, specifically phase, of rainfall over India and noted that it can be predicted well from previous year's attributes. We also studied relations between attributes at different scales. Such predictions can help in policy-framing and climate simulations.

Acknowledgement This work was partially funded by Airbus India.

REFERENCES

[1] Gadgil S (2003), The Indian Monsoon and its Variability, *Annual Review of Earth and Planetary Sciences*, *31*, 429-467.

[2] Gadgil Sulochana and Gadgil Siddhartha (2010), The Indian Monsoon, GDP and Agriculture, *Economic and Political Weekly*, *41*, 4887-4895.

[3] Kavi Kumar K S (2011), Climate sensitivity of Indian agriculture: do spatial effects matter?, *Cambridge Journal of Regions, Economy and Society*, *4*, 221-235.

[4] Mall R K, R Singh R, Gupta A., Srinivasan G. and Rathore L. S. (2006), Impact of climate change on Indian agriculture: a review, *Climatic Change*, *78*, 445-478.

[5] Iyengar, RN and Kanth, STG Raghu (2005), Intrinsic Mode Functions and a Strategy for Forecasting Indian Monsoon Rainfall, *Meteorology and Atmospheric Physics*, *90*, 17-36

[6] Gadgil, S., Rajeevan, M., and Nanjundiah, R. (2005). Monsoon prediction-Why yet another failure?, *Current science, 88(9)*, 1389-1400.

[7] Rajeevan M, Pai D S, R Anil Kumar and B Lal (2007), New statistical models for long-range forecasting of southwest monsoon rainfall over India, *Climate Dynamics*, *28*, 813-828.

[8] Sahai A K, Grimm A M, Satyan V and Pant G B (2003), Long-lead prediction of Indian summer monsoon rainfall from global SST evolution, *Climate Dynamics*, *20*, 855-863.

[9] M. Saha, P. Mitra, and R.S.Nanjundiah (2016). Autoencoder-based identification of predictors of Indian monsoon, *Meteorology and Atmospheric Physics, 128(5)*, 613-628.

[10] Ghosh S, Luniya V, Gupta A (2009), Trend analysis of Indian summer monsoon rainfall at different spatial scales, *Atmospheric Science Letters*, *31*, 429-467.

[11] Kashid, S S and Maity, R (2012), Prediction of monthly rainfall on homogeneous monsoon regions of India based on large scale circulation patterns using Genetic Programming, *Journal of Hydrology*, 454, 26–41.

[12] R.S.Nanjundiah, P.A.Francis, M.Ved, and S.Gadgil (2013), Predicting the extremes of Indian Summer Monsoon rainfall with coupled ocean-atmosphere models, *Current Science*, *104 (10)*, 1380-1393.

[13] Hans von Storch and Francis W. Zwiers, Statistical Analysis in Climate Research, Cambridge University Press, 1999

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

MASSIVE SCALE DEEP LEARNING FOR DETECTING EXTREME CLIMATE EVENTS

Soo Kyung Kim[1], Sasha Ames[1], Jiwoo Lee[1], Chengzhu Zhang[1],Aaron C. Wilson[1], Dean Williams[1]

***Abstract*—Conventional extreme climate event detection relies on high spatial resolution climate model output for improved accuracy. It often poses significant computational challenges due to its tremendous iteration cost. As a cost-efficient alternative, we developed a system to detect and locate extreme climate events by deep learning. Our system can capture the pattern of extreme climate events from pre-existing coarse reanalysis data, corresponds to only 16 thousand grid points without expensive downscaling process with less than 5 hours to training our dataset, and less than 5 seconds to testing our test set using 5-layered Convolutional Neural Networks (CNNs). As the use case of our framework, we tested tropical cyclones detection with labeled reanalysis data and our cross validation results show 99.98% of detection accuracy and the localization accuracy is within 4.5 degrees of longitude/latitude (which is around 500 km, and is 3 times of data resolution).**

I. MOTIVATION

Deep learning, a subset of machine learning, is the latest iteration of neural network approaches that model intricate structure in large datasets. [1], [2] These models control how a machine should change its internal parameters that are used to compute the representation in each layer from the representation in the previous layer. Recent advances in deep learning have led to groundbreaking results in several domains with complex, nonlinear prediction functions, such as speech recognition, computer vision, etc. Due to its ability to capture the latent abstraction of massive scale complex data, deep learning is making similar inroads in scientific data analysis [3], and already have shone light on recent projects in laboratories, such as satellite image classification at NASA AMES [4], and analysis of human brain records at LBNL [5].

The large volume and complex nature of climate data pose many challenges to traditional analysis: there is already too much available data, namely peta-bytes of general calculation model (GCM) output than what can be analyzed efficiently. The data analysis processes using conventional tools is done essentially by hand, and requires considerable time from trained experts. [6], [7], [8], [9] Moreover, conventional climate research relies on Earth System Modeling, which increases the spatial resolution up to 1.3km through downscaling process with regional climate models (RCMs) started from the input of 100 to 300km scaled coarse resolution of GCM [10], [11]. The need to increase the spatial resolution of models for improved accuracy poses significant computational challenges due to its tremendous iteration cost. For instance, to produce regional climate model output at 1.3 km resolution, corresponding to the calculation of 256.8 million grid points per vertical level, requires the full-time execution of the model on a supercomputer with 130,000 cores (at NOAA). Exacerbating the problem, the computing power needed to run a model with N-times finer resolution increases by factor of 2^N. As a cost-efficient alternative, we developed event detection and localization system by deep learning which can capture the pattern of events from pre-existing coarse GCM scaled reanalysis data, corresponds to only 16 thousand grid points without expensive downscaling process with less than 10 hours to training our dataset.

As the used case of our framework, we tested tropical cyclones detection with labeled reanalysis data and achieve 99.98% of detection accuracy and 4.5 degree of localization accuracy. To our knowledge, these techniques have not been explored at the scales of data available in the climate domain, and thus, presents us a unique opportunity to break ground in that regard.

II. METHOD

Our system is consisted of two modules, such as detection CNNs and localization CNNs. Figure 1 shows schematic description of our framework.

In first stage, our in-house python program collects and adds labels indicating presence and location of extreme climate events on climate dataset. Labeling has been guided by historical report of specific extreme

[1]Lawrence Livermore National Laboratory, Livermore, CA

climate event. As one of the use cases, tropical cyclones has been chosen as extreme climate event. This is because: (1) there are well-constructed historical reports of tropical cyclones informing exact date and location of tropical cyclones [12], [13], and (2) It is believed that the climatic pattern of tropical cyclones is relatively simple than other extreme events [14]. Collected climate dataset is consisted of multi-channeled reanalysis data, and each channel represents different climate variables. Reanalysis data are global datesets that are generated by incorporating observations into a stable data assimilation system within climate models and are routinely used to study climate processes. Guided by the historical tropical cyclones reports, reanalysis data of five climate variables, - cloud fraction (clt), precipitation (pr), surface level pressure (psl), eastward near-surface wind (uas), northward near-surface wind (vas) - which are pertinent to a tropical cyclone diagnosis, have been collected, and labeled grid box as $true$ instance where and when the exact tropical cyclones occurred, and $false$ otherwise. Data with $true$ instance has been collected three times repeatedly on the same climate data map by randomly shifting grid box to sustain transient invariance. Data with $true$ instance has additional label representing relative (x, y) location of tropical cyclones inside of the bounding box which will be reconstructed as exact location of longitude and latitude later by combining with the location of its bounding box. The grid box has size of 20°(longitude)×20°(latitude), and resolution of collected reanalysis data is around 1.25°×1.25°. Therefore, the collected dataset is 5 channeled 2-D feature with size of 16×16. Figure 2 visualizes the exact location of tropical cyclones in our dataset overlaid with grid box. We have collected 109,000 incidents of tropical cyclone according to the JTWC tropical cyclone best track data [15] from 1979 to 2016. JMA reanalysis data (JRA-55) of different climate variables [16] has been downloaded through ESGF [17], [18], and selectively collected bounding boxes those containing tropical cyclones, utilizing CDMS module [19]. The total size of constructed tropical cyclone dataset is 218,000 with 50% of $true$ instances and 50 % of $false$ instance.

Once a labeled dataset is constructed, in second stage, the first convolutional neural networks, so called 'Detection CNNs', is trained for the binary classification of $true$ and $false$ instance of tropical cyclones. We have used 5-layered CNNs started by using a simple MNIST structure to begin with that uses all of the elements for state of the art results [20], [21], then optimized for our use case of tropical cyclone detection. As the optimizer,

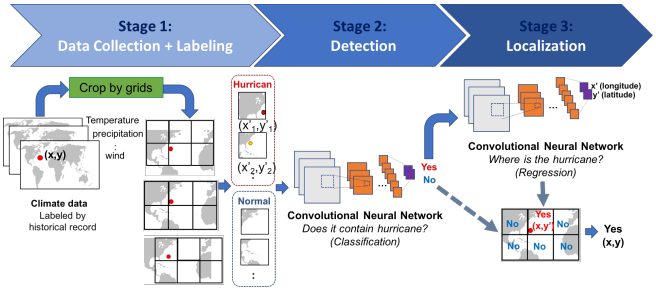


Fig. 1. Systemic framework for detection and localization of extreme climate event.

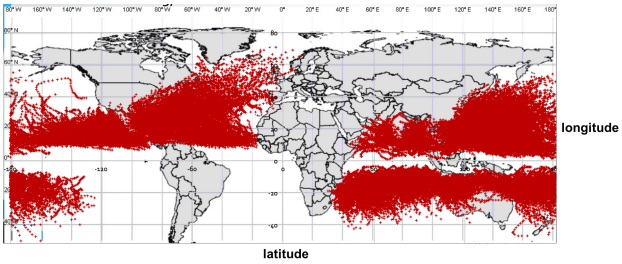


Fig. 2. Dataset: Visualization of historic cyclones from JWTC hurricane report from 1979 to 2016.

stochastic gradient descent [22] has been chosen with batch size of 32. We used Tensorflow-r12.0 to design and develop code [23]. The detailed architecture of detection CNNs is as follows:

1) Each layer contains one convolutional layer and pooling layer. Each convolutional layer is called a Convolution2D. All five convolutional layers are designed to have 64 feature maps, each with the size of 5×5 and a rectifier activation function.
2) Next we defined a pooling layer that takes the max called MaxPooling2D [24]. It is configured with a pool size of 2×2.
3) After passing five convolutional layers as described (1) and (2), the regularization layer followed using Dropout [25], which is configured to randomly exclude 20% of neurons in the layer in order to reduce overfitting.
4) Finally, the output layer has 2 neurons for the 2 classes, $true$ and $false$ instance, and a softmax activation function [26] to output probability-like predictions for each class.

To consider input with multiple channels, we used the late fusion [27] techniques which different CNNs take inputs of different channels, and combines outputs from different CNNs as single feature. Then, the final readout layer takes the combined feature to classify $true$ and $false$ instance. Our comparison test between late fusion

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

Fig. 3. Test accuracy of the first CNNs for Cyclone detection task.

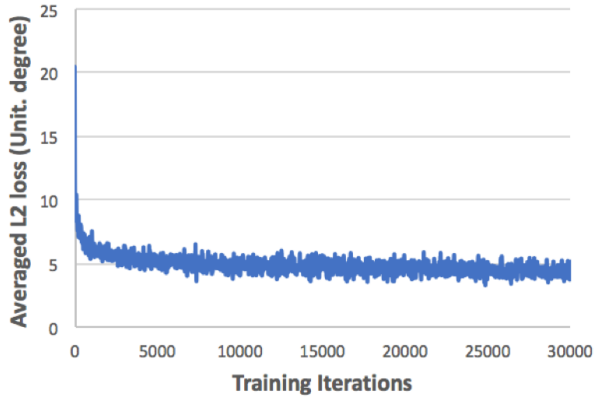


Fig. 4. Localization accuracy of the second CNNs for Cyclone localization task.

with early fusion, which uses input combined with multiple channels at the early stage and then use single CNNs to classify instance, shows late fusion performs better to detect tropical cyclones.

Lastly, in third stage, the second convolutional neural networks, so called 'Localization CNNs', is trained to solve the regression problem to predict location, (x, y), inside of $20° \times 20°$sized bounding box. We have used same CNNs architecture with 'Detection CNNs', except we are solving a regression problem rather than a classification problem. For localization CNNs, we only use the $true$ instance datasets (those with (x, y) labels) in training.

After finishing the procedure of training of two CNNs, the whole framework takes global scaled five-channeled reanalysis data as input, and global scaled input (global climate map) divided as multiple gridded boxes. Then, detection CNNs takes those gridded boxes and classify $true$ and $false$ instance. Localization CNNs can locate (x, y) of a tropical cyclone only for the gridded dataset classified as a $true$ instance within the detection CNNs. Therefore, our final output generated from the framework is the detection result of each grid box and location (x, y) in the $true$ grid box. In test procedure, $false$ grid box can generate (x, y) and we attempt not to update network weight when It's $false$ instance from detection CNN by multiplying detection result(1 for $true$ and 0 for $false$) to accuracy loss.

III. EVALUATION

We have used 80% of our datasets (with size of 872,000 data) for training CNNs and used the remaining 20%(with size of 21,800 data) hold-out for testing performance [28]. For each iteration, we shuffled whole data set and randomly selected 20% of test set. We iterate to training network and evaluating accuracy with different training set and test set until 200,000 time-steps after we achieved perfect convergence at 2000 steps. Therefore, detection and localization accuracy we got here is the cross-validation results. Figure 3 shows the accuracy of 'Detection CNNs' on our test set measured by training iterations. Accuracy has been calculated using reduced mean average of mini batch [29]. As Figure 3 shown, we achieved more than 99.98 % of detection accuracy of tropical cyclones after 2000 iteration.

Figure 4 shows the $l2$ loss with the unit of degree of 'Localization CNNs' on our test set measured by training iterations. As Figure 4 shows, we reduced $l2$ loss [28] of accurate localization of event down to 4.5 degrees after 10000 iterations. Considering one degree is around 111 km, our designed 'Localization CNNs' can locate tropical cyclone around 500 km off from the true center of tropical cyclones. Considering the fact that (1) the diameter of tropical cyclone is around 100 to 2000 km and (2) resolution of input is quite low as 1.25 degree (138km), our localization error is low enough which is capable to predict only 2.5 pixels offs from the true location.

Our results showing high detection accuracy and good localization performance, even with low quality (resolution) input, demonstrate that (1) there exists the specific pattern of tropical cyclones in multi-variabled climate model output, and (2) this pattern is distinguishable even in low resolution around 1.25 degrees. Our results suggest that its is possible to detect extreme climate events in low resolution model output, which can potentially save computational cost for conventional expensive downscaling process. Also, our approach can

be applied without scientific or algorithmic definition but only using data. Our model can be more adaptive to general purpose and massive data rather than classical method to detect climate events based on scientific knowledge.

ACKNOWLEDGMENTS

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. (LLNL-CONF-734529)

REFERENCES

[1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[2] J. Johnson, P. MacKeen, A. Witt, E. Mitchell, G. Stumpf, M. Eilts, and K. Thomas, "The storm cell identification and tracking algorithm: An enhanced WSR-88D algorithm," *Weather and Forecasting*, vol. 13, no. 2, pp. 263–276, 1998.

[3] C. Chin and D. E. Brown, "Learning in science: A comparison of deep and surface approaches," *Journal of research in science teaching*, vol. 37, no. 2, pp. 109–138, 2000.

[4] S. Basu, S. Ganguly, S. Mukhopadhyay, R. DiBiano, M. Karki, and R. Nemani, "Deepsat: a learning framework for satellite imagery," in *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems*, p. 37, ACM, 2015.

[5] S. Murugesan, K. Bouchard, E. Chang, M. Dougherty, B. Hamann, and G. H. Weber, "Multi-scale visual analysis of time-varying electrocorticography data via clustering of brain regions," *BMC bioinformatics*, vol. 18, no. 6, p. 236, 2017.

[6] T. R. Knutson, J. L. McBride, J. Chan, K. Emanuel, G. Holland, C. Landsea, I. Held, J. P. Kossin, A. Srivastava, and M. Sugi, "Tropical cyclones and climate change," *Nature Geoscience*, vol. 3, no. 3, pp. 157–163, 2010.

[7] M. Beniston, D. B. Stephenson, O. B. Christensen, C. A. Ferro, C. Frei, S. Goyette, K. Halsnaes, T. Holt, K. Jylhä, B. Koffi, *et al.*, "Future extreme events in european climate: an exploration of regional climate model projections," *Climatic change*, vol. 81, pp. 71–95, 2007.

[8] E. Sánchez, C. Gallardo, M. Gaertner, A. Arribas, and M. Castro, "Future climate extreme events in the mediterranean simulated by a regional climate model: a first approach," *Global and Planetary Change*, vol. 44, no. 1, pp. 163–180, 2004.

[9] O. Rübel, S. Byna, K. Wu, F. Li, M. Wehner, W. Bethel, *et al.*, "Teca: A parallel toolkit for extreme climate analysis," *Procedia Computer Science*, vol. 9, pp. 866–876, 2012.

[10] B. Hewitson and R. Crane, "Climate downscaling: techniques and application," *Climate Research*, pp. 85–95, 1996.

[11] 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.

[12] M.-C. Wu, K.-H. Yeung, and W.-L. Chang, "Trends in western north pacific tropical cyclone intensity," *Eos, Transactions American Geophysical Union*, vol. 87, no. 48, pp. 537–538, 2006.

[13] J.-H. Chu, C. R. Sampson, A. S. Levine, and E. Fukada, "The joint typhoon warning center tropical cyclone best-tracks, 1945–2000," *Ref. NRL/MR/7540-02*, vol. 16, 2002.

[14] V. F. Dvorak, *Tropical cyclone intensity analysis using satellite data*, vol. 11. US Department of Commerce, National Oceanic and Atmospheric Administration, National Environmental Satellite, Data, and Information Service, 1984.

[15] J. H. Chu, " JTWC Tropical Cyclone Best Track Data Site." http://www.usno.navy.mil/NOOC/nmfc-ph/RSS/jtwc/best_tracks/, 2015. [Online; accessed 9-July-2015].

[16] A. Ebita, S. Kobayashi, Y. Ota, M. Moriya, R. Kumabe, K. Onogi, Y. Harada, S. Yasui, K. Miyaoka, K. Takahashi, *et al.*, "The japanese 55-year reanalysis jra-55: an interim report," *Sola*, vol. 7, pp. 149–152, 2011.

[17] D. N. Williams, "Earth system grid federation (esgf): Future and governance world climate research programme (wcrp)," *Working Group on Coupled Modelling (WGCM)Stakeholders and ESGF*, pp. 1–17, 2012.

[18] D. N. Williams, "Earth System Grid Federation ." https://esgf.llnl.gov, 2017. [Online; accessed 3-July-2017].

[19] D. N. Williams, R. S. Drach, P. F. Dubois, C. Doutriaux, C. J. OConnor, K. M. AchutaRao, and M. Fiorino, "Climate data analysis tool: An open software system approach," in *13th Symp. on Global Change and Climate Variations*, 2002.

[20] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, *et al.*, "Learning algorithms for classification: A comparison on handwritten digit recognition," *Neural networks: the statistical mechanics perspective*, vol. 261, p. 276, 1995.

[21] D. C. Ciresan, U. Meier, L. M. Gambardella, and J. Schmidhuber, "Convolutional neural network committees for handwritten character classification," in *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, pp. 1135–1139, IEEE, 2011.

[22] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of COMPSTAT'2010*, pp. 177–186, Springer, 2010.

[23] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[24] J. Nagi, F. Ducatelle, G. A. Di Caro, D. Cireşan, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella, "Max-pooling convolutional neural networks for vision-based hand gesture recognition," in *Signal and Image Processing Applications (ICSIPA), 2011 IEEE International Conference on*, pp. 342–347, IEEE, 2011.

[25] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[26] D. Heckerman and C. Meek, "Models and selection criteria for regression and classification," in *Proceedings of the Thirteenth conference on Uncertainty in artificial intelligence*, pp. 223–228, Morgan Kaufmann Publishers Inc., 1997.

[27] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732, 2014.

[28] E. Alpaydin, *Introduction to machine learning*. MIT press, 2014.

[29] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE transactions on information theory*, vol. 14, no. 1, pp. 55–63, 1968.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

SENSITIVITY OF GLOBAL ECOSYSTEMS TO CLIMATE ANOMALIES IN OBSERVATIONS AND EARTH SYSTEM MODELS

Matthias Demuzere[1], Stijn Decubber[2], Diego Miralles[1], Christina Papagiannopoulou[2], Willem Waegeman[2], Niko Verhoest[1], Wouter Dorigo[3]

***Abstract*—Vegetation is a key player in the climate system, constraining atmospheric conditions through a series of feedbacks. This fundamental role highlights the importance of understanding regional drivers of ecological sensitivity and the response of vegetation to climatic changes. While nutrient availability and short-term disturbances can be crucial for vegetation at various spatiotemporal scales, natural vegetation dynamics are overall driven by climate. At monthly scales, the interactions between vegetation and climate become complex: some vegetation types react preferentially to specific climatic changes, with different levels of intensity, resilience and lagged response. For our current Earth System Models (ESMs) being able to capture this complexity is crucial but extremely challenging. This adds uncertainty to our projections of future climate and the fate of global ecosystems. Here, following a Granger causality framework based on a random forest (RF) predictive model, we exploit the current wealth of satellite data records to uncover the main climatic drivers of monthly vegetation variability globally. Results based on three decades of satellite data indicate that water availability is the most dominant factor driving vegetation in over 60% of the vegetated land. These observation-based results will then used to benchmark ESMs on their representation of vegetation sensitivity to climate and climatic extremes.**

I. MOTIVATION

Vegetation takes on a central position in the climate system, affecting atmospheric conditions through a series of positive and negative feedbacks. Plants regulate water, energy and carbon cycles, through their transfer of vapour from land to atmosphere (i.e. transpiration, interception loss), effects on the surface radiation budget (e.g. albedo, surface temperature, emission of volatile organic compounds), exchange of carbon dioxide with the atmosphere (i.e. photosynthesis, respiration), and influence on wind circulation ([1], [2], [3], [4]). Simultaneously, vegetation dynamics and the distribution of ecosystems are largely driven by the availability of light, temperature, and water; thus, they are mostly sensitive to climate conditions ([5], [6], [7]). Because of the strong two-way relationship between terrestrial vegetation and climate variability, predictions of future climate can be improved through a better understanding of the vegetation response to past climate variability.

The current wealth of earth observation data can be used for this purpose. Nowadays, independent sensors on different platforms collect optical, thermal, microwave, altimetry, and gravimetry information, with the longest composite records of environmental and climatic variables spanning up to 35 years. This enables the study of multidecadal climatebiosphere interactions. Simple correlation statistics and multi-linear regressions using some of these data sets have led to important steps forward in understanding the links between vegetation and climate (e.g. [5], [8], [9]). However, these methods in general are insufficient when it comes to assessing causality, particularly in systems like the land-atmosphere continuum in which complex feedback mechanisms are involved. Therefore, as an extension of linear Granger-causality analysis, this work presents a novel non-linear framework consisting of several components, such as data collection from various databases, time series decomposition techniques, feature construction methods, and predictive modelling by means of RFs.

II. METHOD

Given a particular target time series, one speaks of the existence of 'Granger causality' if the prediction of this target variable improves when information from

Corresponding author: M. Demuzere, matthias.demuzere@ugent.be [1]Laboratory of Hydrology and Water Management, Ghent University, Ghent, Belgium [2]Depart. of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium [3]Depart. of Geodesy and Geo-Information, Vienna University of Technology, Vienna, Austria

other time series is taken into account in this prediction ([10]). Here, we quantify the extent to which a variable $\boldsymbol{x}$ (i.e. a predictive feature) is 'Granger-causing' a target variable $\boldsymbol{y}$ (i.e. a residual vegetation index such as Normalized Difference Vegetation Index (NDVI) or Leaf Area Index (LAI)) by computing the increase in the variance of $\boldsymbol{y}$ that is explained by the non-linear random forest ([11]) model predictions when $\boldsymbol{x}$ is included in the set of predictive features used by the model (this set also includes past values of $\boldsymbol{y}$ to conform to the definition of Granger causality). In the remainder of the paper, the 'baseline' model refers to including only past values of $\boldsymbol{y}$, while the 'full model' incorporates all available information. The explained variance used here is defined as $R^2 = 1 - \frac{RSS}{TSS}$, with RSS being the sum of squared errors of the predictions (relative to the true target residuals), and TSS being the sum of the squared differences between the true values and their long-term mean.

In our experiments, we treat each continental pixel as a separate problem for the RF regressor implementation, with the number of trees equal to 100 and the maximum number of predictor variables per node equal to the square root of the total number of predictor variables. Changes in these parameters or in the randomness of the algorithm do not cause substantial changes in the results. Model performance is assessed by means of 5-fold cross-validation ([12]). The window length is fixed to 12 months because initial experimental results revealed that longer time windows did not lead to improvements in the predictions. For more details on the methodology, please refer to [13], [7].

Input features to the RF framework used to predict NDVI/LAI residuals have been selected from the current pool of satellite and in situ observations on the basis of meeting a series of spatiotemporal requirements: (a) expected relevance of the variable for driving vegetation dynamics, (b) multidecadal record and global coverage available, and (c) adequate spatial and temporal resolution. The selected data sets can be classified into three different categories: water availability (including precipitation, snow water equivalent, and soil moisture data sets), temperature (both for the land surface and the near-surface atmosphere), and radiation (considering different radiative fluxes independently). Rather than using a single data set for each variable, all data sets meeting the above requirements have been selected, leading to a total of 21 different data sets ([13], [7]). They span the study period 19812010 at the global scale and have been converted to a common monthly temporal resolution and 1° x 1° latitudelongitude spatial resolution. On top of considering raw and anomaly time series, also 'higher-level' features are taken into account, consisting of 1) lagged variables (with monthly lags up to six months into the past), 2) cumulative variables (corresponding to the cumulative mean over the antecedent one to six months), and 3) extreme indices (including the maximum and minimum of a variable per month, number of days per month exceeding a given threshold, values of specific percentiles, etc.). To conclude, as a proxy for the state and activity of vegetation, we use the third-generation (3G) Global Inventory Modelling and Mapping Studies (GIMMS) satellite-based NDVI and LAI datasets ([14]).

The observation-based results are then used to benchmark Earth System Model's on their representation of vegetation sensitivity to climate and climatic extremes. ESMs are selected from the Coupled Model Intercomparison Project Phase 5 (CMIP5) based on their availability of daily output for all variables of interest. This resulted in a subset of six ESMs: BCC-CSM1 ([15]), GFDL-ESM2G ([16]) and MIROC-ESM ([17]), BNU-ESM (http://esg.bnu.edu.cn), CAN-ESM2 ([18]) and INM-CM4 ([19]). Note that the first three ESMs include a dynamic global vegetation model (DGVM), which calculates interactive vegetation variation (biomass and coverage) due to climate change simulated by the atmospheric model component. Similar as in [3], robustness is increased by selecting 50 years of data for the model analysis (1956-2005) rather than the shorter 35-year period used for the observational analysis.

III. EVALUATION

The main results on detecting linear and non-linear Granger-causality relationships targeting NDVI residuals are provided in [13]. Yet, since ESMs do not model vegetation greenness expressed by NDVI, the methodology has been repeated using GIMMS' LAI data (Figure 1). The results are in line with those from [13], providing confidence in both the methodology as well as the use of the LAI residuals as a target variable.

Figure 1 depicts strong regional differences in the non-linear relationships between vegetation and climate. Especially in semi-arid regions such as eg. Australia, Africa, and Central and North America, which are frequently exposed to water limitations. In those regions, more that 40% of the variance of LAI anomalies can be explained by antecedent climate variability (see also [13]). On the other hand, the variance of LAI explained in other areas, such as the Eurasian taiga, tropical rainforests, or China, is often below 10%. We hypothesise two potential reasons: (a) the

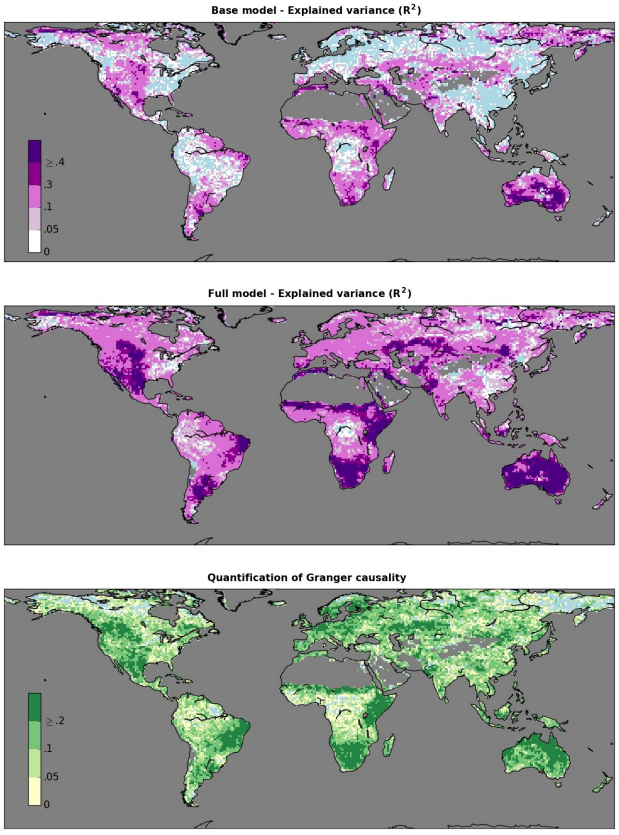


Fig. 1. Non-linear Granger causality of climate on vegetation. (Top) Explained variance (R^2) of LAI residuals based on the baseline RF model. (Middle) Explained variance (R^2) of LAI residuals based on a full RF model in which all climatic variables are included as predictors. (Bottom) Improvement in terms of R^2 by the full RF model with respect to the baseline RF model that uses only past values of LAI residuals as predictors; positive values indicate (non-linear) Granger causality. Blue shades refer to regions where the RF models have no skill ($R^2 < 0$).

uncertainty in the observations used as target and predictors are typically larger in these regions (especially in tropical forests and at higher latitudes), and (b) these are regions in which vegetation anomalies are not necessarily primarily controlled by climate but may be predominantly driven by phenological and biotic factors ([20]), occurrence of wildfires ([21]), limitations imposed by the availability of soil nutrients ([22]), or agricultural practices ([23]).

In order to further investigate the impact of replacing NDVI with LAI as a target value, the results are clustered regionally according to the International Geosphere-Biosphere Program (IGBP) land cover classification ([24]). Figure 2 shows both the full RF results for NDVI, taken from [13], as well as the results for the full RF model targeting LAI. First of all, the scatter plot reveals good correspondence between targeting both vegetation indices. In addition, one can clearly see that the full model outperforms the baseline model in all IGBP land cover classes, i.e. that Granger causality exists for all these biomes. Moreover, the error bars indicate that the variances of the two models are analogous; i.e. they are low or high in both models in the same land cover class.

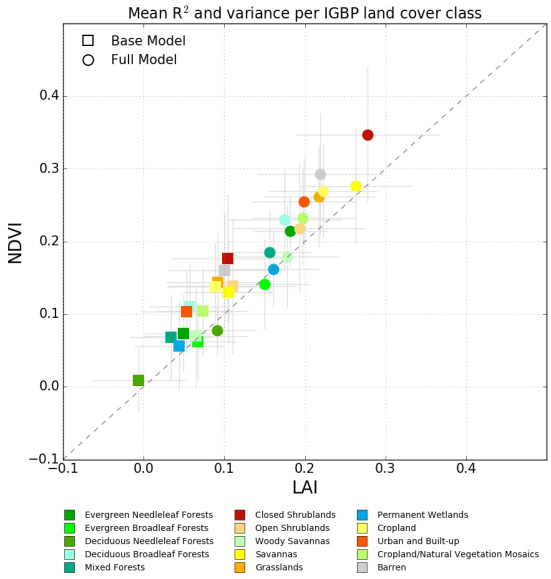


Fig. 2. Mean R^2 and variance per IGBP land cover class for the baseline (squares) and full (circles) RF model, targeting both NDVI and LAI residuals. Note that the NDVI results correspond to those presented in [13].

Based on the non-linear Granger-causality framework, [7] indicate that water availability is the most dominant factor driving vegetation globally: about 61% of the vegetated surface was primarily water-limited during 1981-2010 (results not shown here). This dependency of global vegetation on water availability is substantially larger than previously reported. For more details on the results, please refer to [7]. In a following (future) step, an identical set of features will be extracted from the previously defined ESM subset, and will serve as input to the same non-linear Granger-causality framework. This will allow the exploration of the sensitivity of vegetation to climatic anomalies in an ESM framework.

ACKNOWLEDGEMENTS

This work is funded by the Belgian Science Policy Office (BELSPO) in the framework of the STEREO III programme, project SAT-EX (SR/00/306). D. G. Miralles acknowledges support from the European

Research Council (ERC) under grant agreement no. 715254 (DRY-2-DRY). W. Dorigo is supported by the TU Wien Wissenschaftspreis 2015, a personal grant awarded by the Vienna University of Technology. We thank the individual developers of the wide range of global data sets used in this study. Finally, we acknowledge the World Climate Research Programmes Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modelling groups for producing and making available their model output.

REFERENCES

[1] G. B. Bonan, "Forests and climate change: forcings, feedbacks, and the climate benefits of forests.," *Science*, vol. 320, no. 5882, pp. 1444–1449, 2008.

[2] Z. Zeng, S. Piao, L. Z. X. Li, L. Zhou, P. Ciais, T. Wang, Y. Li, X. Lian, E. F. Wood, P. Friedlingstein, J. Mao, L. D. Estes, R. Myneni, S. Peng, X. Shi, S. I. Seneviratne, and Y. Wang, "Climate mitigation from vegetation biophysical feedbacks during the past three decades," *Nature Climate Change*, 2017.

[3] J. K. Green, A. G. Konings, S. H. Alemohammad, J. Berry, D. Entekhabi, J. Kolassa, J.-E. Lee, and P. Gentine, "Regionally strong feedbacks between the atmosphere and terrestrial biosphere," *Nature Geoscience*, no. May, 2017.

[4] A. J. Teuling, C. M. Taylor, J. F. Meirink, L. A. Melsen, D. G. Miralles, C. C. van Heerwaarden, R. Vautard, A. I. Stegehuis, G.-J. Nabuurs, and J. V.-G. de Arellano, "Observational evidence for cloud cover enhancement over western European forests," *Nature Communications*, vol. 8, p. 14065, 2017.

[5] R. R. Nemani, C. D. Keeling, H. Hashimoto, W. M. Jolly, S. C. Piper, C. J. Tucker, R. B. Myneni, and S. W. Running, "Climate-driven increases in global terrestrial net primary production from 1982 to 1999.," *Science (New York, N.Y.)*, vol. 300, no. 5625, pp. 1560–3, 2003.

[6] A. W. Seddon, M. Macias-Fauria, P. R. Long, D. Benz, and K. J. Willis, "Sensitivity of global terrestrial ecosystems to climate variability," *Nature*, vol. 531, no. 7593, pp. 229–232, 2016.

[7] C. Papagiannopoulou, D. G. Miralles, W. A. Dorigo, N. E. Verhoest, M. Depoorter, and W. Waegeman, "Vegetation anomalies caused by antecedent precipitation in most of the world," *Environmental Research Letters*, vol. 12, 2017.

[8] J. Barichivich, K. R. Briffa, R. Myneni, G. van der Schrier, W. Dorigo, C. J. Tucker, T. J. Osborn, and T. M. Melvin, "Temperature and snow-mediated moisture controls of summer photosynthetic activity in northern terrestrial ecosystems between 1982 and 2011," *Remote Sensing*, vol. 6, no. 2, pp. 1390–1431, 2014.

[9] D. Wu, X. Zhao, S. Liang, T. Zhou, K. Huang, B. Tang, and W. Zhao, "Time-lag effects of global vegetation responses to climate change," *Global Change Biology*, vol. 21, no. 9, pp. 3520–3531, 2015.

[10] C. Granger, "Investigating Causal Relations by Econometric Models and Cross-spectral Methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[11] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[12] H. von Storch and F. W. Zwiers, *Statistical analysis in climate research*. Cambridge University Press, 2001.

[13] C. Papagiannopoulou, D. G. Miralles, S. Decubber, M. Demuzere, N. E. C. Verhoest, W. A. Dorigo, and W. Waegeman, "A non-linear Granger causality framework to investigate climate–vegetation dynamics," *Geoscientific Model Development*, vol. 10, pp. 1945–1960, 2017.

[14] Z. Zhu, J. Bi, Y. Pan, S. Ganguly, and A. Anav, "Global data sets of vegetation leaf area index (LAI)3g and fraction of photosynthetically active radiation (FPAR)3g derived from global inventory modeling and mapping studies (GIMMS) normalized difference vegetation index (NDVI3G) for the period 1981 to 2," *Remote sensing*, pp. 1–16, 2013.

[15] T. Wu, W. Li, J. Ji, X. Xin, L. Li, Z. Wang, Y. Zhang, J. Li, F. Zhang, M. Wei, X. Shi, F. Wu, L. Zhang, M. Chu, W. Jie, Y. Liu, F. Wang, X. Liu, Q. Li, M. Dong, X. Liang, Y. Gao, and J. Zhang, "Global carbon budgets simulated by the Beijing Climate Center Climate System Model for the last century," *Journal of Geophysical Research Atmospheres*, vol. 118, no. 10, pp. 4326–4347, 2013.

[16] J. P. Dunne, J. G. John, S. Shevliakova, R. J. Stouffer, J. P. Krasting, S. L. Malyshev, P. C. D. Milly, L. T. Sentman, A. J. Adcroft, W. Cooke, K. A. Dunne, S. M. Griffies, R. W. Hallberg, M. J. Harrison, H. Levy, A. T. Wittenberg, P. J. Phillips, and N. Zadeh, "GFDL's ESM2 global coupled climate-carbon earth system models. Part II: Carbon system formulation and baseline simulation characteristics," *Journal of Climate*, vol. 26, no. 7, pp. 2247–2267, 2013.

[17] S. Watanabe, T. Hajima, K. Sudo, T. Nagashima, T. Takemura, H. Okajima, T. Nozawa, H. Kawase, M. Abe, T. Yokohata, T. Ise, H. Sato, E. Kato, K. Takata, S. Emori, and M. Kawamiya, "MIROC-ESM 2010: Model description and basic results of CMIP5-20c3m experiments," *Geoscientific Model Development*, vol. 4, no. 4, pp. 845–872, 2011.

[18] V. K. Arora, J. F. Scinocca, G. J. Boer, J. R. Christian, K. L. Denman, G. M. Flato, V. V. Kharin, W. G. Lee, and W. J. Merryfield, "Carbon emission limits required to satisfy future representative concentration pathways of greenhouse gases," *Geophysical Research Letters*, vol. 38, no. 5, pp. 3–8, 2011.

[19] E. M. Volodin, N. a. Dianskii, and a. V. Gusev, "Simulating present-day climate with the INMCM4.0 coupled model of the atmospheric and oceanic general circulations," *Izvestiya, Atmospheric and Oceanic Physics*, vol. 46, no. 4, pp. 414–431, 2010.

[20] L. R. Hutyra, J. W. Munger, S. R. Saleska, E. Gottlieb, B. C. Daube, A. L. Dunn, D. F. Amaral, P. B. de Camargo, and S. C. Wofsy, "Seasonal controls on the exchange of carbon and water in an Amazonian rain forest," *Journal of Geophysical Research: Biogeosciences*, vol. 112, no. 3, pp. 1–16, 2007.

[21] G. R. Van Der Werf, J. T. Randerson, L. Giglio, G. J. Collatz, M. Mu, P. S. Kasibhatla, D. C. Morton, R. S. Defries, Y. Jin, and T. T. Van Leeuwen, "Global fire emissions and the contribution of deforestation, savanna, forest, agricultural, and peat fires (1997-2009)," *Atmospheric Chemistry and Physics*, vol. 10, no. 23, pp. 11707–11735, 2010.

[22] J. B. Fisher, G. Badgley, and E. Blyth, "Global nutrient limitation in terrestrial vegetation," *Global Biogeochemical Cycles*, vol. 26, no. 3, pp. 1–9, 2012.

[23] Y. Liu, Z. Pan, Q. Zhuang, D. G. Miralles, A. J. Teuling, T. Zhang, P. An, Z. Dong, J. Zhang, D. He, L. Wang, X. Pan, W. Bai, and D. Niyogi, "Agriculture intensifies soil moisture decline in Northern China.," *Scientific reports*, vol. 5, p. 11261, 2015.

[24] T. R. Loveland and a. S. Belward, "The IGBP-DIS global 1km land cover data set, DISCover: First results," *International Journal of Remote Sensing*, vol. 18, no. 15, pp. 3289–3295, 1997.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

EXTRACTING MODES OF VARIABILITY AND CHANGE FROM CLIMATE MODEL ENSEMBLES

Robert C. Wills[1], David S. Battisti[1], Dennis L. Hartmann[1], Tapio Schneider[2]

***Abstract*—Ensembles of climate model simulations are commonly used to separate externally forced climate change from internal climate variability. However, much of the information gained from running large ensembles is lost in traditional methods of data reduction such as linear trend analysis or large-scale averaging. In this paper, we describe a statistical method to extract patterns of low-frequency variability and change from large ensembles. We demonstrate how this method characterizes modes of forced climate change (e.g., global warming) and low-frequency internal variability (e.g., the Pacific decadal oscillation) in the CESM large ensemble.**

I. MOTIVATION

Internal climate variability gives rise to uncertainty in long-term climate predictions [1]. Ensembles of climate model simulations are often used to quantify this uncertainty and to better understand the average response to external forcing [2], [3], [4]. Separating the forced response from the internal variability also helps to understand multi-decadal internal variability [5], which may lead to better decadal climate predictions [6]. However, most climate studies diagnose the spatial pattern of climate change by computing linear trends and/or diagnose the temporal behavior of climate variability by studying large-scale spatial averages. These methods of dimension reduction loose valuable information about the complex spatiotemporal structure of climate variability and change.

Principal component analysis (PCA) provides spatiotemporal information about the modes of variability that explain the most variance in a dataset. However, by maximizing variance, PCA can mix together physically distinct modes of variability such as global warming and the El Niño–Southern Oscillation (ENSO). One method to correct for this mode mixing is to look for linear combinations of the empirical orthogonal functions (EOFs) that maximize a particular type of variance representing a "signal" compared to "noise" that exists within internal variability or amongst realizations, so-called optimal filtering or signal-to-noise maximizing EOF analysis [7], [8], [9], [10], [11]. These methods take advantage of any spatial structure in the "noise" to optimally filter it out. Here, we use low-frequency component analysis (LFCA, [12]) to find patterns with the maximum ratio of low-frequency (signal) to high-frequency (noise) variance, correcting for mode mixing based on differences in time scale between physically plausible modes of variability.

Corresponding author: R. C. Wills, rcwills@uw.edu [1]Department of Atmospheric Sciences, University of Washington, Seattle, WA [2]Department of Environmental Sciences and Engineering, California Institute of Technology, Pasadena, CA

II. METHOD

The basic assumption behind our approach is that externally forced climate change operates on longer time scales than most internal variability. We can thus isolate patterns of climate change by solving for spatial patterns that describe variability with the maximum ratio of low-frequency to total variance, where low-frequency variance is defined as the variance remaining after application of a lowpass filter. LFCA provides an algorithm to find such spatial patterns for a truncated basis of EOFs. This method orders modes by their ratio of low-frequency to total variance, providing orthogonal indices of climate variability that tend to be ordered by time scale. For example, it separates global warming, the Pacific Decadal Oscillation (PDO), and ENSO in observed Pacific SSTs [12]. Here, we generalize LFCA for application to climate model ensembles. Our *ensemble LFCA* method is as follows:

1. **Compute ensemble covariance matrix.** For an ensemble of n_E climate model simulations, each with $n \times p$ data matrix X_i, we compute the $p \times p$ covariance matrix C_i with respect to either (a) the ensemble-mean climatology vector $\mathbf{x}_E$ or (b) the individual-ensemble-member climatology vector $\mathbf{x}_i$. Option (b) discards differences in climatology between ensemble members, while option (a) does not.

2. **Ensemble EOF analysis.** We compute the EOFs $\mathbf{a}_k$, which are the eigenvectors of the ensemble covari-

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

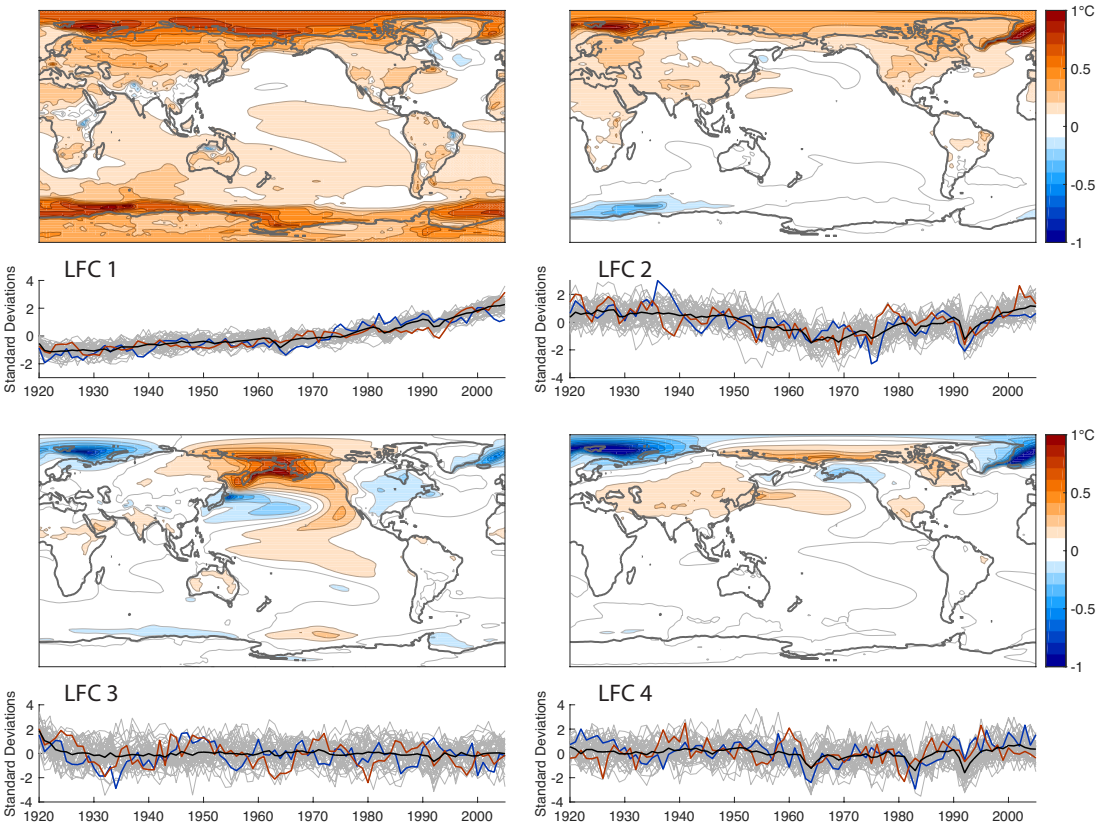


Fig. 1. Low-frequency patterns (LFPs) and components (LFCs) of the CESM large-ensemble historical simulation, with $N = 25$ EOFs retained and maximization of the variance remaining after application of Lanczos lowpass filter with cutoff $\tau = 10$ years. Orange (blue) lines show the ensemble member with the most (least) change in LFC 1 over the last 20 years. All other ensemble members are shown with grey lines. A black line shows the average of the LFC over all ensemble members.

ance matrix

$$C_E = n_E^{-1} \Sigma_{i=1}^{n_E} C_i. \tag{1}$$

The EOFs are normalized $||\mathbf{a}_k|| = 1$, such that the principal components have unit variance and the corresponding eigenvalue $\sigma_k^2 = \mathbf{a}_k^T C_e \mathbf{a}_k$ gives the variance associated with the kth EOF.

3. **Low-frequency component analysis.** We apply the LFCA algorithm [12] (see also [10]) to find the linear combination of the first N EOFs,

$$\mathbf{u}_k = \begin{bmatrix} \frac{\mathbf{a}_1}{\sigma_1} & \frac{\mathbf{a}_2}{\sigma_2} & \dots & \frac{\mathbf{a}_N}{\sigma_N} \end{bmatrix} \mathbf{e}_k, \tag{2}$$

such that the ratio of low-frequency to total variance

$$r_k = \frac{\left(\widetilde{X}_E \mathbf{u}_k\right)^T \widetilde{X}_E \mathbf{u}_k}{\left(X_E \mathbf{u}_k\right)^T X_E \mathbf{u}_k} \tag{3}$$

is maximized when the data is projected onto it. Here, X_E is the full-ensemble data matrix, obtained by concatenating individual-ensemble-member data matrices $X_E = [X_1^T \; X_2^T \; \dots \; X_{n_E}^T]^T$, and $\widetilde{X}_E$ is the lowpass-filtered full-ensemble data matrix, obtained by concatenating lowpass-filtered data matrices (i.e., we do not filter over the discontinuities between ensemble members). In practice, the linear combination coefficients $\mathbf{e}_k$ are computed as the eigenvectors of the covariance matrix of the first N lowpass-filtered principal components, such that filtering only needs to be applied to an $n \cdot n_E \times N$ matrix of the leading principal components (see derivation in [12]).

4. **Visualizing results.** The result is low-frequency components (LFCs) given by

$$\text{LFC}_k = X_E \mathbf{u}_k \tag{4}$$

and low-frequency patterns (LFPs) given by

$$\mathbf{v}_k = X_E^T \text{LFC}_k = [\sigma_1 \mathbf{a}_1 \; \sigma_2 \mathbf{a}_2 \; \dots \; \sigma_N \mathbf{a}_N] \, \mathbf{e}_k. \tag{5}$$

These are analogous to principal components and EOFs in PCA. The linear coefficients are normalized $||\mathbf{e}_k|| = 1$, such that the LFCs have unit variance and the LFPs show the spatial pattern associated with a 1-standard-deviation anomaly in the corresponding LFC.

III. RESULTS

We demonstrate our method by applying it to annual-mean surface temperatures from a 40-member ensemble of "historical" simulations with the Community Earth System Model (CESM) [4]. These simulations simulate climate from 1920 to 2005 based on historical forcing by greenhouse gasses, anthropogenic and volcanic aerosols, and ozone. The ensemble members differ only by machine-precession perturbations in their atmospheric initial condition in 1920, such that their climatology vectors $\mathbf{x}_i$ differ only as a result of internal variability. We include these climatology differences by using option (a) in step 1. We retain $N = 25$ EOFs in the LFCA and use a Lanczos filter with lowpass cutoff $\tau = 10$ years to focus on multi-decadal variability. The results are insensitive to the choice of cutoff for $\tau > 5$ years. However, there is no good criterion for choosing N, so in practice one must look for results that are robust across parameters (see discussion in [12]).

The first LFP shows a global warming pattern, with amplified warming over land and at high latitudes (Fig. 1). The associated LFC increases by 3 standard deviations from 1920 to 2005, emerging well beyond the ensemble spread. The second LFP/LFC shows cooling of the North Atlantic, Arctic, and Northern Hemisphere land through the 1950s and 60s and a subsequent recovery (Fig. 1), corresponding roughly to the time series of anthropogenic aerosol radiative forcing [13]. It also shows a large excursion in the ensemble mean due to the Mt. Pinatubo eruption in 1991 and a large variance amongst ensemble members, perhaps related to Atlantic multi-decadal variability [14]. The third LFP/LFC shows low-frequency internal variability associated with the PDO ([15], Fig. 1). There is only a small excursion in the ensemble mean, before 1930, related to spin-up from initial conditions. The fourth LFP/LFC shows low-frequency variability over the Barents-Kara Sea, Eurasia, and the North Pacific (Fig. 1). It shows primarily internal variability (i.e., there is little agreement amongst ensemble members), but there is a response to the volcanic eruptions in 1982 and 1991. The remaining LFCs 5-25 show internal variability with increasingly shorter time scales. One interesting aspect of these results is that modes can be a combination of internal variability and forced responses (e.g., LFC 2-4), whereas most other analysis methods assume that modes are either one or the other.

To quantify the number of ensemble members needed to obtain these results, we compute the pattern correlation of LFPs obtained from analyses with fewer ensemble members, with those of the 40-member analysis

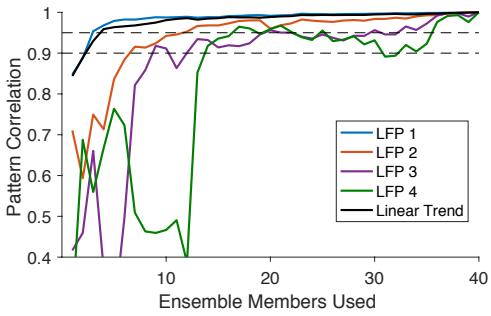


Fig. 2. Convergence of the first 4 LFPs as the ensemble size is increased, based on pattern correlation with results of the 40-ensemble-member analysis; comparison to convergence of the ensemble-mean linear trend. To make sub-ensembles, we pick the first n_E ensemble members from the 40-member ensemble. All ensemble members are identical in design, so we do not expect that our conclusions are sensitive to this sampling method.

(Fig. 2). For LFP 1, we find a pattern correlation > 0.95 using only 3 ensemble members. This is slightly less than the 4 ensemble members needed for a pattern correlation > 0.95 between the ensemble-mean linear trend and the 40-member ensemble-mean linear trend. Few ensemble members are needed for a robust estimate of global warming. The second, third, and fourth LFPs take longer to converge, requiring 7, 9, and 14 ensemble members, respectively, to reach a pattern correlation > 0.9. Only with large ensembles (or long control runs) can we understand these higher-order LFCs.

IV. SUMMARY AND OUTLOOK

We have demonstrated that *ensemble LFCA* can identify modes of low-frequency variability and change that are robust across climate model ensembles. It needs only 3 ensemble members to identify the forced global warming pattern, and makes no assumptions on the linearity of the warming response.

We have also applied our method to multi-model ensembles, where it is beneficial to discard the large ensemble-member differences in climatology (option b in step 1). This method provides particular utility when there are multiple time scales of forced response, such as in simulations of the response to abrupt 4xCO_2 forcing. This provides a powerful tool to visualize the dominant modes of low-frequency variability and change in large climate datasets.

ACKNOWLEDGMENTS

R.C.W. and D.L.H. acknowledge support from the National Science Foundation (Grant AGS-1549579). R.C.W. and D.S.B. acknowledge support from the Tamaki Foundation.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] D. W. Thompson, E. A. Barnes, C. Deser, W. E. Foust, and A. S. Phillips, "Quantifying the role of internal climate variability in future climate trends," *J. Climate*, vol. 28, no. 16, pp. 6443–6456, 2015.

[2] C. Deser, R. Knutti, S. Solomon, and A. S. Phillips, "Communication of the role of natural variability in future North American climate," *Nat. Clim. Change*, vol. 2, no. 11, pp. 775–779, 2012.

[3] C. Deser, A. S. Phillips, M. A. Alexander, and B. V. Smoliak, "Projecting North American climate over the next 50 years: Uncertainty due to internal variability," *J. Climate*, vol. 27, no. 6, pp. 2271–2296, 2014.

[4] J. Kay *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," *Bull. Am. Meteorol. Soc.*, vol. 96, no. 8, pp. 1333–1349, 2015.

[5] L. M. Frankcombe, M. H. England, M. E. Mann, and B. A. Steinman, "Separating internal variability from the externally forced climate response," *J. Climate*, vol. 28, no. 20, pp. 8184–8202, 2015.

[6] G. A. Meehl *et al.*, "Decadal prediction: can it be skillful?," *Bull. Am. Meteorol. Soc.*, 2009.

[7] M. R. Allen and L. A. Smith, "Optimal filtering in singular spectrum analysis," *Phys. Lett. A*, vol. 234, no. 6, pp. 419–428, 1997.

[8] S. Venzke, M. R. Allen, R. T. Sutton, and D. P. Rowell, "The atmospheric response over the North Atlantic to decadal changes in sea surface temperature," *J. Climate*, vol. 12, no. 8, pp. 2562–2584, 1999.

[9] T. Schneider and S. M. Griffies, "A conceptual framework for predictability studies," *J. Climate*, vol. 12, no. 10, pp. 3133–3155, 1999.

[10] T. Schneider and I. M. Held, "Discriminants of twentieth-century changes in earth surface temperatures," *J. Climate*, vol. 14, no. 3, pp. 249–254, 2001.

[11] M. Ting, Y. Kushnir, R. Seager, and C. Li, "Forced and internal twentieth-century SST trends in the North Atlantic," *J. Climate*, vol. 22, no. 6, pp. 1469–1481, 2009.

[12] R. C. Wills, T. Schneider, J. M. Wallace, D. S. Battisti, and D. L. Hartmann, "Disentangling global warming, multidecadal variability, and El Niño in Pacific temperatures," *Proc. Natl. Acad. Sci, submitted*, 2017.

[13] D. T. Shindell *et al.*, "Radiative forcing in the ACCMIP historical and future climate simulations," *Atmos. Chem. Phys*, vol. 13, no. 6, pp. 2939–2974, 2013.

[14] D. B. Enfield, A. M. Mestas-Nuñez, and P. J. Trimble, "The Atlantic multidecadal oscillation and its relation to rainfall and river flows in the continental US," *Geophys. Res. Lett.*, vol. 28, no. 10, pp. 2077–2080, 2001.

[15] N. J. Mantua, S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, "A Pacific interdecadal climate oscillation with impacts on salmon production," *Bull. Am. Meteorol. Soc.*, vol. 78, no. 6, pp. 1069–1079, 1997.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

UNCERTAINTY QUANTIFICATION FOR STATISTICAL DOWNSCALING USING BAYESIAN DEEP LEARNING

Thomas Vandal[1], Auroop R Ganguly[1]

***Abstract*—Global Climate Models (GCMs) contain uncertainty from internal variability of the system, parameterizations, and constraints from unknown physical processes. Furthermore, statistical downscaling of GCMs exasperates the uncertainty further. As downscaled datasets are often leveraged for climate change adaptation and planning, quantifying the uncertainty of our statistical model is crucial. In this work, we show how uncertainty quantification from Bayesian deep learning approaches can be leveraged in climate applications. We review and compare three approaches, Monte-carlo dropout, Concrete dropout, and Alpha-divergence based dropout, for statistically downscaling precipitation in Orlando Florida. In our experiments, we find that Concrete and Monte-carlo dropouts perform well but Alpha-divergence based dropout is less effective. While further experimentation is needed, Concrete dropout is a promising approach which extends Monte-carlo dropout by optimizing dropout probabilities.**

I. MOTIVATION

Downscaling is a process of enhancing the spatial (or temporal) resolution of global climate models (GCMs) for the purpose of climate change adaptation and planning. The statistical approach to downscaling aims to learn a functional mapping between low- and high-resolution historical datasets, which can then be applied directly to coarse resolution GCM outputs. A wide range of methods have been explored for statistical downscaling, ranging from bias correction techniques [1] and nearest neighbor approaches [2], [3] to sparse linear regressions [4] and neural networks [5], [6]. Intercomparison studies have shown that each method tends to perform well for certain regions, seasons, and climate variables but have difficulty generalizing [4]. The lack of consensus between statistical downscaling approaches highlights the difficulty of the problem and the need for uncertainty quantification in downscaled projections.

Given the difficulty of credibly downscaling GCMs, it is crucial to provide stakeholders with a range possible events. For example, a stakeholder may be interested in various metrics of extreme precipitation, including maximum daily return values. These metrics can be extracted from downscaled precipitation from GCM outputs under a set of scenarios. However, we know that these downscaled values have some amount of uncertainty which should be quantified appropriately. The concept of uncertainty quantification applied to statistical downscaling can be leveraged to compute uncertainties on the daily scale which can then be aggregated when computing a chosen metric. In this work, we refer to uncertainty quantification as the uncertainty over our prediction (rather than uncertainty over the parameters).

Bayesian models are well suited for uncertainty quantification and can be leveraged in statistical downscaling. Given the recent hype in deep learning as well as a recent publication using convolutional neural networks for statistical downscaling [6], we begin by studying the use of Bayesian deep learning (BDL) for uncertainty quantification in the climate domain.

Bayesian neural networks, originally studied in the early 1990's [7], applies a prior over the weights and biases of the network and aims to learn the posterior distribution given the data. As is common in many Bayesian modeling problems, the formulation is intractable and unable to scale to multiple hidden layers. BDL has attempted to solve this problem by estimating the posterior distribution using approximate variational inference and leveraging stochastic regularization techniques such as dropout [8]. Similar to more basic deep learning architectures, the tuning of hyper-parameters can greatly influence a model's results. In the remainder of this paper, we discuss the only three approaches which have been proposed for BDL and experiment in

Corresponding author: Thomas Vandal, vandal.t@husky.neu.edu
[1]Sustainability and Data Science Lab, Northeastern University

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

downscaling daily precipitation in Orlando Florida.

II. METHOD

The technique of using dropout as a stochastic regularization technique has been crucial for the major advances in deep learning. Dropout is implemented by randomly setting weights in the neural network to zero during training, usually sampled as a Bernoulli random variable. This process then reduces overfitting by more evenly spreading out information at each layer. The further study of BDL has provided a basic theoretical understand of why dropout effectively reduces overfitting of deep networks and how it can provide robust uncertainty estimates [9]. In this section, we review the fundamental concepts of BDL as well as two variants.

We define an L-layered neural network with inputs $\boldsymbol{X} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$, label $\boldsymbol{Y} = \{\boldsymbol{y}_1, ..., \boldsymbol{y}_N\}$, and weights $\omega = \{\boldsymbol{W}_1, ..., \boldsymbol{W}_L\}$ such that

$$\begin{bmatrix} \boldsymbol{f}^\omega(\boldsymbol{x}) \\ \boldsymbol{g}^\omega(\boldsymbol{x}) \end{bmatrix} = \sqrt{\frac{1}{K_L}} \boldsymbol{W}_L \sigma\Big(...\sqrt{\frac{1}{K_1}}\sigma(\boldsymbol{W}_1 x)\Big) \tag{1}$$

where K_l are the number of units in layer l, as defined in [10]. The predictive probability of a Bayesian neural network can be written as

$$\begin{aligned} p(\boldsymbol{y}^*|\boldsymbol{x}^*, \boldsymbol{X}, \boldsymbol{Y}) &= \int p(\boldsymbol{y}^*|\boldsymbol{x}^*, \omega) p(\omega|\boldsymbol{X}, \boldsymbol{Y}) \\ p(\boldsymbol{y}|\boldsymbol{x}, \omega) &= \mathcal{N}(\boldsymbol{y}; \boldsymbol{y}(\boldsymbol{x}, \omega), \boldsymbol{g}^\omega(\boldsymbol{x})^{-1}) \end{aligned} \tag{2}$$

An approximate distribution $q(\omega)$ is defined as:

$$\begin{aligned} \boldsymbol{W}_i &= \boldsymbol{M}_i * \text{diag}([z_{i,j}]_j^{K_i}) \\ z_{i,j} &= \text{Bernoulli}(p_i) \text{ for } i = 1, ...L, j = 1, ..., K_{i-1} \end{aligned}$$

given some probabilities p_i and variational parameters M_i. The Bernoulli random variable $z_{i,j}$ performs dropout at unit j for layer $i-1$ resulting in a stochastic approximation of W_i. We then minimize the Kullback−Leibler divergence between the posterior and it's variational approximation giving us the optimization objective:

$$\hat{\mathcal{L}}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \text{log}p(\boldsymbol{y}_i|\boldsymbol{x}_i, \omega) + \frac{1}{N}\text{KL}(q_\theta(\omega)||p(\omega)) \tag{3}$$

where θ represents the variational parameters and the term $\text{KL}(q(\omega)||p(\omega))$ is a regularization term, assuring that our approximate distribution does not deviate too far from the prior. Plugging our prior distribution $p(\boldsymbol{y}|\boldsymbol{x}, \omega)$ into 3, we obtain our regression objective:

$$\begin{aligned} \hat{\mathcal{L}}_{reg}(\theta) =& \frac{1}{2N}\boldsymbol{g}(\boldsymbol{x}_i)^{1/2}||\boldsymbol{y}_i - \boldsymbol{f}(\boldsymbol{x}_i)||^2 + \log\, \boldsymbol{g}(\boldsymbol{x}_i)^{-1} \\ &+ \frac{1}{N}\text{KL}(q_\theta(\omega)||p(\omega)) \end{aligned} \tag{4}$$

A. Monte Carlo Dropout

Monte Carlo Dropout (MC-Dropout) is the fundamental technique used in Bayesian deep learning for inference. Given a neural network trained with dropout minimizing 4, for a given new example we can use dropout to sample the posterior distribution to estimate the predictive mean and variance.

$$\begin{aligned} E[\boldsymbol{y}] &= \frac{1}{T}\sum_{i=1}^{T} \hat{\boldsymbol{f}}^\omega(\boldsymbol{x}) \\ Var[\boldsymbol{y}] &= \frac{1}{T}\sum_{i=1}^{T} \boldsymbol{g}^\omega(\boldsymbol{x})^{-1} + \sum_{i=1}^{T} \boldsymbol{f}^\omega(\boldsymbol{x})^2 - E[\boldsymbol{y}]^2 \end{aligned} \tag{5}$$

In this approach, the probability of dropout, p, is held constant which can cause under- or over-estimations of uncertainty as well as less than optimal predictions. The following two approaches, Concrete Dropout [11] and Dropout with Alpha-Divergence [12], attempt to improve the model's robustness and credibility of uncertainty estimates.

B. Concrete Dropout

The first approach we will discuss to learning a more robust uncertainty estimate is through Concrete Dropout [11]. Rather than doing a grid-search over all dropout probabilities (for each layer), which is computationally infeasible, Concrete dropout optimizes this probability. By parameterizing the variational approximation with probabilities such that $\theta = \{\boldsymbol{M}_i, p_i\}_{i=1}^L$, we can rewrite the KL objective in 4 as:

$$\text{KL}(q_\theta(\omega)||p(\omega)) \propto \sum_{l=1}^{L} \frac{l^2(1-p_l)}{2}||\boldsymbol{M}_l||^2 - K\mathcal{H}(p) \tag{6}$$

where

$$\mathcal{H}(p) = -p\text{log}p - (1-p)\text{log}(1-p) \tag{7}$$

is the entropy of the Bernoulli random variable with probability p [11]. The term $\mathcal{H}(p)$ acts as a regularizer and enforces $p \leq 0.5$. The authors show that the dropout probabilities in the lower layers tend to approach zero during training, which, in the past, has been shown to work well in various applications.

C. Dropout with Alpha-Divergences

A second approach to learning uncertainty estimates is by using Dropout with Alpha-Divergences (Dropout-AD) [12]. Using black-box α-divergence minimization as an extension to variational inference, which tends to under-estimate uncertainty, to penalize the function in the domain where examples have not been seen by the model. As shown by generative adversarial examples, small distortions on a trained example can easily fool the neural network. As shown in [12], Black-box α-divergence can reduce this effect. Li and Gal, using a re-parameterization trick, derive a new minimization objective:

$$\tilde{\mathcal{L}}_\alpha(\theta) = -\frac{1}{\alpha}\sum_{n=1}^{N} \text{log-sum-exp}[-\alpha p(\boldsymbol{y}_i|\boldsymbol{x}_i, \hat{\omega}_k)] + \text{KL}[q_\theta(\omega)||p_0(\omega)] \quad (8)$$

where q is the approximated posterior distribution parameterized by α and the log-sum-exp operating over K samples from the approximate posterior $\hat{\omega}_k \sim q(\omega)$. In this formulation the dropout probabilities are held constant.

Experimentations in [12] suggest that $\alpha = 0.5$ provides a good balance between test accuracy and robustness to adversarial examples. However, outside of using adversarial examples, the authors did not explicitly test for robust uncertainty quantification.

III. EXPERIMENTS

To begin to understand the applicability of Bayesian deep learning to climate applications, we experiment on a statistical downscaling dataset. Our dataset, extracted from PRISM[1], contains daily precipitation from years 1981 to 2015 in Orlando Florida. The features are a low-resolution 9x9 patch (1.25°) of precipitation and the label is precipitation at 1/16°. Training and test data are split before and after year 2005, respectively.

Three Bayesian neural networks are then trained, MC-Dropout, Concrete Dropout, and Dropout with Alpha-Divergences, each consisting of 2 convolutional layers of 128 hidden units each and a fully-connected output layer. For MC-Dropout and Dropout with Alpha-Divergences, the dropout rate is set to 0.25, sufficiently large for regularization and enforcing reasonably wide predictive uncertainty. Ten Monte-Carlo samples are used during training of Alpha-divergence dropout and set $\alpha = 0.5$. 100 Monte Carlo samples are used to estimate the mean and variance of the learned distributions.

[1]PRISM Climate Group, Oregon State University

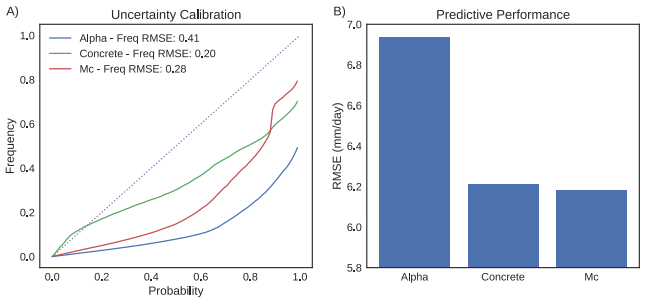


Fig. 1. (A) Comparison of uncertainty calibration between each BDL model. A perfect calibration corresponds to $y = x$. (B) Predictive performance of the test set measured by RMSE in mm/day between observed precipitation and $E[X]$.

The models were trained with a batch size of 128 for 100 epochs using the Adam optimizer.

To evaluate the uncertainty estimates, we quantify what ratio of the samples fall within a certain interval around the prediction. As presented in Figure 1, we find that concrete dropout's prediction intervals are more aligned with the ideal calibration ($y = x$) with MC-dropout performing better at wider intervals. Similar results are found for predictability, with MC-dropout's daily RMSE being just slightly higher than Concrete dropout. Alpha-divergence based dropout has both a larger RMSE and lower calibration performance relative to the others.

When training Concrete-dropout, we find that the learned dropout probabilities are 0.1 and 0.14 for each layer, respectively. This results in slightly more overfitting to the training set and reduce generalizability on the test set. To minimize this effect, one often includes an l_1 or l_2 regularization of the weight parameters.

IV. CONCLUSION

The three approaches tested have shown impressive results when tested on out-of-sample data as well as adversarial examples [12] however the predictive uncertainty on our statistical downscaling dataset is relatively poor. Architecture selection is generally a challenging task requiring many experiments, especially in the case with no other literature on our application. Furthermore, this is the only Alpha-divergence based dropout applied on a regression problem that the authors are aware of.

At it's current state, the method provides value in estimating uncertainty but should be used cautiously in practice as more experimentation is needed. The current formulations of these methods assume the output variable follows a normal distribution, but in many climate applications we know this assumption will

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

not hold. The authors hypothesize that an adaption to a well-suited distribution and network architecture will improve performance, especially at the extremes. Furthermore, the relative uncertainty estimates between multiple samples may provide value in understanding where the model is less confident, which aid in understanding where the performs poorly. Also, as is shown in the literature, these methods can provide more robust predictions when compared to vanilla dropout techniques. For these reasons, we will follow up this work with experiments in statistical downscaling with a modified likelihood distribution.

ACKNOWLEDGMENTS

This work was supported by NASA Earth Exchange (NEX), National Science Foundation CISE Expeditions in Computing under grant number:1029711, National Science Foundation CyberSEES under grant number:1442728, and National Science Foundation BIGDATA under grant number:1447587. The authors acknowledge Upmanu Lall for his valuable comments.

REFERENCES

[1] 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, no. D20, 2002.

[2] H. Hidalgo, M. Dettinger, and D. Cayan, "Downscaling with constructed analogues: Daily precipitation and temperature fields over the united states," 2008.

[3] D. W. Pierce, D. R. Cayan, and B. L. Thrasher, "Statistical downscaling using localized constructed analogs (loca)*," *Journal of Hydrometeorology*, vol. 15, no. 6, pp. 2558–2585, 2014.

[4] T. Vandal, E. Kodra, and A. R. Ganguly, "Intercomparison of machine learning methods for statistical downscaling: The case of daily and extreme precipitation," *arXiv preprint arXiv:1702.04018*, 2017.

[5] A. J. Cannon, "Quantile regression neural networks: Implementation in r and application to precipitation downscaling," *Computers & geosciences*, vol. 37, no. 9, pp. 1277–1284, 2011.

[6] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, "Deepsd: Generating high resolution climate change projections through single image super-resolution," in *23rd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2017.

[7] D. J. MacKay, "A practical bayesian framework for backpropagation networks," *Neural computation*, vol. 4, no. 3, pp. 448–472, 1992.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[9] Y. Gal, *Uncertainty in Deep Learning*. PhD thesis, Ph. D. thesis, University of Cambridge, 2016.

[10] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, pp. 1050–1059, 2016.

[11] Y. Gal, J. Hron, and A. Kendall, "Concrete dropout," *arXiv preprint arXiv:1705.07832*, 2017.

[12] Y. Li and Y. Gal, "Dropout inference in bayesian neural networks with alpha-divergences," *arXiv preprint arXiv:1703.02914*, 2017.

Improving Spatiotemporal Skill Assessment of Climate Field Reconstructions

Sooin Yun, Bo Li, Jason E. Smerdon and Xianyang Zhang

***Abstract*—Paleoclimate reconstructions that target spatiotemporal climate fields have been used to estimate climate conditions of the last several millennia from networks of heterogeneously distributed networks of multiple climate proxies. Assessing the skill of the methods used for these reconstructions is critical as a means of understanding the spatiotemporal uncertainties in the derived reconstructions products. Li et al. [2016] developed a skill assessment metric that evaluates the difference of the mean and covariance structure between spatiotemporal fields. We apply Li et al. [2016] to results from synthetic reconstruction experiments based on multiple climate model simulations from Smerdon et al. [2015] to assess the skill of four reconstruction methods, and further interpret and understand the comparison results using analysis of Empirical Orthogonal Functions that represent the noise filtered climate field.**

I. Introduction

Climate field reconstructions (CFRs) target hemispheric or global patterns of spatiotemporal temperature changes. Among the different uncertainties that have been addressed in CFR work, methodological assessments have become a recent focus. In particular, the performance of CFR methods have been tested using synthetic experiments called pseudoproxy experiments (PPEs; Smerdon [2012]). The basic premise of PPEs is to subsample the surface temperature field from a last-millennium simulation derived from a general circulation model (GCMs) in a way that mimics the data available for deriving real-world CFRs. This subsampled data then forms the input data for a CFR algorithm, which is used to generate an estimate of the excluded data in the spatiotemporal complete climate field from the last-millennium simulation. The derived CFR can then be compared to the known values of the simulated climate field as a means of evaluating the CFR skill.

We use the multi-model PPE results from Smerdon et al. [2015] and apply the set of robust Skill Assessment comparisons proposed by Zhang and Shao [2015] and Li et al. [2016] that are based on the functional principal components inherent in the data. This method has significant strong points: it evaluates whether the spatially-varying mean surfaces and covariance structures of two climate fields exhibit similar patterns, it is completely non-parameric and thus is free of the risk of model misspecification, and it assess which direction the differences lie.

Despite the above listed benefits, it can be difficult to interpret and determine why some CFRs methods perform differently and why they do so within PPEs based on different climate model output. In this paper we explore several important factors that underlie the skill assessment metrics to more clearly articulate the reasons why the applied CFR methods perform differently within simulation-specific PPEs and across PPEs based on different last-millennium simulations.

II. Data

The PPEs are based on concatenated last-millennium and historical simulations from modeling centers as configured and implemented in the Coupled Model Intercomparison Project Phase 5 and the Paleoclimate Modeling Intercomparison Project Phase 3 (CMIP5/PMIP3). Simulations from the following models are employed: the Beijing Climate Center CSM1.1 model (BCC), the National Center for Atmospheric Research Community Climate System version 4 model (CCSM), the Goddard Institue for Space Studies E2-R model (GISS), the Institue Pierre-Simon Laplace CM5A-LR model (IPSL) and the Max -Plank Institute ESM-LR model (MPI). In all cases, annual means from the surface temperature fields of the models are used and all fields are interpolated to 5-degree latitude-longitude grids from which all sampling is performed (Smerdon et al. [2015]). These fields are subsampled to approximate available instrumental temperature grids in the Brohan et al. [2006] surface temperature dataset (1,732 grid cells in total) and the multi-proxy network used in Mann et al. [2009] that yield 283 proxy locations.

Four CFR methods commonly employed in the literature are used, including two versions of the regularized expectation maximization (RegEM) method that both employ truncated total least squares (RegEM-TTLS) for regularization (Schneider [2001]; Mann et al. [2007]).

The first is a non-hybrid version of RegEM-TTLS (TTLS) as originally adopted by Schneider [2001] and the second is the hybrid version (TTLH) applied in Mann et al. [2009]. Additionally, we apply standard ridge regressions (Hoerl and Kennard [1970]) and canonical correlation analysis (CCA) as described in Smerdon et al. [2010]. All CFR methods use a calibration from 1850-1995 C.E. and a reconstruction interval from 850-1850 C.E.

III. Method

The methods of comparing two spatiotemporal random fields developed in Zhang and Shao [2015] and Li et al. [2016] based on a functional data analysis approach provide useful tools for CFR skill assessments. The basic idea is to perform the comparison in subspaces that are of much lower dimension but preserve a large portion of the variability.

Let $\{X_t(s)\}_{t=1}^N$ and $\{Y_t(s)\}_{t=1}^N$ be two spatiotemporal random fields observed over spatial locations, $s \in D$, and time points, $t = 1, \ldots, N$. We define the mean and covariance function of each spatial process as follows: $\mu_X(s) = E\{X_t(s)\}$ and $\mu_Y(s) = E\{Y_t(s)\}$ are mean functions over $s \in D$ and $C_X(s, s') = cov\{X_t(s), X_t(s')\}$ and $C_Y(s, s') = cov\{Y_t(s), Y_t(s')\}$ are the covariance functions of $X_t(s)$ and $Y_t(s)$ over $s, s' \in D$, respectively. To compare the mean and covariance functions of two spatiotemporal random fields, we consider the following two hypotheses:

(i) $H_0 : \mu_X = \mu_Y \quad vs. \quad H_a : \mu_X \neq \mu_Y$,

(ii) $H_0 : C_X = C_Y \quad vs. \quad H_a : C_X \neq C_Y$.

We denote the eigenvalues and eigenfunctions, also called empirical orthogonal functions (EOFs), corresponding to $\hat{C}_X$ by $\{\hat{\lambda}_X^j\}$ and $\{\hat{\phi}_X^j\}$, where $\hat{C}_X$ denotes the sample covariance function. Then we define a sequence of vectors consisting of the projected mean differences on the first K eigenfunctions: $\hat{\psi}_k = (< \hat{\mu}_{X,k} - \hat{\mu}_{Y,k}, \hat{\phi}_X^1 > \cdots < \hat{\mu}_{X,k} - \hat{\mu}_{Y,k}, \hat{\phi}_X^K >)^T$ for $1 \leq k \leq N$, where $< x, y >= x^T y$, and $\hat{\mu}_{X,k}$ ($\hat{\mu}_{Y,k}$) denotes the sample mean based on the recursive subsamples $\{X_t(s)\}_{t=1}^k$ ($\{Y_t(s)\}_{t=1}^k$). Our test statistic for hypothesis (i) is $TS1(K) = N\hat{\psi}_N^T V_\psi^{-1}(K)\hat{\psi}_N$ where $V_\psi = \frac{1}{N^2}\sum_{k=1}^N k^2(\hat{\psi}_k - \hat{\psi}_N)(\hat{\psi}_k - \hat{\psi}_N)^T$. The parameter K is user chosen and determines how many eigenfunctions are to be used in the test. Similarly, to test the covariance function, we define a sequence of matrices by the projected covariance differences, $C_k = [c_k^{i,j}]$, where $c_k^{i,j} =< (\hat{C}_{X,k} - \hat{C}_{Y,k})\hat{\phi}_X^i, \hat{\phi}_X^j >$, $1 \leq k \leq N$, $1 \leq i, j \leq K$, where $\hat{C}_{X,k}$ ($\hat{C}_{X,k}$) denotes the sample covariance function based on $\{X_t(s)\}_{t=1}^k$ ($\{Y_t(s)\}_{t=1}^k$). Let $\hat{\alpha}_k$ be the vectorized C_k, that contains the elements on and below the main diagonal of C_k. Test statistic for hypothesis (ii) is $TS2(d) = N\hat{\alpha}_N^T V_\alpha^{-1}(d)\hat{\alpha}_N$, where $d = K(K+1)/2$ and $V_\alpha(d) = \frac{1}{N^2}\sum_{k=1}^N k^2(\hat{\alpha}_k - \hat{\alpha}_N)(\hat{\alpha}_k - \hat{\alpha}_N)^T$. Again K is chosen by the user and can be determined by the cumulative percentage of total variation. The pivotal distributions of $TS1(K), TS2(d)$ under certain regularity conditions are given in Zhang and Shao [2015] and Li et al. [2016]. For more details and generalized methods, see Zhang and Shao [2015] and Li et al. [2016].

An important control on the skill of CFRs is tied to how well the leading EOFs of climate in the reconstruction period represent those of the calibration period. We use the inner product of the leading EOFs from those two time periods to compare the similarity of the two EOFs. For instance, if the absolute value of the inner product is close to 0, it suggests that the two EOFs are very different, while if the inner product is close to 1, it implies that they are equivalent. The p-values to test how significantly those inner products are different from zero can be derived using bootstrap analysis.

The leading EOFs only show the direction or the pattern of the variation but is uninformative of the total variability of the temperatures or the magnitude of the covariance structure. To measure the latter, we compute the eigenvalues. Let $\hat{\phi}_i^X(s), \hat{\phi}_i^X(s')$ be the i-th eigenfunction of C_X over $s, s' \in D$. The eigenvalue is defined as $\eta_{X,i}$ in $C_X(s, s') = \sum_{i=1}^k \eta_{X,i}\hat{\phi}_i^X(s)\hat{\phi}_i^X(s')$. The coefficient $\eta_{X,i}$ measures the variability of temperatures on the ith leading EOF direction.

IV. Results

We test the equivalence of mean structures between the CFRs and their target fields for all combinations of the five climate model simulations and the four CFR methods. Fig 1 shows p-values of the tests at each principal component. To understand the comparison results, we examine the key features of climate models and use them to interpret the skill of the CFRs in conjunction with a particular climate model.

The inner product of leading EOFs of the calibration period and of the reconstruction period demonstrates how strong the leading EOFs in the reconstruction period represent those in the calibration period (Table I). The CCSM and MPI models have significantly larger inner product values on their diagonal, implying that the calibration period well preserves the dominant pattern of leading EOFs in the reconstruction period. Moreover, for those two climate models, the order of the modes are preserved as well. In contrast, the BCC model reveals very weak associations between the calibration and the reconstruction period, and IPSL only displays strong a association for the first EOF.

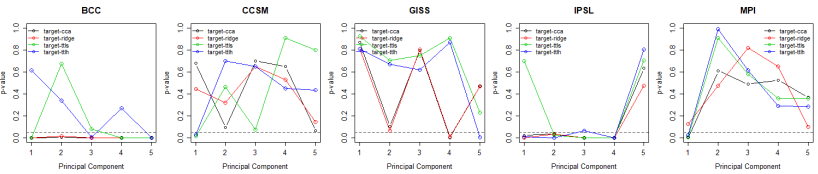


Fig. 1: Mean Comparison at each Principal Components

TABLE I: Inner Product of Reconstruction and Calibration period EOFs

INNER PRODUCT		EOF of (Reconstruction Period) 1	2	3
EOF	1	0.550	0.042	0.160
of	2	$0.761^{\dagger}$	0.084	0.334
BCC (Calibration)	3	0.020	0.436	0.605
EOF	1	$0.966^{\ddagger}$	0.002	0.098
of	2	0.093	$0.872^{\ddagger}$	0.320
CCSM (Calibration)	3	0.057	0.280	$0.820^{\dagger}$
EOF	1	$0.861^{\ddagger}$	0.194	0.098
of	2	0.164	$0.895^{\ddagger}$	0.071
GISS(Calibration)	3	0.027	0.213	0.225
EOF	1	$0.886^{\ddagger}$	0.223	0.222
of	2	0.378	0.643	0.616
IPSL(Calibration)	3	0.096	0.627	0.496
EOF	1	$0.976^{\ddagger}$	0.077	0.072
of	2	0.084	$0.943^{\ddagger}$	0.227
MPI (Calibration)	3	0.092	0.179	$0.844^{\dagger}$

Note : Significances of inner products are denoted by ‡, † for 5% and 10% levels respectively.

As indicated in Table I, the skill of CFRs is highly related to the stationarity of the leading EOFs in the calibration and the reconstruction periods. Additionally, if a large fraction of the variability in the climate field is represented by a few leading EOFs, and this feature is similar in the calibration and reconstruction periods, the CFRs tend to recover the true mean structure well. Because BCC and IPSL simulations violate either or both of the two conditions, CFRs based on BCC and IPSL have reduced skill in this sense. The performance of TTLS and TTLH largely depend on how well the first EOF of the reconstruction period represents the dominant EOF patterns in the calibration period, and CCA and RIDGE usually outperform the other methods when the leading variation in the reconstruction period is well preserved in the first few EOFs with the same order as the calibration period.

The total variability measured by the eigenvalues ($\eta_{X,i}$) is shown in Fig 2. This plot indicates that CCSM and MPI exhibit larger variability whereas BCC exhibits much smaller variability. We find among the CFRs, TTLS and TTLH tend to better represent the variability of the target field compared to the CCA and RIDGE methods.

Overall, the covariance comparisons between the CFRs and their target fields are poor if comparisons are limited to all 1,732 grid points. However, if we only focus on the covariance structure where the El Nino-Southern Oscillation (ENSO) is dominant, we identify stronger associations between the CFRs and their target. This is because the CFRs may fail to recover covariance structure at local scales, but are skillful in recovering the large-scale covariance structure in the climate field.

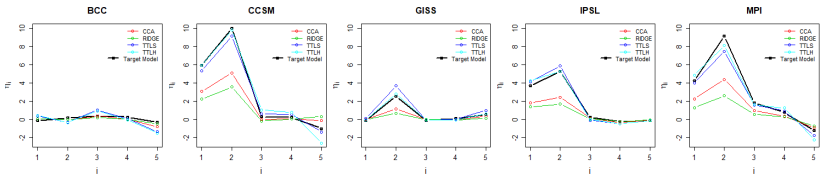


Fig. 2: Eigenvalues on ENSO teleconnection field

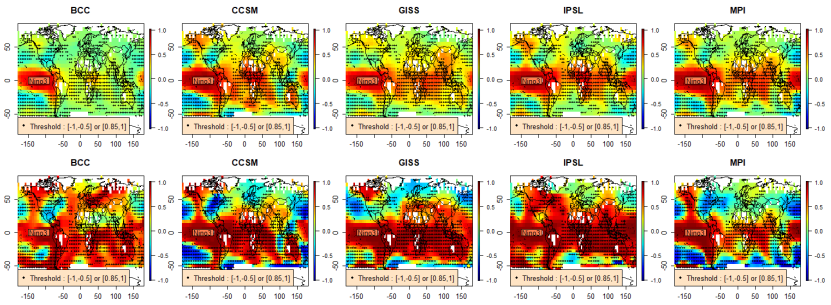


Fig. 3: Correlation at ENSO teleconnection region

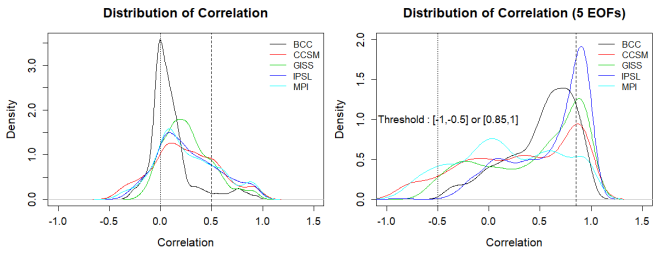


Fig. 4: Correlation distribution

Another means of investigating ENSO dependencies are through the correlation between the mean temperature time series within the Nino3 region (5N-5S, 150W-90W) and all other grid points. The upper panel of Fig 3 displays this correlation and it is seen that this correlation structure is somewhat unclear due to the noise in each climate field. In order to clearly show the dominant correlation pattern, we compute the correlation of the first five leading EOFs instead of the original temperature as seen in the bottom panel of Fig 3. The bottom panel exhibits a more pronounced correlation pattern. This is corroborated by the correlation distribution plot in Fig 4, which shows that the

correlation between temperatures has a long right tail whereas the correlation of the leading five EOFs (right) has a long left tail.

To compare the covariance over the ENSO teleconnection region, we only compare the covariance of points where the teleconnection is evident. (EOF correlation is greater than 0.85 or less than -0.5). Fig 5 shows the results of the covariance comparison at Nino3 and selected regions.

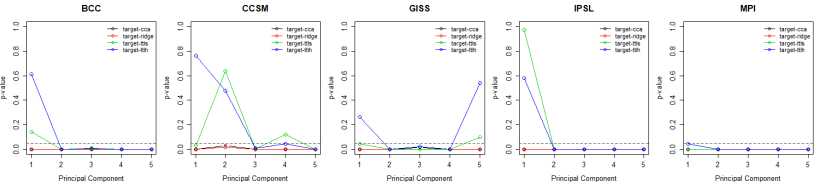


Fig. 5: Skill Assessment Comparison of Covariance

Similar to results in Fig 2, TTLS and TTLH well depict the variability of covariance compared to CCA and RIDGE throughout all climate models. Among all five models, CFRs of CCSM exhibit the true covariance structure the best. This is true not only because the large variability reflected in the leading EOFs was well depicted in TTLS and TTLH, but also because the correlation pattern of the ENSO teleconnection region was very pronounced after filtering out the noise in Fig 3 (bottom panel). However, the MPI model does not imply any common covariance structure due to the weak teleconnection signal on the leading EOFs (Fig 4).

V. Conclusion

Skill assessment comparisons of the mean can depend on many factors such as the sample network variability represented in the climate models; however, the major factor appears to be how well the calibration period represents the reconstruction period in terms of preserving the variability in their leading EOFs. Additionally, the covariance comparison accounts for the variability of covariance reflected in the leading EOFs as well as the magnitude of the teleconnection on the dominant modes. Different CFRs perform differently based on the specific characteristics of the climate model simulations.

References

P. Brohan, J. J. Kennedy, I. Harris, S. F. Tett, and P. D. Jones. Uncertainty estimates in regional and global observed temperature changes : A new data set from 1850. Journal of Geophysical Research : Atmoshperes (1984-2012), 111(D12), 2006.

A. E. Hoerl and R. W. Kennard. Ridge regression : Biased estimation for nonorthogonal problems. Technometrics, 12(1), pages 55–67, 1970.

B. Li, X. Zhang, and J. E. Smerdon. Comparison between spatio-temporal random processes and application to climate model data. Environmetrics, 27(5), 27:267–279, 2016.

M. E. Mann, S. Rutherford, E. Wahl, and C. Ammann. Robustness of proxy-based climate field reconstruction methods. Journal of Geophysical Research : Atmospheres (1984-2012), 112(D12), 2007.

M. E. Mann, Z. Zhang, S. Rutherford, R. S. Bradley, M. K. Hughes, D. Shindell, C. Ammann, G. Faluvegi, and F. Ni. Global signatures and dynamical origins of the little ice age and medieval climate anomaly. Science, 326(5957), pages 1256–1260, 2009.

T. Schneider. Analysis of incomplete climate data : Estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate, 14(5), pages 853–871, 2001.

J. E. Smerdon. Climate models as a test bed for climate reconstruction methods : pseudoproxy experiments. Wiley interdisciplinary Reviews : Climate Change, 3:63–77, 2012.

J. E. Smerdon, A. Kaplan, D. Chang, and M. N. Evans. A pseudoproxy evaluation of the cca and regem methods for reconstructing climate fields of the last millennium. Journal of Climate, 23(18), pages 4856–4880, 2010.

J. E. Smerdon, S.Coats, and T. R. Ault. Model-dependent spatial skill in pseudoproxy experiments testing climate field reconstruction methods for the common era. Climate Dynamics, pages 1–22, 2015.

X. Zhang and X. Shao. Two sample inference for the second-order property of temporally dependent functional data. Bernoulli, 21(2), pages 909–929, 2015.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

ANALOG NOWCASTING OF SOLAR IRRADIANCE FROM GEOSTATIONARY SATELLITE IMAGES

Alex Ayet[1,2] and Pierre Tandeo[3]

***Abstract*—Accurate forecasting of Global Horizontal Irradiance (GHI) is essential for the integration of the solar resource in an electrical grid. We implement a novel data-driven model for up to 6h probabilistic forecasting of GHI. Cloud dynamics are emulated using an analog method on a geostationary satellite database (herein 5 years of hourly images). It contains both the images to be compared to the current meteorological conditions and their successors at one or more hours of interval. No approximation is thus made on the physics of the system, unlike numerical weather forecast. The algorithm is computationally efficient and requires no tuning. It is designed to be easily used on different locations, requiring only GHI satellite images.**

I. INTRODUCTION

In the context of a growing need for sustainable energy, the solar resource ranks among the most promising solutions to meet this upcoming demand. However, the intermittent nature of the production makes its integration into an electrical gird challenging. The main input for most solar power generation systems is Global Horizontal Irradiance (GHI), and its accurate probabilistic and deterministic forecasting is thus essential.

Depending on the forecast horizon, different approaches are used for GHI forecasting (see the reviews [1], [2]). Satellite images have proven to be efficient for the intra-day horizon (up to six hours). The popular cloud motion vector methods ([3], [4]) estimate a motion field from successive cloud satellite images, to then advect the clouds, producing the forecast. The main drawback of this methods is the need for post-processing to take into account the cloud dissipation and deformation. Analog methods are also used, e.g. [5] that combines outputs from numerical models and in-situ data as features of a k-nearest neighbors algorithm. However, to satisfy the forecasting demand for a big amount of different sites, in locations where numerical models and in-situ observations are sparse, a robust and easy to use method is still needed.

For precipitation nowcasting, atmospheric analogs have become an important topic (e.g. [6], [7]) due to the availability of huge radar datasets. Analogs represent two atmospheric states closely resembling each other [8], with the hypothesis that these states evolve similarly. The forecast is thus issued by finding similar states in an historical database (the analogs), and considering how the atmosphere evolved following these states (the successors). The whole physics of the system is thus contained in the analog-successor pair.

The aim of this paper is to present an operational method to forecast GHI over a *precise* solar energy source (e.g. a solar photovoltaic panel). It uses only one source of data: hourly satellite images of GHI, which are easily accessible for different locations. Finally, the method needs no tuning, meaning that it can be easily applied to forecast the irradiance over sites in different locations, with different climatic conditions.

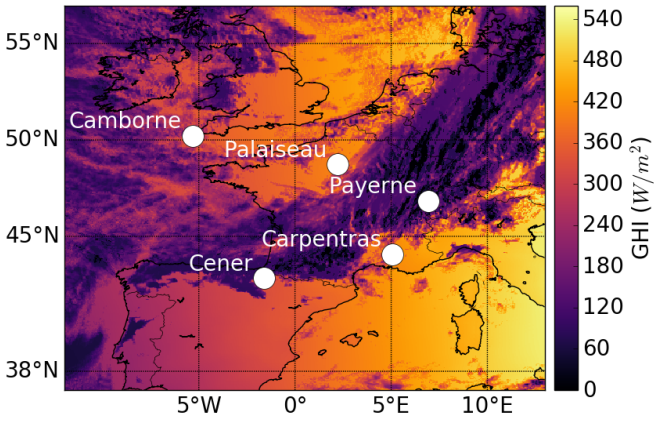


Fig. 1: OSI SAF satellite GHI image with selected BSRN stations in Europe on July 2nd, 2016.

Corresponding author: A. Ayet, alex.ayet@ens.fr [1]Elum Energy, Paris, France [2]École Normale Supérieure, Paris [3]Institut Mines-Télécom Atlantique, Brest, France

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

II. DATA AND SETUP

To demonstrate the method, we use an archive of $18,521$ GHI images obtained from the geostationary satellite Meteosat and processed by the Ocean and Sea Ice Satellite Application Facility (OSI SAF, [9]) covering western Europe and Africa. The images are remapped on a regular grid of $0.05°$, and interpolated to produce hourly maps. The archive extends from Sept. 6th, 2011 to Dec. 31st, 2016. We use the 2016 year as a test year and the rest of the archive as the training set. The method is tested at the location of five stations of the Baseline Surface Radiation Network (BSRN, see [10]) where in-situ pyrgeometer data is available. The stations, shown in Fig. 1, cover a wide range of climatic situations (oceanic, mountain, continental) and constitute a good framework to test the robustness of the method.

The variability of GHI, hereinafter noted G, is due both to the daily and seasonal solar cycle (the clear sky contribution G_{clr}) and to the cloudiness. Since clouds only reduce GHI, the clear sky at a location (x, y) is obtained as

$$G_{clr}(t, x, y) = \max_{t' \in \mathbb{S}(t)} \mathrm{G}(t', x, y), \tag{1}$$

with $\mathbb{S}(t)$ a 3-month interval around time t, with constant hour. This is a general expression that does not require any clear sky model (e.g. [11]). The cloud index c (between zero and one) is then defined as

$$G = (1 - c)G_{clr}. \tag{2}$$

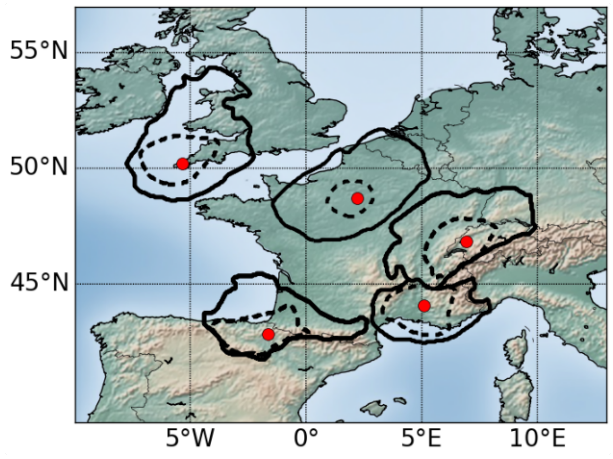


Fig. 2: Correlation masks for different BSRN sites (red dots) for Jan. 1st (full line) and July 1st (dotted line).

III. METHODOLOGY

The main GHI variability being due to change in cloudiness, the analog method presented herein forecasts the cloud index, which is then converted to GHI using Eq. (2).

A. Correlation Mask

For a given site of coordinates (x_s, y_s), it is crucial to automatically select the zone in which the analogs are looked for. A daily correlation map $\mathcal{C}^m$ (for a day d) between the pixel of interest and the surrounding region is computed. We use a metric inspired by [12] that measures the average spatial extension of cloud structures around the site

$$\mathcal{C}^m(d, x, y) = \frac{\overline{c(t, x_s, y_s,)c(t, x, y)}^d}{\left[\overline{c(t, x_s, y_s)^2}^d \overline{c(t, x, y)^2}^d\right]^{1/2}}, \tag{3}$$

where the averages $\overline{\cdot}^d$ are temporal within a 3-month interval around d. The region where the correlation is higher than 0.9 is then selected. Examples of masks are presented in Fig. 2. In the following, all computations are performed considering only image pixels in the mask corresponding to the site and day of forecast.

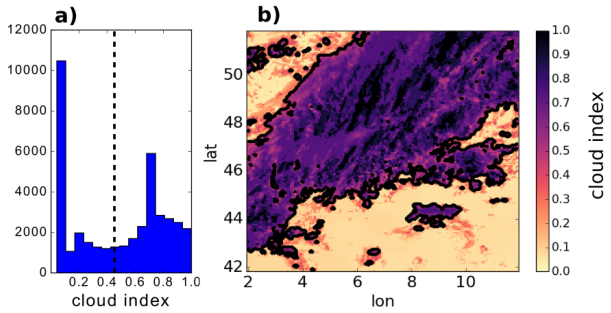


Fig. 3: (a) Histogram of cloud index from image (b) on the 2nd of July 2016 at Payerne. The threshold in (a) corresponds to the thick black line in (b).

B. Analog forecasting

The forecast consists in two steps: the analogs selection and aggregation. To avoid overfitting and for computational efficiency, the analogs selection is done considering images compressed in a four dimensional space. For a given cloud index image, Otsu's method [13] (similar to a bimodal Fisher's discriminant analysis on a histogram of cloud index) is used to obtain a threshold

separating the clear sky and the cloud pixels (c below and above the threshold respectively, see Fig. 3). The images (observed and database) are compressed into four features between 0 and 1:

1) the *cloud fraction*: number of cloud pixels over the total number of pixels in the mask
2) the *cloud spread*: number of cloud pixels over the number of pixels in the convex hull of the clouds. It tends to one when there is only one cloud (its convex hull is nearly identical to the cloud itself) and to zero when there are many separate clouds
3) the *clear sky intensity*: the mean cloud index of the clear sky pixels
4) the *cloud intensity*: the mean cloud index of the cloud pixels.

Following [7], the database is first shrinked by considering images only within a time of the year (3-month window) and time of the day ($\pm$3h) interval with respect to the date at which the forecast is to be issued. This increases the likelihood of finding similar convective and advective patterns. Then, the k-nearest neighbors of the observed image are selected using the Euclidean distance in the four-dimensional features space. For the BSRN sites of this study, the optimal number of neighbors is close to $k = 90$.

Next, the selected analogs are aggregated to produce a probabilistic forecast. For a given k-th analog, we determine the optimal spatial translation $\boldsymbol{\delta}$ such that the correlation C_k^s between the analog cloud index c_k^a and the observed cloud index c^o

$$\mathcal{C}_k^s(\mathcal{T}_{\boldsymbol{\delta}}) = \frac{< c^o(x,y)\mathcal{T}_{\boldsymbol{\delta}} c_k^a(x,y) >}{\left[< c^o(x,y)^2 >< \mathcal{T}_{\boldsymbol{\delta}} c_k^a(x,y)^2 >\right]^{1/2}} \quad (4)$$

is maximal ($\mathcal{T}_{\boldsymbol{\delta}}\cdot$ is the translation operator).

A local linear operator (see [14] or [15] for more details) is then applied on the translated images: it consists in fitting a linear regression between the translated analogs and successors, taking into account the weights computed in Eq. (4). The regression operator is then applied to the current observation to provide the analog nowcasts.

IV. Evaluation

The deterministic forecast of GHI (the mean of the predicted Gaussian) is evaluated with the normalized Root Mean Squared Error (RMSE) for a set of validation observations $\mathcal{S}$

$$\text{RMSE} = \frac{\sqrt{\sum_{s\in\mathcal{S}}(c_s^o - \hat{c}_s)^2}}{\sum_{s\in\mathcal{S}} c_s^o} \quad (5)$$

with c_s^o the observed cloud index from a satellite image at the BSRN sites, and $\hat{c}_s$ the corresponding forecast. The analog forecast is compared to an Eulerian persistence (keeping the last cloud index observation frozen), and a hourly climatology, obtained by taking the hourly average of GHI in the train dataset for days in a 2-weeks interval around the forecasted day. Results are given in Fig. 4 and indicate good performance and robustness to different locations. In all cases, the analog nowcasting procedure reaches better performances than the persistence and climatology method.

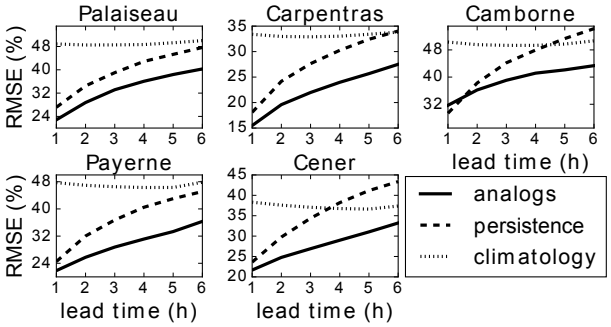


Fig. 4: Normalized RMSE for the five BSRN sites.

V. Conclusion and Perspectives

We have presented a computationally efficient method for GHI analog nowcasting on a particular site. The method uses a k-nearest neighbors algorithm on a four-dimensional feature-space of cloud index to then apply a local regression between selected analogs and successors. The methodology has proven to be robust to different geographical locations, and requires no tuning, no in-situ data nor a numerical weather model.

The method will be extended by downscaling predictions on a particular site. The analogs times will be used to select historical in-situ data, using the BSRN data also available during the period 2011-2016. An aggregation operator will then be applied to forecast the in-situ production.

Acknowledgments

The first author acknowledges the funding provided by Elum Energy as well as the supports from IMT-Atlantique and Elum R&D teams.

REFERENCES

[1] D. Heinemann, E. Lorenz, and M. Girodo, "Forecasting of solar radiation," *Solar energy resource management for electricity generation from local level to global scale. Nova Science Publishers, New York*, 2006.

[2] M. Diagne, M. David, P. Lauret, J. Boland, and N. Schmutz, "Review of solar irradiance forecasting methods and a proposition for small-scale insular grids," *Renewable and Sustainable Energy Reviews*, vol. 27, pp. 65–76, 2013.

[3] A. Hammer, D. Heinemann, E. Lorenz, and B. Lückehe, "Short-term forecasting of solar radiation: a statistical approach using satellite data," *Solar Energy*, vol. 67, no. 1, pp. 139–150, 1999.

[4] H. Escrig, F. Batlles, J. Alonso, F. Baena, J. Bosch, I. Salbidegoitia, and J. Burgaleta, "Cloud detection, classification and motion estimation using geostationary satellite imagery for cloud cover forecast," *Energy*, vol. 55, pp. 853–859, 2013.

[5] S. Alessandrini, L. Delle Monache, S. Sperati, and G. Cervone, "An analog ensemble for short-term probabilistic solar power forecast," *Applied energy*, vol. 157, pp. 95–110, 2015.

[6] L. Panziera, U. Germann, M. Gabella, and P. V. Mandapaka, "Nora – nowcasting of orographic rainfall by means of analogues," *Q. J. R. Meteorol. Soc.*, vol. 137, no. 661, pp. 2106–2123, 2011.

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

[8] E. N. Lorenz, "Atmospheric predictability as revealed by naturally occurring analogues," *Journal of the Atmospheric sciences*, vol. 26, no. 4, pp. 636–646, 1969.

[9] P. Le Borgne, G. Legendre, and A. Marsouin, "Meteosat and goes-east imager visible channel calibration," *Journal of Atmospheric and Oceanic Technology*, vol. 21, no. 11, pp. 1701–1709, 2004.

[10] A. Ohmura, H. Gilgen, H. Hegner, G. Müller, M. Wild, E. G. Dutton, B. Forgan, C. Fröhlich, R. Philipona, A. Heimo, *et al.*, "Baseline surface radiation network (bsrn/wcrp): New precision radiometry for climate research," *Bulletin of the American Meteorological Society*, vol. 79, no. 10, pp. 2115–2136, 1998.

[11] D. Cano, J.-M. Monget, M. Albuisson, H. Guillard, N. Regas, and L. Wald, "A method for the determination of the global solar radiation from meteorological satellite data," *Solar Energy*, vol. 37, no. 1, pp. 31–39, 1986.

[12] I. I. Zawadzki, "Statistical properties of precipitation patterns," *Journal of Applied Meteorology*, vol. 12, no. 3, pp. 459–472, 1973.

[13] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE transactions on systems, man, and cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[14] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American statistical association*, vol. 74, no. 368, pp. 829–836, 1979.

[15] R. Lguensat, P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, "The analog data assimilation," *Monthly Weather Review*, 2017, published online, https://doi.org/10.1175/MWR-D-16-0441.1.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

ROBUST COPULA DEPENDENCE FOR CLIMATE NETWORK ANALYSIS

Yi Li[1], Adam Ding[1]

***Abstract*—Climate system exhibits highly complex, nonlinear and inter-connective phenomenon. Complex network theory has been recently proposed to discover various climate properties of the earth system. While some interesting results have been proposed, there are still much unknown of the sophisticated mechanism of the nature. In this paper, we propose to use a nonlinear dependence measure, the robust copula dependence measure (RCD), in the climate network study, in comparison with the traditional Pearson's correlation coefficient which has been the major tool used in the current climate network literatures. We compare different statistical properties of the networks constructed with these two types dependence measure. Moreover, we show that with the help of the nonlinear dependence measure, we could have some interesting finding in the teleconnection exploration.**

I. INTRODUCTION

Ever since the identification of certain types of the small world network models [1], the complex network has been a mature field with a wide range of applications, ranging from the modeling of social network, the structure of the World Wide Web (WWW), to the gene network. Recently, the complex network theory has been brought to the climate community as the climate network [2][3][4][5], which aims to view and model the climate data from the complex network point of view.

Most of the current research in climate network is based the Pearson's correlation coefficient (*cor*), which is an excellent tools for detecting linear signals. However, as is known to us all, the climate system involves a great amount of nonlinearity. Thus, we attempt to apply a cutting edge nonlinear dependence measure, the robust copula dependence measure (RCD) [6][7], to help constructing the climate network. This could potentially enable us to include more nonlinear features or information based on our data, which may be helpful for us in understanding more about the nature.

On the other hand, teleconnection [8][9] has been a traditional but still on-going research topic in climate science. Detecting relatively weak but informative information of the teleconnection from the highly correlated nearest-neighbour effect is a challenging problem. The usage of nonlinear dependence measure in climate network could help to discover novel and interesting teleconnection phenomenon.

The rest of the paper is organized as follow: in the next section, we describe the data set, the RCD, the constructed climate networks using *cor* and RCD. In section III, we discuss the application of the climate network in teleconnection discovery. We conclude this paper in section IV.

II. THE CLIMATE NETWORK

A. Climate Data

In this paper, we consider the reanalysis monthly mean temperature data (1948 - 2016) from the National Oceanic & Atmospheric Administration (NOAA) [1]. The resolution of this data is $5^\circ \times 5^\circ$. We normalize the raw data (temperature) by removing the historical long range mean for each month. Namely, for a given climate variable $x(y, m)$ at year y and month $m \in \{1, 2, \cdots, 12\}$, the anomaly data $\tilde{x}(y, m) = x(y, m) - \langle x(y, m)\rangle_y$, where $\langle\cdot\rangle_y$ is the average over y.

B. The Robust Copula Dependence Measure

The robust copula dependence is a nonlinear dependence measure that is recently proposed in [6][7]. Given two random variable X and Y with the CDF $F_X(x)$ and $F_Y(y)$, the RCD between X and Y is defined as the following quantity:

$$RCD(X, Y) = \frac{1}{2}\int_{[0,1]^2} |c(u, v) - 1|dudv, \quad (1)$$

where u, v correspond to the copula transformed variables $U = F_X(X)$ and $V = F_Y(Y)$ and $c(\cdot, \cdot)$ is a copula density function [10]. RCD is an equitable

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

[1]https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html

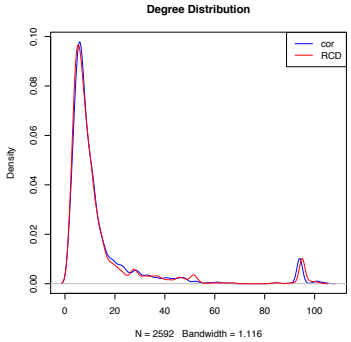


Fig. 1. Degree distribution of the networks based on the two dependence measure.

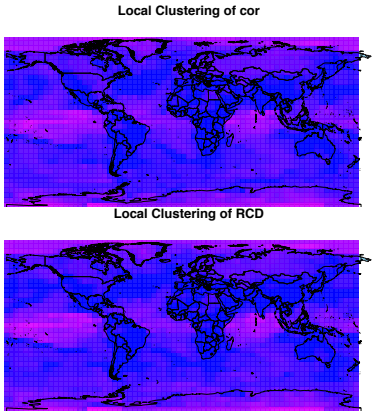


Fig. 2. Local clustering of the networks based on the two dependence measure (Redder is higher value, and bluer is lower value).

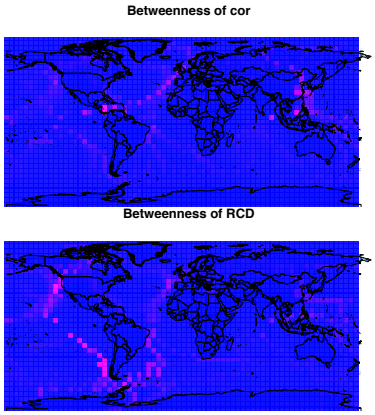


Fig. 3. Betweenness of the networks based on the two dependence measure (Redder is higher value, and bluer is lower value).

dependence measure, that is, ranking the dependence strength purely by the noise level regardless the functional shape (see [6][7] for details). Therefore, it detects nonlinear relationship as well as linear relationship, unlike the cor. RCD allows us to detect more nonlinear but interesting relationships hidden in the high dimensional data.

C. Climate Network

A network, or graph, is a combination of vertexes and edges which is usually denoted by $G = (V, E)$, where V is the collection of all the vertexes and E is the set of all the edges with the endpoints from V. Moreover, the climate network models the climate data in the way that the vertexes are the geographical locations while the edges represent the possible relationship between each pair. In our current situation, the vertexes are the spatial grid points from the reanalysis model. While cor is often used in describing the relationships along each edge, we apply RCD as a nonlinear proxy instead.

To analyze the climate network, several statistics can be used to summarize the characteristics of the intrinsic structure. Here, we focus on four major aspects: the degree distribution, clustering coefficient, betweenness, and community detection. The networks are constructed based on the top 0.5 percentile of the edge weight, although other testing based method is also possible.

The degree of a vertex is the total number of edges correspond to this vertex. The overall distribution of the degree can provide us a general information about the connectivity of the network. As we can see from Figure 1, the two network based on Pearson's correlation coefficient and RCD have similar degree distribution.

Local clustering effect is a way trying to delineate, for a given vertex, whether the connected points are also connected. It is defined as the fraction of connected triples through each vertex that are closed. As we can see from both plots in Figure 2, the clustering effect are more likely to occur around the equator and the poles.

As another important summary statistics, the betweenness is a way to measure the centrality of a network. A vertex with high betweenness plays a significant role to bridge two sub-network together. It is the number of geodesic paths that pass through the corresponding vertex. Figure 3 shows the relatively different results based on cor and RCD. As we can see from the plots, the nonlinear RCD based network indicates the grid points around the America continent has higher betweenness.

Community detection is a traditional way to find homogeneous regions based on the relationship/distance between each pair of the vertexes. Figure 4 presents the hierarchical clustering results based on the two distance matrices (here, we define $1 - |cor(x, y)|$ and $1 - RCD(x, y)$ as a semi-distance). Results shows that

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

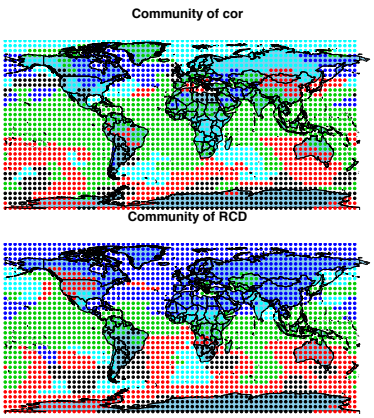


Fig. 4. Communities of the networks based on the two dependence measure based on the hierarchical clustering algorithm.

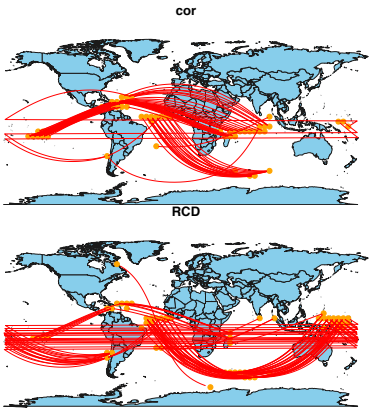


Fig. 5. Visualization of the network structure that are constructed by the remotely highly correlated edges base on their spatial location on map.

both of the communities are distributed according to their latitude.

III. TELECONNECTION

It is of great interest and significance to study the teleconnection, the dependence relationship between faraway locations, of the earth system. The application of the nonlinear dependence measure in the climate network could cast light on finding novel information from climate data set. To this end, we first filter out the neighbourhood points within five grid units, and carefully examine the networks constructed based on the top 100 dependence scores.

Figure 5 shows that the basic structure of the two networks. The vertexes are geographical grids and the edges are the cords. While most of the relationships are around the equator, we notice that the nonlinear dependence RCD detect an arc with long latitude distance. This arc is further analyzed in Figure 6.

Figure 6 is the scatter plot that is uniquely identified with the nonlinear dependence measure RCD, which is between the grid point $(65°W, 60°N)$ and $(30°E, 65°S)$. As we can see from the Figure 6, the anomaly data in these two positions exhibit a bilinear relationship, which cannot be detected by linear dependence measure like Pearson's correlation coefficient. This bilinear relationship of the anomaly data indicates some possible interesting finding: if the air temperature of one place is in its average value, the other place is more likely in its extreme case, and vise versa. This may help to guide the further research in studying the climate extremes.

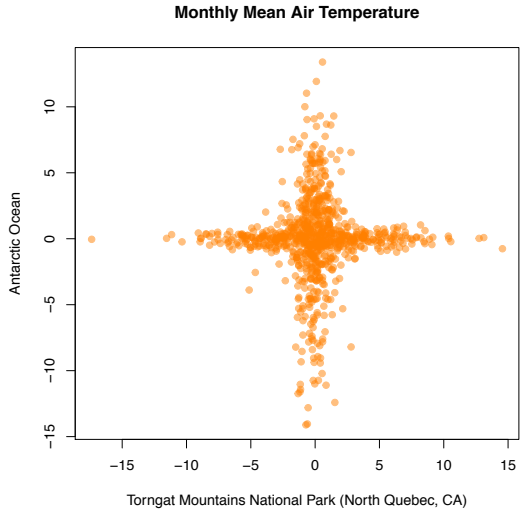


Fig. 6. Interesting possible teleconnection that is uniquely detected by nonlinear dependence measure.

IV. CONCLUSION

In this paper, we apply an equitable nonlinear dependence measure RCD to the climate network. We compare the networks constructed based on the linear Pearson's coefficient coefficient and the nonlinear RCD and use climate network as a tool to study the teleconnection. Result shows that some interesting nonlinear relationship could be found by RCD.

ACKNOWLEDGMENTS

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

REFERENCES

[1] D. J. Watts and S. H. Strogatz, "Collective dynamics of'small-world'networks.," *Nature*, vol. 393, no. 6684, pp. 409–10, 1998.

[2] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths, "Complex networks in climate dynamics," *The European Physical Journal Special Topics*, vol. 174, pp. 157–179, Jul 2009.

[3] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, "An exploration of climate data using complex networks," *SIGKDD Explor. Newsl.*, vol. 12, pp. 25–32, Nov. 2010.

[4] K. Steinhaeuser, N. V. Chawla, and A. R. Ganguly, "Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science," *Stat. Anal. Data Min.*, vol. 4, pp. 497–511, Oct. 2011.

[5] K. Steinhaeuser, A. R. Ganguly, and N. V. Chawla, "Multivariate and multiscale dependence in the global climate system revealed through complex networks," *Climate Dynamics*, vol. 39, pp. 889–895, Aug 2012.

[6] Y. Chang, Y. Li, A. Ding, and J. Dy, "A robust-equitable copula dependence measure for feature selection," in *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Statistics (forthcoming)*, Citeseer, 2016.

[7] A. Ding, J. Dy, Y. Li, and Y. Chang, "A robust-equitable measure for feature ranking and selection (forthcoming)," *Journal of Machine Learning Research*, 2017.

[8] I. Choi, B. Li, H. Zhang, and Y. Li, "Modelling spacetime varying enso teleconnections to droughts in north america," *Stat*, vol. 4, no. 1, pp. 140–156, 2015. sta4.85.

[9] A. A. Tsonis, K. L. Swanson, and G. Wang, "On the role of atmospheric teleconnections in climate," *Journal of Climate*, vol. 21, no. 12, pp. 2990–3001, 2008.

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

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

A STUDY OF CAUSAL LINKS BETWEEN THE ARCTIC AND THE MIDLATITUDE JET-STREAMS

Savini Samarasinghe[1], Marie McGraw[2], Elizabeth A. Barnes[2], Imme Ebert-Uphoff[1]

***Abstract*—This paper investigates causal links between Arctic temperatures and the jet-streams. We apply two different frameworks for this application based on the concepts of (1) *Granger causality* and (2) *Pearl causality*. Both methods show that Arctic temperature and jet speed each exhibit strong auto-correlation (as expected), and that jet speed drives Arctic temperature at timescales of 5-15 days, while Arctic temperature drives jet speed at timescales of up to 5 days, in the North Pacific. A positive feedback loop is also identified and discussed, among additional findings. This study is only the beginning of a larger effort to apply and compare different causality methods in order to gain a deeper understanding of the causal connections between the Arctic and weather at lower latitudes.**

I. MOTIVATION

Arctic amplification–that is, the phenomenon of Arctic temperatures rising much faster than the global mean ([1])–and its present and future effects on midlatitude weather and climate have received substantial attention in recent years. While it is well known that the midlatitude circulation can drive changes in Arctic temperatures and sea ice, it is unclear how and to what extent the Arctic influences midlatitude weather ([2]). Some argue that Arctic amplification is already influencing midlatitude weather (e.g. [3], [4], [5], [6]), while others state that any possible signal is too small to have been observed amidst the background of atmospheric variability (e.g. [7], [8], [9]). Regarding Arctic influence on midlatitudes under climate change, idealized and fully-coupled climate model simulations have shown an equatorward shift of the jet-stream and weakening of the zonal winds in response to Arctic warming and sea ice loss (e.g. [10], [11], [12], [13]), but little is understood about the underlying dynamics behind this response in models or whether the models can adequately simulate the processes involved. Making progress requires that we study the two-way causal connections between Arctic temperatures and the midlatitude circulation, and place the different pathways in context of one another and the background of atmospheric variability.

The typical approach for assessing causal links in climate dynamics (including studying the links between the jet-streams and Arctic warming/sea ice loss) is targeted modeling studies. While incredibly useful for understanding the physical mechanisms at play, this approach only allows for studying cause and effect in isolation, and does not allow for the feedbacks to fully develop. In addition, we have entered a period where atmospheric science tends to be "data rich" both in observations and model output [14]. There is great need for additional tools that can aid scientists in identifying and extracting signals. Causal discovery techniques provide (1) robust definitions of causality, (2) can have direct ties to forecasting/prediction, (3) augment targeted model studies, (4) place pathways in context relative to other drivers and feedbacks, and (5) allow for a direct comparison of results from observations and models.

Here we use two different frameworks to learn about causal relationships for this system. The first framework uses vector autoregression (VAR) type models (plain VAR and LASSO), combined with the concept of *Granger causality*. The second framework is based on the concept of *Pearl causality*. We apply both frameworks to the study of causal links between the Arctic and midlatitude jet-streams. The purpose is two-fold: (1) by comparing the results of two very different frameworks we hope to obtain robust results; (2) we hope to make more geoscientists aware of the different types of causal analysis tools.

II. RELATED WORK

In recent years significant work has been done on using causal reasoning for climate applications, including [15] [16] [17] [18] [19] [20] [21], on developing tools for that purpose [22] [23], and on causal attribution of climate events [24]. Of highest relevance to this work are causality studies specifically for the Arctic: Strong and Magnusdottir [25]; Kretschmer et al. [26]. These studies demonstrate the utility of causality techniques for studying Arctic-midlatitude connections, however, each employs a different approach. Thus, it is unclear whether different causality approaches would produce

Corresponding author: Imme Ebert-Uphoff, iebert@colostate.edu.
[1]Electrical & Computer Engineering, Colorado State University, Fort Collins, CO, USA. [2]Dept. of Atmospheric Science, Colorado State University, Fort Collins, CO, USA.

similar results, or whether a particular technique is best suited for this topic. In addition, neither study investigates the relationship between Arctic temperatures and the jet-streams - the focus of this work.

III. DATA

We use daily data from the NCAR CESM1 Large Ensemble Control run. We use Years 402 to 2,200, resulting in 656,634 days (1,799 years) of data. For our analysis we use only one season per year, either DJF or JJA, roughly dividing the number of data samples for each experiment by four. We focus on the North Pacific (120°E - 230°E) and the following three indices: jet latitude, $\mathcal{L}$; jet speed, $\mathcal{S}$; Arctic temperature (averaged over 70°N-90°N), $\mathcal{T}$. For each time series the seasonal cycle was subtracted in order to focus on anomalies, then it was averaged into non-overlapping chunks of 5 days to smooth out *weather noise*. Then we extract the values corresponding to the season of choice. Finally, each time series is standardized, i.e. we subtract its mean and divide by its standard deviation.

IV. METHODS BASED ON GRANGER CAUSALITY

We first explain two closely related models, *VAR* models and *LASSO* models, then discuss how they can each be linked to the concept of *Granger causality*.

A. Vector Autoregression (VAR) model

A VAR(p) model estimates vector $\mathbf{y}_t$ in terms of its p lags as follows:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{A}_1\mathbf{y}_{t-1} + ... + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{e}_t, \qquad (1)$$

where p denotes the number of lags considered; vector $\mathbf{y}_t$ contains the values of k considered variables at time t; $\mathbf{c}$ is a coefficient vector; $\mathbf{A}_i$ are the $k \times k$ coefficient matrices (for $i = 1, \ldots, p$); and $\mathbf{e}_t$ is the vector of error terms (residuals). Eq. (1) is a standard regression problem and a standard least-squares approach is used to calculate the model parameters [27], vector $\mathbf{c}$ and matrices $\mathbf{A}_i$. We derive such a VAR model for several different values of p, then look at convergence characteristics to choose the smallest p for which the model no longer changes significantly.

B. LASSO model (Regularized Regression)

For an interpretation based on Granger causality we need to distinguish which of the coefficients, a_{ij}^k, of matrices $\mathbf{A}_i$ are non-zero. (The reasons will become apparent in the next subsection.) For a standard VAR model that requires using a cut-off value, since, due to noise and numerical accuracy, none of the coefficients is likely to be *exactly* zero. The LASSO (least absolute shrinkage and selection operator) [28], [29] approach solves this problem in a more elegant and robust way. It adds constraints, namely it limits the sum of the magnitude of the elements of all $\mathbf{A}_i$ $(i = 1, \ldots, 10)$ matrices to be below a chosen threshold [30]. This forces small coefficients to become exactly zero, while the remaining coefficients compensate for that change. As such it performs *variable selection* along with prediction, i.e. it tells us which input variables (and at which lags) are actually important in the model. LASSO results in a model of the exact same form as Eq. (1), but where many coefficients are exactly zero, which makes the subsequent Granger analysis more straightforward.

C. Connection to Granger causality

Once a model of the form in Eq. (1) is obtained, we perform validation tests to assure the model is stable [27],[31], then apply the concept of Granger causality by inspection of the coefficients in $\mathbf{A}_i$. Let a_{ij}^k denote the element of row i and column j of matrix $\mathbf{A}_k$. Then a_{ij}^k denotes the effect of $y_{j,(t-k)}$ (the jth variable, lagged by k) on $y_{i,t}$ (the ith variable, without lag). Furthermore, since the data was normalized, a_{ij}^k indicates for a change of one standard deviation of $y_{j,(t-k)}$ how much change to expect (approximately) in $y_{i,t}$. (This quantitative interpretation should be used with caution, as many geophysical relationships are non-linear, and the model is thus only a rough approximation.) Then, for $i \neq j$, we see in this model that $y_{j,(t-k)}$ is useful for the prediction of $y_{i,t}$, if and only if $a_{ij}^k \neq 0$. Consequently, the jth variable, y_j, is said to *Granger-cause* the ith variable, y_i, if and only if at least one of the coefficients $a_{ij}^k \neq 0$ for any lag $k = 1, \ldots, p$.

V. METHOD BASED ON PEARL CAUSALITY

The concept of *Granger causality* is related to *predictability*. In contrast Pearl and Rebane developed the framework of *causal calculus* [32] based on the concept of *intervention*, which forms the basis for graphical models and for the concept of *Pearl causality* (see also [33], [34]). The method we use builds on the fact that it is impossible to *prove* a cause-effect relationship between two variables based on just observations, but that one can nevertheless *disprove* such relationships based on observations. We thus use an elimination method that first assumes that all variables have cause-effect connections to each other (for all lags), then uses conditional independence tests to *eliminate* the great majority of these connections. This method usually yields a small set of *potential* cause-effect relationships, each of which may or may not be a true causal relationship. Nevertheless, the sets of actual causal relationships is a subset of the resulting set. The specific method used is the temporal version [15], [18] of the *PC stable* algorithm [35], which is a variant of the classic *PC* algorithm [36] (so named after the first names of the two authors, i.e. no relation to PCA). For more information, see [18]. For brevity, we refer to PC stable as simply PC in the remainder of this document.

VI. RESULTS AND INTERPRETATION

Our primary focus for now is on the boreal winter (DJF) results, and the relationship between jet speed ($\mathcal{S}$) and Arctic temperature ($\mathcal{T}$). The results of the LASSO model run for a maximum lag of 25 days (p = 5) is shown in Figure 1a, while the results of the PC model run using 11 time slices is shown in Figure 1b. To create the time slices, we used the original variables (y) and 10 time shifted versions of y, namely shifted by -25, -20, ..., -5, +5, ..., +25 days [18]. All three methods–VAR (not shown), LASSO, and PC–agree quite well with each other.

The LASSO model (Figure 1a) shows both the magnitudes and the signs of the jet speed-Arctic temperature ($\mathcal{S}$-$\mathcal{T}$) relationship. First, we note that both $\mathcal{S}$ and $\mathcal{T}$ are autocorrelated (curved arrows), with coefficients that decay over the 25 day period but remain non-zero. Second, $\mathcal{T}$ drives $\mathcal{S}$ 5 days earlier (as well as 15, 20, and 25 days earlier), with the positive coefficient indicating that warmer temperatures drive a faster jet in the North Pacific. $\mathcal{S}$ also drives $\mathcal{T}$ at a lag of 5 days, with the negative coefficient indicating that faster jets are associated with a colder Arctic. However, at a lag of 15 days and beyond, the relationship between $\mathcal{S}$ and $\mathcal{T}$ changes–$\mathcal{S}$ drives $\mathcal{T}$ with positive LASSO coefficients, indicating that a stronger North Pacific jet drives warmer Arctic temperatures. **Collectively, the LASSO results indicate that there is a positive feedback loop between Arctic temperature and North Pacific jet speed–a warmer Arctic drives a stronger North Pacific jet, and the stronger jet drives further Arctic warming.**

The PC model (Figure 1b) agrees quite well with the results of the LASSO model (although its formulation does not provide the magnitudes or signs of the relationships). In the PC model, we did not allow instantaneous connections between variables to make it easier to compare results with the VAR and LASSO models. The autocorrelated relationships (curved arrows) in the PC model are quite similar to those in the LASSO model. In the PC model, $\mathcal{T}$ drives $\mathcal{S}$ at a lag of 5 days only, and $\mathcal{S}$ drives $\mathcal{T}$ at lags of 15 and 20 days. These are the lags with the strongest coefficients in the LASSO model. So, the PC model and the LASSO model show very similar results, with the lags with the strongest LASSO coefficients also showing significant relationships in the PC model.

Jet latitude, $\mathcal{L}$, also shows evidence of a causal relationship with $\mathcal{T}$ in both the LASSO and PC models (not shown). The influence of $\mathcal{T}$ on $\mathcal{L}$ is not strong, with both PC and LASSO showing few significant relationships. However, the influence of $\mathcal{L}$ on $\mathcal{T}$ is stronger. The LASSO model shows that $\mathcal{L}$ drives $\mathcal{T}$ with negative coefficients at most lags, indicating that a more poleward jet drives colder Arctic temperatures (and vice versa). The PC model shows a very similar relationship to the LASSO model, with $\mathcal{L}$ driving $\mathcal{T}$ at similar lags.

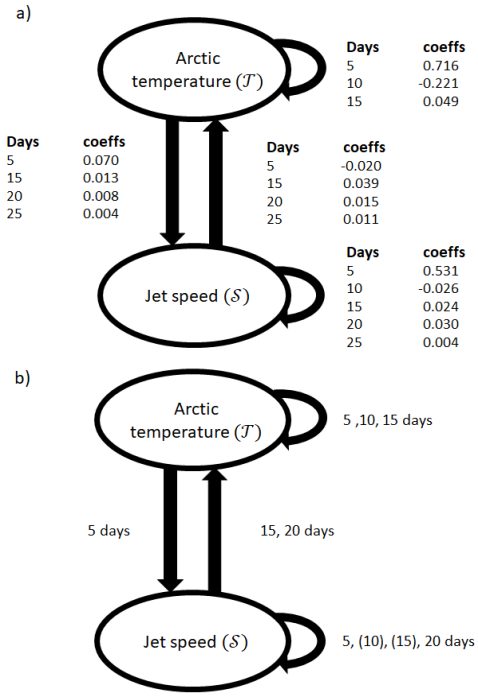


Fig. 1. Arctic temperature ($\mathcal{T}$) and jet speed ($\mathcal{S}$) relationships as described by (a) LASSO ($\lambda = 0.005, p = 5$) and (b) PC (11 time slices, $\alpha = 0.05$) models. Non-zero LASSO regression coefficients are shown next to their corresponding arrows in (a). (b) shows lags at which a significant relationship was present according to PC stable.

VII. CONCLUSIONS AND FUTURE WORK

Using VAR, LASSO, and PC models, we have demonstrated that Arctic temperature drives jet speed at timescales of 5-15 days in the North Pacific. This relationship is positive, with warmer Arctic temperatures driving a stronger jet, and a stronger jet driving warmer Arctic temperatures. The work described here is only the beginning of a larger study. Future steps include: (1) expansion of these methods to reanalysis, 2-D spatial fields, and inclusion of additional variables such as sea ice extent; (2) providing results from significance testing by comparing unrestricted and restricted VAR models; (3) quantifying the strength of causal relationships beyond use of regression coefficients.

ACKNOWLEDGMENTS

Support for this work was provided by NSF grants AGS-1545675 (Barnes) and AGS-1445978 (Ebert-Uphoff) under the CLD program.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] M. C. Serreze and R. G. Barry, *The Arctic Climate System*. Cambridge, UK: Cambridge University Press, 2005.

[2] E. A. Barnes and J. A. Screen, "The impact of Arctic warming on the midlatitude jet-stream: Can it? Has it? Will it?," *WIREs Clim. Change*, vol. 6, pp. 277–286, 2015.

[3] J. E. Overland and M. Wang, "Large-scale atmospheric circulation changes are associated with the recent loss of Arctic sea ice," *Tellus*, vol. 62, 2010.

[4] J. A. Francis and S. J. Vavrus, "Evidence linking Arctic amplification to extreme weather in mid-latitudes," *Geophys. Res. Lett.*, vol. 39, 2012.

[5] J. Liu, J. A. Curry, H. Wang, M. Song, and R. M. Horton, "Impact of declining Arctic sea ice on winter snowfall," *Proc. Natl. Acad. Sci. (USA)*, vol. 109, pp. 4074–4079, 2012.

[6] Q. Tang, X. Zhang, X. Yang, and J. A. Francis, "Cold winter extremes in northern continents linked to Arctic sea ice loss," *Environ. Res. Lett.*, vol. 8, 2013.

[7] J. A. Screen and I. Simmonds, "The central role of diminishing sea ice in recent Arctic temperature amplification," *Nature*, vol. 464, pp. 1334–1337, 2010.

[8] E. A. Barnes, "Revisiting the evidence linking Arctic amplification to extreme weather in midlatitudes," *Geophys. Res. Lett.*, vol. 40, 2013.

[9] E. A. Barnes, E. Dunn-Sigouin, G. Masato, and T. Woollings, "Exploring recent trends in Northern Hemisphere blocking," *Geophys. Res. Lett.*, vol. 41, 2014.

[10] G. Magnusdottir, C. Deser, and R. Saravanan, "The effects of North Atlantic SST and sea ice anomalies on the winter circulation in CCM3. Part I: Main features and storm track characteristics of the response," *J. Climate*, vol. 17, pp. 857–876, 2004.

[11] A. H. Butler, D. W. J. Thompson, and R. Heikes, "The steady-state atmospheric circulation response to climate change-like thermal forcings in a simple general circulation model," *J. Climate*, vol. 23, pp. 3474–3496, 2010.

[12] C. Deser, R. A. Tomas, M. Alexander, and D. Lawrence, "The seasonal atmospheric response to projected Arctic sea ice loss in the late twenty-first century," *J. Clim*, vol. 23, pp. 333–351, 2010.

[13] Y. Peings and G. Magnusdottir, "Response of the wintertime Northern Hemisphere atmospheric circulation to current and projected Arctic sea ice decline: A numerical study with CAM5," *J. Climate*, vol. 27, pp. 244–264, 2014.

[14] J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling, "Climate data challenges in the 21st century," *Science*, vol. 331, no. 6018, pp. 700–702, 2011.

[15] T. Chu, D. Danks, and C. Glymour, "Data driven methods for nonlinear granger causality: Climate teleconnection mechanisms," tech. rep., Carnegie Mellon University, Department of Philosophy, 2005.

[16] X. Chen, Y. Liu, H. Liu, and J. G. Carbonell, "Learning spatial-temporal varying graphs with applications to climate data analysis.," in *AAAI*, 2010.

[17] M. T. Bahadori and Y. Liu, "Granger causality analysis with hidden variables in climate science applications," in *Climate Informatics workshop (CI 2011)*, 2011.

[18] 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.

[19] T. Zerenner, P. Friederichs, K. Lehnertz, and A. Hense, "A gaussian graphical model approach to climate networks," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 24, no. 2, p. 023103, 2014.

[20] D. Hammerling, A. H. Baker, and I. Ebert-Uphoff, "What can we learn about climate model runs from their causal signatures," in *Proceedings of the Fifth International Workshop on Climate Informatics*, 2015.

[21] M. C. McGraw and E. A. Barnes, "Memory matters: A case for Granger causality in climate variability studies," *Journal of Climate*, 2017, under review.

[22] A. Arnold, Y. Liu, and N. Abe, "Temporal causal modeling with graphical granger methods," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 66–75, ACM, 2007.

[23] J. Runge, *Detecting and quantifying causality from time series of complex systems*. PhD thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät, 2014.

[24] 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.

[25] C. Strong, G. Magnusdottir, and H. Stern, "Observed feedback between winter sea ice and the north atlantic oscillation," *Journal of Climate*, vol. 22, no. 22, pp. 6021–6032, 2009.

[26] M. Kretschmer, D. Coumou, J. F. Donges, and J. Runge, "Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation," *Journal of Climate*, vol. 29, no. 11, pp. 4069–4081, 2016.

[27] H. Lütkepohl, *New Introduction to Multiple Time Series Analysis*. Springer, corr. 2nd printing ed., 2007.

[28] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[29] W. B. Nicholson, D. S. Matteson, and J. Bien, "VARX-L: Structured regularization for large vector autoregressions with exogenous variables," *International Journal of Forecasting*, vol. 33, no. 3, pp. 627–651, 2017.

[30] Wikipedia, "Lasso (statistics)." https://en.wikipedia.org/wiki/Lasso_(statistics), as of Sept 2017.

[31] B. Pfaff *et al.*, "VAR, SVAR and SVEC models: implementation within R package vars," *Journal of Statistical Software*, vol. 27, no. 4, pp. 1–32, 2008.

[32] G. Rebane and J. Pearl, "The recovery of causal poly-trees from statistical data," in *Proceedings of the Third Conference on Uncertainty in Artificial Intelligence*, pp. 222–228, AUAI Press, 1987.

[33] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufman Publishers, revised second printing ed., 1988.

[34] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT Press, 2nd ed., 2000.

[35] D. Colombo and M. H. Maathuis, "Order-independent constraint-based causal structure learning," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3741–3782, 2014.

[36] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social science computer review*, vol. 9, no. 1, pp. 62–72, 1991.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017 Hosted by the National Center for Atmospheric Research in Boulder, CO

A VISION FOR THE DEVELOPMENT OF BENCHMARKS TO BRIDGE GEOSCIENCE AND DATA SCIENCE

Imme Ebert-Uphoff[1], David R. Thompson[2], Ibrahim Demir[3], Yulia R. Gel[4], Mary C. Hill[5], Anuj Karpatne[6], Mariana Guereque[7], Vipin Kumar[6], Enrique Cabral-Cano[8], Padhraic Smyth[9].

***Abstract*— The massive surge in the amount of observational field data demands richer and more meaningful collaboration between data scientists and geoscientists. This document was written by members of the Working Group on Case Studies of the NSF-funded RCN on *Intelligent Systems Research To Support Geosciences (IS-GEO, https://is-geo.org/)* to describe our vision to build and enhance such collaboration through the use of specially-designed benchmark datasets. Benchmark datasets serve as summary descriptions of problem areas, providing a simple interface between disciplines without requiring extensive background knowledge. Benchmark data intend to address a number of overarching goals. First, they are concrete, identifiable, and public, which results in a natural coordination of research efforts across multiple disciplines and institutions. Second, they provide multi-fold opportunities for objective comparison of various algorithms in terms of computational costs, accuracy, utility and other measurable standards, to address a particular question in geoscience. Third, as materials for education, the benchmark data cultivate future human capital and interest in geoscience problems and data science methods. Finally, a concerted effort to produce and publish benchmarks has the potential to spur the development of new data science methods, while providing deeper insights into many fundamental problems in modern geosciences. That is, similarly to the critical role the genomic and molecular biology data archives serve in facilitating the field of bioinformatics, we expect that the proposed geosciences data repository will serve as "catalysts" for the new discicpline of geoinformatics. We describe specifications of a high quality geoscience benchmark dataset and discuss some of our first benchmark efforts. We invite the Climate Informatics community to join us in creating additional benchmarks that aim to address important climate science problems.**

Corresponding author: Imme Ebert-Uphoff, iebert@colostate.edu. [1]Electrical & Computer Engineering, Colorado State University, Fort Collins, CO. [2]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA. [3]Civil & Environmental Engineering, University of Iowa, Iowa City, IA. [4]Mathematical Sciences, University of Texas at Dallas, Richardson, TX. [5]Geology, University of Kansas, Lawrence, KS. [6]Computer Science & Engineering, University of Minnesota, Minneapolis, MN. [7]Geological Sciences, University of Texas, El Paso, TX. [8]Instituto de Geofisica, Universidad Nacional Autónoma de México, Coyoacán, CDMX, Mexico. [9]Computer Science, University of California, Irvine, CA..

I. MOTIVATION

For decades there has been a strong trend in the geosciences in the direction of larger, more diverse datasets that demand sophisticated mathematical and computer science expertise [1]. This is a consequence of improvements in computing power, which permit far more sophisticated physical modeling; improvements in measurement technology, which permit acquisition of high-resolution large-scale datasets; and demands of challenging problems such as measuring the planet's response to a changing climate. Meeting these challenges requires rich communication between the geoscientists familiar with the application domain and data scientists that could bring novel computational methods to the field. Closing this gap is a primary goal of the Climate Informatics workshop series, and is shared by this benchmark development effort. This document describes the manner in which benchmark standard datasets (or simply, benchmarks) of typical geoscience data analysis problems can bridge the two communities.

A. Relation to Existing Efforts

Benchmarks can be seen as an extension of classic data repositories. In particular, classic data repositories, such as those maintained by *NCAR* [2], *NOAA* [3], *NASA* [4], *USGS* [5], and related repositories [6], [7], provide vast amounts of data, but only domain scientists would know how to use the data efficiently, which questions to ask, and how to set up an interesting analysis [8]. The same holds for repositories maintained by journals, such as the *Geoscience Data Journal* [9] and *Nature Scientific Data* [10]. We seek to bridge the gap between the geoscientist and the data scientist by having geoscientists preselect and preprocess interesting data, couple them with interesting and unsolved science questions, and add data documentation and background explanations suitable for non-domain scientists. The key to the benchmarks is this *packaging* of existing data with science questions suitable for data scientists.

The proposed benchmarks can also be seen as an extension of existing efforts originating from the data science community, such as the CI Hackathon events [11], [12], the UIOWA Midwest Big Data Hackathon

[13], the Challenges in Machine learning events [14], the Kaggle data science platform [15], and the UCI Machine Learning Repository [16]. (In fact, we may incorporate the CI 2016 Hackathon topic [12] and topics from the UIOWA Midwest Big Data Hackathon [13] as benchmarks.) The following aspects distinguish our benchmark datasets from such data science competitions: (1) benchmarks tend to be more open-ended, i.e. there might be no pre-defined performance measure; (2) benchmark data sets are meant to initiate and stimulate interdisciplinary discussion, and in turn to faciliate *long-term* collaborations between data scientists and domain scientists; (3) the benchmark data in their current form do not aim to focus on comparison among different participant groups (i.e., no competition); however, the benchmarks can also be utilized for various data science contests and challenges involving analysis, modeling, validation, and prediction.

B. Specific Goals

The benchmarks are intended to serve several goals, including:

(1) A **means for two-way communication** to connect the two disciplines, geoscience and data science: (i) data scientists learn about typical data analysis tasks in the geosciences, including typical properties of geoscience data, the types of science questions geoscientists are interested in, and existing approaches for data analysis in the geosciences; (ii) geoscientists learn about potential new methods, tools, and services for data analysis.

(2) Benchmarks seek to **stimulate new collaborations**, which may lead to discovery of new approaches and methods for data analysis in the geosciences; science advances by gaining new insights from geo data using the new approaches; formation of new collaboration teams that will work together in the future.

(3) A **permanent repository of complex data sets, representative of geoscience problems** is an important resource for education, research, and to promote an emergent coordination of research activities and conversations in the research literature that build off each other (not possible when every lab has an independent data set with its own idiosyncrasies).

II. Desired Benchmark Characteristics

Given the list of goals in the preceding section, what are the key characteristics and elements of an ideal benchmark set? Below is a list of properties that we believe make a data set particularly suitable as a benchmark in this context. An outstanding benchmark is expected to satisfy many, but usually not all of these characteristics.

High Impact: A problem with high potential impact should be chosen, and the connection to that impact clearly spelled out, namely how will the proposed tasks contribute to advances in science or benefit society?

Active Geoscience research area: To stimulate long-term interactions between geoscientists and data scientists, the benchmark should come from an active research area, i.e. a group of geoscientists should be eager to continue to work on the topic, to answer questions from the data scientist(s), and to help him/her interpret any analysis results.

Challenge for data science: The problem should be challenging for the data scientists in some way. This is almost a given for geoscience applications, because 1) if the analysis was straightforward the geoscientist would have done the analysis him/herself; 2) geoscience data by their very nature tend to pose several challenges for standard data analysis methods (see Section II-A below). Known challenges should be spelled out for each benchmark, but some challenges will only become apparent during the analysis.

Data science generality and versatility: Ideally, solutions generated from the data set analysis and proposed models and algorithms will not only help to address a stated set of problems in geosciences, but also will be applicable to a broad range of other settings and possibly other disciplines, i.e. will stimulate and facilitate development of new methodology in data science.

Rich information content: Ideally the data set provides stimulus for analysis at many different levels, i.e. it lends itself to answering more than one science question. If so, one can gain a lot from a single data set.

Hierarchical problem statement: Each benchmark should include a data set and a clear description of what types of analyses are suggested. Ideally, there is a hierarchy of analysis tasks, ranging from relatively straight-forward tasks to more open-ended tasks.

A means for evaluating success: Data scientists need some kind of means to evaluate whether their algorithms are successful in solving the problem. Ideally, some kind of performance measure should thus be included for at least some of the tasks. However, in very open-ended applications, the performance measure might be developed during the collaboration.

Quick start guide: It should be as easy as possible for data scientists to start working with the data. Data scientists focus on data first, so the data needs to be easily accessible, and ideally there would be quick-start instructions on how to explore them. We seek to include for each benchmark a data use tutorial, consisting of (1) code snippets in a well-known framework (e.g., Matlab, Python, R) that illustrate how to read and visualize the data, potentially also illustrating some sample analysis steps; (2) plots generated from the code snippets that illustrate some of the data properties, so that data scientists can get a better feeling for the problem before even touching the data.

Understandable geoscience context: Geoscience data generally has a rich background, ranging from the motivation for collecting the data in the first place (science question), to the instruments used to take it,

the pre-processing that has already taken place, and the science questions it seeks to answer. Providing a brief summary of this background, in a way that is easy for someone outside the field to understand (no jargon), results in more efficient collaboration and may yield a more meaningful analysis of the data.

Citability: As discussed in the article on the *Geoscience Paper of the Future* [17], it is crucial to provide for each data set (1) a license specifying conditions for use, and (2) a unique and persistent identifier to make it citable, and later allow search engines to easily find all research papers using it. Both criteria can be met by using *Zenodo* (https://zenodo.org/) to host the data sets. Zenodo is a data repository run by CERN, that provides free hosting of data sets up to 50 GB, provides a selection of license terms and assigns a unique DOI number to each data set.

Communication between researchers: A public Google document provides both an FAQ and a communication channel for the domain experts and anyone working with the data. Researchers may use it to ask questions, exchange experiences and discuss results.

A. Suitable data science methods

Many data science methods cannot be directly applied to geoscience data, because of the challenging properties of such data. Karpatne et al. [18] categorize the most important challenging properties as follows: spatiotemporal structure; high dimensionality; heterogeneity in space and time; existence of objects with amorphous spatial/temporal boundaries; multi-scale/multi-resolution data; low sample size; paucity or absence of ground truth; noise, incompleteness, and uncertainty in data. It will be useful to identify for each benchmark which challenging properties are present.

Furthermore, it is difficult to convince geoscientists to use any method they do not understand. In fact geoscientists strongly prefer *transparent* methods, which allow them to follow the basic reasoning and generate novel scientific insights, over *black box* methods [19].

B. Distribution and Advertisement of benchmarks

All benchmarks will be featured on the IS-GEO website, and advertised through papers (such as this one), talks, and mailing lists, and we will reach out personally to data scientists through the IS-GEO members to make them aware of these benchmarks.

III. Sample Benchmarks

The IS-GEO benchmark project was created in Spring 2017. To date we have created one benchmark and are working on two more.

The first benchmark was developed in collaboration with researchers at the Jet Propulsion Laboratory (JPL), and deals with the automatic analysis of their imaging spectrometer data in order to detect significant sources of methane in the atmosphere [20], [21]. Methane (CH4) is a powerful Greenhouse Gas in the atmosphere and it is essential to determine its most important sources in the environment, such as geologic seeps, animal husbandry, decomposition in landfills, and oil and gas extraction and production. The ultimate goal of this benchmark is to develop methods for the reliable detection, and potentially classification, of methane sources from imaging spectrometer data. The key challenge is to distinguish methane sources from background noise in the spectrometer images. Domain experts currently perform this task manually by visual inspection of the imaging spectrometer data.

With regard to the requirements from Section II, this benchmark satisfies many of them. Namely, it is a high impact application, as it has the potential to reduce greenhouse gas emissions, and thus global warming; it is an active research area of research institutions such as JPL; it provides a rich, multi-layered and challenging playground for data scientists, because the data includes high levels of noise, as well as artifacts from roads and buildings that may be addressed through a variety of sophisticated statistical and image processing techniques; the problem statement consist of a hierarchy of tasks of increasing difficulty; we developed a quick start tutorial with Matlab code examples and visualization of sample data; there are manually labeled results for methane detection that can be used to evaluate performance for the simpler tasks, while the remaining tasks are more open-ended.

We are currently working with the team of the 2016 Climate Informatics Hackathon event to extend their challenge, prediction of sea ice cover based on several atmospheric variables [12], to a benchmark. We also collected a list from the IS-GEO community containing ten additional benchmark topics to consider.

IV. An Invitation to the CI Community

We invite the members of the Climate Informatics community to get involved in this effort. In particular, we would appreciate feedback on the general vision presented here, and any collaboration for the creation of additional benchmarks. Furthermore, we invite you to find out more about the general IS-GEO initiative (https://is-geo.org/) and to become a member of that community as well.

V. Acknowledgments

We are grateful to the ChaLearn organization for sharing resources and giving helpful advice. Their guidelines for setting up Challenges in Machine Learning [14] served as a great starting point. This activity is part of the IS-GEO Research Collaboration Network funded by the NSF (Award #1632211, EarthCube RCN IS-GEO: Intelligent Systems Research to Support Geosciences). A portion of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

REFERENCES

[1] I. Demir, H. Conover, W. F. Krajewski, B.-C. Seo, R. Goska, Y. He, M. F. McEniry, S. J. Graves, and W. Petersen, "Data-enabled field experiment planning, management, and research using cyberinfrastructure," *Journal of Hydrometeorology*, vol. 16, no. 3, pp. 1155–1170, 2015.

[2] National Center for Atmospheric Research (NCAR), "NCAR Community Data Portal (CDP)." http://cdp.ucar.edu/.

[3] National Oceanic and Atmospheric Administration (NOAA), "National Centers for Environmental Information (NCEI) - Data Access." https://www.ncdc.noaa.gov/data-access.

[4] National Aeronautics and Space Administration (NASA), "NASA's Open Data Portal." https://data.nasa.gov/.

[5] U.S. Geological Survey , "U.S. Geological Survey Science Data Catalog." https://data.usgs.gov.

[6] Interdisciplinary Earth Data Alliance (IEDA), "Observational solid earth data from the ocean, earth, and polar sciences - data repositories." http://app.iedadata.org/compliance/dmp/replist.php.

[7] U.S. Government, "The home of the U.S. Governments open data." https://www.data.gov/.

[8] I. Ebert-Uphoff and Y. Deng, "Three steps to successful collaboration with data scientists," *Earth and Space Science (EOS)*, vol. 98, 2017. https://doi.org/10.1029/2017EO079977.

[9] Royal Meteorological Society, "Geoscience data journal." An Open Access Journal, published by Wiley. http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)2049-6060.

[10] Macmillan Publishers (part of Springer Nature), "Scientific Data." An open-access journal for descriptions of scientifically valuable datasets, and research that advances the sharing and reuse of scientific data. https://www.nature.com/sdata/.

[11] Paris Saclay Center for Data Science, "Climate informatics hackathon 2015," 2015. https://www.lri.fr/~kegl/Ramps/edaElNino.html.

[12] B. Kégl and A. Rhines, "Climate informatics hackathon 2016," in *Proceedings of the 6th International Workshop on Climate Informatics (CI 2016)*, 2016. http://dx.doi.org/10.5065/D6K072N6.

[13] University of Iowa, "UIOWA Midwest Big Data Hackathon." http://bigdata.uiowa.edu.

[14] ChaLearn, "Challenges in machine learning." http://www.chalearn.org/.

[15] Kaggle, "Kaggle: Your home for data science." https://www.kaggle.com/datasets.

[16] UC Irvine, Center for Machine Learning and Intelligent Systems, "UCI machine learning repository." https://archive.ics.uci.edu/ml/datasets.html.

[17] Y. Gil, C. H. David, I. Demir, B. T. Essawy, R. W. Fulweiler, J. L. Goodall, L. Karlstrom, H. Lee, H. J. Mills, J.-H. Oh, *et al.*, "Toward the geoscience paper of the future: Best practices for documenting and sharing research from data to software to provenance," *Earth and Space Science*, vol. 3, no. 10, pp. 388–415, 2016.

[18] A. Karpatne, H. A. Babaie, S. Ravela, V. Kumar, and I. Ebert-Uphoff, "Machine learning for the geosciences - opportunities, challenges, and implications for the ML process," in *Workshop on Mining Big Data in Climate and Environment (MBDCE 2017), 17th SIAM International Conference on Data Mining (SDM 2017)*, April 2017.

[19] A. Karpatne, G. Atluri, J. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, "Theory-guided data science: A new paradigm for scientific discovery from data," *IEEE Transactions on Knowledge and Data Engineering*, 2017.

[20] C. Frankenberg, A. K. Thorpe, D. R. Thompson, G. Hulley, E. A. Kort, N. Vance, J. Borchardt, T. Krings, K. Gerilowski, C. Sweeney, *et al.*, "Airborne methane remote measurements reveal heavy-tail flux distribution in four corners region," *Proceedings of the National Academy of Sciences*, p. 201605617, 2016.

[21] D. Thompson, I. Leifer, H. Bovensmann, M. Eastwood, M. Fladeland, C. Frankenberg, K. Gerilowski, R. Green, S. Kratwurst, T. Krings, *et al.*, "Real-time remote detection and measurement for airborne imaging spectroscopy: a case study with methane," *Atmospheric Measurement Techniques*, vol. 8, no. 10, pp. 4383–4397, 2015.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

FIRE EVENT PREDICTION FOR IMPROVED REGIONAL SMOKE FORECASTING

Zachary Butler[1], Yang Chen[2], James Randerson[2], and Padhraic Smyth[1]

Abstract— **Smoke from wildfires is a significant public health concern with over 300,000 people dying annually worldwide. Given these large health impacts an important goal is to forecast fire emissions on multi-day time scales, for example, to provide higher quality forecasts for operational smoke forecasting systems. In this paper we describe initial work on statistical predictive modeling techniques that use historical satellite and weather data to predict fire activity on daily time-scales and for a regional spatial domain. Prediction results from 10 years of wildfire data in Alaska illustrate how local weather information can be used to improve the quality of multiday fire forecasts.**

I. MOTIVATION AND BACKGROUND

Fire is an important and dynamic ecosystem process that responds to climate change and human modification of the land surface [1]. Fire emissions of greenhouse gases, ozone precursors, and black carbon aerosols have a warming effect on climate, whereas emissions of organic carbon aerosols and post-fire changes in species composition (and surface biophysics) may have an opposing effect. In concert, human health impacts from fire aerosols are widespread and significant [2]. Smoke impacts on health are amplified in regions downwind of large regional fire complexes [3,4]. For example, exceptionally large fire complexes in Alaska and Canada in June of 2015 generated smoke plumes that traveled thousands of miles, and significantly reduced air quality in cities across the central U.S.

To help mitigate these impacts, several federal agencies have created smoke forecasting systems, such as the European Union's Monitoring Atmospheric Composition and Climate System (MACC) [5], the U.S. Navy's Fire Locating and Monitoring of Burning Emissions (FLAMBE) Project [6], NOAA's Smoke Forecasting System [7], the U.S. Forest Service's BlueSky smoke modeling framework [8], and NASA's GEOS-5 Forward Processing (FP) system [9]. These systems often use near real-time satellite observations of fire radiative power to estimate the spatial pattern of fire emissions. The fire emissions, in turn, are introduced into an atmospheric model that uses weather forecasts of winds and other meteorological variables to transport the smoke into downwind areas. Most of these systems assume that the spatial structure and intensity of fire emissions remain constant over the duration of the forecast. Thus, while the evolving impact of fires on atmospheric composition is determined by the influence of meteorology on aerosol transport and loss processes, increases in burned areas or modification of fire behavior due to changing weather are not considered.

In this paper we describe our work on developing models that can predict daily fire emissions over the course of a weather forecast. In our approach, we draw upon satellite data streams and online archives of weather forecasts, with a specific focus on Alaska (the methodology is, however, more broadly applicable). A primary goal of this work is to understand limits to fire prediction originating from uncertainties in weather forecasting and from our ability to model fire behavior. Unlike most existing fire prediction systems that use physically-based fire behavior models to predict spread rates of individual fires and which do not track fires locally [e.g., 10, 11, 12, 13], our approach is designed to track multiple fires simultaneously and to predict new ignitions. Fire emissions forecasting at this larger spatial scale represents a novel prediction challenge and is needed for operational smoke forecasting systems operating at regional or global scales. In subsequent work we plan to couple our fire prediction algorithms to smoke forecasting systems with the goal of improving the quality of aerosol predictions from existing approaches. We expect that by allowing emissions to evolve dynamically over the duration of a forecast, we will be able to considerably improve the accuracy and value of smoke forecasting systems for public health and air quality applications.

Corresponding author: P. Smyth, smyth@ics.uci.edu.
[1]Department of Computer Science, University of California, Irvine
[2]Department of Earth System Science, University of California, Irvine.

II. DATA

By exploiting the strong emission of mid-infrared radiation from fires, Moderate Resolution Imaging Spectroradiometer (MODIS) instruments onboard NASA's Terra and Aqua satellites detect active fires at a ~1 km spatial resolution using a contextual algorithm [14]. Here we used the daily active fire count in the MODIS Fire Location Product (MCD14ML, from http://modis-fire.umd.edu/) as our target variable y for fire prediction – see Figure 1 for an example below.

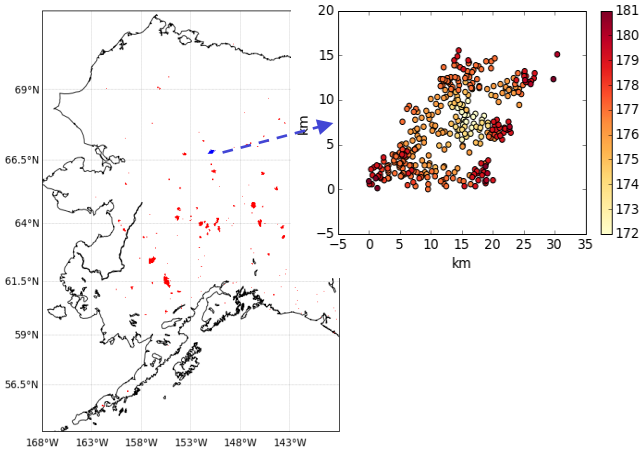


Figure 1: Left: Active fire detections in Alaska during the 2013 fire season. Right: The spatio-temporal evolution of the blue detections (which form a fire cluster – see Section III), color-coded by day of detection.

The weather variables we used were from the NOAA Global Forecast System (GFS). GFS is a global numerical weather prediction system, which is run four times a day, and produces forecasts for up to 16 days in advance. Here we used gridded (0.5º) surface temperature, surface humidity, surface wind, and precipitation rate data from the GFS analysis [15]. These weather variables were converted to daily averages (daily accumulations for precipitation) (Figure 2). We collected both data sources for each fire season in Alaska from 2007-2016. We defined the fire season as May 14-Aug 31 - most fire detections occur within this window.

III. METHODS

The modeling problem of interest is to predict the total number of fire detections in a fixed spatial area on day $t+k$ given information on day t, with $k = 1, 2, 3, \ldots$ We decompose the overall problem into two parts: (1) spatially local predictions for clusters of fires, and (2) global prediction of new ignition events. (Only local prediction is described here given space limitations). We represent fires as spatio-temporal clusters of fire detections (e.g, Fig. 1). We assume two fire pixels belong to the same fire cluster if they are within 5 km of each other (or connected through a chain of points, each within 5 km of the next). This clustering yielded 1335 fire clusters over the period 2007-2016 with each cluster persisting for a mean duration of 9.7 days.

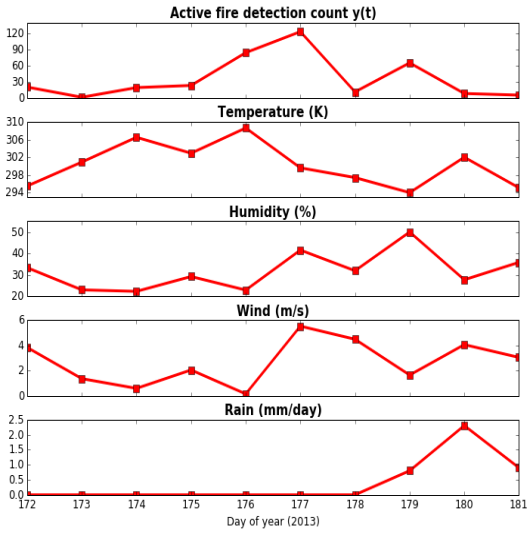


Figure 2: Time series of active fire counts for the cluster from Figure 1 with various weather variables used as model drivers. The increase in rain and humidity results in the fire dying out.

We predict the number of detections for active fires on day $t+k$ via Poisson regression [e.g., 16]. Specifically, we model the log of the Poisson rate on day $t+k$ (i.e., the expected number of detections on that day) for fire cluster i as

$$\log(\, E[y^i_{t+k}]\,) = B_0 + B_y \log\left(y^i_t + 1\right) + \sum_w B_w x^i_{t+k,w}$$

Here $\log y^i_{t+k}$ is the number of fire detections for cluster i on day $t+k$, $x^i_{t+k,w}$ represents different weather variables w for cluster i on day $t+k$, and the B's are global model coefficients (not dependent on t or on cluster i). The B's are estimated by maximizing the log-likelihood, defined as the sum of the log probability of the observed data on day $t+k$ (under a Poisson model), over clusters and days for each cluster, and where the Poisson mean is a function of the model parameters (the B's), conditioned on the covariates y^i_t and $x^i_{t+k,w}$. The weather variables $x^i_{t+k,w}$ are estimated at the spatial centroid of fire i on day $t+k$ (e.g., see Figure 1 (right)). Thus, predictions are made locally in time and space for each fire cluster: metrics for assessing performance are then aggregated over fire clusters and over days when each fire is active. The results in this abstract are based on having "perfect weather forecasts", using the actual future

weather data as a proxy for forecasted weather, realizing that actual forecasts will be noisier.

IV. RESULTS

Our experiments addressed three questions:

1. Can we predict fire detections more accurately than the baseline of $y(t+k) = y(t)$?
2. To what extent can weather covariates improve predictions beyond autoregression?
3. How does predictive performance decrease in quality as we predict further into the future?

To answer these questions we conducted two experiments. In the first experiment we fit a model to all years to predict on each day, and for each fire cluster, the number of detections on day $t+k$ given covariates defined on day t. The resulting regression coefficients (the B's) for the log of the Poisson rate are shown in Table 1 below.

	Intercept	Counts y(t)	Temp	Hum	Wind	Rain
Normalized	1.138	0.935	0.217	-0.043	0.033	-0.556
Unnormalized	-9.473	0.717	0.035	-0.002	0.022	-0.163

Table 1: Regression coefficients for the cluster Poisson regression model.

The normalized coefficients above are for a model fit with standardized inputs (with a mean of zero, standard deviation of 1) and the unnormalized results are for a model without standardization. The model coefficients agree with physical intuition. The coefficients for temperature and wind are positive and those for humidity and rain are negative. For example, a unit increase in temperature (one degree Kelvin) for the unnormalized model results in a multiplicative increase in expected number of detections by a factor of exp(0.035) = 1.036. Similarly a unit increase in rainfall (1 mm/day) corresponds to a multiplicative decrease of exp(-0.163) = 0.850 in expected count of fire detections per cluster.

The second experiment was designed to evaluate the relative accuracy of different models via cross-validation at the yearly level. We trained models on every year but one and evaluated performance on the held-out year, then aggregating the results across all held-out years. We fit instances of each model to make predictions at day $t+k$, k=1,2,3,4, and 5, conditioned on (a) fire detections at day t, and (b) weather covariates defined on day $t+k$. Since many existing smoke forecasting models assume that fire detections are constant over the duration of the forecast, i.e., $y(t+k) = y(t)$, we compared against this as a baseline. The three types of models we trained are:

1. Autoregression (with lag 1): the only covariate is the number of detections on day t, $y(t)$.
2. Temp/Hum: the covariates include autoregression, $y(t)$, and temperature and humidity on day $t+k$.
3. All weather: This adds rain and humidity to the Temp/Hum model.

Figure 3 below shows the mean absolute error (MAE) of predicted detections compared to actual number of detections, as a function of k. Models with weather covariates outperform the baseline and autoregression across all values of k, indicating that regression models built on historical data can provide systematic improvements over current forecasting practices.

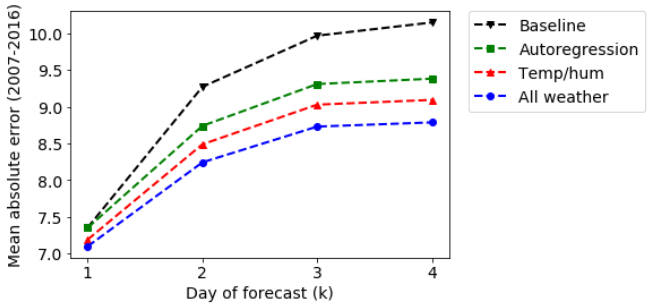


Figure 3: Cross-validated MAE from 2007-2016. The x-axis is the day we are predicting: at x=3, we are using the counts at time *t* and weather at time *t*+3 to predict counts at time *t*+3.

V. CONCLUSIONS

We investigated the use of statistical methods for predicting fire growth over time using patterns of historical fire and weather data in Alaska. We find that the incorporation of weather variables allows for more accurate prediction compared to models solely based on temporal autoregression. Under the assumption of perfect weather forecasts the relative improvements became larger the further the model forecasted into the future, suggesting that accurate weather forecasts have the potential to significantly improve the quality of smoke forecasts. Ongoing work and future directions include: measuring the degradation in accuracy from actual weather forecasts relative to perfect weather information; incorporating additional local variables such as vegetation, elevation, topological features such as rivers, lakes, and roads, history of prior burned areas; adding spatial context to the models; and adding longer term memory through fire weather indices to capture ground moisture and drying effects.

ACKNOWLEDGEMENTS

The work in this paper was supported in part by the National Science Foundation under award DGE-1633631.

REFERENCES

[1] Andela, N. et al.. A human-driven decline in global burned area. *Science*. 356 (2017), 1356–1362.

[2] Johnston, Fay H., et al. Estimated global mortality attributable to smoke from landscape fires. *Environmental Health Perspectives* 120.5 (2012): 695.

[3] Aouizerats, B., et al. Importance of transboundary transport of biomass burning emissions to regional air quality in Southeast Asia during a high fire event. *Atmospheric Chemistry and Physics* 15.1 (2015): 363-373.

[4] Marlier, Miriam E., et al. El Niño and health risks from landscape fire emissions in southeast Asia. *Nature Climate Change* 3 (2013): 131.

[5] Kaiser, J. W., et al. Biomass burning emissions estimated with a global fire assimilation system based on observed fire radiative power. *Biogeosciences* 9.1 (2012): 527.

[6] Reid, Jeffrey S., et al. Global monitoring and forecasting of biomass-burning smoke: Description of and lessons from the Fire Locating and Modeling of Burning Emissions (FLAMBE) program. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 2.3 (2009): 144-162.

[7] Rolph, Glenn D., et al. Description and verification of the NOAA smoke forecasting system: the 2007 fire season. *Weather and Forecasting* 24.2 (2009): 361-378.

[8] Larkin, Narasimhan K., et al. The BlueSky smoke modeling framework. *International Journal of Wildland Fire* 18.8 (2010): 906-920.

[9] Darmenov, A., and A. da Silva (2015). The Quick Fire emissions dataset (QFED): Documentation of versions 2.1, 2.2 and 2.4", Rep. NASA/TM–2015–104606/Vol. 38, Greenbelt, Maryland.

[10] Andrews Patricia L. Current status and future needs of the BehavePlus fire modeling system. *International Journal of Wildland Fire,* 23 (2013), 21-33.

[11] Sullivan, A. L. (2009). Wildland surface fire spread modelling, 1990–2007. 2: Empirical and quasi-empirical models. *International Journal of Wildland Fire*, *18*(4), 369-386.

[12] Altintas, I., et al., Towards an integrated cyberinfrastructure for scalable data-driven monitoring, dynamic prediction and resilience of wildfires, *Procedia Computer Science*, *51* (2015), 1633-1642.

[13] Taylor, S. W., Woolford, D. G., Dean, C. B., & Martell, D. L. (2013). Wildfire prediction to inform management: statistical science challenges. *Statistical Science*, 586-615.

[14] Giglio, L., J. D. Kendall, and R. Mack. A multi-year active fire dataset for the tropics derived from the TRMM VIRS. *International Journal of Remote Sensing* 24.22 (2003): 4505-4525.

[15] https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs

[16] Cameron, A. Colin, and Pravin K. Trivedi. *Regression Analysis of Count Data*. Cambridge University Press (2013).

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017 Hosted by the National Center for Atmospheric Research in Boulder, CO

LONG-RANGE FORECASTING USING COMPASS MACHINE LEARNING

Alison O'Connor[1], Ray Bell[2], Ben Kirtman[2], Joe Gorman[1]

Abstract—**The Climatological Observations for Maritime Prediction and Analysis Support Service (COMPASS) uses machine learning to create long-range forecasts of the probability that future conditions will differ from average climatology or mission-specific thresholds. COMPASS learns over multi-model forecast data to generate skillfully-superior forecasts to improve mission readiness and effectiveness; ensure safety; as well as reduce cost, labor, and resource requirements. Furthermore, COMPASS enables Navy operational planners and decision makers to use more reliable long-range forecasting capabilities to improve current forecasting systems and mission-planning tools.**

I. MOTIVATION

Current US Navy forecasting systems and mission-planning tools cannot easily incorporate long-range forecasts that can improve mission readiness and effectiveness; ensure safety; as well as reduce cost, labor, and resource requirements. If Navy operational planners and decision makers had tools and systems that incorporated these long-range forecasts, they could plan missions using more reliable long-range weather and climate predictions.

Recently, large research efforts have focused on improving the quality of long-range forecasts. Specifically, long-range multi-model forecast ensembles have been developed with higher predictive performance compared to individual forecast models [1], [2], [3], [4], [5], [6]. Multi-model ensembles are forecasting systems that consist of several forecast models from different modeling centers. Numerous studies have shown that multi-model ensemble approaches based on dynamic predictions increase the accuracy and tractability goals of long-range forecasting due to error cancellation and non-linear interaction of diagnostics [5]. At the forefront of these efforts are the North American Multi-Model Ensemble (NMME) [3] and the SubSeasonal (SubX) Multi-Model Ensemble Prediction System (EPS).

The NMME and SubX EPS offer predictions for environmental parameters (e.g., temperature and wind speed) for any region of the world, ranging from one week to a year in advance. Each model within the prediction systems consists of x ensembles ($x >= 10$) with varying initial conditions and model physics to produce a total of x predictions for every environmental parameter. The data is provided at daily frequency and a spatial resolution of 100 km^2.

The NMME and SubX EPS are ideal for system integration because of their skilled predictions; however, the ensemble techniques can be improved. Using forecast models from the NMME and SubX, higher skill predictions can be produced if forecast models are *combined effectively*.

Currently, combining forecast models is a cumbersome process. While many methods for weighting models exist [7], [8], [9], the best method for a given mission requires expert knowledge of both the forecast models and possible combination methods. When weights are assigned, they often apply to models as a whole. This is problematic because each forecast model differs in performance under different circumstances. For example, some models excel in certain regions, such as the tropics; others excel during specific time periods, such as winter. Currently, models are not combined using fine-grained weight assignments based on region, time period, and parameter. A fine-grained approach to weighting would reflect the forecast skill of models in various predictive environments, leveraging the strengths of each model as it pertains to a specific mission.

We are developing a robust service known as the Climatological Observations for Maritime Prediction and Analysis Support Service (COMPASS). COMPASS accurately and efficiently combines long-range forecast model ensemble data by learning fine-grained weight assignments via machine learning to

Corresponding author: A. O'Connor, aoconnor@cra.com
[1]Charles River Analytics, Inc. [2]University of Miami

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

greatly improve the operational plans produced by forecasting systems and mission-planning tools.

II. METHOD

Under the COMPASS effort, Charles River is developing a service that uses machine learning to successfully combine long-range data from multi-model forecast ensembles. Given a desired region, time period, and environmental parameters for an upcoming mission, COMPASS automatically evaluates, selects, and combines data from long-range forecast models. The result is a single mission-specific forecast of the probability that future conditions will differ from average climatology or mission-specific thresholds.

To describe COMPASS' forecast generation method, let us assume in July 2017 a planner is interested in the probability that wind speeds will be below, near, or above a threshold of 20 mph during a mission set to take place in October 2017 (3 month lead-time). The mission is a 10-day aircraft carrier transit in the Pacific Ocean from San Diego, California to Honolulu, Hawaii. Using the specifications of this mission, COMPASS generates its forecast using the following process:

COMPASS begins by clustering all Octobers from historical years with multi-model ensemble data available (1981-2016) according to the Multi-Variate El Niño Southern Oscillation Index (MEI). For the clusters output by k-means clustering [10] (k = 2), COMPASS identifies which cluster is most similar to October 2017. The years in this cluster are used for training. Because the Octobers within this cluster are most representative of the test period, we expect improved results by learning over these years only. Learning over years that do not have similar context are likely to deteriorate results. However, learning using too few years may also deteriorate results. Therefore, if a cluster with less than five years is chosen, COMPASS learns over all years.

To build the training data set, COMPASS obtains all July forecasts of the Pacific region in October for the training years in the identified cluster. The forecasts obtained are generated by different models within the NMME. COMPASS also obtains the observed data specifying the conditions that actually occurred in the region and time period in each training year. Assuming there are 17 Octobers in the cluster and four NMME models with daily data (CanCM3, CanCM4, CCSM4, and CFSv2), there are a total of 68 historic forecasts for learning. Each model has at least 10 ensembles which increases the number of predictions by at least an order of magnitude.

Using the training data and observed data, COMPASS learns patterns of forecast model successes and failures, without user involvement. Specifically, COMPASS creates several machine-learned weighted forecast models for each of the training years by combining individual NMME forecast models. To create these weighted models, COMPASS applies equal weighting (EQ), ridge regression (RR) [11], and Bayesian model averaging (BMA) [8], creating three weighted models for each training year. Under the assumption that there are 17 training years, then there are 51 (17 years x 3 models) weighted models.

Next, COMPASS evaluates each weighted model using another algorithm that learns their historic successes for the mission's region, environmental parameters, and month in historic years. The reason for having this layer of learning is due to the fact that that weighted forecasts are often superior, in terms of skill, to individual ones; however, which weighted combination to use differs between missions. Due to the vast amount of data and expertise required of each NMME and weighted model, selecting the best weighted method is an intractable problem for humans. Therefore, by creating three weighted models using machine-learning techniques that have proven to be successful in weather and climate forecasting [7], [8], [11], we can then select the weighting method that best suites the specific mission.

The weighted models are assessed in terms of rank probability skill score (RPSS), which compares the rank probability score (RPS) of a prediction in the forecast to the RPS of the constant climatology forecast, using a tercile-based system (above, near, or below the threshold) [12]. The superior weighted method is determined based on the weighted method with the largest number of predictions that yielded the highest RPSS across all training years as compared to the other two methods. Once we decide on the superior weighted method, COMPASS creates the weighted model accordingly for the test period (October 2017) using all NMME forecasts from the training years.

Finally, for each prediction in the final forecast, COMPASS obtains a probabilistic distribution of whether the environmental parameter will be below, near, or above the average climatology and mission-specific threshold.

III. EVALUATION

Early experimental results and analysis demonstrate the benefits of our approach to produce long-range forecasts of the probability that future conditions will differ from average climatology or mission-specific

thresholds that are superior, in terms of skill, to individual NMME forecast models.

To come to this conclusion, we evaluated two use cases. The first use case is similar to the example described in this paper: a Pacific transit from San Diego to Honolulu. However, in the use case we evaluated, the forecast was made in May 2010 for December 2010. The second use case is an Atlantic transit during December 2010 from Norfolk, VA to the Mediterranean Sea using the forecast from October 2010.

Our results are consistent between both use cases. Specifically, the COMPASS forecast outperforms all individual forecasts and is equal to the best performing weighted model. This can be seen in Fig. 1 for the temperature parameter for the Pacific use case.

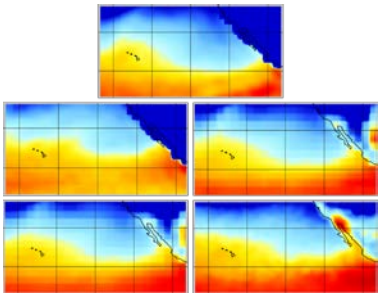


Fig. 1. (top): observed forecast over Pacific region; (middle-left) COMPASS forecast; (middle-right) CanCM3; (bottom-left) CCSM4; (bottom-right) CanCM4

The COMPASS forecast (middle-left), which is the BMA forecast, most similarly resembles the observed forecast (top) of the conditions that actually happened compared to the three individual NMME forecasts. Further, the BMA forecast is superior to the EQ and RR model for December 2010 as generated using all the training years.

For the second use case, the COMPASS forecast is the EQ forecast. To compare forecasts, we calculated the RPSS for each prediction in every individual NMME forecast and machine-learned model. For each prediction, we assigned the rank of 1 to the forecast with the highest (best) RPSS, 2 to the forecast with the second highest RPSS, and so on. We show these rankings for the second use case in Table 1 for the wind speed parameter.

Table 1: Rank Probability Skill Score Results

Model	Rank Score	No. 1 Ranks	No. 2 Ranks	No. 1 & 2 Ranks	No. 7 Ranks
COMPASS-(EQ)	1482	538	420	958	14
BMA	1349	463	435	898	12
CanCM4	531	298	222	520	287
CFSv2	487	237	187	424	174
RR	451	306	181	487	342
CCSM4	430	237	161	398	205
CanCM3	417	294	127	421	298

***Rank score: 2 * No. 1 Ranks + No. 2 Ranks – No. 7 Ranks**

There were over 1,500 predictions (one for every day and 100 km^2 region) in the entire October 2017 forecast. The COMPASS forecast is the best forecast, in terms of the rank score metric, compared to any other forecast (individual or weighted). Furthermore, it is ranked first more often than any other forecast and ranked 7th (last) for only 14 predictions in the entire forecast. It is a vast improvement over the individual NMME models across all metrics. We compared the forecasts on several other metrics not shown in Table 1, including Brier skill score [12], Heidke hit rate [13], and false alarm rate [14]. The COMPASS (EQ) forecast had the highest Brier skill score, lowest false alarm rate, and a near equal Heidke hit rate to the BMA forecast, which was not selected as the best weighted method.

Currently, we are continuing to refine the COMPASS machine-learning approach and compare it to related work [15]. As we continue to compare several procedural variations, such as number of years in the training data and the use of forecast scoring criteria (in addition to or in place of RPSS), to discover which forecast generation approach yields weighted forecasts that consistently out-perform individual ones. Some of our preliminary findings include:

- The best weighted forecast is strictly better than all individual models
- Performance increases with more training years but only to a point
- Performances increases with clustering
- Two layers of learning, in which we learn weighted models from individual models then select the best one, consistently out-performs: (1) using only one weighted method to learn a single forecast directly from individual NMME forecasts; and (2) combining all three weighted combinations into a single forecast

ACKNOWLEDGMENTS

This work was performed under the US Office of Navy Research (ONR) contract number N00014-15-P-1067. This work was funded in its entirety by ONR. The authors would like to thank Dr. Daniel Eleuterio for his technical support and engagement on this project.

REFERENCES

[1] Goddard, L., and co-authors. "Current approaches to seasonal-to-interannual climate predictions," International Journal of Climatology, 21, 1111–1152, 2001.

[2] Kirtman, B. P. "The COLA anomaly coupled model: Ensemble ENSO prediction," Monthly Weather Review, 131, 2324-2341, 2003.

[3] Kirtman, B. P., Min, D., Infanti, J. M., Kinter III, J. L., Paolino, D. A., Zhang, Q., van den Dool, H., Saha, S., Mendez, M. P., and Becker, E. "The North American Multi-Model Ensemble (NMME): Phase-1 seasonal to interannual prediction, phase-2 toward developing intra-seasonal prediction," Bulletin of the American Meteorological Society, 2013.

[4] Palmer, T. N., Brankovic, C., and Richardson, D.S. "A probability and decision-model analysis of PROVOST seasonal multimodel ensemble integrations," Quarterly Journal of the Royal Meteorology Society, 126, 2013–2034, 2000.

[5] Palmer, T.N., and co-authors. "Development of a European multi-model ensemble system for seasonal-to-interannual prediction (DEMETER)," Bulletin of American Meteorology Society, 85, 853-872, 2004.

[6] Hagedorn, R., Doblas-Reyes, F. J., and Palmer, T. N. "The rationale behind the success of multi-model ensembles in seasonal forecasting–I", Basic concept. *Tellus A*, 57, 219-233, 2005.

[7] Montgomery, J. M., Hollenbach, F. M., and Ward, M. D. "Improving predictions using ensemble Bayesian model averaging," Political Analysis, 20, 271-291, 2012.

[8] Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. "Using Bayesian model averaging to calibrate forecast ensembles," Monthly Weather Review, 133, 1155-1174, 2005.

[9] Tippett, M. K., Barnston, T., Goddard, L., Mason, S., Mendez, M. P., and van den Dool, H. "Recalibrating and Combining Ensemble Predictions," (2011).

[10] Lee, Howard B., and James B. Macqueen. "A K-Means cluster analysis computer program with cross-tabulations and next-nearest-neighbor analysis," *Educational and Psychological Measurement* 40.1: 133-138, 1980.

[11] Pena, Malaquias, and Huug van den Dool. "Consolidation of multimodel forecasts by ridge regression: Application to Pacific sea surface temperature," *Journal of Climate* 21.24: 6521-6538, 2008.

[12] Weigel, Andreas P., et al. "The Discrete Brier and Ranked Probability Skill Scores," Monthly Weather Review, 135, 1, 118-124, 2007.

[13] Barnston, Anthony G. "Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score," *Weather and Forecasting* 7.4: 699-709, 1992.

[14] Mason, S. J., & Graham, N. E. "Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels," *Weather and Forecasting*, *14*(5), 713–725: 1999.

[15] [15] Monteleoni, C., Schmidt, G. A., Saroha, S., & Asplund, E. "Tracking climate models," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *4*(4), 372-392, 2011.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

WASSERSTEIN k-MEANS++ FOR CLOUD REGIME HISTOGRAM CLUSTERING

Matthew Staib[1], Stefanie Jegelka[1]

***Abstract*—Much work has sought to discern the different types of cloud regimes, typically via Euclidean k-means clustering of histograms. However, these methods ignore the underlying similarity structure of cloud types. Wasserstein k-means clustering is a promising candidate for utilizing this structure during clustering, but existing algorithms do not scale well and lack the quality guarantees of the Euclidean case. We resolve this by generalizing k-means++ guarantees to the Wasserstein setting and providing a scalable minibatch algorithm for Wasserstein k-means. Our methods empirically perform well and lead to new, different cloud regime prototypes.**

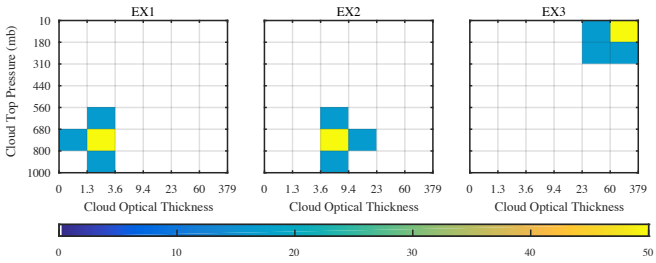


Fig. 1. In Euclidean distance, EX1 is equally far from EX2 and EX3. In the specific Wasserstein distance defined in Section IV, EX3 is over 20 times farther from EX1 than EX2 is from EX1, because EX1 and EX2 are concentrated on similar PC-TAU cells.

I. MOTIVATION

Given the climatic importance of clouds, much recent work has focused on identifying and then analyzing the main cloud regimes [1], [2], [3], [4], [5], [6], [7]. Once determined, these regimes are used in many settings, e.g., assessing general circulation models [5], [8], and therefore accurately identifying these regimes is crucial to understanding the climate system.

The vast majority of work applies k-means clustering to joint histograms of cloud top pressure (PC) and optical depth (TAU) (henceforth PC-TAU histograms of "cloud types"), e.g. [1], [2], [6]. Histograms are treated as vectors and compared via the Euclidean distance between them. This approach scales well to large datasets but ignores the latent structure of the data, in particular the similarity between different cloud types. Moreover, the clustering problem is solved via Lloyd's algorithm [9], which is empirically effective but gives no guarantees about the cluster quality.

Instead, we apply histogram clustering techniques based on Wasserstein distance [10], a metric between probability distributions (or histograms) that respects the underlying geometry of the space, in this case the similarity structure of cloud types. As illustrated in Figure 1, histograms with similar frequencies for similar cloud types are close in this metric, in contrast to Euclidean distance, which ignores cloud type similarity. We further 1) show that k-means++ seeding [11], which gives provably good cluster seedings in the Euclidean case, yields the same guarantee for the Wasserstein metric, 2) provide an efficient minibatch algorithm for Wasserstein k-means that scales to climate data, and 3) show histogram clustering can yield notably different cloud regimes than identified via Euclidean k-means.

Corresponding author: M Staib, mstaib@mit.edu [1]Computer Science and Artificial Intelligence Laboratory, MIT

II. THEORETICAL BACKGROUND

Given a set of points $\{x^i\}_{i\in\mathcal{I}}$, metric k-means clustering seeks to find a set of centroids $\mathcal{C} = \{c^j\}_{j=1}^k$ in a convex set K (e.g. the probability simplex) minimizing

$$\phi(\mathcal{C}) = \sum_{i\in\mathcal{I}} \min_{j=1,\dots,k} d(x^i, c^j)^2. \tag{1}$$

Typically d is taken to be the Euclidean distance, $d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$. In this setting, Lloyd's algorithm [9], which alternates between assigning points x^i to the closest cluster centroid c^j and replacing c^j with the mean of the points assigned to it, converges to a local optimum but lacks other guarantees: in fact, finding an optimal set $\mathcal{C}$ of centroids is NP-hard [12]. The k-means++ seeding algorithm alleviates this problem: this efficient, randomized algorithm produces an $O(\log k)$-optimal clustering in expectation [11]. This solution can then be fine-tuned by Lloyd's algorithm. This result has been extended to the case when $d(x,y)^2$ is replaced by a Bregman or total Jensen divergence [13], [14].

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

Algorithm 1 Minibatch metric k-means

Input: point set X, parameter k
$\{c^j\}_{j=1}^k \leftarrow k$-MEANS++INITIALIZATION$(X, k)$
$n_j \leftarrow 0$ for $j = 1, \ldots, k$ ▷ Cluster sizes
loop
  Draw $x^1, \ldots, x^m \sim X$
  $s_i \leftarrow \text{argmin}_{j=1,\ldots,k} d(x^i, c^j)^2$ for $i = 1, \ldots, m$
  for $i = 1, \ldots, m$ **do**
    $j \leftarrow s_i$ ▷ Cluster index assigned for x^i
    $\gamma \leftarrow 1/n_j$
    $c_j \leftarrow \text{proj}_K(c_j - \gamma \nabla_c [d(x^i, c^j)^2])$
    $n_j \leftarrow n_j + 1$
  end for
end loop

Results for general metrics exist for other seeding algorithms, e.g. [15], but these scale poorly and are hence impractical; to our knowledge, the general metric case has not yet been addressed for k-means++.

In contrast to Euclidean distance, Wasserstein distance between distributions μ and ν on points $\{y^i\}_{i=1}^n$ accounts for the "cost" C_{ij} of moving y^i to y^j. Viewing μ and ν as two piles of dirt, we can define a notion of distance between them: *how much* dirt must we move *how far* to transform one pile into the other, moving dirt as efficiently as possible? Formally, if $C_{ij} = g(y^i, y^j)^p$ for a distance metric g, the p-Wasserstein distance $W_p(\mu, \nu)$ is defined as the value of the linear program

$$\begin{array}{ll} \min & \langle C, T \rangle^{1/p} \equiv \min \left(\sum_{ij} C_{ij} T_{ij}\right)^{1/p} \\ \text{s.t.} & 1^T T = \mu,\ 1^T T^T = \nu,\ T \geq 0. \end{array} \tag{2}$$

The joint distribution T is a "transport plan" that moves mass from μ to ν. A full discussion of Wasserstein distance and optimal transport is outside the scope of this paper; we refer the reader to [10], [16] for theoretical foundations, and [17], [18], [19] for computing Wasserstein distance. In our clustering formulation, we use $d(x^i, c^j) = W_p(x^i, c^j)$.

Wasserstein distance has been applied to a limited extent to histogram clustering [20], [21]. The main computational challenge is computing the centroid, i.e., the Wasserstein barycenter of the measures in one cluster, in place of the Euclidean mean. Reasonably efficient barycenter algorithms exist [22], [21], [23] but scaling to large datasets remains an active research area.

III. Theory and Algorithm

We sample initial cluster centroids via k-means++ seeding where we replace the Euclidean by Wasserstein distance. Then we fine-tune the seeding with a stochastic minibatch k-means algorithm suitable for large scale climate data. Our Theorem III.1 states an approximation guarantee for our method; the seeding guarantee is proved by building on results from [24, Theorem 2]:

Theorem III.1. *Suppose centroids $\mathcal{C}$ are chosen via k-means++ seeding applied to any metric d (e.g. $d = W_p$). Then the objective function $\phi(\mathcal{C})$ satisfies*

$$\mathbb{E}[\phi(\mathcal{C})] \leq 8(\ln k + 2) \min_{\mathcal{C}^*} \phi(\mathcal{C}^*). \tag{3}$$

Once an initial seeding is selected, Lloyd's algorithm can be applied to fine-tune the clustering, and can only improve the objective value. However, updating the centroids requires expensive full passes over the dataset.

A more scalable alternative is a variant of online or minibatch gradient descent applied to Problem (1). In particular, we generalize an algorithm from [25] to the Wasserstein case. The result is our algorithmic contribution: Algorithm 1 enjoys the guarantees of Theorem III.1, and efficiently fine-tunes the clusters without many expensive passes over the entire dataset. In particular, for W_p distances, we can compute the required gradients $\nabla_c [d(x^i, c^j)^2]$ via linear programming and the chain rule, and project efficiently onto the simplex K [26], [27], [28]. Note that we accomplish this without ever needing to compute a Wasserstein barycenter, in contrast to past work on histogram clustering.

IV. Experiments

a) Experimental setup: We applied our clustering framework to PC-TAU histograms from the International Satellite Cloud Climatology Project (ISCCP) [29]. We focused specifically on data from the tropical region within 15° of the equator as in [2], in 3 hour increments from 1994-2009.

Wasserstein distances depend on a "ground" distance metric g between points: we built the ground metric g by mapping the cloud top pressure and optical depth pairs to an equally-spaced grid in $\mathbb{R}^2$ and using Euclidean distance. An extra "no cloud" state is added with constant distance $0.5D$ to each other state as in [17], where D is the maximum distance otherwise. We ran Algorithm 1 for 20 iterations with minibatch sizes of $m = 1000$. Gradients $\nabla_c [d(x^i, c^j)^2]$ were computed using Gurobi [30], and each outer iteration took about 10 seconds on a modern 8-core desktop computer. The initial k-means++ seeding was approximated using the algorithm from [31], with 2000 burn-in steps.

Both Euclidean and Wasserstein-based clustering were tested. Prior work had carefully determined the number of clusters k by analyzing correlations between cluster centroids [2], [3], [4], [6]. In the Euclidean

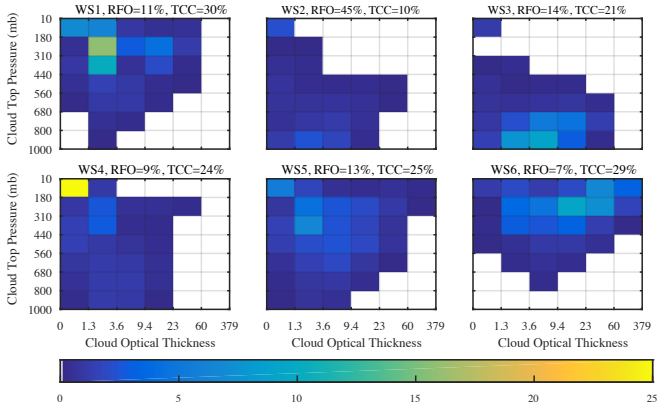


Fig. 2. Weather states (cluster centroids) produced by Algorithm 1 applied to Euclidean distance. Note the similarity to those from [2]. RFO is relative frequency of occurence; TCC is total cloud cover.

k	4	5	6	7	8
$W_1(c^i, c^j)^2$	0.283	0.244	0.168	0.141	0.174
$\phi(\mathcal{C})/\lvert\mathcal{I}\rvert$	0.107	0.098	0.086	0.078	0.074

Fig. 3. Minimum squared Wasserstein distance $W_1(c^i, c^j)^2$ between cluster centroids and the scaled k-means objective value $\phi(\mathcal{C})$, as the number of clusters k varies. Note that the nearest distance drops considerably from $k = 5$ to $k = 6$.

case, we chose $k = 6$ to match [2]. In the Wasserstein case, we instead analyzed the minimum W_p distance between cluster centroids, seeking a balance between a low objective value and spread out centroids.

b) Results: First, we applied Algorithm 1 to the standard Euclidean setting, producing cluster centroids (weather states) as shown in Figure 2. We essentially reproduce the same weather states as in [2] for the same tropical region.

We then clustered with respect to W_p distance, for $p \in \{1, 2\}$. Qualitatively, $p = 2$ led to centroids that are more spread out, as W_2 induces a lower penalty for moving mass between very close points. Hence, we focus on $p = 1$ in this paper. Table 3 shows the minimum W_1 distances between cluster centroids, together with estimates of $\phi(\mathcal{C})$. There is a notable dropoff in minimum distance after $k = 5$ without great improvement in the objective, so 5 clusters were chosen.

The resulting $k = 5$ weather states (WS) are shown in Figure 4. For each point in the tropical region, we give in Figure 5 a visual breakdown of how frequently that point belongs to each weather state (c.f. [2, Figure 2]). There are clear correspondences between the Euclidean-derived weather states and the Wasserstein ones. Note that Euclidean WS1, WS4, and WS5 split into Wasserstein WS3 and WS5. These Euclidean weather states

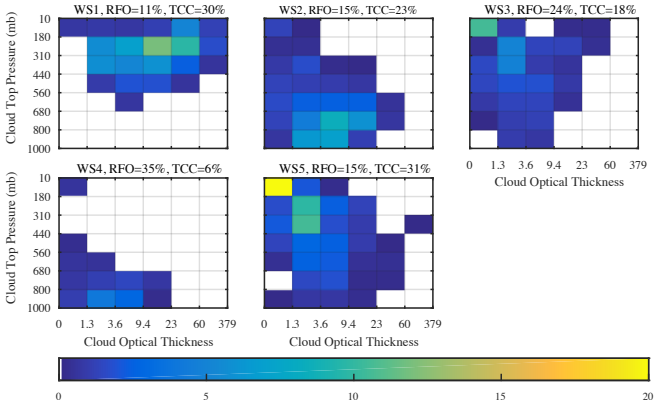


Fig. 4. Weather states from Algorithm 1 applied to W_1 distance. RFO is relative frequency of occurence; TCC is total cloud cover.

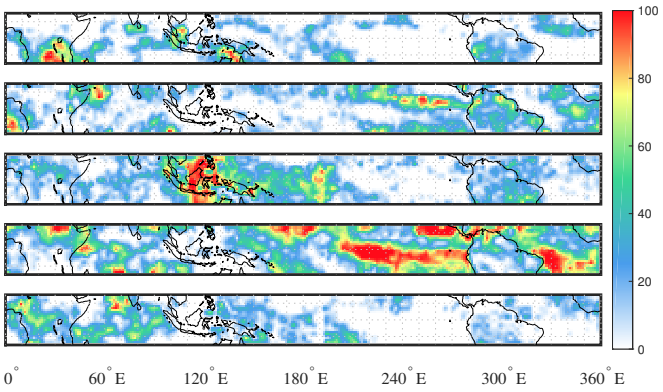


Fig. 5. Heatmaps for weather states 1 (top) through 5 (bottom). On the heatmap for one weather state, each point is colored according to how often it belongs to that state.

are more muddled, having very similar total cloud cover and concentration (under g); in contrast, their Wasserstein counterparts have similar concentration, but notably different total cloud cover.

V. DISCUSSION

We propose Wasserstein histogram clustering as a way to leverage prior knowledge about similarity and geometry in learning from climate datasets. We demonstrate that Wasserstein k-means++ clustering is achievable at large scale and with provable guarantees. Applying these techniques to cloud regimes yields different inferred weather states than Euclidean clustering.

For determining cloud regimes, we still need a principled way to select the ground distance metric between cloud types, perhaps via metric learning. Further in-depth analysis of these new, different weather states is needed, and of cloud regimes beyond the tropics considered here. More generally, identifying new geometry-aware clustering tasks in climate science is fertile ground for future work.

ACKNOWLEDGMENTS

This research was conducted with Government support under and awarded by DoD, Air Force Office of Scientific Research, National Defense Science and Engineering Graduate (NDSEG) Fellowship, 32 CFR 168a, and also supported by NSF CAREER award 1553284.

REFERENCES

[1] C. Jakob and G. Tselioudis, "Objective identification of cloud regimes in the Tropical Western Pacific," *Geophys. Res. Lett.*, vol. 30, p. 2082, Nov. 2003.

[2] W. B. Rossow, G. Tselioudis, A. Polak, and C. Jakob, "Tropical climate described as a distribution of weather states indicated by distinct mesoscale cloud property mixtures," *Geophys. Res. Lett.*, vol. 32, p. L21812, Nov. 2005.

[3] W. B. Rossow, Y. Zhang, and J. Wang, "A Statistical Model of Cloud Vertical Structure Based on Reconciling Cloud Layer Amounts Inferred from Satellites and Radiosonde Humidity Profiles," *J. Climate*, vol. 18, pp. 3587–3605, Sept. 2005.

[4] K. D. Williams and G. Tselioudis, "GCM intercomparison of global cloud regimes: Present-day evaluation and climate change response," *Clim Dyn*, vol. 29, pp. 231–250, Aug. 2007.

[5] K. D. Williams and M. J. Webb, "A quantitative performance assessment of cloud regimes in climate models," *Clim Dyn*, vol. 33, pp. 141–157, July 2009.

[6] G. Tselioudis, W. Rossow, Y. Zhang, and D. Konsta, "Global Weather States and Their Properties from Passive and Active Satellite Cloud Retrievals," *J. Climate*, vol. 26, pp. 7734–7746, May 2013.

[7] A. J. McDonald, J. J. Cassano, B. Jolly, S. Parsons, and A. Schuddeboom, "An automated satellite cloud classification scheme using self-organizing maps: Alternative ISCCP weather states," *J. Geophys. Res. Atmos.*, vol. 121, p. 2016JD025199, Nov. 2016.

[8] S. Mason, J. K. Fletcher, J. M. Haynes, C. Franklin, A. Protat, and C. Jakob, "A Hybrid Cloud Regime Methodology Used to Evaluate Southern Ocean Cloud and Shortwave Radiation Errors in ACCESS," *J. Climate*, vol. 28, pp. 6001–6018, Apr. 2015.

[9] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, pp. 129–137, Mar. 1982.

[10] C. Villani, *Optimal Transport: Old and New*. No. 338 in Grundlehren der mathematischen Wissenschaften, Berlin: Springer, 2009. OCLC: ocn244421231.

[11] D. Arthur and S. Vassilvitskii, "K-means++: The Advantages of Careful Seeding," in *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, (Philadelphia, PA, USA), pp. 1027–1035, Society for Industrial and Applied Mathematics, 2007.

[12] D. Aloise, A. Deshpande, P. Hansen, and P. Popat, "NP-hardness of Euclidean sum-of-squares clustering," *Mach Learn*, vol. 75, pp. 245–248, May 2009.

[13] S. Sra, S. Jegelka, and A. Banerjee, "Approximation algorithms for Bregman clustering, co-clustering and tensor clustering," tech. rep., 2008.

[14] F. Nielsen and R. Nock, "Total Jensen divergences: Definition, Properties and k-Means++ Clustering," *arXiv:1309.7109 [cs, math]*, Sept. 2013.

[15] S. Ahmadian, A. Norouzi-Fard, O. Svensson, and J. Ward, "Better Guarantees for k-Means and Euclidean k-Median by Primal-Dual Algorithms," in *Foundations of Computer Science*, 2017.

[16] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, vol. 87 of *Progress in Nonlinear Differential Equations and Their Applications*. Cham: Springer International Publishing, 2015.

[17] O. Pele and M. Werman, "Fast and robust Earth Mover's Distances," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 460–467, Sept. 2009.

[18] M. Cuturi, "Sinkhorn Distances: Lightspeed Computation of Optimal Transport," in *Advances in Neural Information Processing Systems 26* (C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, eds.), pp. 2292–2300, Curran Associates, Inc., 2013.

[19] A. Genevay, M. Cuturi, G. Peyré, and F. Bach, "Stochastic Optimization for Large-scale Optimal Transport," in *Advances in Neural Information Processing Systems 29* (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds.), pp. 3440–3448, Curran Associates, Inc., 2016.

[20] J. Li and J. Z. Wang, "Real-time computerized annotation of pictures," *IEEE Trans Pattern Anal Mach Intell*, vol. 30, pp. 985–1002, June 2008.

[21] J. Ye, P. Wu, J. Z. Wang, and J. Li, "Fast Discrete Distribution Clustering Using Wasserstein Barycenter With Sparse Support," *IEEE Trans. Signal Process.*, vol. 65, pp. 2317–2332, May 2017.

[22] M. Cuturi and A. Doucet, "Fast Computation of Wasserstein Barycenters," in *Proceedings of The 31st International Conference on Machine Learning*, pp. 685–693, 2014.

[23] M. Staib, S. Claici, J. Solomon, and S. Jegelka, "Parallel Streaming Wasserstein Barycenters," in *Neural Information Processing Systems*, 2017.

[24] F. Nielsen and K. Sun, "Clustering in Hilbert simplex geometry," *arXiv:1704.00454 [cs]*, Apr. 2017.

[25] D. Sculley, "Web-scale K-means Clustering," in *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, (New York, NY, USA), pp. 1177–1178, ACM, 2010.

[26] M. Held, P. Wolfe, and H. P. Crowder, "Validation of subgradient optimization," *Mathematical Programming*, vol. 6, pp. 62–88, Dec. 1974.

[27] C. Michelot, "A finite algorithm for finding the projection of a point onto the canonical simplex of ∝n," *J Optim Theory Appl*, vol. 50, pp. 195–200, July 1986.

[28] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l 1-ball for learning in high dimensions," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 272–279, ACM, 2008.

[29] W. B. Rossow and E. N. Dueñas, "The International Satellite Cloud Climatology Project (ISCCP) Web Site: An Online Resource for Research," *Bull. Amer. Meteor. Soc.*, vol. 85, pp. 167–172, Feb. 2004.

[30] I. Gurobi Optimization, "Gurobi Optimizer Reference Manual," 2016.

[31] O. Bachem, M. Lucic, S. H. Hassani, and A. Krause, "Approximate K-Means++ in Sublinear Time," in *Thirtieth AAAI Conference on Artificial Intelligence*, Feb. 2016.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

NON-UNIFORM SPATIAL DOWNSCALING OF CLIMATE VARIABLES

Soukayna Mouatadid[1], Steve Easterbrook[1], Andre Erler[2]

***Abstract*—The goal of this study is to present a scalable and robust approach to spatial downscaling of climate variables. We explore the ability of artificial neural networks (ANN) to downscale a climate variable to a given location of interest. We illustrate our proposed method in a downscaling application of monthly mean air temperature and precipitations at twelve stations located across the topographically complex province of British Columbia, Canada. Our method generalizes well to different locations and leads to high downscaling accuracy. The performance of the models is measured based on four statistical metrics, including the coefficient of determination, and the root mean square error.**

I. INTRODUCTION

Complete and accurate climate datasets are not readily available in many regions around the world. They are especially lacking in the areas most sensitive to climate change [1], due, in part, to the complex topography of such regions, where it is difficult to install and maintain weather stations. As a result, some of the regions most affected by climate change are unable to obtain detailed climate data needed to understand impacts and develop adaptation plans for future climate change [2].

To address this problem, scientists often rely on *gridded reanalysis products* as a replacement for observational data [3]. These datasets are produced by using the available station observations to constrain a physics-based simulation that then fills in the missing data points to provide a complete, physically realistic gridded data product [4]. However, gridded products for remote areas are typically coarse resolution, and do not capture small-scale climatic characteristics associated with regional topographic features, such as mountain ranges or lakes. For this reason, it is usually necessary to re-process these data sets to a finer scale, in a way that accounts for such features, but does not introduce additional errors and biases. This process is referred to as *downscaling*. This can be done using a high resolution regional dynamical model, but is computationally demanding. *Statistical downscaling* instead relies on statistical or empirical relationships between the large-scale predictor field from the model simulations and the variables of interest, at the location of interest. Statistical downscaling is challenging where there is insufficient historical data to derive robust relationships. Several recent papers review the spatial interpolation methods used for downscaling in meteorology and climatology [5], [6], [7].

We present a novel statistical downscaling method that learns from gridded reanalysis data and local station data. Our method learns a mapping between a low-resolution reanalysis dataset and the climate at specific locations, using an ANN model. It can be used for locations with available historical time-series (task 1) as well as locations where no historical data is available (task 2), a case where existing downscaling methods perform poorly.

II. MODEL DEVELOPMENT

For each task, we investigate the use of an ANN model. The theoretical background for the algorithm is provided in [8]. The predictand of our models is the expected value of a given climate variable at a specific location and time. We have tested the method for two variables: monthly mean temperature and monthly mean precipitation. Our predictors from the reanalysis dataset include monthly means of: cloud forcing net longwave flux; upward and downward solar radiation fluxes; u-wind and v-wind; relative humidity; and sea level pressure.

In the first task, we downscale the gridded reanalysis data to a location for which past observations are available. In this scenario, the historical values recorded at the station were used as the predictand, and the reanalysis data at 16 grid points around the station were used as model predictors, selected such that the location of the station of interest is at the center grid cell of a 4 x 4 sub-grid or square. We refer to these 16 grid points as the station's neighborhood. The studies in [9],

Corresponding author: S Mouatadid, soukayna@cs.toronto.edu
[1]Department of Computer Science, University of Toronto
[2]Department of Physics, University of Toronto

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

[4] showed that the sixteen grid points around a station of interest all supply relevant information to the model.

In the second task, the goal was to explore how a gridded dataset can be downscaled to locations where no past observational record is available. The methodology used here is similar to [9], where the focus was on predicting solar energy over a spatial grid by developing a support vector machine model for each individual cell of a gridded dataset. We develop a model for a location of interest, using the information available from that location's neighbourhood. Again, we use the square formed by the nine grid cells (i.e., 16 grid points) as the location's neighborhood. As there is no data for the location of interest, we use other stations within the given neighborhood. For the training set, the input variables are the reanalysis values from the sixteen grid points surrounding these stations along with each stations' coordinates (i.e., latitude, longitude and elevation), and the output variables are the observations recorded at the stations that fall within the neighborhood.

To test the method, we select one station as the location of interest, and exclude its data from the training set. The output variable in our tests corresponds to the observational data recorded at this location, and the input variables are the reanalysis values from the sixteen grid nodes around said location, and the location coordinates. During the training phase, the model has not been fed any value related to the location of interest, and during the testing phase, the model's only input is the information from the reanalysis dataset, and the location's coordinates. Figure 1 illustrates the construction of a test set for a model used to downscale to a location of interest (s1) with three neighbouring stations (s9, s11 and s12) used for training. Following this methodology, the models can be used to downscale to any location (any latitude, longitude, elevation), whether or not it is in the testing set.

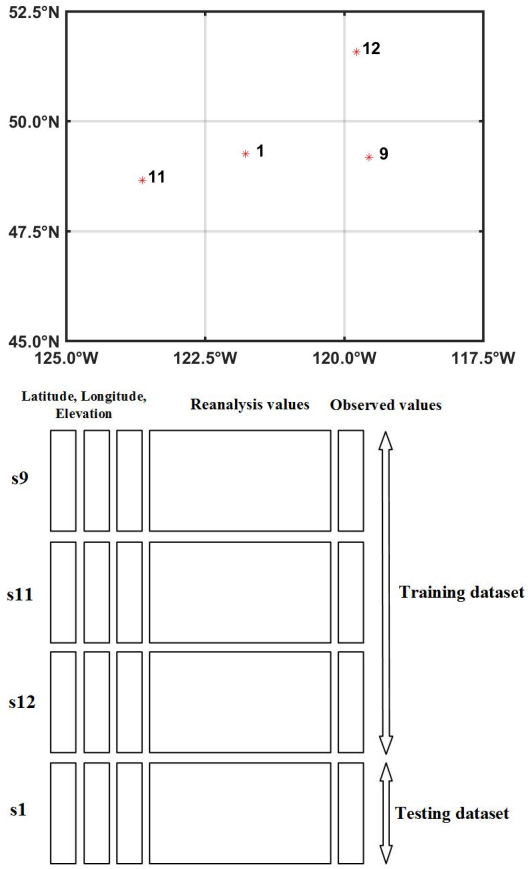


Fig. 1: Development of models' datasets for task 2.

III. APPLICATION AND EXPERIMENTS

This section presents the experimental results when applying our method on monthly mean air temperature and precipitation datasets for British Columbia. The station data used as target in our study consists of the observed values of monthly mean air temperatures and precipitation. These were obtained for twelve stations that are part of the Environment and Climate Change Canada network [10]. The reanalysis data used as the models' predictors (inputs) are from the NCEP/NCAR (National Centers for Environmental Prediction/National Center for Atmospheric Research) reanalysis dataset. NCEP/NCAR dataset is a combination of physical process and model forecast gridded data at the 2.5°x 2.5°spatial resolution. Details regarding this dataset's development can be found in [11]. The data used extended over a 56-year period from 1960 to 2015.

The predictand and predictor data were standardized to fall within a range of [0, 1]. By standardizing the variables and recasting them into dimensionless units, the arbitrary effect of similarity between objects is removed. The data was partitioned into a training and testing set. Parameter tuning was achieved through cross-validation. In all cases, 10% of the available data was used to test the models. In order to compare the developed models' performance, the following measures of goodness of fit were used: the root mean square error ($RMSE$), the mean absolute error (MAE), the mean absolute deviation (MAD) as well as the coefficient of determination (R^2).

A. Results and discussion

The results show that overall, the monthly mean air temperature and precipitations were predicted with high accuracy. In general, the results for the monthly air temperature models are more accurate than the precipitation

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

results. In fact, for the monthly air temperature target, the R^2 values, at test time, range between 0.980 and 0.998 for the first task and between 0.987 and 0.997 for the second task. When it comes to the monthly precipitations variable, the R^2 values, at test time, range between 0.616 and 0.893 and between 0.390 and 0.916 for the first and second tasks respectively.

Regarding the first task, downscaling to locations where past observations are available, the results in table I show that the stations where the downscaling accuracy was highest are station 8 for the monthly temperature variable and station 7 for the monthly precipitation target. The worst performance was obtained for station 2 and station 5 for the monthly air temperature and monthly precipitations respectively. The relatively lower R^2 values for both stations 2 and 5 can be explained by their proximity to large bodies of water (i.e., Atlin Lake and Stuart Lake).

When it comes to the second task, where the objective is to downscale to locations with no past observational records, the stations with the highest downscaling accuracy are station 11 for the monthly air temperature and station 4 for the monthly precipitation (see Table II). These results confirm our intuition that one of the key factors impacting the downscaling accuracy is the number of stations in the neighbourhood or square surrounding the station of interest. In fact, station 11 is surrounded by three neighbouring stations (i.e., stations 1, 3 and 4) and station 4 is surrounded by stations 3 and 11. It's also interesting to look at how the performance changes with respect to elevation. Interestingly, the best downscaling accuracy, with respect to each task and climate variable, was obtained for stations 4, 7, 8 and 11 which are located at low elevation at 7, 6, 41 and $18m$ respectively. The worst performance was obtained at stations 2, 5 and 9 located at higher elevations of 674, 686 and $297m$.

Finally, when it comes to the impact of the models' structure on the performance of the machine learning techniques, we noticed that the performance only slightly changes as the number of hidden neurons varied (the results for all the developed models are not shown here due to space constraints). In general, networks with a smaller number of hidden neurons gave poorer performance, and so did networks with a high number of hidden neurons, as they resulted in underfitting and overfitting respectively. Overall, the best performances were obtained when the number of hidden neurons varied between a minimum of 7 and a maximum of 17.

TABLE I: Results of the best models at test time for task 1.

Station	Neurons in layers	Air temperature		Precipitation	
		$RMSE$	R^2	$RMSE$	R^2
s1	(144-17-1)	1.119	0.991	0.576	0.860
s2	(144-7-1)	0.327	0.980	1.299	0.827
s3	(144-17-1)	1.814	0.987	0.538	0.749
s4	(144-17-1)	1.949	0.997	0.199	0.856
s5	(144-17-1)	0.494	0.994	0.673	0.616
s6	(144-7-1)	0.498	0.995	0.625	0.619
s7	(144-17-1)	0.725	0.994	0.335	0.893
s8	(144-17-1)	0.945	0.998	0.160	0.807
s9	(144-7-1)	0.427	0.995	0.647	0.624
s10	(144-7-1)	0.848	0.988	0.896	0.649
s11	(144-7-1)	0.810	0.997	0.299	0.888
s12	(144-7-1)	0.329	0.992	0.750	0.829

TABLE II: Results of the best models at test time for task 2.

Station	Neurons in layers	Air temperature		Precipitation	
		$RMSE$	R^2	$RMSE$	R^2
s1	(144-17-1)	0.444	0.994	1.051	0.888
s2	(144-7-1)	1.093	0.987	0.456	0.581
s3	(144-17-1)	0.371	0.994	1.427	0.877
s4	(144-17-1)	0.328	0.991	1.664	0.916
s5	(144-17-1)	1.018	0.988	0.594	0.490
s6	(144-7-1)	0.707	0.994	0.397	0.752
s7	(144-17-1)	0.318	0.995	0.972	0.831
s8	(144-17-1)	0.290	0.994	1.163	0.775
s9	(144-7-1)	0.737	0.993	0.497	0.390
s10	(144-7-1)	0.629	0.993	0.651	0.701
s11	(144-7-1)	0.286	0.997	1.046	0.869
s12	(144-7-1)	0.649	0.994	0.489	0.637

IV. CONCLUSIONS AND FUTURE WORK

This study presented a new downscaling method for two specific tasks: downscaling at locations where past observations are available to train the models, and downscaling for locations where there is no past record, using neighbouring stations to train the models. We explored the ability of artificial neural networks to downscale monthly mean temperatures and precipitations for selected stations in British Columbia. The results showed that using artificial neural networks to learn from reanalysis gridded data and station observations can lead to accurate downscaling results. In further work, we plan to test the application of these methods for downscaling additional climate variables, including climate extremes as these are important for assessing climate change impacts, and for planning adaptation strategies for future climate change.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] L. Candela, K. Tamoh, G. Olivares, and M. Gomez, "Modelling impacts of climate change on water resources in ungauged and data-scarce watersheds. Application to the Siurana catchment (NE Spain)," *Science of the Total Environment*, vol. 440, pp. 253–260, 2012.

[2] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. Ganguly, "DeepSD: Generating high resolution climate change projections through single image super-resolution," *arXiv preprint arXiv:1703.03126*, 2017.

[3] D. T. Price, D. W. McKenney, I. A. Nalder, M. F. Hutchinson, and J. L. Kesteven, "A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data," *Agricultural and Forest Meteorology*, vol. 101, pp. 81–94, 2000.

[4] J. Mahfouf, B. Brasnett, and S. Gagnon, "A Canadian precipitation analysis (CaPA) project: Description and preliminary results," *Atmosphere-Ocean,*, vol. 45, no. 1, pp. 1–17, 2007.

[5] O. Tveito, M. Wegehenkel, F. VanDerWel, and H. Dobesch, "Spatialisation of climatological and meteorological information with the support of GIS (working group 2)," *The Use of Geographic Information Systems in Climatology and Meteorology, Final Report*, pp. 37–172, 2006.

[6] K. Stahl, R. Moore, J. Floyer, M. Asplin, and I. McKendry, "Comparison of approaches for spatial interpolation of daily air temperature in a large region with complex topography and highly variable station density," *Agricultural and Forest Meteorology*, vol. 193, no. 3, pp. 224–236, 2006.

[7] N. Hofstra, M. Haylock, M. New, P. Jones, and C. Frei, "Comparison of six methods for the interpolation of daily European climate data," *Journal of Geophysical Research*, vol. 113, 2008.

[8] C. Bishop, *Neural networks for pattern recognition.* Oxford: Clarendon Press, 1995.

[9] R. Martin, R. Aler, J. M. Valls, and I. M. Galvan, "Machine learning techniques for daily solar energy prediction and interpolation using numerical weather models," *Concurrency and Computation: Practice and Experience*, vol. 28, no. 4, pp. 1261–1274, 2016.

[10] Environment and Climate Change Canada, "Adjusted and homogenized Canadian climate data - daily temperature and precipitation (AHCCD - daily T&P)," 2017.

[11] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, and D. Joseph, "The NCEP/NCAR 40-year reanalysis project," *Bulletin of the American Meteorological Society*, vol. 77, no. 3, pp. 437–471, 1996.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

GLOBENET: CONVOLUTIONAL NEURAL NETWORKS FOR TYPHOON EYE TRACKING FROM REMOTE SENSING IMAGERY

Seungkyun Hong*[,1,2], Seongchan Kim[2], Minsu Joh[1,2], Sa-kwang Song†[,1,2]

***Abstract*—Advances in remote sensing technologies have made it possible to use high-resolution visual data for weather observation and forecasting tasks. We propose the use of multi-layer neural networks for understanding complex atmospheric dynamics based on multi-channel satellite images. The capability of our model was evaluated by using a linear regression task for single typhoon coordinates prediction. A specific combination of models and different activation policies enabled us to obtain an interesting prediction result in the northeastern hemisphere (ENH).**

I. INTRODUCTION

Recent decades have seen significant efforts by meteorologists to develop numerical weather prediction (NWP) models such as Weather Research&Forecasting (WRF) to predict and produce rich atmospheric metrics such as the air pressure, temperature, and wind speed. The purpose of these processes is to predict extreme weather events capable of causing severe damage to human society. Typhoons, i.e., mature tropical cyclones known to commonly develop in the northwestern Pacific Basin, are one of the targets of atmospheric dynamics modeling. However, these models require considerable computational resources and processing time.

Previous studies [1]–[3] have shown that deep neural networks yield reliable results for weather-related problems and are computationally less intensive compared to large NWP models. Recent weather research comprising 3D data, such as weather simulation results or radar reflectivity datasets, involved the application of convolutional neural networks (CNN) which are known to be capable of extracting rich regional features from multi-dimensional data.

Meanwhile, advances in satellite equipment have made it possible to accumulate extensive global observations than was previously the case. Modern satellite imaging sensors collect visual global observations coupled with infrared (IR) and visible (VIS) wavelength and have capabilities of 0.25~4km at s.s.p. (Spatial resolution) and multiple channels from 5ch (MI) to 36ch (MODIS). Considering that high-resolution global observations exceeding 1 TB in size are collected by several weather research centers daily, it has become possible to use sophisticated visual information from massive datasets.

Nevertheless, an approach for typhoon eye tracking based on bare remote sensing images has not been reported yet. Moreover, high-resolution imagery itself has never been utilized extensively without any modification on information. Our solution to these problems was to focus on the utilization of the entire visual context from large-scale global observation to yield models for typhoon eye tracking based on deep CNNs. We first present two discrete neural networks based on multiple convolutional layers to develop an understanding of complex atmospheric dynamics, and then discuss the prediction results.

II. RELATED WORK

In recent years, many researchers have investigated the use of deep neural networks to solve various weather-related problems. Kordmahalleh et al. [1] explored a model to predict cyclone tracks from the large NOAA best track database. Racah et al. [2] suggested a semi-supervised model for extreme weather events from long-term CAM5 climate simulation datasets. Xingjian et al. [3] suggested a complex CNN-LSTM network (ConvLSTM) for the prediction of future precipitation rates from reflectivity data recorded by ground-based radar stations. However, not many investigations involved processing high-dimensional imagery data of weather phenomena.

Several complex network topologies proved capable of high-accuracy prediction or classification such as image classification from multi-layer inception followed

[1]Korea University of Science and Technology (UST) [2]Korea Institute of Science and Technology Information (KISTI)
†Corresponding Author

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

by repeated convolution-pooling steps [4] and symmetric skip connections for clear image restorations [5]. Unlike other CNN surveys, our research focuses on the characteristics of high-resolution remote sensing images containing vast and detailed descriptions of typhoon and cloud movement. In other words, our models utilize satellite images by preserving their details without significant modification due to processing such as image resizing or cropping.

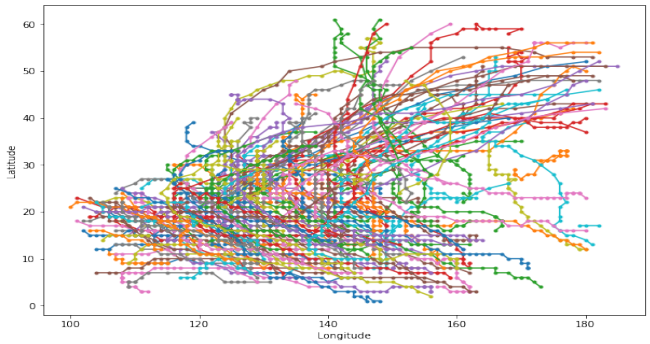


Fig. 1. Typhoon trajectories spanning 6 years (2011~2016) from the JMA RMSC-Tokyo Best Track Data

III. METHOD

A. Data

Predicting a typhoon eye coordinate requires two detailed types of information: a trajectory point consisting of the latitude and longitude and a single satellite image. The trajectory dataset we used was the Japan Meteorological Agency (JMA)'s official best track information **(Figure 1)**, which was used with a decimal precision of 1 latitudinal/longitudinal degree. The satellite image set comprised images acquired with the COMS-1 [6] MI of the Korea Meteorological Agency (KMA). The MI covers five channels including four IR images and one VIS image; however, because VIS imaging cannot be used for observation of the area of interest at midnight, only the 4-ch IR images **(Figure 2)** were chosen for the image dataset.

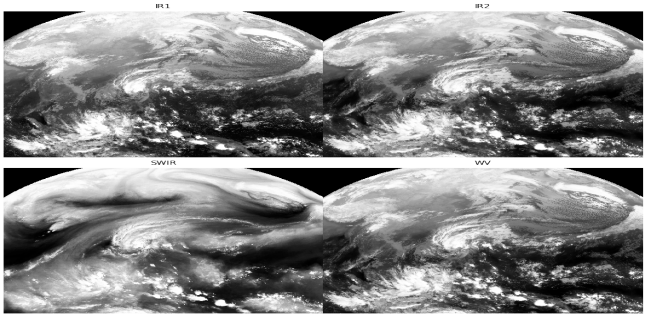


Fig. 2. Normalized COMS-1 MI 4ch IR images with CDF

We used 2,674 satellite images collected from 2011 until the end of 2016, covering nearly six years and 152 occurrences of single typhoons. The scope of the scale level between the latitude (80°) and longitude (150°) was matched by normalizing both the input (satellite image) and output (typhoon track with latitude and longitude) to a value between 0 and 1.

Furthermore, both the image and track data are randomly sampled for mini-batch training and shapes of arrays are reshaped to fitting on neural networks. Therefore, the dimensionality of an image becomes four, denoted by the $NumSamples \times Height \times Width \times Channels$ (NHWC) format, and that of the track becomes three ($NumSample \times Latitude \times Longitude$).

B. Models

The model receives 3D satellite images denoted by NHWC format as inputs. Our research surveyed two discrete network topologies for extracting rich features of cloud shapes. After an input image passes the multiple convolutional layers, each network has the exact fully connected dense layers for linear metrics regression. In the regression step, any values related to the weather event can be trained and predicted as a target value. Our networks are developed to achieve fast examination; thus, they only fit a point of the typhoon from the input image. The overall prediction process can be described as follows:

1) An input image is used for feature extraction by cascaded convolutional filters and max-pooling.
2) Fully connected layers following flattening of the filtered images builds a nonlinear connection for predicting the point of a single typhoon eye.
3) A smaller dense layer is used to ensure that the model fits its final outcome and ground truth, stimulated by linear activation.

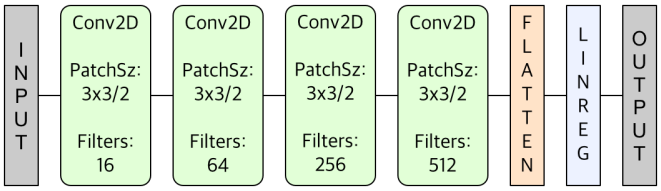


Fig. 3. Network Topology of Simple CNN with Fast Striding

1) Simple CNN with Fast Striding: Our first model **(Figure 3)** uses four conv. layers in conjunction with max pooling, which is similar to the basic CNN named LeNet-5 [7]. We applied 2-pixel strides and max-pooling technique on filter convolutions due to the high dimensionality of the input data. These optimizations resulted in each layer extracting fewer sparse features, further minimizing computational resources.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

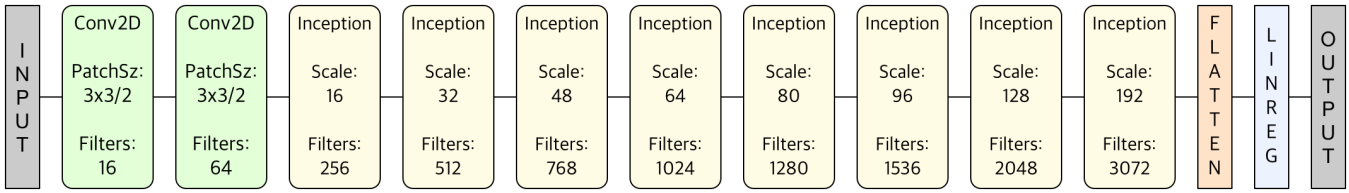


Fig. 4. Network Topology of Complex CNN with Inception Units

2) Complex CNN with Joint Inception Units: Our second model **(Figure 4)** starts by jointly using two conv. layers & max pooling steps for image reduction and as candidates to prepare for feature extraction. Then, eight deep-and-complex inception layers follow previously convolved images similar to the approach used in GoogLeNet [4]. In each inception unit, the images pass four different filter policies involving various types of feature extraction. After each convolutional procedure in the inception unit has been completed, all the filtered images are concatenated to produce the same image dimension with increased depth.

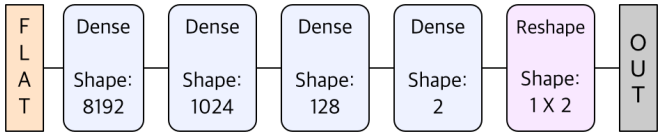


Fig. 5. Fully connected layers for linear regression process after flattening the convolutional layer

3) Linear metrics regression: After the CNN produces low-dimensional images with greater channel depth, these features are flattened to build fully connected (FC) layers. **(Figure 5)** Three fixed FC layers develop a nonlinear connection for obscure value prediction. Finally, a single FC layer coupled with a linear activation unit produces the predicted coordinate of typhoon eye.

IV. EVALUATION

The accuracy of typhoon eye tracking models is defined as (1), where P (2) is a point consisting of the latitude and longitude of the ground truth and $\hat{P}$ (3) is a point from the prediction result:

$$RMSE_{Prediction} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(P-\hat{P})^2} \quad (1)$$

$$P = (Lat_{gt}, Long_{gt}) \quad (2)$$

$$\hat{P} = (Lat_{pred}, Long_{pred}) \quad (3)$$

We examined the accuracy of our models for linear regression by conducting multiple experiments with several different configurations. Each network can set two different activation functions for each convolutional step (ReLU/LeakyReLU [8]/ELU [9]) and fully connected step (Sigmoid/Tanh). The Adam optimizer [10] is used for gradient optimization with an initial learning rate 1e-5, which is known to achieve fast optimization. The entire dataset is divided into training and testing sets in the ratio 9:1.

All our models use the toolkit named Keras [11] with the TensorFlow [12] backend as a neural network framework.

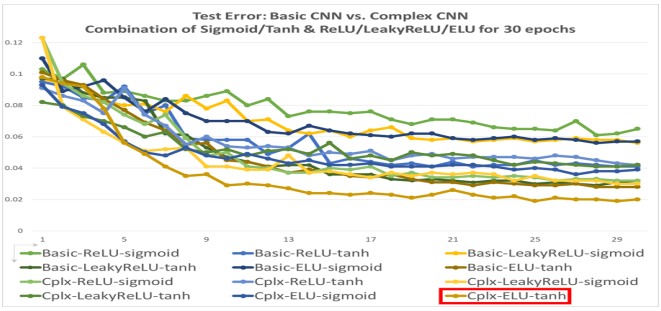


Fig. 6. Test Error: Basic CNN vs. Complex CNN - Combination of ReLU/LeakyReLU/ELU & Sigmoid/Tanh

Figure 6 shows the overall test error during 30 epochs of training process. The best prediction was achieved with Complex CNN with ELU/Tanh activation policy, with an RMSE of 0.02, about 40.24 nmi (74.53 km) in great circle distance. In contrast, the worst prediction models was Basic CNN with ReLU/sigmoid activation, characterized by an RMSE of 0.065 which is about 195.96 nmi (362.91 km) in great circle distance. Every result obtained with Complex CNN was more accurate than with Simple CNN with the same activation policy.

V. CONCLUSION AND FUTURE WORKS

We proposed two different neural networks to develop an understanding of atmospheric dynamics and weather events, for typhoons in particular. We studied typhoon tracking by attaching the observation images after evaluating the capability of the model to extract the topology features. This produced an RMSE of approximately 0.02 to guess the center of a single typhoon in a space of 80-degrees latitude by 150-degrees longitude.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

However, our model only focused on single-event typhoons in remote sensing images. Hence, we also plan to survey multiple occurrences in a single image that would require simultaneous tracking, as well as predicting additional weather metrics such as the air pressure, and wind speed for the simulation of complex atmospheric circulation and long-term global climate modeling.

ACKNOWLEDGMENTS

This work formed part of research projects carried out at the Korea Institute of Science and Technology Information (KISTI) (Project No. K-17-L05-C08, Research for Typhoon Track Prediction using an End-to-End Deep Learning Technique)

REFERENCES

[1] M. M. Kordmahalleh, M. G. Sefidmazgi, A. Homaifar, and S. Liess, "Hurricane trajectory prediction via a sparse recurrent neural network," in *Proceedings of the 6th International Workshop on Climate Informatics: CI2016*, 2016.

[2] E. Racah, C. Beckham, T. Maharaj, C. Pal, *et al.*, "Semi-supervised detection of extreme weather events in large climate datasets," *arXiv preprint arXiv:1612.02095*, 2016.

[3] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, pp. 802–810, 2015.

[4] 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*, pp. 1–9, 2015.

[5] X. Mao, C. Shen, and Y.-B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Advances in Neural Information Processing Systems*, pp. 2802–2810, 2016.

[6] M.-L. Ou, S.-R. C. Jae-Gwang-Won, *et al.*, "Introduction to the coms program and its application to meteorological services of korea," in *Proceedings of the 2005 EUMETSAT Meteorological Satellite Conference, Dubrovnik, Croatia*, pp. 19–23, 2005.

[7] Y. LeCun, L. Jackel, L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, *et al.*, "Learning algorithms for classification: A comparison on handwritten digit recognition," *Neural networks: the statistical mechanics perspective*, vol. 261, p. 276, 1995.

[8] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, 2013.

[9] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

[10] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[11] F. Chollet *et al.*, "Keras." https://github.com/fchollet/keras, 2015.

[12] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning.," in *OSDI*, vol. 16, pp. 265–283, 2016.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

DETECTING PRECURSORS OF TROPICAL CYCLONE USING DEEP NEURAL NETWORKS

Daisuke Matsuoka[1], Masuo Nakano[1], Daisuke Sugiyama[1], Seiichi Uchida[2]

Abstract— **Predicting tropical cyclonegenesis areas before their generation has tremendous social and academic significance. In this work, we investigate the predictability of the generation of tropical cyclones and their precursors based on one million cloud images visualized from 30-year simulation data using deep neural networks. Tropical cyclogenesis areas are predicted by ensemble learning using ten weak-classifiers, each possessing an accuracy of 85.0-88.0%. We succeeded in predicting the precursors of tropical cyclones seven and five days before their formation with a *Recall* of 88.6% and 89.6% (average of *Precision* is 11.4%), respectively, from only cloud images.**

I. MOTIVATION

Large scale tropical cyclones (TCs) such as typhoons, cyclones, and hurricanes affect significant damage to human life, agriculture, forestry and fisheries, and infrastructure. In Japan, Typhoon Lionrock in 2016 dealt severe damage to building and left many dead as a consequence. Predicting TC formation as soon as possible is important not only from an academic perspective, but also in disaster mitigation.

Thus far, TCs have been predicted using various methods that incorporate numerical simulation and/or satellite observation data. Dvorak proposed T-number to estimate the strength of TCs from satellite imagery data [1], [2]. The early stage Dvorak Analysis (EDA) developed by Japan Meteorological Agency (JMA) has been utilized in operational forecast since 2001 [3]. National Hurricane Center (NHC) and the Central Pacific Hurricane Center (CPHC) also use the advanced Dvorak method for TCs prediction in 48 hours lead time with an accuracy of 15-57% [4]. Yamaguchi et al. demonstrated that combined use of the Dvorak method and multi-model ensemble forecasts improved predictability to 35-79% [5]. However, if numerical models try to simulate farther into the future, the predicted error of simulated results increases due to the initial value dependency. This is the limitation of the deductive method in the weather forecasting. Therefore, we adopt machine learning, an inductive approach, to extract features of clouds from past data, before TCs occurrence.

In recent years, deep learning, one of the machine learning methods based on neural networks, has been increasing attraction and being applied to pattern recognition [6], [7]. In meteorology, several proposals for studies using deep neural networks for existing TCs detection [8], tornado prediction [9], hurricane pathway prediction [10] and extreme rain fall prediction [11]. Comparatively fewer studies have been proposed for predicting TCs, because the precursor of TCs, "the Eggs" of them, are difficult to identify even by meteorological experts.

In the present study, we investigate the probability of predicting TCs 7 days prior from long-term global atmospheric simulation data using deep neural networks.

II. DATA

The atmospheric simulation data used in this study is 30-year data produced by the NICAM with a 14-km horizontal resolution [12]. This model employs fully compressible nonhydrostatic equations and guarantees the conservation of mass and energy. Equations are discretized by the finite volume method. One characteristic feature of this model is that it explicitly calculates deep convective circulations without using any cumulus parameterizations. This model is suitable for the reproduction of tropical phenomena such as TCs [13] and the Madden-Julian Oscillation (MJO) [14]. For additional details on this model, please see the original and survey papers [15], [16].

To generate supervised image data, we employed a TCs tracking algorithm [13], [17] to NICAM simulation data. In the first step, candidate grid points at the center of TCs are selected from local extrema of sea level

Corresponding author: D. Matsuoka, daisuke@jamstec.go.jp
[1]Japan Agency for Marine-Earth Science and Technology (JAMSTEC) [2]Kyushu University

pressure (SLP). In the second step, candidate points, which satisfy some criteria related to wind speed, relative vorticity, temperature, duration, and range are combined temporarily and spatially into the TC track. By adapting this algorithm, existing TCs as well as those yet to be formed are detected with their center point, elapsed time, maximum wind speed, and minimum SLP. TCs are defined by a threshold value of maximum wind speed; it is difficult to distinguish TCs before and after their formation. In this work, we classify the precursor (10 days before their formation) and developing TCs (7 days after their formation) under one category (Fig. 1. (a)). In addition, low pressure clouds that were candidates for TCs but do not satisfy the criteria of duration are labeled as "not TCs" (Fig. 1.(b)). These images are visualized from Outgoing Long Radiation (OLR) and their horizontal sizes are 800-1000 km (64x64 pixels). We generate approximately one million images of TCs, including their Egg and four million images of not TCs from 30-year simulation data.

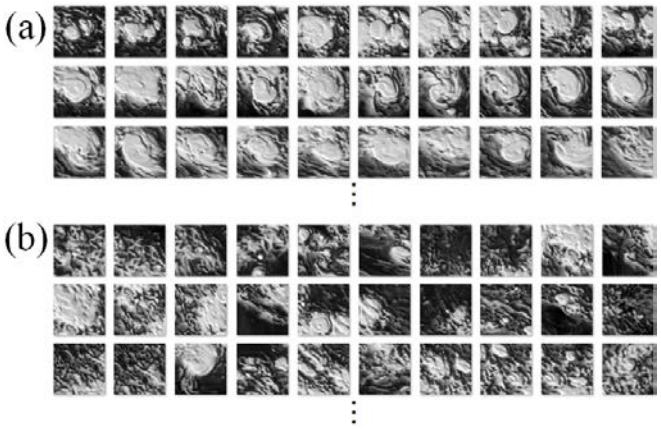


Fig. 1. Supervised image data. (a) Sequential images of a life cycle of a TC and (b) images of cloud with low pressure (not TCs).

III. METHOD AND RESULTS

The spatial distribution of clouds is important for atmospheric pattern recognition and therefore we adopt a 2D convolutional neural network for image classification described in Table 1. The architecture includes four convolutional layers, three pooling layers, and two fully connected layers. Input data is grayscaled image data of 64x64 pixels and output is generated in two classes (0 or Negative: not TCs, 1 or Positive: TCs). Hyper parameters are optimized by sensitivity study which explores 216 settings of the number of the convolutional layer (1-5) and pooling layer (1-5), number of parameter in the fully-connected layer (100, 300, 500, 1000, 2000), drop out ratio (0.2, 0.3, 0.4, 0.5), size of convolutional filter (3x3, 5x5, 7x7), and number of feature maps (16, 32, 64); the architecture with the highest performance is adopted accordingly.

Table 1. The architecture of our CNN.

Input	64x64
Convolution 1	3x3@16
Convolution 2	3x3@32
Pooling	2x2
Convolution 3	3x3@64
Pooling	2x2
Convolution 4	3x3@64
Pooling	2x2
Fully-connected	500
Fully-connected	2

The accuracy using 100,000 images (50,000 for each of two classes) for training and 5,000 images for cross-validation test is 87.61%. In order to improve the accuracy, ensemble learning that fuses the plurality of an individual trained weak-classifier is effective way. Although Bagging [18], Boosting [19] and Random Forest [20] are well known methods of ensemble learning, we employ a simplified version of AdaBoost [21] which is one of a Boosting method. In our method, ten weak-classifiers (*Classifier 1*, *2*, …, *10*) are generated by learning ten sets of randomly selected training data on the same neural network. One percent of erroneous predicted data in *Classifier i* (i=1, 2, …, 9) is used again in *Classifier i*+1 as training data (original AdaBoost repeats this process). The accuracy of ten classifiers is listed in Table 1. Here, accuracy indicates that all the predictions including positive (TCs) and negative (not TCs) are correct.

Table 2. Accuracy of ten weak-classifiers.

Model number	Accuracy (%)
Classifier 1	87.61
Classifier 2	86.22
Classifier 3	87.12
Classifier 4	86.30
Classifier 5	86.20
Classifier 6	87.46
Classifier 7	87.06
Classifier 8	85.78
Classifier 9	85.86
Classifier 10	86.42

Our method outputs the ensemble average using the accuracy of each weak-classifier. The final probability p for predicting the presence of TCs in an arbitrary region is defined as follows using a weighted average:

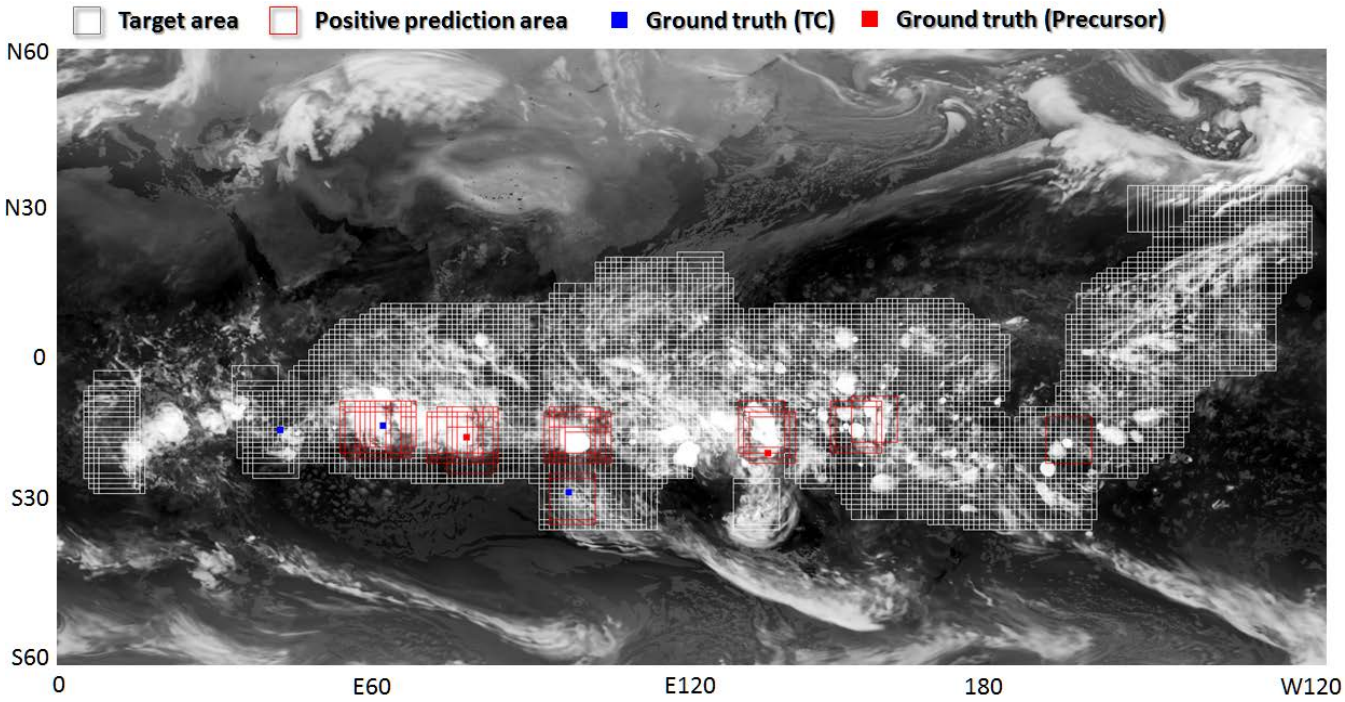


Fig. 2. Example of predicting result in a snapshot image. Seven positive prediction areas (overlapped areas are regarded to one area), three TCs, and two precursors are included.

$$p = \sum_{i=1}^{n} \frac{w_i x_i}{w_i} \qquad (1).$$

Here, w_i is the weight value of classifier i, and x_i is output value by *Classifier i* (0: not TCs, 1: TCs).

IV. Evaluation

Generated classifiers are applied to untrained global simulation data. We clip rectangular 64x64 pixels images with cloud cover of 30-80% (this range covers 92% of TCs) from global scale images as the prediction target area. Fig. 2. shows an application result of ensemble classifier to one snapshot image. Target areas and predicted areas of TCs are represented by white and red boxes, respectively. Real TCs and precursors (ground truth) calculated by the TC track algorithm are represented by blue and red dots, respectively. In this example, four TCs including two precursors of five ground truth can be predicted in a snapshot image (*Recall* is 4/5=80%); however, three areas of seven positive prediction areas are mispredicted (*Precision* is 4/7=57.1%). Here, *Precision* is the proportion of all positive prediction (TCs) that are correct, and *Recall* is the proportion of all real positive (ground truth is positive) that are correct [22].

Recall of each elapsed time-frame using 1-year of untrained data are shown in Fig. 3. The precursors of TCs seven days before their formation have a *Recall* of 88.6% and five days before have a *Recall* of 89.6%.

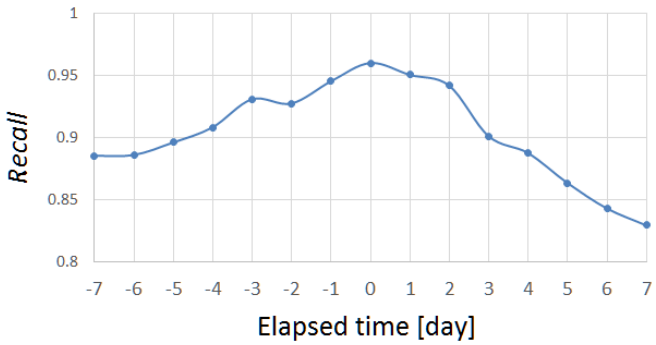


Fig. 3. *Recall* of each elapsed time-frame. Negative value indicates before TCs generation.

Meanwhile, average of *Precision* is 11.4% (minimum is 4.7% and maximum is 57.1%); this means 88.6% of positive prediction is incorrect. In order to improve total prediction performance, it is necessary to improve both *Recall* and *Precision*. The future work includes to investigate the reason for mispredicted cases and understand the contents of convolutional layer and weight.

Acknowledgments

We are grateful to Dr. C. Kodama for simulation data production and Dr. Y. Yamada for production of best track data of tropical cyclones. This work is partially supported by KAKENHI (16K13885) Grant-in-Aid for Challenging Exploratory Research, KAKENHI (26700010) Grant-in-Aid for Young Scientists (A) and

KAKENHI (17K13010) Grant-in-Aid for Young Scientists (B).

REFERENCES

[1] V. G. Dvorak, "Tropical cyclone intensity analysis and forecasting from satellite imagery," *Monthly Weather Review*, vol. 103, pp. 420–430, 1975.

[2] V. G. Dvorak, "Tropical cyclone intensity analysis using satellite data," *NOAA Technical Report NESDIS 11*, pp. 1–47, 1984.

[3] A. Tsuchiya, T. Mikawa, and A. Kikuchi, "Method of Distinguishing Between Early Stage Cloud Systems that Develop into Tropical Storms and Ones that Do Not", *Geophysical Magazine Series 2*, Nos. 1–4, pp. 49–59, 2001.

[4] J. Cossuth, R. D. Knabb, D. P. Brown, and R. E. Hart, "Tropical Cyclone Formation Guidance Using Pregenesis Dvorak Climatology. Part I: Operational Forecasting and Predictive Potential," *Weather and Forecasting*, vol. 28, pp. 100–118, 2013.

[5] M. Yamaguchi, U. Shimada, T. Iriguchi, M. Sawada, H. Owada, N. Koide, K. Yamashita, K. Ito, and Y. Miyamoto, "Recent Research and Development at MRI/JMA to Improve Typhoon Forecasts," In *the 71st Interdepartmental Hurricane Conference*, 2017.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[7] K. Simonyan, and Z. Zisserman, "Very deep convolutional networks for large-scale image recognition," In *International Conference on Learning Representation (ICLR) 2015*, 2015

[8] Y. Liu, E. Racah, S. Prabhat, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, and W. Collins, "Application of deep convolutional neural networks for detecting extreme weather in climate datasets," *arXiv reprint arXiv:1605.01156*, 2016.

[9] T. Trafalis, I. Adrianto, M. Richman, and S. Lakshmivarahan, "Machine-learning classifiers for imbalanced tornado data," *Computational Management Science*, Vol. 11, pp. 403–418, 2014.

[10] M. M. Kordmahalleh, M. G. Sefidmazgi, A. Homaifar, and S. Liess, "Hurricane Trajectory Prediction via a Sparse Recurrent Neural Network," In *Proceedings of the 5th International Workshop on Climate Informatics*, 2015.

[11] S. Gope, S. Sarkar, and P. Mitra, "Prediction of Extreme Rainfall using Hybrid Convolutional-Long Short Term Memory Networks," In *Proceedings of the 6th International Workshop on Climate Informatics*, 2016.

[12] C. Kodama, Y. Yamada, A. T. Noda, K. Kikuchi, Y. Kajikawa, T. Nasuno, T. Tomita, T. Yamaura, H. G. Takahashi, M. Hara, and Y. Kawatani, "A 20-year climatology of a NICAM AMIP-type simulation," *Journal of the Meteorological Society of Japan*, vol. 93, no. 4, pp. 393–424, 2015.

[13] M. Nakano, M. Sawada, T. Nasuno, M. Satoh, "Intraseasonal variability and tropical cyclonenesis in the western north pacific simulated by a global nonhydrostatic atmospheric model", *Geophysical Research Letter*, vol. 42, issue 2, pp. 565–571, 2015.

[14] H. Miura, M. Satoh, T. Nasuno, A. T. Noda, and K. Oouchi, "A madden-julien oscillation event realistically simulated by a global cloud-resolving model," *Science*, Vol. 318, pp. 1763–1765, 2007

[15] S. Tomita, and M. Sato, "A new dynamical framework of nonhydrostatic global modeling using the icosahedral grid," *Fluid Dynamics*, vol. 1, no. 8, pp. 357–400, 2004.

[16] M. Sato, H. Tomita, H. Yashiro, H. Miura, C. Kodama, T. Seiki, A. T. Noda, Y. Yamada, D. Goto, M. Sawada, T. Miyoshi, Y. Niwa, M. Hara, T. Ohno, S. Iga, T. Arakawa, T. Inoue, and H. Kubokawa, "The Non-hydrostatic icosahedral atmospheric model: Description and development," *Progress in Earth and Planetary Science*, vol. 1, pp. 1–32, 2014.

[17] M. Sugi, A. Noda, and N. Sato, "Influence of the Global Warming on Tropical Cyclone Climatology: An Experiment with the JMA Global Model", *Journal of the Meteorological Society of Japan*, Vol. 80, No. 2, pp. 249–272, 2002.

[18] L. Breiman, "Bagging Predictors," *Machine Learning*, Vol. 24, pp. 123–140, 1996.

[19] M. Kearns, and L. Valiant "Crytographic limitations on learning Boolean formulae and finite automata," In *Proceedings of the 21st annual ACM Symposium on Theory of Computing*, pp. 433–444, 1989.

[20] L. Breiman, "Random Forests," *Machine Learning*, Vol. 45, Issue 1, pp. 53–2, 2001.

[21] Y. Freund, and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", *Journal of Computer and System Sciences*, Vol. 55, pp. 119–139, 1997.

[22] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," *Journal of Machine Leaning Technologies*, vol. 2, no. 2, pp. 37–63, 2011.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017 Hosted by the National Center for Atmospheric Research in Boulder, CO

NEMR PREDICTABILITY ASSESSMENT OVER INDIAN PENINSULA USING ELM

Yajnaseni Dash[1], Saroj K. Mishra[1], B.K. Panigrahi[2]

Abstract—**This study proposed an Extreme Learning Machine based approach to evaluate the potential predictability of Northeast Monsoon Rainfall (NEMR) over the Indian southern peninsular region by using Sea Surface Temperature (SST) and Sea Level Pressure (SLP) as predictors. Performance of six different activation functions of ELM such as hardlim, radbas, sigmoid, sine, tansig and tribas was investigated. It is observed that among all activation functions, radbas performs better with minimal error scores. Thus, using global SST and SLP as predictors for NEMR prediction, radbas activation function of ELM gives optimal outcome.**

I. MOTIVATION

The northeast monsoon season is also known as retreating southwest monsoon, post monsoon season [1] or winter season [2]. This monsoon season occurs over the Indian southern peninsula during the period of October, November and December (OND). The contribution of this season to annual rainfall in the east coast of the Indian peninsula is 50% [3]. Western Ghats plays a significant role in determining the climate of the southern part of India. The west coast of India does not receive much rainfall during the southwest monsoon due to Western Ghats. Reverse wind distribution begins during OND period, causing extensive rainfall due to the movement of maritime air of the equator towards southern India [4].

Prediction of Northeast Monsoon Rainfall (NEMR) is a difficult task due to the dynamic nature of the monsoon. However, its prediction is an essential task as it has socioeconomic impact on the country. Artificial neural networks are among one of the most admired mathematical technique for classification, pattern recognition, prediction etc [5]. Back-propagation neural network (BPNN) has been used for prediction of Indian summer monsoon [6, 7]. However, its limitations include lengthy computational time due to the iterative weight update mechanism, stopping criteria, learning rate etc. [8, 9]. Extreme Learning Machine (ELM) was proposed to overcome the various issues of BPNN [10].

Northeast monsoon depends on several tele-connection parameters, but our key goal of this study is to assess the predictability of NEMR based on global predictors like Sea Surface Temperature (SST) and Sea Level Pressure (SLP). Two previous studies [11, 12] have used SST anomaly as a predictor for predicting monsoon rainfall using artificial neural network (ANN). A previous study by Nair et al. has shown that SST over the Indian Ocean (equatorial), the Bay of Bengal, the Pacific Ocean (central) and the Atlantic Ocean (north and south) have an effect upon northeast monsoon rainfall during OND [13]. This signifies that SST all of the major oceans have considerable influence over NEMR. SLP is an influencing factor for monsoon over the Indian subcontinent. SLP-based predictors influencing the NEMR include the regions of Southern Greenland and North Pacific Ocean; similarly, for the ISMR, the regions include the Tibetan plateau, Bering Sea near Alaska and Southern Atlantic Ocean, and north-west Europe. Even studies have considered the combination of SLP of North America, South America and a region of the Southern Ocean below Australia as a predictor [14]. So, we have considered global data for SST and SLP to check the predictability of NEMR.

This study used global climatic predictors such as SST and SLP for analyzing the predictability of northeast monsoon over the southern Indian peninsula using an ELM based approach.

II. METHOD

In ELM, Moore-Penrose generalized inverse method is used to accomplish the learning task by following a non-iterative mechanism. It is computationally much faster than BPNN and there is no need to tune the hidden layer [10]. Thus, we have employed this ELM

Corresponding author: Yajnaseni Dash, yajnasenidash@gmail.com [1]Centre for Atmospheric sciences, IIT Delhi [2]Department of Electrical Engineering, IIT Delhi

technique for the current study. The following steps have been carried out in this study.

1) *Data:* The input data comprised of Sea Surface Temperature (SST) and Sea Level Pressure (SLP) which were obtained from NCEP/NCAR reanalysis monthly mean data provided by the NOAA ESRL Physical Sciences Division (PSD) [15, 16]. The time series of area weight grids, data has been taken into consideration with coverage of 90°N-90°S and 0°E-358°E. The peninsular region rainfall dataset of post monsoon (OND i.e. October +November + December) season for the period 1948-2014 years were obtained from Indian Institute of Tropical Meteorology (IITM), Pune to examine the Northeast Monsoon Rainfall (NEMR) predictability. The peninsular region covers Coastal Andhra Pradesh, Rayalaseema, Tamil Nadu and Pondicherry, Coastal and South Interior Karnataka, and Kerala subdivisions. For operational purposes, the peninsular region consisting of these meteorological subdivisions is considered by Indian Meteorological Department (IMD). The OND mean rainfall time series was taken as the target (observation) dataset [17]. In fig.1 homogenous monsoon regions are presented from which peninsular region is distinguishable.

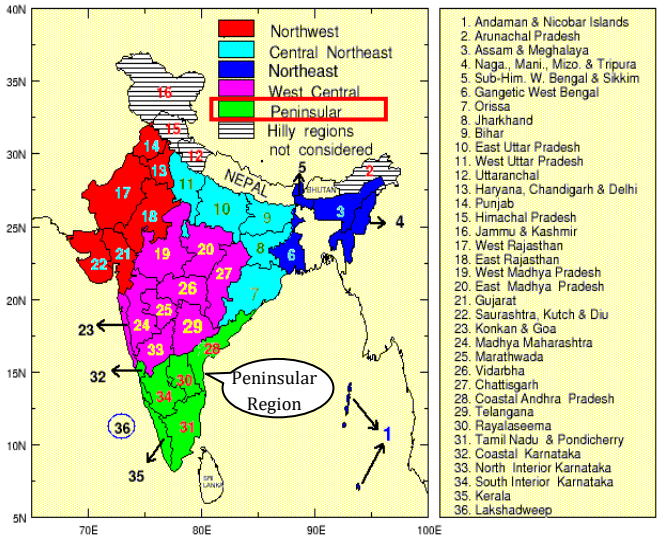


Fig.1. Map showing the Indian peninsular region (Source: IITM, Pune)

2) *Scaling:* Min-Max normalization method was used for scaling the data in the range [-1, 1] using the equation (1) given below.

$$X = \frac{X - Min_q}{Max_q - Min_q} \qquad (1)$$

3) *Activation function:* The key role of activation function is very crucial as it elucidate the non-linear relationship between input and output. We have used six different activation functions namely, hardlim, radbas, sigmoid, sine, tansig and tribas.

4) *ELM model:* ELM is the improved version of single layer feed forward neural network [10]. The entire dataset was divided into training (1948-2009) and testing (2010-2014). The training dataset was used to train the ELM model. Then, the trained model has been applied to the independent test dataset. Finally, the obtained outcomes were compared with the original observations using Mean Absolute Relative Error (MARE) and Root Mean Square Relative Error (RMSRE).

III. EVALUATION AND DISCUSSION

The present study has been carried out to observe the potential predictability of the northeast monsoon rainfall over the Indian southern peninsula by using SST and SLP as predictors based on ELM approach. SST anomaly has been used as a predictor for monsoon rainfall by two previous studies [11, 12] ANN. We have used SST and SLP as global predictors as input for this prediction task of NEMR.

The correlation between SST anomaly and observed rainfall was shown to be very low in linear as well as in polynomial trend equations [11]. In our study SST has shown a low positive correlation and low coefficient of determination (R^2). We have also studied another predictor SLP to find out its impact on northeast monsoon. But, it is found that SLP has a very low negative correlation with northeast monsoon. The correlation of SST is higher than SLP; this indicates that SST may have greater impact in climate variability of the southern peninsular India during northeast monsoon period.

In this study, the effect of the different activation functions in climate data has been examined. Prediction errors (observed - predicted) in fig. 2 has been presented to show the error score obtained by different activation function for the independent test period (2010-2014). Though there is no statistically significant difference among the six activation functions corresponding to the observed and predicted value, the average prediction error (observed - predicted) is lowest for radbas depicting its superiority among all. The higher prediction error in 2010 may be

due to the excess rainfall during the northeast monsoon in this year. In 2010, the rainfall over the peninsular India has been recorded excess than normal (nearly one and a half times) leading to several damages and loss of life i.e. more than 150 people died alone from the Tamil Nadu subdivision [18].

It can be clearly visible from fig. 3 that for radbas activation function both MARE and RMSRE scores are low as compared to other activation functions i.e. hardlim, sigmoid, sine, tansig and tribas. Based on MARE scores, the superiority of activation functions follows the order as given below:

radbas > tansig > sigmoid > sine > hardlim > tribas

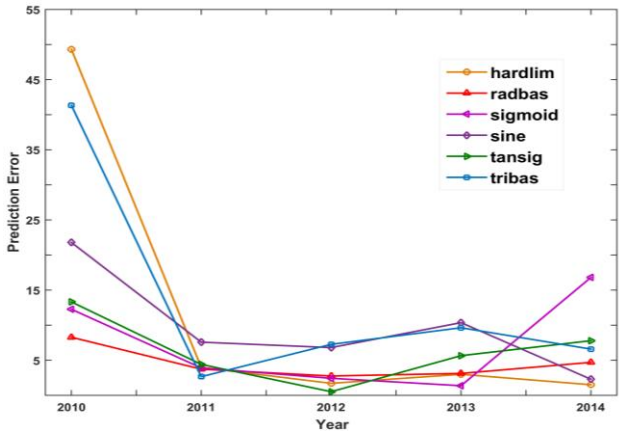


Fig. 2. Prediction error (Observed - predicted) over period (2010-2014)

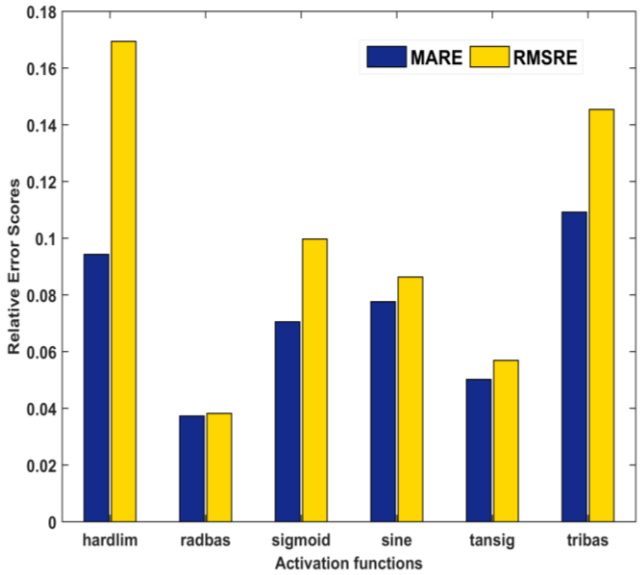


Fig. 3. Relative error scores of different activation functions

So, it is observed in this study that radbas has better performance among all activation functions of ELM. This optimal radbas activation function has also been used by a standard neural network to have a comparison with ELM. The error scores have shown that radbas activation function using ELM (MARE 0.0373; RMSRE 0.0382) is performing better than standard neural network (MARE 0.1128; RMSRE 0.1518).

Thus, from this study, it is found that ELM has the potential predictability to predict the intricate monsoon by using SST and SLP as input predictors. Among the six activation functions of ELM, radbas activation function has shown better performance.

IV. CONCLUSION

Prediction of NEMR in the Peninsular Indian region has been considered for this study using ELM approach. Global SST and SLP data have been used to check the predictability of NEMR. SST has a better correlation than SLP for northeast monsoon. It is found that radbas has better performance than other activation functions such as hardlim, sigmoid, sine, tansig and tribas using ELM. Comparison of MARE and RMSRE scores among ELM and standard neural network has shown better performance of radbas using ELM. Thus, for prediction of NEMR, radbas activation function of ELM gives optimal outcome using global predictors like SST and SLP.

ACKNOWLEDGMENTS

This work was partially supported by the Department of Science and Technology, Centre of Excellence in Climate Modelling at Indian Institute of Technology Delhi, New Delhi, India through the project number RP03350. We are also thankful to Indian Institute of Technology Delhi (IITD) for providing computational facility and financial support.

REFERENCES

[1] N. Singh, N. A. Sontakke, “On the variability and prediction of the rainfall in the post-monsoon season over India,” Int. J. Climatol., vol.19, pp.300-309, 1999.

[2] R. G. Nageswara, “Variations of the SO relationship with summer and winter monsoon rainfall over India,” J. Climate, vol. 12, pp. 3486-3495,1999.

[3] P. Kumar, K. Rupa Kumar, M. Rajeevan, A. K. Sahai, “On the recent strengthening of the relationship between ENSO and northeast monsoon rainfall over south Asia,” Climate Dynamics, vol. 28, pp. 649-660, 2007.

[4] R. H. Kripalani, P. Kumar, “Northeast monsoon rainfall variability over south peninsular India vis-`a-vis the Indian Ocean dipole mode,” Int. J. Climatol., vol.24, pp.1267-1282, 2004.

[5] Y. Dash, S. K. Dubey, “Quality Prediction in Object Oriented System by Using ANN: A Brief Survey”, International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 2, 2012.

[6] A.K. Sahai, M.K. Soman, V. Satyan, All India summer monsoon rainfall prediction using an artificial neural network, Clim. Dyn. Vol.16, pp.291-302, 2000.
[7] P. Singh, B. Borah, Indian summer monsoon rainfall prediction using artificial neural network, Stoch. Environ. Res. Risk. Assess, vol.27, pp.1585-1599, 2013.
[8] Y. Dash, S. K. Mishra, B. K. Panigrahi "Rainfall Prediction of a Maritime State (Kerala), India using SLFN and ELM Techniques," in Proceedings of the International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kerala, Accepted manuscript and received best award, 2017.
[9] Y. Dash, S. K. Mishra, S. Sahany, B.K. Panigrahi, "Indian Summer Monsoon Rainfall Prediction: A Comparison of Iterative and Non-Iterative Approaches," Applied Soft Computing. In Press, Accepted manuscript, 2017.
[10] G.B. Huang, M.B. Li, L. Chen, C.K. Siew, "Incremental extreme learning machine with fully complex hidden nodes," Neurocomputing, vol. 71(x), pp. 576-583, 2008.
[11] G. Chattopadhyay, S. Chattopadhyay, R. Jain, "Multivariate forecast of winter monsoon rainfall in India using SST anomaly as a predictor: neurocomputing and statistical approaches," Comptes Rendus Geosci. vol. 342, pp.755-765, 2010.
[12] N. Acharya, S. Chattopadhyay, M. A. Kulkarni, U.C. Mohanty, "A neurocomputing approach to predict monsoon rainfall in monthly scale using SST anomaly as a predictor," Acta Geophys. vol. 60, pp.260-279, 2012.
[13] A. Nair, N. Acharya, A. Singh, U. C. Mohanty, T. C. Panda, "On the Predictability of Northeast Monsoon Rainfall over South Peninsular India in General Circulation Models," Pure Appl. Geophys. vol.170, pp. 1945-1967, 2013.'
[14] M. Saha, P. Mitra, R. S. Nanjundiah, "Deep learning for predicting the monsoon over the homogeneous regions of India," J. Earth Syst. Sci. vol.126:54, 2017. DOI 10.1007/s12040-017-0838-7.
[15] E. Kalnay, M. Kanamitsu, R. Kistler, et al., "The NCEP/NCAR 40-year reanalysis project," Bull. Am.Meteorol. Soc., vol. 77, no. 3, pp.437-471, 1996.
[16] The NOAA ESRL Physical Sciences Division (PSD). www.esrl.noaa.gov/psd/.
[17] IITM Data Archival. http://www.tropmet.res.in/.
[18] Annual Climate Summary 2010. Available at: http://imdpune.gov.in/Clim_RCC_LRF/Annual_Climate_Summary/annual_summary_2010.pdf. (Accessed on 25. 08.2017).

THE ADVANCED CLIMATE ANALYSIS AND FORECASTING – DECISION SUPPORT SYSTEM (ACAF-DSS)

Bruce Ford[1], Herbert Dawkins[2], and Tom Murphree[3]

Abstract: We are developing a system for improving the operational climate services provided by government agencies, businesses, and other organizations. The Advanced Climate Analysis and Forecasting – Decision Support System (ACAF-DSS) provides users with three primary types of climate information via a unified web application: (1) multiple petabytes of reanalysis and other climate data; (2) real-time outputs from intraseasonal to interannual (I2I) forecasting systems; and (3) real-time assessments of forecast system skill. ACAF-DSS provides access to the datasets and forecasts, along with analysis and visualization tools, in a clustered computing environment. The system provides users with an array of options for using information about the climate system (including atmosphere, ocean, land, and ice components) to develop probabilistic decision support products, such as alternative courses of action (COA) products and measures of product performance. ACAF-DSS was designed to support operational decision making by national security organizations. But the basic approach, data, and methods also apply to climate research, climate support services, and climate related decision making by a wide range of users in government, business, and other sectors of the economy.

I. MOTIVATION

In 2008, we initiated the Advanced Climate Analysis and Forecast (ACAF) project to give operational climate support providers access to climate reanalysis datasets, and to advanced tools for developing analysis and I2I forecasting products based on those datasets. ACAF is similar in several ways to the data access and visualization applications provided by the National Oceanic and Atmospheric Administration Earth System Research Laboratory – Physical Sciences Division (NOAA ESRL-PSD). However, ACAF provides users with a larger number, and a greater range, of climate datasets, as well as a number of additional capabilities for data selection, processing, analysis, and visualization. These include the ability for users to: (1) rapidly select and analyze data, and create conditional composites of data, based on the intensity and phase of climate variations (e.g., El Nino-La Nina, Madden-Julian Oscillation (MJO)); (2) identify and select data according to user specified thresholds (e.g., significant wave heights exceeding 5 meters); and (3) analyze data according to user specified probabilities of occurrence (e.g., probability of ocean current speeds greater than user specified value; Figure 1).

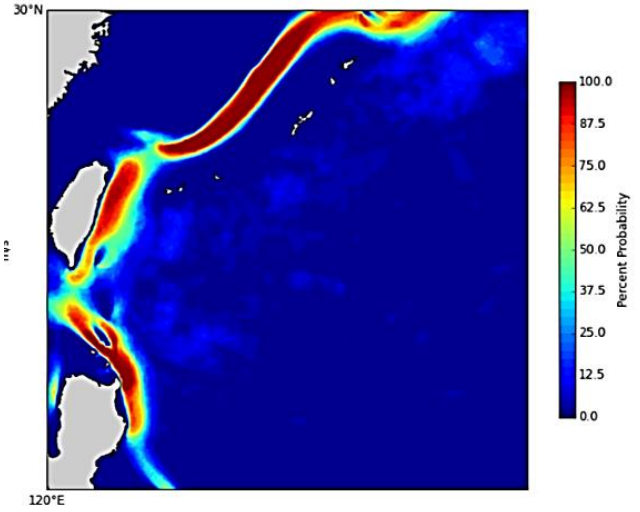


Figure 1. Example ACAF product: percent probability of upper ocean current speeds exceeding 1 m/s in January-March.

The development of ACAF was driven by the needs of US national security organizations for climate dataset access, and I2I analysis and forecasting capabilities, that were not provided by NOAA or other organizations. As an example, national security organizations planning operations several months or seasons in the future need

Authors: [1]B. Ford, bruce@clearscienceinc.com Clear Science, Inc , [2]H. Dawkins, drdawkins@clearscienceinc.com, Clear Science, Inc. [3]T. Murphree, murphree@nps.edu, Naval Postgraduate School

credible assessments of the probabilities of environmental conditions exceeding or falling within specific limits (e.g., the probability of air temperature or ocean surface winds and waves exceeding safe operating limits).

One of our main objectives in developing ACAF was to provide rapid access to, and processing of, large climate datasets by operational climate scientists and other climate support providers working at US national security organizations. We achieved that objective, in large part, by providing an extensive range of: (1) high priority datasets in formats that facilitate rapid data processing; and (2) software tools that eliminate the need for extensive scripting to extract, analyze, and visualize data, and to generate products ready for operational use.

The primary initial users of the ACAF system were scientists in the climate support division of the Fleet Numerical Meteorology and Oceanography Center (FNMOC) who quickly gained proficiency in using ACAF to develop climate analysis and forecasting products to support the planning and analysis of national security operations. The range of ACAF users, as well as ACAF datasets and capabilities, have greatly expanded as the demand for I2I decision support has grown [1]. This has included, for example, demand for datasets with higher temporal and spatial resolution and extent, dataset formatting to increase data processing speeds, tools for dataset comparisons, more advanced statistical and dynamical analysis tools, and a greater range of visualization options.

But the rapid emergence of new and much larger climate datasets and I2I forecasting outputs, and extensive feedback from ACAF users and their customers, led us to begin development in 2016 of a successor system, ACAF-DSS [2]. The primary motivations for this new system were user demands for a much greater range of: (1) new and forthcoming large climate data sets and I2I forecast products; (2) more complex analyses of datasets and forecast products; and (3) improved decision support capabilities.

ACAF-DSS will expand the information used by the ACAF system by including additional high resolution reanalysis data sets (e.g., mesoscale eddy resolving ocean reanalyses) and I2I ensemble based dynamical forecasting products (e.g., those from NOAA's Climate Forecast System (CFS) [3] and the North American Multi-Model Ensemble (NMME) program [4]; Figure 2). In addition, the ACAF-DSS system will provide real-time information on forecast skill and confidence. The system will provide users with access to multiple petabytes of reanalysis and other datasets, I2I forecast outputs, and forecast performance metrics, along with graphical interfaces with which to interrogate, visualize, and analyze the data and other information. Users will be able to generate analysis and forecast products, and corresponding decision support products, including probabilistic analyses, forecasts, and alternative courses of action (COAs).

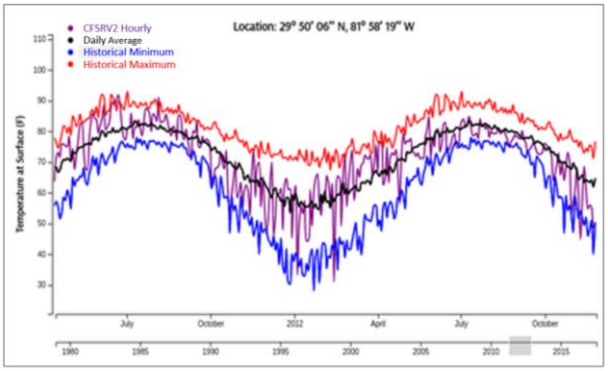


Figure 2. Example ACAF-DSS product: I2I forecasts from CFS of surface air temperature and corresponding historical means and extremes.

II. METHOD

ACAF-DSS is being constructed by using a clustered computing environment coupled with advanced data storage management techniques. This system has two main components: (1) an advanced version of the ACAF system; and (2) the Meteorology and Oceanography (METOC) Store.

Advanced ACAF Component

The advanced ACAF component contains all user (human or machine) interfacing functions, and all data presentation and communication functions. This component allows human and machine users to answer complex planning and long range forecast questions in a multi-dimensional, flexible, and intuitive manner. Users will be able to apply a wide range of statistical and dynamical analysis tools to, for example: (a) analyze correlations, teleconnections, and planetary wave dynamics associated with multiple simultaneous climate variations (Figure 3); (b) compare the forecasts and forecast skill of different forecasting systems by variable, location, time of year, forecast launch time, and forecast lead time; and (c) develop probabilistic decision support products based on user operating limits, risk thresholds, cost constraints, and planning lead time.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

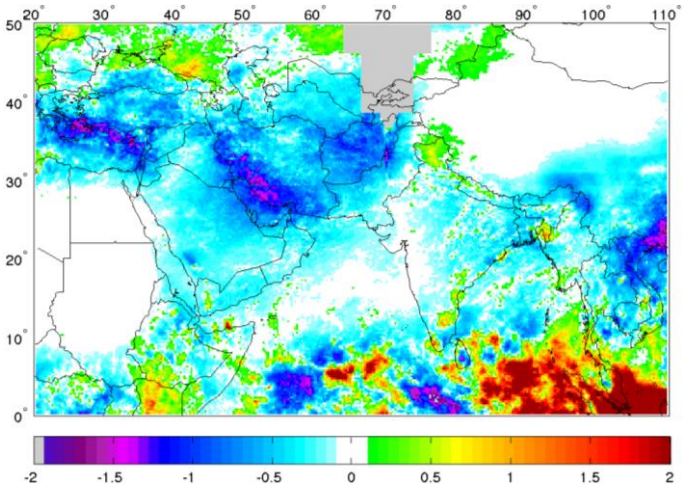


Figure 3. Example ACAF-DSS product: conditional composite precipitation rate anomalies for simultaneous occurrence of El Nino and MJO phase 4 during January-March.

METOC Store Component

The METOC Store is a knowledge store capable of warehousing and analyzing climate system data using multiple instance data managers with access to petabytes of data, coupled with in-memory data cubes (populated with high interest data) and rapid access to other data sources. The METOC Store allows other 'front ends' or interpreters to request METOC Store data for other applications, and to potentially act as a cloud peer fusion platform participating in track management or within an Object Management as a Service (OMaaS) environment.

The METOC Store employs pipeline processes that ingest, composite, evaluate, and stage data to ready it for rapid recall should a user request it. The pipeline processes include the preparation and constant update of forecast performance measures. The clustered computing environment constantly:

- Evaluates new data quality
- Post-processes model runs
- Computes model performance
- Updates existing model performance values
- Extracts temporally corresponding statistical information
- Stages high interest data within in-memory cubes

The data-intense analyses that must take place to answer user requests is performed using the clustered resources of the METOC Store and sends forward very small amounts of post-calculation aggregated data in user specified formats. The METOC Store conducts data extractions and calculations to support statistical and dynamical analyses (e.g., percentiles; probabilities of occurrence; conditional composites; optimal climate normal; correlations and regressions; teleconnections; vertical profiles; horizontal and vertical cross sections, time-space cross sections; fluxes, transports, and divergence of mass, energy and momentum; wave dynamics).

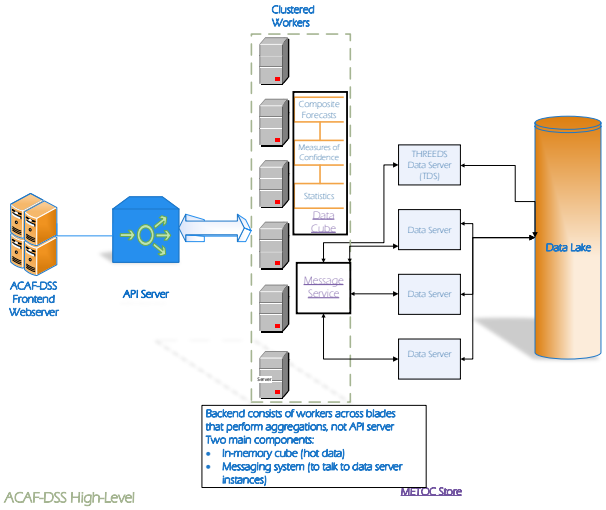


Figure 4. ACAF-DSS high level functional diagram

ACAF-DSS applies a single-map, geo-referenced approach to display interpretive products that allows users to: (a) layer color-fills, iso-lines, vectors, wind barbs, streamlines, time series, and other visualizations; and (b) select from a range of colormaps, line thickness and color, still and animated images, etc. (Figure 5) Multiple output formats are available. including image, GIS formats, and other custom formats, to support user applications of the ACAF-DSS outputs.

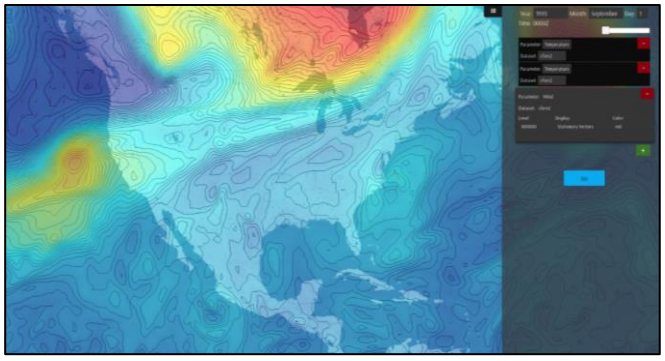


Figure 5. ACAF-DSS single map interface example.

III. EVALUATION

Each version of ACAF and ACAF-DSS is evaluated by the development team, and by independent climate

scientists and operational support providers. ACAF and ACAF-DSS products are assessed by comparisons to independently generated products, when those products are available. With each successive version, new tests are added and previous tests are repeated to ensure system integrity.

To be included in ACAF and ACAF-DSS, a dataset must first be validated through formal scientific validation testing and/or evaluations completed as part of research and operational applications of the dataset. Testing of each dataset within ACAF and ACAF-DSS is focused on verifying that the data processing correctly represents the original data (e.g., correct dataset has been accessed; correct selection and display of date, time, and location; correct analysis of data, etc.). This verification includes comparisons of ACAF and ACAF-DSS products to comparable products from independent sources, such as products generated at the ESRL-PSD sites or products created via independent scripting and data processing. Additional verification is conducted for products that involve derived quantities, such as those that result from statistical or dynamical analyses or extensive data manipulations (e.g., correlations, conditional probabilities, conditional composites, fluxes, transports).

Each version of the system is also beta tested by climate scientists and support providers at FNMOC, the host for the operational version of the system. Problems identified in this operational testing are corrected until the system meets FNMOC standards for operational use. Once the system is operational, system performance is routinely evaluated by FNMOC staff and other users. User feedback is explicitly solicited and analyzed to identify and correct problems, and to plan the next versions of the system. User testing and evaluation have been ongoing since 2008, which has allowed us and FNMOC to amass extensive information on use patterns, user needs, user requests, and system strengths and limitations. This information has been a major factor in designing and developing each new version, and was the primary basis for setting the design requirements for ACAF-DSS.

ACKNOWLEDGMENTS

Funding for this work is provided by the Earth System Prediction Capability Program and the Small Business Innovation Research Program of the Office of Naval Research.

REFERENCES

[1] Ramsaur, D., M. Hutchins, United States Navy Climatology Support Services, FLENUMMETOC N33 Climatology, Monterey, CA, 7 January 2013. https://ams.confex.com/ams/93Annual/webprogram/Paper211375.html.

[2] Ford, Bruce W., "SBIR Phase I Final Report for N142-121 (Navy)", Clear Science, Inc., 27 April, 2015. http://www.clearscienceinc.com/sbir/n142-121/phase1/N0001415P1055Phase1FinalReport.pdf

[3] Saha, Suranjana and Coauthors, 2014: The NCEP Climate Forecast System Version 2 Journal of Climate J. Climate, 27, 2185–2208. doi: http://dx.doi.org/10.1175/JCLI-D-12-00823.1

[4] Kirtman, B., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.,* **95,** 585–601, doi:https://doi.org/10.1175/BAMS-D-12-00050.1.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

GRAPH CONVOLUTIONAL AUTOENCODER WITH RECURRENT NEURAL NETWORKS FOR SPATIOTEMPORAL FORECASTING

Sungyong Seo[1], Arash Mohegh[2], George Ban-Weiss[2], Yan Liu[1]

***Abstract*—The importance of spatiotemporal data mining has been growing with the increasing number of heterogeneous datasets and importance of large datasets such as climate measurements, geographic information systems, virtual globes, the decennial census, and collections of traffic trajectories. For several decades, numerous studies have been done for analyzing time series, however, these traditional models often perform poorly when applied to multi-source datasets such as spatiotemporal data because of its heterogeneity. In this paper, we propose the novel method that combines spatial features with temporal dynamics based on the graph autoencoder (GAE). Specifically, an unsupervised model based on GAE for learning spatial representations is incorporated with Long short-term memory (LSTM) which captures temporal dynamics. We apply our model (GAE-LSTM) for climate applications, the spatial features such as geographical proximity and terrain similarity, and meteorological measurements like temperature, pressure, and precipitation.**

I. MOTIVATION

The recent advances of GPS-equipped technology have enabled to collect of large scale spatial and temporal datasets easily. Along with the increasing volume of such data and high update rate, it is naturally required to create automatic mining models extracting valuable information that is extremely difficult to be described by hand-made features. Furthermore, the special characteristics of the data domain make traditional techniques ineffective. Specifically, most previous models have been oriented for particular domains, respectively and not for multiple correlated domains, such as spatiotemporal data.

For example, Bayesian models, Hidden Markov Model (HMM) [1], [2], [3] and Kalman filter [4], [5] are used to predict time series. Recently, deep neural networks have been studied for general time series prediction. Particularly, recurrent neural networks (RNNs) are suitable for the prediction task. Long short term memory (LSTM) [6] has been widely used because it can handle long term dependency by minimizing vanishing gradient. Thus, many prediction works based on LSTM [7], [8], [9], [10], [11] have been studied. However, these works have only focused on extracting patterns in time series or sequences without combining additional information from other domains.

It is extremely important to consider the two different domains coincidentally because there are strong dependencies between the heterogeneous features. For example, change of meteorological measurements such as temperature is significantly dependent on where the measurements have been collected. If a climate class (tropical or dry climates) is given, it can provide more hints for forecasting measurements correctly. Similarly, temporal dynamics have also implicitly reflected spatial difference or similarity. If we track a traffic of a certain mobile application over time, the temporal pattern can easily provide the spatial information such as urban/rural area.

In this work, we propose a single model which can learn through spatial and temporal observations. Specifically, spatial features are extracted from a graph convolutional autoencoder [12] and the extracted features are incorporated with a temporal state from a recurrent neural network to predict future sequence.

II. METHOD

A. Graph Convolutional Networks

Convolutional neural networks (CNNs) are versatile networks which can extract translationally invariant features with fewer parameters. CNN is particularly effective on images (or videos) that are represented on grid structures (which are same as regular graphs) because it is easy to apply kernels on to the grid structures

Corresponding author: Sungyong Seo, sungyons@usc.edu
[1]Computer Science Department, University of Southern California
[2]Department of Civil and Environmental Engineering, University of Southern California

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

directly. However, it is much harder to generalize the convolutional operation on to general graph structures due to the irregularity. To alleviate the issue, many works [13], [14] have tried to develop the generalized convolutional neural networks called graph convolutional networks (GCNs) in the spectral domain.

Given an undirected weighted graph $G = (\mathcal{V}, \mathcal{E}, \mathcal{W})$ with N vertices, the normalized graph Laplacian is defined as $L = I_N - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ where A is an adjacency matrix and D is a degree matrix $D_{ii} = \sum_j A_{ij}$. Fourier transform is an expansion of a given signal $f(t)$ on the spectral domain (e.g., frequency) based on the eigenfunctions (e.g. $e^{-j2\omega ft}$) of the Laplacian operation (Δ). Similarly, the eigenvectors of the graph Laplacian can be defined as $L = U\Lambda U^\top$ and thus, the graph Fourier transform of a given graph signal x can be defined as:

$$\hat{x} = \sum_{j=0}^{N-1} u_j^\top x = U^\top x, \tag{1}$$

where u_j is an eigenvector corresponding to jth eigenvalue λ_j. The inverse transform is $x = U\hat{x} = UU^\top x$.

Since the transform is defined, the convolution of a given kernel g_θ and a graph signal x can be defined as the inverse transform of a product of g_θ and $\hat{x}$ on the spectral domain:

$$g_\theta * x = U g_\theta U^\top x, \tag{2}$$

where g_θ is a function of the eigenvalues Λ of L as $g_\theta(\Lambda)$.

It is important to note that the Eq. 2 is computationally expensive due to multiplication of the eigenvector matrix and the eigenvector decomposition in the first place. The cost problem is alleviated by the approximation of the Chebyshev polynomials which are defined recursively ($T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ with $T_0(x) = 1$ and $T_1(x) = x$). The expansion up to Kth order is:

$$g_\theta(\Lambda) \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{\Lambda}), \tag{3}$$

where $\tilde{\Lambda} = \frac{2}{\lambda_{\max}}\Lambda - I_N$ and $\lambda_{\max}$ is the largest eigenvalue of L. $\theta = (\theta_0, \cdots, \theta_K)$ is a vector of Chebyshev coefficients.

Since $g_\theta(\Lambda)$ is efficiently computable (Eq. 3), the convolution of the kernel and the signal is defined as:

$$g_\theta * x \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{L})x, \tag{4}$$

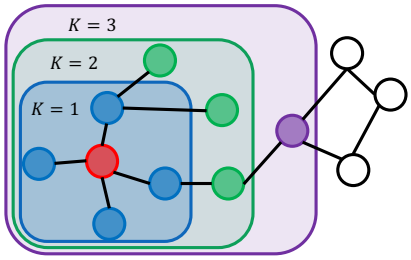


Fig. 1: K order localized filtering on a graph.

with $\tilde{L} = \frac{2}{\lambda_{\max}}L - I_N$. Eq. 4 is K order polynomial in the Laplacian L and it provides localized kernels since $(L^K)_{ij}$ is zero if the node i and j are not connected in K steps in the graph. In Figure 1, the receptive fields are shown for different K order polynomials at the center node (red node).

Since we have the form of the spectral convolution on graphs (Eq. 4), a neural network based model can be obtained by stacking multiple graph convolutional layers (Eq. 4). Furthermore, it is easy to extend to the graph convolutional autoencoder as [15] with the definition of the graph convolution. Details of building neural networks are covered in [13], [12].

B. GAE-RNN

In the previous sections, it is shown that the graph convolution autoencoder extracts latent representations of a given input feature. Furthermore, it is well-known that a recurrent neural network (e.g., LSTM) can be trained to learn temporal dependencies. Since these two modules have their own purposes, it is required to integrate the different modules as a single model. As shown in Figure 2, we concatenate the latent representations, β_i, with the output of the last cell of LSTM and feed it to one shared fully connected layer to make prediction. Thus, the complete loss function of the integrated model is written as:

$$\mathcal{L} = \frac{1}{2}\sum_{i=1}^{N} ||\mathbf{x}_{i,T+1} - \hat{\mathbf{x}}_{i,T+1}||_2^2 + \frac{\alpha}{2}\sum_{i=1}^{N} ||\mathbf{X}_i - \hat{\mathbf{X}}_i||_2^2 + \lambda R(\Theta) \tag{5}$$

where $\mathbf{x}_{i,\cdot}$ is a series of measurements and $\mathbf{X}_i$ is a feature vector at the node i, respectively. The regularizer weights, α, λ are adjustable to control the importance of the regularization terms.

III. EVALUATION

A. Datasets

In order to have a fair comparison, we use real world meteorological measurements from two commer-

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

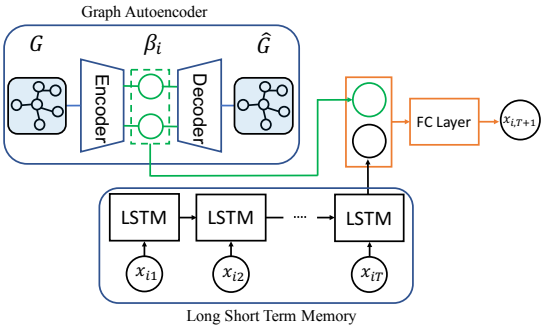
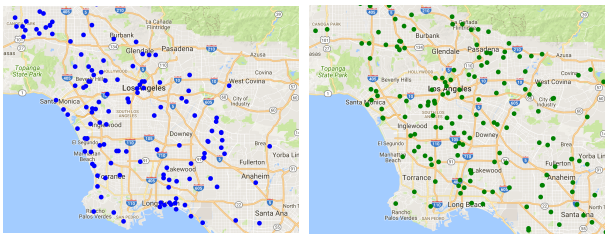


Fig. 2: GAE-RNN model architecture



(a) Weather Underground (b) WeatherBug

Fig. 3: Personal weather stations distributed over Los Angeles area

cial weather service providing real-time weather information, *Weather Underground, WU*[1] and *WeatherBug, WB*[2]. Both services use observations from automated personal weather stations (PWS). The PWSs are illustrated in Figure 3.

In the dataset, each station is distributed around Los Angeles County and land characteristics where the station is located at are provided. The list of the static characteristics, $\mathbf{X}_i$, is; *Latitude, Longitude, Elevation, Tree fraction, Vegetation coverage fraction, Albedo, Distance from the coast, Impervious fraction, Canopy width, Canopy height*, and *Building height*.

At each station, a number of weather data are observed through the installed instruments and recorded. The types of measurements are *Temperature, Solar radiation, Pressure, Precipitation, Relative humidity, Wind speed*, and *Wind direction*. In this experiments, observations from June/2015 to July/2015 are used.

B. Experimental results

In the datasets, while the spatial features are provided, the actual graph structures are not given. Therefore, it is required to create a graph based on distances between the features. We use thresholded cosine similarity to build the graph. By adjusting the threshold, it is possible to tune the size of the receptive fields.

For forecasting experiments, one meteorological measurement is chosen and aggregated over a given temporal granularity (e.g., hour). After setting the length of historical measurements (past 6 hours measurements), the graph structure, spatial features of weather stations, temporal measurements, and temporal gap between measurements are fed into our model and next measured value is predicted. We compare with two baselines, 1) LSTM only (LSTM) and 2) LSTM with graph Laplacian regularization (Lap-LSTM). LSTM has same as GAE-LSTM excluding GAE module. In Lap-LSTM, the graph autoencoder is replaced with a simply stacked autoencoder and a graph Laplacian regularization is assigned to consider the spatial similarity. Mean squared error is used to the comparison and all hyperparameters are chosen by cross validations.

	LSTM	Lap-LSTM	GAE-LSTM
WB	0.893	0.863	0.853
WU	1.537	1.501	1.440

TABLE I: MSE for temperature forecasting

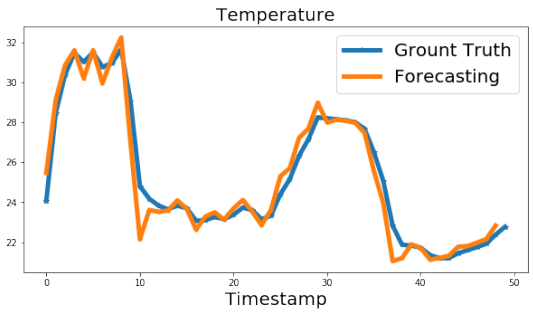


Fig. 4: Forecasting of temperature

Table I shows that GAE-LSTM outperforms other models. Especially, it is notable that GAE-LSTM has better spatial representations than that of Lap-LSTM which is only dependent on the similarity of spatial nodes. We also count how many weather stations have shown less MSE when GAE-LSTM is used compared to results of LSTM. In the time period, total 48 stations in *WU* have observed meteorological measurements. GAE-LSTM outperforms in the 34 stations out of the 48 stations compared to LSTM. For *WB* stations, we have better forecasting results in the 106 stations out of the 158 stations. It clearly shows that the spatial features are useful and able to improve forecasting quality.

In Figure 4, we sample a time period and plot the forecasting of temperatures with the ground truth.

[1]https://www.wunderground.com/
[2]http://weather.weatherbug.com/

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] L. Rabiner and B. Juang, "An introduction to hidden markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[2] M. R. Hassan and B. Nath, "Stock market forecasting using hidden markov model: a new approach," in *Intelligent Systems Design and Applications, 2005. ISDA'05. Proceedings. 5th International Conference on*, pp. 192–196, IEEE, 2005.

[3] Y. Qi and S. Ishak, "A hidden markov model for short term prediction of traffic conditions on freeways," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 95–111, 2014.

[4] A. C. Harvey, *Forecasting, structural time series models and the Kalman filter*. Cambridge university press, 1990.

[5] J. Z. Sun, K. R. Varshney, and K. Subbian, "Dynamic matrix factorization: A state space approach," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 1897–1900, IEEE, 2012.

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

[7] R. Akita, A. Yoshihara, T. Matsubara, and K. Uehara, "Deep learning for stock prediction using numerical and textual information," in *Computer and Information Science (ICIS), 2016 IEEE/ACIS 15th International Conference on*, pp. 1–6, IEEE, 2016.

[8] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, pp. 802–810, 2015.

[9] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[10] 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.

[11] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," in *Advances in neural information processing systems*, pp. 2953–2961, 2015.

[12] T. N. Kipf and M. Welling, "Variational graph auto-encoders," *arXiv preprint arXiv:1611.07308*, 2016.

[13] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[14] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," in *Advances in Neural Information Processing Systems*, pp. 3844–3852, 2016.

[15] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," *Artificial Neural Networks and Machine Learning–ICANN 2011*, pp. 52–59, 2011.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

DEEPRAIN: CONVLSTM NETWORK FOR PRECIPITATION PREDICTION USING MULTICHANNEL RADAR DATA

Seongchan Kim[1], Seungkyun Hong[1,2], Minsu Joh[1,3], Sa-kwang Song[1,2]

***Abstract*—Accurate rainfall forecasting is critical because it has a great impact on people's social and economic activities. Recent trends on various literatures shows that Deep Learning (Neural Network) is a promising methodology to tackle many challenging tasks. In this study, we introduce a brand-new data-driven precipitation prediction model called *DeepRain*. This model predicts the amount of rainfall from weather radar data, which is three-dimensional and four-channel data, using convolutional LSTM (ConvLSTM). ConvLSTM is a variant of LSTM (Long Short-Term Memory) containing a convolution operation inside the LSTM cell. For the experiment, we used radar reflectivity data for a two-year period whose input is in a time series format in units of 6 min divided into 15 records. The output is the predicted rainfall information for the input data. Experimental results show that two-stacked ConvLSTM reduced RMSE by 23.0% compared to linear regression.**

I. INTRODUCTION

Precipitation prediction is an essential task that has great influence on people's daily lives as well as businesses such as agriculture and construction. Acknowledging the importance of this task, meteorologists have been making great efforts to build advanced forecasting model of weather and climate, mainly focusing on Modeling & Simulation based on HPC (High Performance Computing).

In recent years, studies using deep learning techniques have been drawing attention to improve prediction accuracy [1], [2], [3], [4]. The Convolution Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are necessary techniques for the prediction of weather-related tasks. Several studies [4], [2] have employed each technique for precipitation prediction, and others [1], [3] have tried using combinations of these. Primarily, convolutional LSTM (ConvLSTM), which is a variant of LSTM, was devised to embed the convolution operation inside the LSTM cell to model spatial data more accurately by Shi et al. [1]. The authors demonstrated that ConvLSTM works on precipitation prediction in their experiments. However, they utilized three-dimensional and only one-channel data. In our study, we used the ConvLSTM for three-dimensional and four-channel data.

On the other hand, the data types used for rainfall-related prediction using the deep learning method are various. They include radar data [1], [2], past precipitation data, and atmospheric variable observation data such as temperature, wind, and humidity [3], [4]. In general, weather radar observations are the data used as inputs to numerical forecasting and hydrologic models to improve the accuracy of weather forecasts for hazardous weather such as heavy rains and typhoons [5]. Specifically, weather radar data refers to data represented by a radar image that is composed using the moving speed, direction, and strength of a signal transmitted by a radar transmitter into the atmosphere and received after it has collided with water vapor or the like. For example, Figure 1 shows a radar image of the Korean peninsula at 15:00 on April 5, 2017, and shows the rainfall rate in different colors depending on the degree of reflection.

In this study, to estimate the rainfall amount based on the weather radar data, we introduce *DeepRain*, which applies ConvLSTM, one of the variants of LSTM. Our radar data is three-dimensional data (width, height, and depth) consisting of four channels (depth) from four altitudes. The contributions of this study are as follows:

1) We adopt ConvLSTM first for three-dimensional and four-channel radar data to predict the rainfall amount.
2) We stacked the ConvLSTM cells for performance enhancement.

Corresponding author: Sa-kwang Song, esmallj@kisti.re.kr
[1]Disaster Management HPC Technology Research Center, KISTI, Korea; [2]Dept. of Big Data Science, UST-KISTI, Korea; Dept. of [3]S&T Information Science, UST-KISTI, Korea

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

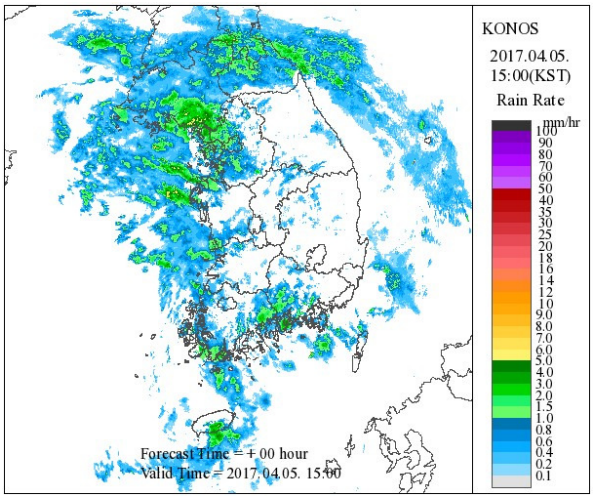


Fig. 1. Examples of radar map for rainfall rate

3) It was confirmed that the proposed method is more effective for predicting rainfall than linear regression and fully connected LSTM (FC-LSTM).

This paper presents the rainfall forecasting model proposed in Chapter 3 and related research in Chapter 2. Section 4 shows the experimental procedure and results, and the paper concludes in Section 5.

II. Related Work

Zhuang and Ding [4] designed a spatiotemporal CNN to predict heavy precipitation clusters from a collection comprising historical meteorological data across 62 years. They used two-convolution, pooling, and fully connected layers. Zhang et al. [2] proposed a 3D-cube successive convolution network for detecting heavy rain. In this study, they cast rainfall detection as a classification problem and identified the presence of heavy rainfall by using radar data of several channels as an input to the convolution network. Gope et al. [3] proposed a hybrid method combining CNN and LSTM. Their model used outputs of CNN as inputs for LSTM. In their model, CNN and LSTM were considered as independent steps. For data, they employed atmospheric variables such as temperature and sea-level pressure. However, they did not consider radar data. Shi et al. [1] devised a ConvLSTM model that enhanced an FC-LSTM model by replacing a fully connected layer with a convolution layer. However, that study was an attempt to generate a radar map for the future based on a past radar map using an many-to-many and end-to-end approach, while in our study, rainfall is forecast using a many-to-one (one-step time series forecasting) approach.

III. Data

The radar rainfall data used in the experiments were distributed by the Shenzhen Meteorological Administration in China for research purposes [6]. The data, which were normalized and anonymized (Details about pre-processing of the data were not publicized.), were radar observations in the Shenzhen area. The data consists of numerical integer values (dBZ). There are 101 * 101 radar reflection values, each representing one cell after modeling a specific area of Shenzhen in grid form (101 * 101 km^2). The 101 * 101 numerical values are grouped into four groups (from an altitude of 3.5 km; 1-km intervals) and 15 intervals (every 6 min) for each altitude. (See Figure 2.) That is, a total of 612,060 (=101 * 101 * 4 * 15) numerical values are listed. The ground truth is the measured rainfall amount (mm^3) from 1 h to 2 h in the area corresponding to the target area of 50 * 50 km^2 from the center of the grid. Therefore, one row of the data set is composed of 612,060 integer values (radar reflectivities) and one float value (ground truth). The complete data set consists of 10,000 rows randomly selected during a two-year period. We randomly divided the data into training (90%), validation (5%), and test data (5%).

IV. Method

ConvLSTM was introduced as a variant of LSTM by Shi in 2015 [1] and it is designed to learn spatial information in the dataset. The main difference between ConvLSTM and FC-LSTM is the number of input dimensions. As FC-LSTM input data is one-dimensional, it is not suitable for spatial sequence data such as our radar data set. ConvLSTM is designed for 3-D data as its input. Further, it replaces matrix multiplication with convolution operation at each gate in the LSTM cell. By doing so, it captures underlying spatial features by convolution operations in multiple-dimensional data. The equations of the gates (input, forget, and output) in ConvLSTM are as follows:

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

$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + b_o) \quad (3)$$

$$C_t = f_t \circ C_{t-1} + \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \quad (4)$$

$$H_t = o - t \circ \tanh(c_t) \quad (5)$$

where i_t, f_t, and o_t are input, forget, and output gate. W is the weight matrix, x_t is the current input data, h_{t-1} is previous hidden output, and C_t is the cell

state. The difference between equations in LSTM is that the convolution operation $(*)$ is substituted for matrix multiplication between W and x_t, h_{t-1} in each gate. By doing this, a fully connected layer is replaced with a convolutional layer, and then the number of weight parameters in the model can be significantly reduced.

In this study, the problem involves predicting rainfall using test radar data with training radar data and its label (in actual fact, the measured amount of rainfall) on a large scale. We solve the problem by utilizing convLSTM. The structure of *DeepRain* using convLSTM is shown in Figure 2. The input data X of the model receives 15 items of data according to the time interval, and the input data of each node is 40,804 (which is reshaped as 101 * 101 * 4; three dimensions with four channels) for radar reflection value (integer) data. The output is the generated value $_O$ of the output gate of the last cell, which is the expected amount of rainfall for the input data. This model configuration (many-to-one) is a result of the data set which has the ground truth label is given to the radar data as a number of rainfall amount falling between 1 hour and 2 hour. Note that our model does not predict next sequences of labeling (many-to-many).

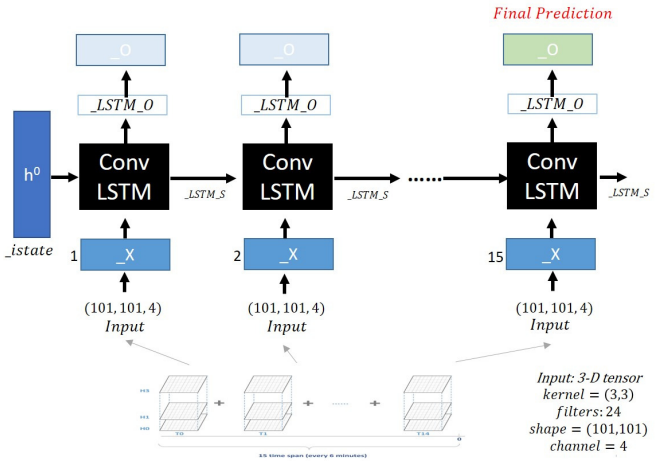


Fig. 2. *DeepRain* architecture using ConvLSTM

V. EXPERIMENT

We trained the proposed model (Figure 2) with the Adam optimizer at a learning rate of 0.001. The input data (originally in txt format at 16 Gb) was transformed into a binary tfrecord (6.4 Gb) to improve the learning speed. We used random shuffling mini-batches for learning; the mini-batch size was set to 30. The training epoch used 50, and it took about 12 h to learn (one-stack ConvLSTM). The Root Mean Square Error (RMSE) was used to measure the prediction accuracy. The testbed environment configuration was as follows:

- CPU: Intel R Xeon R E5-2660v3 @ 2.60GHz
- RAM: 128GB DDR4-2133 ECC-REG
- GPU: NVIDIA R TeslaTM K40m 12GB @ 875MHz (Dual)
- HDD: 4TB 7.2K RPM NLSAS 512n 12Gbps
- Framework: TensorFlow 1.2, Python 3.5.2

Figure 3 shows the learning curves by four different conditions of two models (FC-LSTM and ConvLSTM). ConvLSTM shows significantly better learning performance than FC-LSTM using Adam and Gradient Descendant Optimizer (GDO). A further experiment confirmed that FC-LSTM with GDO required 300 epochs to reduce the loss from 15.5 to 14.4, while ConvLSTM has required only five epochs to reach a loss of 10.0. It seems that the convolution operation efficiently extracted underlying features from the data and enabled quick training.

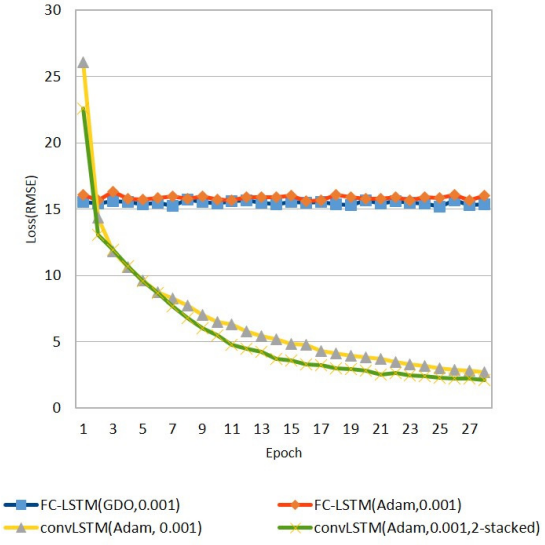


Fig. 3. Learning curves of differently conditioned models

According to Figure 3, the two-stacked ConvLSTM model show more stable performance than the one-stacked ConvLSTM. Then, we decided to find the optimal number of epochs to have a predictable performance that was not overfitted with the two-stacked ConvLSTM. Our further experiments with a validation set indicated that the validation loss increases from epoch 5, as shown in Figure 4. Then, we measured the performance with the test set from the trained model at epoch 5.

Table I lists the results of measuring the error rate (RMSE) for the test data with multiconditioned models

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

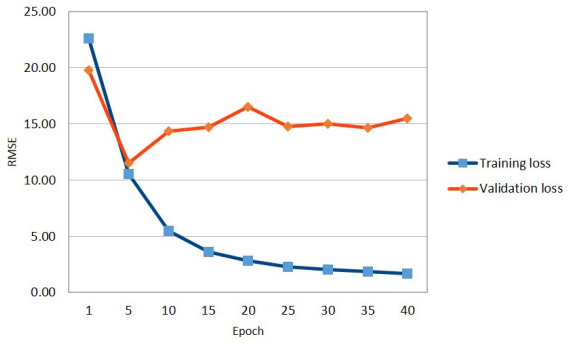


Fig. 4. Training and validation error curve with the two-Stacked ConvLSTM

and a baseline. The result of the two-stacked ConvLSTM is shown to be 11.31, which is 23.0% less than that of the linear regression model used as a control. Furthermore, it is 21.8% lower than that of the FC-LSTM [7]. The poor performance of the FC-LSTM seems to originate from the fact that the FC-LSTM is fed with one-dimensional input data and loses spatial information in the cell.

TABLE I
RMSE OF PREDICTING RAINFALL AMOUNT WITH TEST SET

Model	RMSE	Drop(%)
Linear Regression	14.69	-
DeepRain: FC-LSTM[7]	14.46	1.6
DeepRain: Conv-LSTM(one-Stacked)	11.51	21.6
DeepRain: Conv-LSTM(two-Stacked)	11.31	23.0

VI. CONCLUSION

In this study, we first applied ConvLSTM to three-dimensional and four-channel radar data to predict the rainfall amount between 1 h and 2 h. Experimental results showed the prediction accuracy of the proposed methodology is better than that of the linear regression and the FC-LSTM. Future studies will utilize Convolutional Gated Recurrent Units (ConvGRU) to compare ConvLSTM and expand the data set with several data augmentation techniques to enhance the performance. The augmentation technique will include cropping data of a 50 * 50km^2 area from the center, which is an important consideration in predicting rainfall. In addition, we are devising an effective convolution method on spatial three-dimensional data with multiple variables and channels. Lastly, we have a plan to consider El Nino for our model, which is likely to have an effect on precipitation of the studied area.

ACKNOWLEDGMENTS

This work was supported by research projects carried out by the Korea Institute of Science and Technology Information (KISTI): No. K-17-L05-C08, A Research for Typhoon Track Prediction using End-to-End Deep Learning Technique.

REFERENCES

[1] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in *CVPR*, jun 2015.

[2] W. Zhang, L. Han, J. Sun, H. Guo, and J. Dai, "Application of Multi-channel 3D-cube Successive Convolution Network for Convective Storm Nowcasting," in *CVPR*, 2017.

[3] P. M. Sulagna Gope, Sudeshna Sarkar, "PREDICTION OF EXTREME RAINFALL USING HYBRID CONVOLUTIONAL-LONG SHORT TERM MEMORY NETWORKS," *Proceedings of the 6th International Workshop on Climate Informatics: CI 2016*, 2016.

[4] W. D. Yong Zhuang, "LONG-LEAD PREDICTION OF EXTREME PRECIPITATION CLUSTER VIA A SPATIO-TEMPORAL CONVOLUTIONAL NEURAL NETWORK," in *Proceedings of the 6th International Workshop on Climate Informatics: CI 2016, NCAR Technical Notes NCAR/TN-529+PROC*, 2016.

[5] "Korean National Weather Radar Center, http://radar.kma.go.kr/eng/radar/composition.do."

[6] Shenzhen Meteorological Bureau-Alibab, "Short-Term Quantitative Precipitation Forecasting."

[7] Seongchan Kim, Seungkyun Hong, and Sa-kwang Song, "DeepRain: A Predictive LSTM Network for Precipitation using Radar Data," in *Proceedings of the Korean Computer Congress(KCC 2017)*, 2017.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

MULTIPLE CHANGE DETECTION IN LINEAR TREND OF SERIALLY CORRELATED TIME SERIES

Mohammad Gorji-Sefidmazgi[1], Mina Moradi-Kordmahalleh[1], Abdollah Homaifar[1]

***Abstract*—Trend analysis is important for better understanding of climate change and variability. Since the changes in atmosphere and hydrological variables are not monotonic, a single linear trend is not adequate to represent their changes. Common methods for finding piecewise linear trends are based on some restrictive assumptions, and do not take into account the autocorrelation of the time series. In this study, we propose a multiple change detection based on the Genetic Algorithm and Statistical Inference to find the piecewise linear trend of autocorrelated time series. The Bayesian Information Criterion is used to find an optimal number of change points. The proposed technique is applied on the global surface temperature, and the linear trend and lag-one autoregressive parameters of the time series with multiple change points are estimated.**

I. MOTIVATION

The linear trend is a straightforward assessment of long-term behavior of a time series. But in general, warming trends are nonlinear, especially the warming is accelerated over most of the twentieth century [1]. Seidel et al. showed that modeling of surface temperature trend with a piecewise linear model added to a lag-one autoregressive process (AR(1)) is better than single trend in the sense of *Bayesian Information Criterion* (BIC) [2]. Detection of piecewise linear trend of the trend requires finding the change points in the linear trend and its attribution to a potential cause. Several methods based on statistical tests were used in the literature to find breakpoints in the trend, such as sequential Mann-Kendall and Bai-Perron tests. But these statistical tests are generally developed under the assumption of independence among observations in the time series. However, climate time series, especially with monthly or smaller time scales are autocorrelated. If time series has an autoregressive structure, decision about the null hypothesis in statistical tests can be misleading by inflating/deflating the estimated significance levels and this would lead to false over-rejection/under-rejection of the null hypothesis [3]. One solution is to aggregate the data in order to convert it to a yearly time series and remove the autocorrelation. But this procedure drastically decreases the number of samples and reduces the power of change detection algorithm. On the other hand, the removal of the positive serial correlation component from time series by *pre-whitening* reduces the magnitude of existing trend [4]. In fact, the most efficient method of change detection in serially correlated time series is simultaneous estimation of change point and AR(1) coefficient [5]. In addition, several methods based on Bayesian Inference [6] and Markov model [7] exist in literature for change detection problem. These methods need prior assumptions on probabilistic distribution of the dataset or the change points, while these assumptions may not be generally true.

In this work, we propose a trend analysis algorithm for serially correlated time series based on the *Change Detection based on Genetic Algorithm (GACD)*. Unlike Markov or Bayesian methods, no additional assumption on the time series or change points are necessary. Unlike Bayesian and Markov change detection approaches, it does not need restrictive assumptions on statistical distribution of data or change points. Adding the assumption that the residual follows a Gaussian distribution, the BIC is adopted to find the optimal number of change points. By considering the autocorrelation, the proposed techniques can be applied to the time series with resolution of less than a year, and thus provides a more accurate estimation of trend parameters and change points. Finally, we test the proposed algorithm on global surface and ocean temperature to find a piecewise linear trend.

II. METHOD

The GACD method assumes that the time series has several regimes, while the model of the time series in each regime has constant statistical parameters, and

Corresponding author: A Homaifar, homaifar@ncat.edu
[1]Department of Electrical and Computer Engineering, North Carolina A&T State University, Greensboro, NC.
This work is supported by the National Science Foundation under cooperative agreement No. CCF-1029731

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

the parameters of these regimes are different from each other. Then, the algorithm finds the best values of change points between the regimes, and also the sequence that the regimes appear in the time series [8].

Let $x(t) \in R^n$ be a multidimensional time series over $t = \{1, \ldots, T\}$ with C regimes. The model of the time series in each regime could be a function of time or a probability density function:

• Function f of time and other inputs $u(t) \in R^m$ if they exist), where θ_c is the set of parameters in the c-th regime. Also, p and q are the order of the lagged inputs and the lagged outputs respectively. For this case, the model of time series and the *distance function* $d\left(x(t), \theta_c\right)$ between the time series at time t and the model of the c-th regime is defined as below.

$$x(t) = f(x(t-1), ..., x(t-p), u(t-1), ..., u(t-q), t, \theta_c) \quad (1)$$

$$d(x(t), \theta_c) = \|x(t) - f(x(t-1), ..., x(t-p), u(t-1), ..., u(t-q), t, \theta_c)\|^2 \quad (2)$$

• Probability density function f, where the $u(t) \in R^m$ is the set of covariates. For this case, the model of time series and the distance function using the *negative log-likelihood* can be defined as:

$$P(X = x(t)) = f(x(t)|u(t), \theta_c) \quad (3)$$

$$d\left(x(t), \theta_c\right) = \ell\left(f(x(t)|u(t), t, \theta_c)\right) \quad (4)$$

Then, the problem of time series modeling is defined as a minimization problem in Eq. 5.

$$\min_{\mu,\theta} \sum_{t=1}^{T} \sum_{c=1}^{C} \mu_c(t).d\left(x(t), \theta_c\right) \quad (5)$$

where $\mu_c(t) \in \{0, 1\}$ is the *regime membership function* indicating whether the data at time t belongs to the c-th regime or not. The change points occur at times when the values of $\mu_c(t)$ are changed. For example, if $\mu_2(50) = 0$ and $\mu_2(51) = 1$, then the regime 2 is started at the change point $t = 51$. Since data at each time belongs to only one of the regimes, hence $\sum_{c=1}^{C} \mu_c(t) = 1$ for $t = \{1, ..., T\}$. The problem in Eq. 5 is a non-convex mixed-integer optimization with two sets of unknown parameters, the regime membership function $\mu_c(t)$ and the model parameters θ_c. It is common to add an assumption to the problem that the number of change points is known and equals to W. The optimal value of these hyper-parameters (C and W) can be selected by information theory methods such as BIC [9].

In the GACD, A *population* of *individuals* are generated, while each individual represents a possible solution for the problem of change detection. Each individual is a string of numbers, where the first W strings are change point times and the next $W + 1$ strings represent the order of regimes. For example, the individual of Fig. 1 shows a sample solution for a problem with $T = 100$, and $C = W = 3$. This individual represents that the regime 1 is active in time frames of [1,9] and [40,69]. In a similar way, regime 3 is active in [70,100].

Since each individual shows the starting and ending times of each regime, it is possible to extract a unique regimes membership function from each individual. Thus, the regimes parameters θ_c and error of modeling can be found using the statistical inference. For this aim, the following optimization should be solved for each of the C regimes and $t \in c$-th regime, using maximum likelihood or least-square. The value of E is assumed as *cost* of the corresponding individual.

$$E = \min_{\theta} \sum_{t=1}^{T} d\left(x(t), \theta_c\right) \quad (6)$$

The procedure of optimization using genetic algorithm starts with generating a random population of individuals like Fig. 1. Then, the *crossover* and *mutation* operators are applied to the population members to generate new individuals. The crossover operator selects two members of population, and combines them to generate two new *offsprings* which are then added to the population. The mutation operator, selects an individual, generates a new offspring by applying a slight modification, and then add the offspring to the population. Since each individual represents both the change point times and order of regimes, 50% of crossover and mutation are applied to the change points, and 50% are applied to the regime sequences. For each new member of the population, the value of cost is calculated using Eq. 6. Then, those members of the population with higher (worse) values of cost are deleted from the population. The procedure of {selection, crossover, mutation, deleting} is repeated for enough number of *generations* until the population converges to an individual with a lowest cost. This individual represents the best values for change points and order of regimes [8].

This procedure should be repeated several times for different possible values of number of regimes C and number of change points W. In each case, the value of BIC is found by Eq. 7, where n is the number of

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

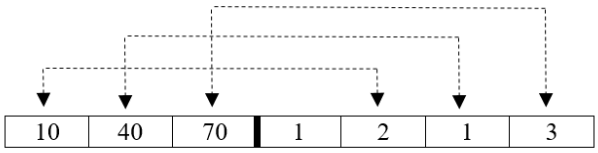


Fig. 1. A sample individual for a time series with length of 100, three change points and three regimes

estimated parameters for each regime. For example if each regimes has a Gaussian model, there are $n = 2$ (mean, variance) estimated parameters for each regime. The model with lowest BIC is selected as the best model.

$$BIC(C, W) = -2 \times ln(E) + ln(T) \times n \quad (7)$$

III. EVALUATION

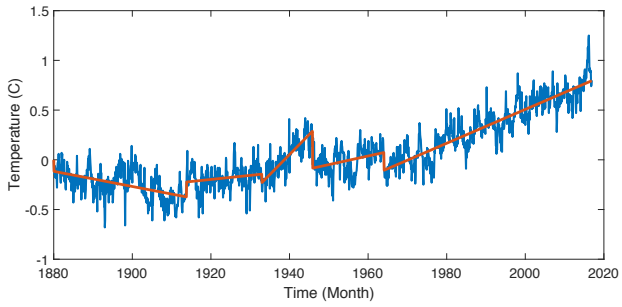
The GACD method is applicable to a wide range of statistical or regression model. In this paper, we test the method for finding the linear trend of global land and ocean temperature anomaly. The monthly data in the time range of [1880, 2016] can be downloaded from https://www.ncdc.noaa.gov/monitoring-references/faq/anomalies.php.

Suppose that the time series $x(t)$ is defined at $t = \{1, \ldots, T\}$ and the linear trend in the c-th regime is in the form of $x(t) = \beta_{0c} + \beta_{1c}t + \epsilon(t)$. Furthermore, let the correlated noise be $\epsilon(t) = \rho_c\epsilon(t-1) + w(t)$, where $w(t) = \mathcal{N}(0, \sigma_c^2)$. The coefficients β_{0c}, β_{1c} and ρ_c are the intercepts, slopes, and AR(1) coefficients of the time series in the c-th regime. The goal of GACD is to find these parameters such that the norm of $w(t)$ is minimized. The Euclidean distance between the time series and the model of the c-th regime is defined as:

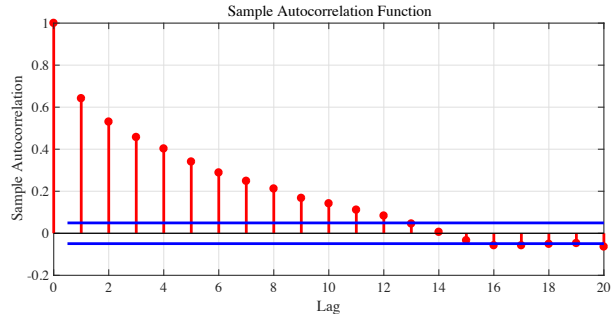
$$d(x(t), \beta_{0c}, \beta_{1c}, \rho_c) = \|[x(t) - (\beta_{0c} + \beta_{1c}.t)] - \rho_c[x(t-1) - (\beta_{0c} + \beta_{1c}.(t-1))]\|^2 \quad (8)$$

If the noise term $\epsilon(t)$ of the time series in each regime is uncorrelated, ordinary least square (OLS) can find the trend parameters in a closed form solution [10]. However, in the case of correlated residuals, other methods need to be utilized. Zhang et al. compared the performance of different methods in estimating the magnitude and statistical significance of trends in time series with AR(1) and showed that the estimation with *Generalized Least Square* (GLS) is better than non-parametric methods for longer time series (i.e. $T > 80$) [11]. Since the maximum likelihood of the linear trend with AR(1) noise doesn't have a closed form solution, the linear trend and AR(1) coefficient are commonly calculated by *Feasible Generalized Least Square* (FGLS). Here, we use the *Prais-Winsten* method [12] to find coefficients of linear regression and AR(1) and calculate the cost of each individual.

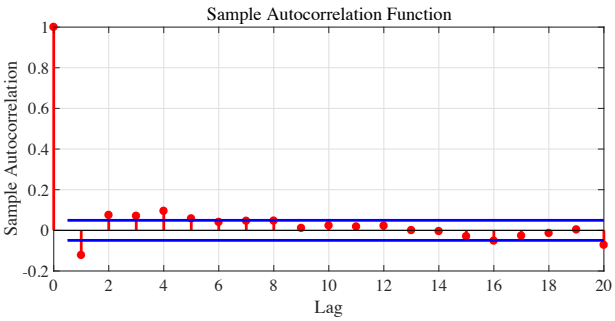
Using BIC, we found that the optimal number of regimes is 5 with 4 change points between them. Fig 2(a) shows the time series with fitted piecewise linear trend. The change points are in 1913, 1933, 1945 and 1964. The results of our analysis can be compared with similar studies [6]. Figure 2(b) and 2(c) show the autocorrelation of $\epsilon(t)$ and $w(t)$. It can be seen that by considering the AR(1) component, the residual of trend detection is almost uncorrelated.



(a)



(b)



(c)

Fig. 2. (a) Global temperature and fitted piecewise linear trend (b) The autocorrelation of residuals $\epsilon(t)$. (b) The autocorrelation of residuals $w(t)$ after fitting the trend.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

REFERENCES

[1] C. L. E. Franzke, "Warming trends: Nonlinear climate change," *Nature Climate Change*, vol. 4, no. 6, pp. 423–424, 2014.

[2] D. J. Seidel and J. R. Lanzante, "An assessment of three alternatives to linear trends for characterizing global atmospheric temperature changes," *Journal of Geophysical Research: Atmospheres*, vol. 109, no. D14, 2004.

[3] 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.

[4] S. Yue, P. Pilon, B. Phinney, and G. Cavadias, "The influence of autocorrelation on the ability to detect trend in hydrological series," *Hydrological Processes*, vol. 16, pp. 1807–1829, June 2002.

[5] F. Serinaldi and C. G. Kilsby, "The importance of prewhitening in change point analysis under persistence," *Stochastic Environmental Research and Risk Assessment*, pp. 1–15, Feb. 2015.

[6] E. Ruggieri and M. Antonellis, "An exact approach to Bayesian sequential change point detection," *Computational Statistics and Data Analysis*, vol. 97, pp. 71–86, 2016.

[7] A. M. Greene, T. Holsclaw, A. W. Robertson, and P. Smyth, "A Bayesian Multivariate Nonhomogeneous Markov Model," in *Machine Learning and Data Mining Approaches to Climate Science* (V. Lakshmanan, E. Gilleland, A. McGovern, and M. Tingley, eds.), pp. 61–69, Springer International Publishing, 2015.

[8] M. Gorji Sefidmazgi, M. Moradi Kordmahalleh, and A. Homaifar, "Identification of Switched Models in Non-Stationary Time Series based on Coordinate-Descent and Genetic Algorithm," in *Annual Conference on Genetic and Evolutionary Computation*, (Madrid), pp. 1399–1400, 2015.

[9] V. Jandhyala, S. Fotopoulos, I. MacNeill, and P. Liu, "Inference for single and multiple change-points in time series," *Journal of Time Series Analysis*, 2013.

[10] 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.

[11] X. Zhang and F. W. Zwiers, "Comment on Applicability of prewhitening to eliminate the influence of serial correlation on the Mann-Kendall test by Sheng Yue and Chun Yuan Wang," *Water Resources Research*, vol. 40, Mar. 2004.

[12] C. W. Ostrom, *Time Series Analysis: Regression Techniques*. Quantitative Applications in the Social Sciences, SAGE Publications, 1990.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

PATTERN EXTRACTION IN DYNAMICAL SYSTEMS USING INFORMATION GEOMETRY: APPLICATION TO TROPICAL INTRASEASONAL OSCILLATIONS

Eniko Székely[1], Dimitrios Giannakis[1]

***Abstract*—Datasets generated by dynamical systems are increasingly large both in sample size and dimensionality, and require new data analysis techniques that incorporate temporal information to process the large amount of observations. The framework that we propose here operates in spaces of probability measures induced by observables of the dynamical system rather than the more conventional machine learning approaches operating in the original data space. This allows us to use techniques from information geometry to study the dynamical evolution of observables. Dimension reduction is further employed to extract meaningful temporal and spatiotemporal patterns from the observations. The method is applied to the Lorenz 63 system and to real-world observations of the realtime multivariate Madden-Julian oscillation (RMM) index.**

I. INTRODUCTION

Dynamical systems are inherent to a vast array of applications, and it is worthwhile looking at the multitude of data analysis techniques available in the data mining and machine learning literature to analyze them. However, most existing machine learning techniques often consider the observations to be independent and identically distributed and do not take into account the temporal information (e.g., the time ordering of the data) which is directly linked to the dynamical evolution of the system.

The most common approach to pattern extraction in dynamical systems is to work in ambient data spaces, and compute the eigenfunctions of a covariance [1], [2], [3] or kernel [4], [5] operator defined on the data manifold. When working with dynamical systems it is important to consider the dynamical evolution of the system, and this is often achieved by embedding the data into the Takens time-lagged embedding (delay coordinate) space [6], [7]. A different approach that has been successfully applied to the analysis of nonlinear dynamical systems extracts the eigenfunctions of the Koopman (shift, composition) operator governing the evolution of observables in the phase space [8], [9], [10], [11], [12], [13]. However, in recent years there has been an increased interest in trying to uncover the dynamical processes by working in probability spaces [14], [15], [16], [17] where techniques from information geometry [18], [19] can be further employed. The current work lies at the intersection of three fields, namely machine learning, information geometry, and dynamical systems theory.

Corresponding author: E. Székely, eszekely@cims.nyu.edu
[1]Courant Institute for Mathematical Sciences, New York University, New York, NY 10012

II. METHOD

Let $(X, \mathcal{B}_X, \Psi_t, \mu)$ be a continuous-time deterministic ergodic dynamical system, where X is a compact topological space equipped with its Borel σ-algebra $\mathcal{B}_X$, $\Psi_t : X \mapsto X$ with $t \in \mathbb{R}$ is the flow map, and μ is a Ψ_t-invariant probability measure. Let also $f : X \mapsto \mathbb{R}^d$ be a d-dimensional vector-valued observable over the state space, and $T = [-\Delta t, 0]$ a closed time interval. We consider that we have at our disposal a time-ordered sequence $\{y_1, y_2, \ldots, y_N\}$ of d-dimensional measurements $y_i = f(x_i)$ of f taken at the states $x_i = \Psi_{(i-1)\,\delta t}(x_1) \in X$ for some initial state x_1 and sampling interval δt. We consider to have access to only partial observations of the dynamical system through the observable f.

Trajectories along the dynamical system induce probability measures over the measurable space defined by f. For every $x \in X$ we can define the probability measure $p_x : \mathcal{B}(\mathbb{R}^d) \mapsto [0, 1]$ such that

$$p_x(S) = \frac{\lambda(\{t \in T \mid f(\Psi_t(x)) \in S\})}{\Delta t}, \quad S \subset \mathbb{R}^d, \quad (1)$$

where λ is the Lebesgue measure over T.

The set $\mathcal{S} = \{p_x \mid x \in X\}$ of these probability measures forms a statistical manifold equipped with a natural Riemannian metric, i.e., the Fisher information

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

metric. This allows us to further use techniques from information geometry to study the evolution of observables of the dynamical system. In practice, we work with discrete probability measures, and make no assumptions on the shape of the probability distributions. This positions the present work in the nonparametric case where we use kernel density estimation (KDE) techniques [20] to estimate the PDFs in eq. (1).

Statistical distances, i.e., divergences, on a statistical manifold, such as the Hellinger distance further used in this paper, measure the amount of information between two probability density functions (PDF). The squared Hellinger distance between two discrete d-dimensional distributions $\hat{p}_{x_i}$ and $\hat{p}_{x_j}$ with densities $\hat{\rho}_{x_i}$ and $\hat{\rho}_{x_j}$, respectively, is defined as

$$\begin{aligned} d_H^2(\hat{p}_{x_i}, \hat{p}_{x_j}) &= \frac{1}{2} \left\| \sqrt{\hat{\rho}_{x_i}} - \sqrt{\hat{\rho}_{x_j}} \right\|_2^2 \\ &= \frac{1}{2} \sum_{l=1}^{Q} \left(\sqrt{\hat{\rho}_i^l} - \sqrt{\hat{\rho}_j^l} \right)^2, \end{aligned} \tag{2}$$

where Q is the number of evaluation points in KDE.

The Hellinger distance is used to define a symmetric and positive definite kernel over the space of these probability measures p_x. The squared Hellinger distance is negative definite [21], and any negative definite kernel can be used to define a positive definite kernel using the Gaussian kernel $k_H = e^{-\frac{d_H^2}{2\varepsilon^2}}$, where ε is the kernel width.

A natural class of scalar-valued observables on $\mathcal{S}$ are eigenfunctions of the Laplace-Beltrami operator associated with the Fisher information metric and the Gaussian kernel, which we approximate here using the Diffusion Maps algorithm of Coifman and Lafon [22]:

$$L\phi_j = \lambda_j \phi_j, \tag{3}$$

where L is a discrete Laplacian operator associated with k_H (constructed using the Diffusion Maps with normalization parameter $\alpha = 1$), ϕ_j are the eigenfunctions, and λ_j are their associated eigenvalues. The Laplace-Beltrami eigenfunctions capture temporal and spatiotemporal patterns of interest of the dynamical system, and are useful for dimension reduction and feature extraction.

III. EXPERIMENTS

A. Lorenz attractor

In this first experiment, we consider the Lorenz 63 mathematical model [23] initially proposed as a simple model for atmospheric convection, which consists of three ordinary differential equations:

$$\frac{d\omega^1}{dt} = \sigma(\omega^2 - \omega^1), \quad \frac{d\omega^2}{dt} = \omega^1(\rho - \omega^3) - \omega^2, \quad \frac{d\omega^3}{dt} = \omega^1\omega^2 - \beta\omega^3,$$

where $\omega^1, \omega^2, \omega^3$ are the system states, t is the time, and σ, ρ, β are the system parameters. We consider here the typical parameter values for the Lorenz system: $\rho = 28, \sigma = 10, \beta = 8/3$. The embedding in $\mathbb{R}^3$ is given by

$$F : X \mapsto \mathbb{R}^3, \quad F = (f^1, f^2, f^3),$$

$$f^1(x) = \omega^1(x), \qquad f^2(x) = \omega^2(x), \qquad f^3(x) = \omega^3(x).$$

We generated $N = 66,828$ points starting at the initial point $(0, 1, 1.05)$ for the time interval $T = [0, 500]$, after having removed the first 150 (transient) points. The probability measures were estimated using KDE with an embedding window of $\Delta t = 30$ timesteps and $q = 50$ evaluation points per dimension. The parameters of the Diffusion Maps algorithm were set to $k = 2,000$ nearest neighbors, and $\varepsilon = 0.4$ the width of the Gaussian kernel. The Lorenz 63 system is highly nonlinear and non-periodic, and feature extraction is therefore a challenging problem.

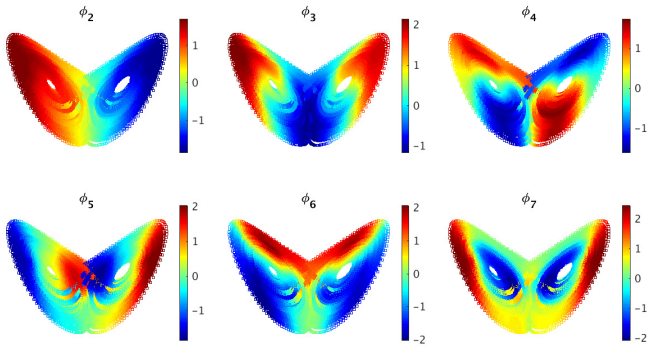


Fig. 1. Leading Diffusion Maps eigenfunctions of the Lorenz 63 system partially observed through $f = (f^1, f^2)$. The first eigenfunction is the constant vector of ones.

The leading Laplace-Beltrami eigenfunctions for the system partially observed through $f = (f^1, f^2)$ are shown in Fig. 1. The eigenfunctions ϕ_j capture different patterns of slowly varying timescales of the system, while faster varying timescales emerge as we go deeper in the eigenfunction spectrum. Figure 2 shows examples of one-dimensional time series and two-dimensional representations of the Diffusion Maps eigenfunctions for the system partially-observed through $f = (f^1, f^2)$. In this case, eigenfunction ϕ_2 represents the two wings of the Lorenz attractor, i.e., positive and negative values correspond to the left and right wing of the attractor, respectively (see also Fig. 1); eigenfunction ϕ_3 represents the variation within each wing with positive values

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

corresponding to points further apart from the intersection of the wings, while negative values correspond to points closer to the intersection; and eigenfunction ϕ_4 represents a switch between the wings. We tested our algorithm for robustness using the following values: $\Delta t \in [30, 50]$, $k \in [1000, 5000]$, $\varepsilon \in [0.2, 1]$, and the results are robust within these ranges.

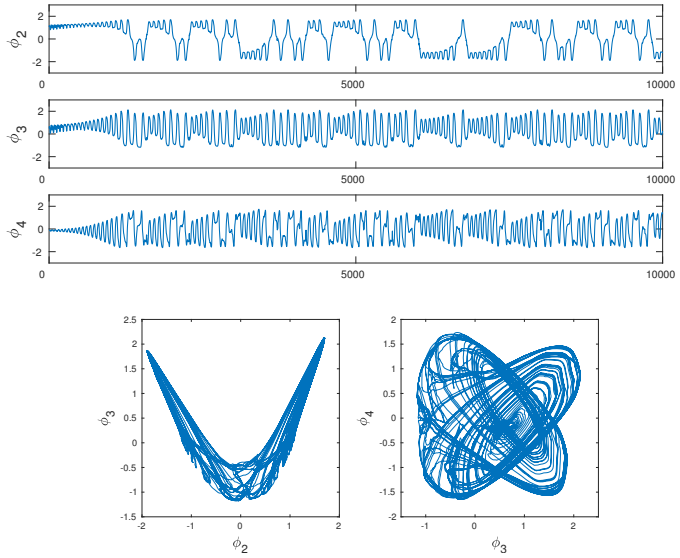


Fig. 2. **Top:** Time series of the Diffusion Maps eigenfunctions of the Lorenz 63 system partially observed through $f = (f^1, f^2)$ used in the spatial reconstructions in Fig. 1 for the first 10,000 samples. The first eigenfunction is the constant vector of ones. **Bottom:** Examples of two-dimensional representations of the eigenfunctions.

B. Realtime multivariate MJO (RMM) index

In the second experiment we extract temporal patterns of the realtime multivariate Madden-Julian oscillation (RMM) index [24]. The dominant boreal winter tropical intraseasonal oscillation (ISO) is the well-known Madden-Julian oscillation (MJO; [25], [26]), a 30-90-day eastward-propagating pattern with zonal wavenumber 1-4. Among the multitude of indices for MJO, the RMM index is the most common measure of ISO activity used all year-round, both for boreal winter and boreal summer activity. RMM is a combined measure of the first two empirical orthogonal functions (EOFs) of bandpass-filtered, and equatorially averaged outgoing longwave radiation (OLR) and 200hPa and 850hPa zonal wind data.

The dataset for the RMM index covers 23 years from September 1983 to June 2006, sampled once a day. The parameters that we use for extracting the eigenfunctions are as follows: $k = 100$ the number of nearest neighbors, $\varepsilon = 0.1$ the width of the Gaussian kernel, and $\Delta t = 60$ days the embedding window. We chose $\Delta t = 60$ days as it represents the average time of an MJO (30-90 days). The correlation between the first non-constant eigenfunction ϕ_2 and the averaged RMM index shown in Fig. 3 is of 0.9359 (both have been normalized). This shows that the eigenfunctions detected using our framework recover intrinsic properties of the statistical manifold, i.e., here ϕ_2 recovers the mean of the PDFs on the manifold. The next two eigenfunctions $\{\phi_3, \phi_4\}$ are shown in Fig. 4. We are working on interpreting the other eigenfunctions in the spectrum [27]. For example, we find that some of the eigenfunctions are strongly correlated ($\approx 0.6 - 0.7$) with the standard deviation of the RMM index over time windows. Having the mean and the standard deviation could be very useful for example for prediction when the past trajectory of the dynamical system is known. We tested our algorithm for robustness using the following values: $\Delta t \in [30, 90]$, $k \in [50, 1000]$, $\varepsilon \in [0.05, 1]$, and the results are very robust within these ranges.

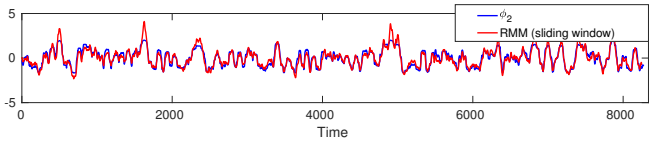


Fig. 3. The second eigenfunction ϕ_2 vs. the averaged RMM index over a sliding window with $\Delta t = 60$ days.

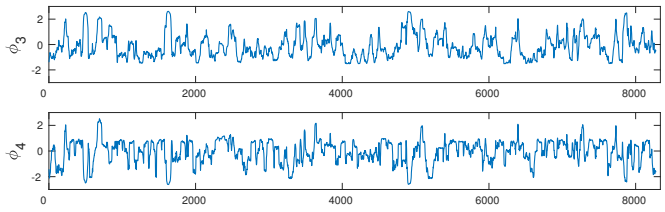


Fig. 4. Eigenfunctions $\{\phi_3, \phi_4\}$ associated with the RMM index.

IV. Conclusion

We introduced a new framework for pattern extraction in dynamical systems that relies on machine learning and new concepts from information geometry on statistical manifolds. We demonstrated on a mathematical model, i.e., the Lorenz 63 system, and on real-world observations of the realtime multivariate MJO index, the efficiency and robustness of our technique. In ongoing work [27], we are interested in better understanding the eigenfunctions and their connection to intrinsic properties of the dynamical system.

ACKNOWLEDGMENTS

This research was supported by ONR MURI grant N00014-12-1-0912 and ONR grant N00014-14-1-0150.

REFERENCES

[1] N. H. Packard *et al.*, "Geometry from a time series," *Phys. Rev. Lett.*, vol. 45, pp. 712–716, 1980.

[2] D. S. Broomhead and G. P. King, "Extracting qualitative dynamics from experimental data," *Phys. D*, vol. 20, no. 2–3, pp. 217–236, 1986.

[3] M. Ghil *et al.*, "Advanced spectral methods for climatic time series," *Rev. Geophys.*, vol. 40, 2002.

[4] D. Giannakis and A. J. Majda, "Nonlinear Laplacian spectral analysis for time series with intermittency and low-frequency variability," *Proc. Natl. Acad. Sci.*, vol. 109, no. 7, pp. 2222–2227, 2012.

[5] T. Berry, R. Cressman, Z. Greguric Ferencek, and T. Sauer, "Time-scale separation from diffusion-mapped delay coordinates," *SIAM J. Appl. Dyn. Sys.*, vol. 12, pp. 618–649, 2013.

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

[7] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *J. Stat. Phys.*, vol. 65, no. 3–4, pp. 579–616, 1991.

[8] I. Mezić and A. Banaszuk, "Comparison of systems with complex behavior," *Physica D-nonlinear Phenomena*, vol. 197, p. 101–133, Oct 2004.

[9] I. Mezić, "Spectral properties of dynamical systems, model reduction and decompositions," *Nonlinear Dynamics*, vol. 41, p. 309–325, Aug 2005.

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

[11] D. Giannakis, J. Slawinska, and Z. Zhao, "Spatiotemporal feature extraction with data-driven Koopman operators," *Journal of Machine Learning Research*, vol. 44, pp. 103–115, 2015.

[12] D. Giannakis, "Data-driven spectral decomposition and forecasting of ergodic dynamical systems," *Appl. Comput. Harmon. Anal.*, 2016. arXiv:1507.02338.

[13] S. Brunton, B. Brunton, J. L. Proctor, and J. N. Kutz, "Koopman invariant subspaces and finite linear representations of nonlinear dynamical systems for control," *PLoS ONE*, vol. 11, no. 2, 2016.

[14] M. Muskulus and S. Verduyn-Lunel, "Wasserstein distances in the analysis of time series and dynamical systems," *Physica D: Nonlinear Phenomena*, vol. 240, no. 1, pp. 45 – 58, 2011.

[15] R. Talmon and R. Coifman, "Empirical intrinsic geometry for nonlinear modeling and time series filtering," *Proceedings of the National Academy of Sciences*, vol. 110, no. 31, pp. 12535–12540, 2013.

[16] W. Lian, R. Talmon, H. Zaveri, L. Carin, and R. Coifman, "Multivariate time-series analysis and diffusion maps," *Signal Process.*, vol. 116, pp. 13–28, Nov. 2015.

[17] C. J. Dsilva, R. Talmon, C. W. Gear, R. R. Coifman, and I. G. Kevrekidis, "Data-driven reduction for a class of multiscale fast-slow stochastic dynamical systems," *SIAM Journal on Applied Dynamical Systems*, vol. 15, no. 3, pp. 1327–1351, 2016.

[18] S. Amari and H. Nagaoka, *Methods of Information Geometry*, vol. 191 of *Translations of Mathematical Monographs*. Providence: American Mathematical Society, 2007.

[19] J. D. Lafferty and G. Lebanon, "Diffusion kernels on statistical manifolds," *J. Mach. Learn. Res.*, vol. 6, pp. 129–163, 2005.

[20] A. Bowman and A. Azzalini, *Applied smoothing techniques for data analysis*. Oxford University Press, 1997.

[21] M. Hein and O. Bousquet, "Hilbertian metrics and positive definite kernels for probability measures," in *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2005.

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

[23] E. N. Lorenz, "Deterministic nonperiodic flow," *J. Atmos. Sci.*, vol. 20, pp. 130–141, 1963.

[24] M. C. Wheeler and H. H. Hendon, "An all-season real-time multivariate MJO index: Development of an index for monitoring and prediction," *Mon. Wea. Rev.*, vol. 132, no. 8, pp. 1917–1932, 2004.

[25] R. A. Madden and P. R. Julian, "Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific," *J. Atmos. Sci.*, vol. 28, no. 5, 1971.

[26] R. A. Madden and P. R. Julian, "Description of global-scale circulation cells in the tropics with a 40–50 day period," *J. Atmos. Sci.*, vol. 29, no. 6, pp. 1109–1123, 1972.

[27] E. Székely and D. Giannakis, "An information-geometric approach for feature extraction in ergodic dynamical systems," *In preparation*, 2017.

7th International Workshop on Climate Informatics
September 20-22, 2017
CI 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

A PHYSICS-BASED APPROACH TO UNSUPERVISED DISCOVERY OF COHERENT STRUCTURES IN SPATIOTEMPORAL SYSTEMS

Adam Rupe[1,2], James P. Crutchfield[1], Karthik Kashinath[2], Mr Prabhat[2]

***Abstract*—Given that observational and numerical climate data are being produced at ever more prodigious rates, increasingly sophisticated and automated analysis techniques have become essential. Deep learning is quickly becoming a standard approach for such analyses and, while great progress is being made, major challenges remain. Unlike commercial applications in which deep learning has led to surprising successes, scientific data is highly complex and typically unlabeled. Moreover, interpretability and detecting new mechanisms are key to scientific discovery. To enhance discovery we present a complementary physics-based, data-driven approach that exploits the causal nature of spatiotemporal data sets generated by local dynamics (e.g. hydrodynamic flows). We illustrate how novel patterns and coherent structures can be discovered in cellular automata and outline the path from them to climate data.**

I. MOTIVATION

Incredibly complex and sophisticated models are currently employed to simulate the global climate system to facilitate our understanding of climate as well as increase our predictive power, most notably in regards to the effects of increased carbon levels. Our ability to simulate however has rapidly outpaced our ability to analyze the resulting data. Often the climate community resorts to rather simplistic data analyses, such as linear decomposition methods like EOF analyses [1], [2] or detecting (linear) trends in climate data time series [3]. Nonlinear and more sophisticated techniques are rarely brought to bear. Here we focus on one particular aspect of nonlinear dynamical systems analysis, the detection and discovery of coherent structures, such as cyclones and atmospheric rivers in climate data.

Coherent structures were introduced in the study of fluid dynamics and were initially defined as regions characterized by high levels of coherent vorticity, *i.e.* regions where instantaneously space and phase correlated vorticity are high. The contours of coherent vorticity constitute an identifier to the structure's boundaries. However, pinning down this concept of coherent structures with rigorous and principled definitions or heuristics which produce consistent results across a wide class of physical systems is a challenging and open problem [4]. Climate practitioners are left with more ad hoc approaches [5], [6], [7] which can make it difficult to draw meaningful conclusions from analysis [8].

Deep learning attempts to sidestep this issue by learning how to identify coherent structures from labeled data [9]. However, we currently can not peer into the box to find out exactly what the defining characteristics a deep net uses to identify structures. Current state of the art achieves semi-supervised bounding box identification [10]. The ultimate goal would be unsupervised segmentation; that is, a pixel-level identification without reliance on labeled training data. It is not yet clear how to achieve this.

Like deep learning, our theory [11] approaches coherent structures from a rather different (and more general) perspective than the original context of Lagrangian coherence principles in fluid flows.

II. METHOD

Starting from basic physics principles, coherent structures can most generally be seen as *localized broken symmetries*. Two questions naturally arise; what are the symmetries which are broken and how can we identify such symmetry in a diverse range of spatiotemporal systems? Coherent structures can be found in a variety of systems with different physical properties. Convection cells in hydrodynamic systems and spiral waves in reaction-diffusion systems, for example. It is clear that the common thread is the underlying nonlinear dynamics of these systems [12], [13], [14].

Corresponding author: A Rupe, atrupe@ucdavis.edu [1]Complexity Sciences Center, Department of Physics, University of California Davis [2]NERSC, Lawrence Berkeley National Laboratory

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

A framework known as *computational mechanics* [15], [16] has been developed to study pattern and structure in this dynamical context. The canonical object of computational mechanics is the *ϵ-machine* [17], a type of stochastic finite-state machine known as a hidden Markov model, which consists of a set of *causal states* and transitions between them. The causal states are constructed from the *causal equivalence relation*.

$$\overleftarrow{x}_i \sim_\epsilon \overleftarrow{x}_j \iff \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}_i) = \Pr(\overrightarrow{X}|\overleftarrow{X} = \overleftarrow{x}_j).$$

In words, two pasts $\overleftarrow{x}_i$ and $\overleftarrow{x}_j$ are *causally equivalent* if and only if they make the same prediction for the future $\overrightarrow{X}$; that is, they have the same conditional distribution over the future. The causal states are the unique *minimal sufficient statistic* of the past to predict the future.

For our application to coherent structures we use a straightforward spatiotemporal generalization known as the *local causal states* [18]. For systems which evolve under some local dynamic and information propagates through the system at a finite speed, it is quite natural to use *lightcones* as local notions of pasts and futures. Formally, the past lightcone of a spacetime point $x(\vec{r}, t)$ is the set of all points at previous times that could possibly influence it. That is,

$$\ell^-(\vec{r}, t) \equiv \left\{x(\vec{r}', t') \mid t' \leq t \text{ and } ||\vec{r}' - \vec{r}|| \leq c(t' - t)\right\}$$

where c is the finite speed of information propagation in the system. Similarly, the future lightcone is given as all the points at subsequent times that could possibly be influenced by $x(\vec{x}, t)$.

$$\ell^+(\vec{r}, t) \equiv \left\{x(\vec{r}', t') \mid t' > t \text{ and } ||\vec{r}' - \vec{r}|| < c(t - t')\right\}$$

The choice of lightcone representations for both local pasts and futures is ultimately a weak-causality argument; influence and information propagate locally through a spacetime site from its past lightcone to its future lightcone.

The generalization of the causal equivalence relation is straightforward. Two past lightcones are causally equivalent if they have the same conditional distribution over future lightcones.

$$\ell_i^- \sim_\epsilon \ell_j^- \iff \Pr\left(\mathrm{L}^+|L^- = \ell_i^-\right) = \Pr\left(\mathrm{L}^+|L^- = \ell_j^-\right)$$

This *local causal equivalence relation* over lightcones is designed around an intuitive notion of *optimal local prediction* [18]. At some site $x(\vec{r}, t)$ in spacetime, given knowledge of all past spacetime points which could possibly affect $x(\vec{r}, t)$, i.e. its past lightcone $\ell^-(\vec{r}, t)$, what might happen at all subsequent spacetime points which could be affected by $x(\vec{r}, t)$, i.e. its future lightcone $\ell^+(\vec{r}, t)$? Local causal states are minimal sufficient statistics for optimal local prediction. Moreover, the particular local prediction done here uses lightcone shapes, which are associated with local causality in the system. Thus it is not direct causal relationships (e.g. learning equations of motion from data) that the local causal states are discovering. Rather, they are exploiting a kind of causality in the system (i.e. that the future follows the past and that information propagates at a finite speed) in order to discover spacetime structure.

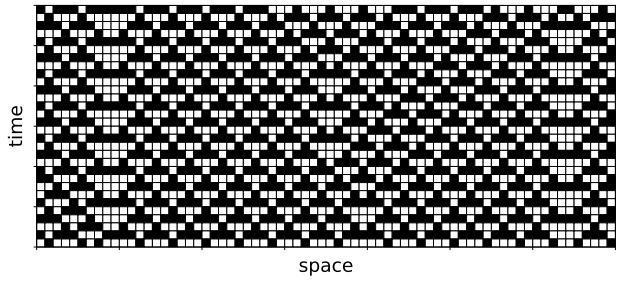
Once local causal states have been inferred from data, each site in a representative spacetime field can be assigned its local causal state label in a process known as *causal filtering* [11]. This is how we achieve unsupervised image segmentation. Though it must be clearly stated that this is a *spacetime segmentation*, and not a general image segmentation algorithm, exactly because it works *only* in systems for which lightcones are well-defined.

Using the local causal states we can, in a general and principled manner, discover dynamical spatiotemporal symmetries in a system from data. These symmetry regions are known as *domains* and are defined as regions where the associated local causal state field, after causal filtering, has spacetime symmetry tilings. A *coherent structure* is then defined as a set of spatially localized, temporally persistent (in the Lagrangian sense) non-domain local causal states.

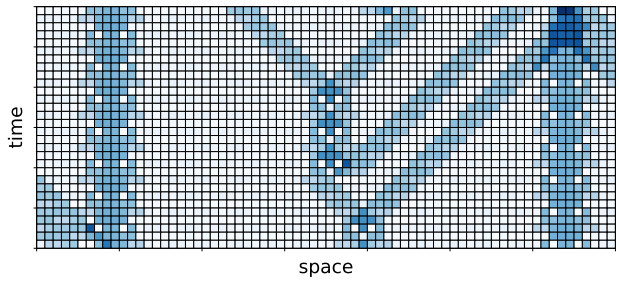
From prior work by Hanson and Crutchfield [19], [20], [21], the domains of 1-D cellular automata are well understood as dynamically invariant sets of homogeneous spatial configurations. There is strong empirical evidence [11] that the domains of cellular automata discovered by the local causal states are *exactly* the domains as described by Hanson and Crutchfield. Therefore the local causal states are discovering spatiotemporally symmetries which are externally well-defined. In turn there is a strong agreement between the description of coherent structures in cellular automata discovered by local causal states and the coherent structures as described by Hanson and Crutchfield.

III. TOWARDS CLIMATE

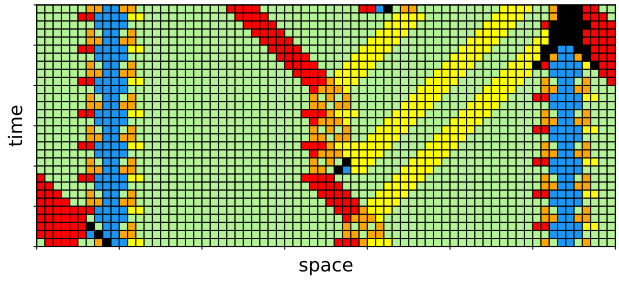
With consistent and readily interpretable results on cellular automata we are now working on generalizing to real-valued spatiotemporal systems, with specific emphasis on canonical fluid flows. Others have done preliminary work on this generalization, where an extra discretization (typically via clustering) step is needed during reconstruction [22], [23].



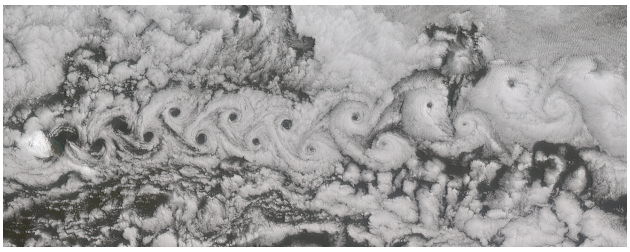
(a) Raw spacetime field of ECA 54



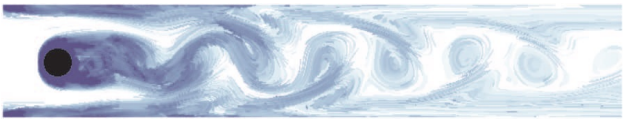
(b) Local statistical complexity filter of (a)



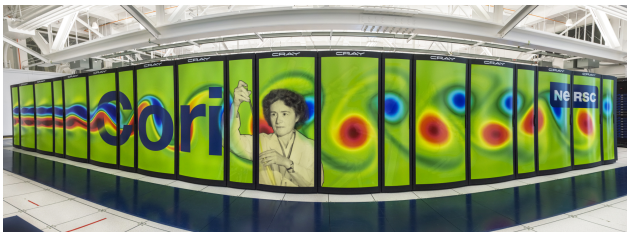
(c) Local causal state coherent structure filter of (a)



(d) Vortex street in atmosphere



(e) Vortex street complexity field



(f) Colored vortices on Cori

Fig. 1. Visualization of results on 1D cellular automata (fully-discrete spatiotemporal models) and projected analogous results for fluid systems. CA results for elementary cellular automaton rule 54 are given in (a)-(c). The raw spacetime field is shown in (a) and a corresponding local statistical complexity field in (b). From the local statistical complexity filter, which is a qualitative information-theoretic "rare event" filter, it is clear there are coherent structures on top of a background domain, but the four different structures can not be explicitly distinguished and identified. Thus a more detailed coherent structure filter using our unsupervised local causal state segmentation analysis is given in (c). Here states participating in the domain spacetime symmetry tiling are colored green, and other non-domain states which satisfy our definition for a coherent structure are colored according to the structure(s) they belong to. Interaction states not associated with domain or a coherent structure are in black. An outline of analogous results for vortex shedding is shown in (d)-(f). (d) A vortex street in the cloud layer over the arctic (Source: https://photojournal.jpl.nasa.gov/catalog/PIA03448). (e) The local statistical complexity of the vorticity field for a canonical vortex street simulation, taken from [22], analogous to the qualitative structure filter of (b). (f) Closer to the more detailed and principled coherent structure filter of (c) are the colored vortices displayed on the cover of the NERSC Cori HPC system. We emphasize the analogy is not that learning about coherent structures in CAs will give insight into fluid and climate structures. Rather, it is to illustrate how we foresee our approach will discover coherent structures in fluids and climate, in much the same way we can currently discover structures in CAs.

These groups have also used local causal states for coherent structure detection, including real-valued applications like fluids and even climate [22]. However, they have all relied on the "local statistical complexity" [24], which is the point-wise entropy over local causal states. At best this is simply a qualitative filtering tool which aides in visual recognition of structures and at worst can give both false positive and false negative misidentification. We are the first to give a principled and rigorous method for coherent structure discovery and description using the local causal states, and are working to generalize this more detailed analysis to real-valued systems. In doing so we hope to move beyond the scope of data visualization these prior groups were working in, and facilitate novel scientific discovery, particularly in climate science.

On the theory side, we must confirm our methods on known fluid structures. As the theory is founded in

basic dynamical principles it is likely to apply without much modification in fluid systems. We will also begin to explore whether our methods can facilitate additional mechanistic insight beyond structure discovery. For example, whether there are any links between the local causal state analysis and thermodynamic considerations.

On the implementation side, the computational costs of local causal state reconstruction in more complex systems will require fully-distributed execution on large HPC machines. This will certainly be the case for TB scale climate data sets we ultimately are interested in. As our primary objective is automated coherent structure discovery, moving from canonical fluid flows to large-scale climate data will largely be a matter of computational scaling. With access to HPC experts from the Intel Big Data Center and the NERSC Cori system at Lawrence Berkeley National Laboratory we feel well-positioned to tackle these computational challenges.

ACKNOWLEDGMENTS

Adam Rupe and Jim Crutchfield would like to acknowledge Intel for supporting the IPCC at UC Davis. Prabhat and Karthik Kashinath were supported by the Intel Big Data Center. This research is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract W911NF- 13-1-0390, and used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

[1] NCAR, "Climate data analysis tools and methods." https://climatedataguide.ucar.edu/climate-data-tools-and-analysis, Accessed: 2017-07-25.

[2] M. Ghil, M. Allen, M. Dettinger, K. Ide, D. Kondrashov, M. Mann, A. Robertson, A. Saunders, Y. Tian, and F. V. P. Yiou, "Advanced spectral methods for climatic time series," *Reviews of Geophysics*, vol. 40, no. 1, 2002.

[3] D. Shea, "Climate data analysis tools and methods - trend analysis." https://climatedataguide.ucar.edu/climate-data-tools-and-analysis/trend-analysis. Accessed: 2017-07-25.

[4] A. Hadjighasem, M. Farazmand, D. Blazevski, G. Froyland, and G. Haller, "A critical comparison of lagrangian methods for coherent structure detection," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 5, p. 053104, 2017.

[5] F. Vitart, J. Anderson, and W. Stern, "Simulation of interannual variability of tropical storm frequency in an ensemble of gcm integrations," *Journal of Climate*, vol. 10, no. 4, pp. 745–760, 1997.

[6] K. Walsh and I. Watterson, "Tropical cyclone-like vortices in a limited area model: Comparison with observed climatology," *Journal of climate*, vol. 10, no. 9, pp. 2240–2259, 1997.

[7] Prabhat, S. Byna, V. Vishwanath, E. Dart, M. Wehner, W. D. Collins, *et al.*, "Teca: Petascale pattern recognition for climate science," in *International Conference on Computer Analysis of Images and Patterns*, pp. 426–436, Springer, 2015.

[8] D. Faranda and D. Defrance, "A wavelet-based approach to detect climate change on the coherent and turbulent component of the atmospheric circulation," *Earth System Dynamics*, vol. 7, no. 2, pp. 517–523, 2016.

[9] Y. Liu, E. Racah, Prabhat, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, and W. Collins, "Application of deep convolutional neural networks for detecting extreme weather in climate datasets," *arXiv preprint arXiv:1605.01156*, 2016.

[10] E. Racah, C. Beckham, T. Maharaj, Prabhat, and C. Pal, "Semi-supervised detection of extreme weather events in large climate datasets," *arXiv preprint arXiv:1612.02095*, 2016.

[11] A. Rupe and J. P. Crutchfield, "Local causal states and discrete coherent structures." http://csc.ucdavis.edu/~cmg/compmech/pubs/dcs.htm, 2017.

[12] M. Cross and H. Greenside, *Pattern Formation and Dynamics in Nonequilibrium Systems*. Cambridge University Press, 2009.

[13] R. Hoyle, *Pattern Formation: An Introduction to Methods*. New York: Cambridge University Press, 2006.

[14] M. Golubitsky and I. Stewart, *The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space*, vol. 200. Birkhäuser, 2003.

[15] J. P. Crutchfield and K. Young, "Inferring statistical complexity," *Phys. Rev. Let.*, vol. 63, pp. 105–108, 1989.

[16] J. P. Crutchfield, "Between order and chaos," *Nature Physics*, vol. 8, no. January, pp. 17–24, 2012.

[17] C. R. Shalizi and J. P. Crutchfield, "Computational mechanics: Pattern and prediction, structure and simplicity," *J. Stat. Phys.*, vol. 104, pp. 817–879, 2001.

[18] C. Shalizi, "Optimal nonlinear prediction of random fields on networks," *Discrete Mathematics & Theoretical Computer Science*, 2003.

[19] J. E. Hanson and J. P. Crutchfield, "The attractor-basin portrait of a cellular automaton," *J. Stat. Phys.*, vol. 66, pp. 1415 – 1462, 1992.

[20] J. P. Crutchfield, "Discovering coherent structures in nonlinear spatial systems," in *Nonlinear Ocean Waves* (A. Brandt, S. Ramberg, and M. Shlesinger, eds.), (Singapore), pp. 190–216, World Scientific, 1992. also appears in Complexity in Physics and Technology, R. Vilela-Mendes, editor, World Scientific, Singapore (1992).

[21] J. E. Hanson and J. P. Crutchfield, "Computational mechanics of cellular automata: An example," *Physica D*, vol. 103, pp. 169–189, 1997.

[22] H. Jänicke, A. Wiebel, G. Scheuermann, and W. Kollmann, "Multifield visualization using local statistical complexity," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1384–1391, 2007.

[23] G. Goerg and C. Shalizi, "Mixed licors: A nonparametric algorithm for predictive state reconstruction," in *Artificial Intelligence and Statistics*, pp. 289–297, 2013.

[24] C. Shalizi, R. Haslinger, J.-B. Rouquier, K. Klinkner, and C. Moore, "Automatic filters for the detection of coherent structure in spatiotemporal systems," *Physical Review E*, vol. 73, no. 3, p. 036104, 2006.

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

TOWARDS A STATISTICAL MODEL OF TROPICAL CYCLONE GENESIS

Arturo Fernandez[1,2], Karthik Kashinath[2], Jon McAuliffe[1], Prabhat[2], Philip B. Stark[1], Michael Wehner[2]

***Abstract*—Tropical Cyclones (TCs) are an important class of extreme weather phenomenawith a high impact on humans. Their formation/genesis, evolution, intensification, and dissipation over land are important problems in climate science. This paper explores how accurately a statistical model can predict TC genesis in numerical models. We use the TECA software to extract TC trajectories from CAM5.1 model output, then apply L1-regularized logistic regression to create a predictive model with interpretable results. The active variables selected by the analysis confirm earlier hypotheses about TC genesis.**

I. MOTIVATION

Numerical global circulation models allow researchers to explore the contributions of environmental variables that are inaccessible to direct measurement. The space-time resolution of CAM5.1 is especially useful for simulating real-world weather dynamics and allowing statistical methods to discern what lead times, time intervals, space resolution, and measured variables are the most useful in predicting genesis events. The goal of this study is to develop accurate TCG forecasts as well as test the relationship between measured variables and TC formation globally. Several studies have investigated TCG. Typically, studies reduce the high-dimensional collection of spatio-temporal and environmental variables to aggregate measures motivated by previous climatological studies. Notably, recent genesis potential metrics [1] have foregone Sea Surface Temperature (SST or TS) in favor of alternative intensity measures [2]. Other studies replace climatological intensity metrics with probabilities from statistical models ([3], [4], [5], [6]).

Analysis of TC activity in a single basin does not provide much insight into the mechanism that leads to TCG. An analysis of the Northwest Pacific (NWP) [6], Australian (AUS) [7], or North Atlantic (NATL) ([8], [9]) basin will often say more about a seasonal weather phenomenon, such as the Madden-Julian Oscillation (MJO) or the African Easterly Waves (AEW), than the actual physical variables that influence genesis. The present study is unique and expands on previous works by developing a single probabilistic model for TCG across the globe. Such a framework makes it possible test the uniformity of the relationships between the physical environment and TCG probabilities across space and time. Previous studies have had a limited ability to discover TCG mechanisms because they have included variables such as the specific time and position. Although these models are predictive, the space-time variables are proxies for important physical features such as temperature and precipitation, and therefore mask their role in TCG. Here, we transform the raw data into a standardized form that retains the interpretability of the original measurements.

Corresponding author: Arturo Fernandez, arturof@berkeley.edu
[1]Department of Statistics, University of California, Berkeley
[2]Lawrence Berkeley National Lab, Berkeley, CA

II. METHOD

We analyze data from two sources. The first is the Community Atmospheric Model (CAM5.1), run at Lawrence Berkeley National Lab's (LBNL) National Energy Research Scientific Computing Center (NERSC) facility. It consists of a 17-year simulation that emulates global climate conditions starting in 1990. This dataset includes 16 state variables (winds, temperature, humidity, etc.) at 3 hour intervals at a spatial resolution of 0.23° by 0.31° or approximately 25 km at the equator. The second data source consists of tropical cyclone trajectories detected by the Toolkit for Extreme Climate Analysis (TECA 1.0) [10]. We define Tropical Cyclone Genesis (TCG) Events to be the first recorded track for a given storm system.

To investigate the physical factors that lead to TCG, we piece together *lag vectors* that correspond to sequential lead times for a 24-hour history. These lag vectors are formed by piecing together d_v *state vectors*. A state vector contains the values of a state variable in a grid of height d_y by width d_x, centered around a TCG event, that has been flattened. Ultimately, the covariate vector used for training the models is of size

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

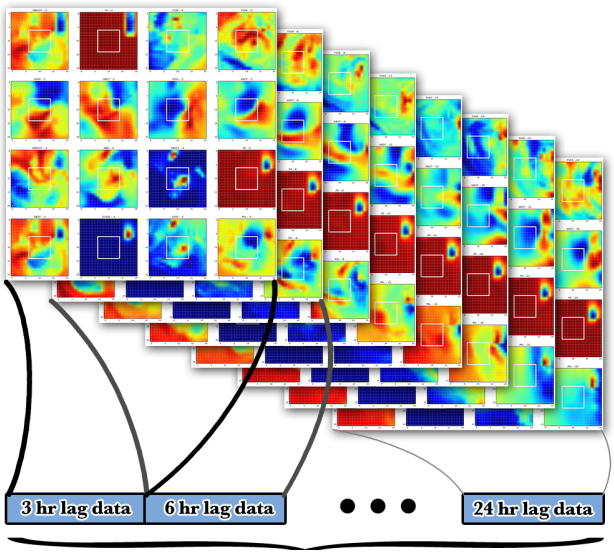


Fig. 1: Converting data to a covariate vector in $\mathbb{R}^d$

$d = d_t \times d_v \times d_y \times d_x = 76880$ since $d_t = 8$, $d_v = 10$ (reduced from 16 due to redundancies), and $d_y = d_x = 31$. This process is explained in Figure 1.

We use ℓ_1-regularized Logistic Regression (L1LR) to predict and describe TCG events. Other tree-based machine learning methods such as Extreme Gradient Boosting (XGBoost) [11] were tried as well. L1LR gave comparable accuracy results to XGBoost and is easier to interpret, so only it is discussed here. This linear, statistical model is used to predict whether an environment will provide favorable conditions for TCG, and not to model the physics of the real system. As seen in Section III, the model's performance justifies its use.

Consider a dataset $\{x_i, y_i\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$ is covariate information (feature vector), $y_i \in \{0, 1\}$ is a class label, and N is the number of observed examples. *Logistic regression* (LR) models the probability distribution of y, given covariate information x, as

$$p(y = 1|x; \theta) = \sigma(\theta^T x) = \frac{1}{1 + \exp(-\theta^T x)} \quad (1)$$

where $\theta \in \mathbb{R}^d$ is the parameter vector for the model and $\sigma(\cdot)$ is the sigmoid function. Such a model is typically fit by maximum likelihood estimation (MLE): $\hat{\theta} = \arg\max_\theta \prod_{i=1}^N p(y_i|x_i; \theta)$. This optimization problem can be regularized by adding an ℓ_1 penalty to bias the estimate towards a *sparse* $\hat{\theta}$. Expressed using the log likelihood, the L1LR estimate is :

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^{N} -\log p(y_i|x_i; \theta) + \lambda\|\theta\|_1 \quad (2)$$

Adjusting λ trades off between sparsity and model accuracy.

Parameters	Min Vorticity	Min TWC	Min Thickness	Min PSL
Strong Storms	1.6×10^{-4}	0.8	50	400
Weak Storms	0.2×10^{-4}	0.1	6.25	50

TABLE I: Storm Parameters

Examples	Positive	Negative
Training (years 1-13)	1149	1333
Test (years 14-17)	355	397

TABLE II: Sample Sizes

Here we define a positive example $y = 1$ as a strong system and $y = 0$ as a weak system. Strong systems are defined to have standard threshold parameters used by TECA, whereas weak storms are captured by relaxed parameters (see Table I).

Lastly, we note that there are clear differences between tropical cyclones in the Northern (NH) and South Hemispheres (SH): their trajectories move away from the equator in opposite directions and their vorticities have opposite signs. We remedy this situation by simply flipping state-variable grids along their vertical axis for TCs in the SH (i.e. working with their mirror image).

III. EVALUATION

To arrive at our final model, we assessed the classification error, Receiver Operating Characteristic (ROC) Area Under the Curve (AUC), and sparsity. First, we compared different lag times and incrementally added lag times to decide how much temporal information to use. As lead times were increased, accuracy went down. When additional lags were included, accuracy metrics did not increase noticeably thus we decided to focus on information available at a 3 hour lag time.

We evaluated our model's accuracy on two test datasets that were not used to fit the model. First, we assessed its accuracy on a test set with strong and weak systems (see Table II) . Secondly, we tested its ability to classify randomly sampled points, in space and time, with no TC track association as negative examples. We call this the *inactive* data, or Test Set 2 in Table III.

Eval Set	n	Error	AUC	# Mis.
Test Set	752	0.165	0.916	124
Test Set 2	13,134	0	-	0

TABLE III: Model Accuracy

7th International Workshop on Climate Informatics
September 20-22, 2017
Hosted by the National Center for Atmospheric Research in Boulder, CO

Table III shows that the genesis of strong versus weak systems can be classified with reasonably high accuracy. Limits are expected since there exist strong systems with no preceding weak tracks. The model has an excellent AUC, which suggests that we can obtain high accuracy for predicting strong systems with relatively low cost in mis-labeling some weak systems as strong. Nonetheless, the evaluation on inactive data shows the model's ability to control false alarms. Out of the inactive evaluation points, none is misclassified. By focusing on a difficult problem (differentiating between strong and weak TCG), inactive points are much less likely to be classified as leading to TCG.

Figure 2 displays the most accurate model on the test set. Although the model has at least one non-zero coefficient for every state variable, U850 (Zonal wind at 850 mbar) and PSL (Sea Level Pressure) have the largest coefficients. This suggests that U850 and PSL are the most predictive state variables for TCG. To confirm this, we developed a variable importance measure by permuting the data spatially and across samples, and then measuring the change in training accuracy. Figure 3 confirms that U850 and PSL are the most predictive variables with TS (Sea Surface Temperature), QREFTH (Reference Height Humidity) and PRECT (Total, convective and large-scale, precipitation rate) also having recognizable importance. The model's wind patterns are related to known mechanisms such as vorticity and the sea surface temperature pattern reflects the warm core we would expect. Notably, stronger systems are preceded by a more pronounced PSL local minima and also have increased levels of humidity and precipitation.

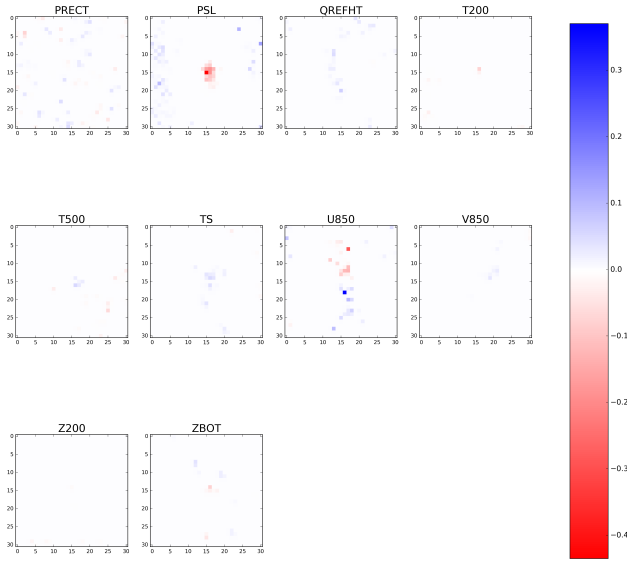


Fig. 2: Final Model

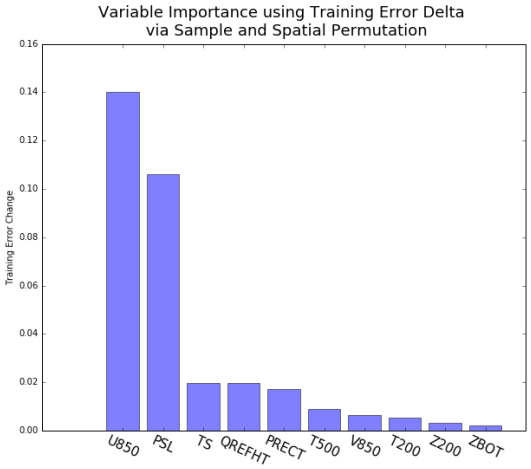


Fig. 3: Variable Importance

IV. CONCLUSION

Our results and the current literature make it clear that much work remains to be done to understand TCG. The current research helps connect global climate models to the physics of TCG and might be useful for developing early warning systems. Of course, the current work addresses only model output and not actual atmospheric data; performance on real-world data remains to be determined. We show that ℓ_1-regularized logistic regression applied to a high-dimensional vector of state variables can accurately differentiate between strong and weak systems and correctly classify inactive data. Future work includes testing on larger datasets and testing more complex statistical methods, such as deep neural nets.

ACKNOWLEDGMENTS

This work was conducted while AF was supported by a UC Berkeley Graduate Diversity Fellowship and an employee of Lawrence Berkeley National Lab.

REFERENCES

[1] K. Emanuel and D. Nolan, "Tropical cyclone activity and the global climate system," in *26th Conference on Hurricanes and Tropical Meteorology*, pp. 240–241, 2004.

[2] K. A. Emanuel, "The maximum intensity of hurricanes," *Journal of the Atmospheric Sciences*, vol. 45, no. 7, pp. 1143–1155, 1988.

[3] M. DeMaria, J. A. Knaff, and B. H. Connell, "A tropical cyclone genesis parameter for the tropical atlantic," *Weather and Forecasting*, vol. 16, no. 2, pp. 219–233, 2001.

[4] A. B. Schumacher, M. DeMaria, and J. A. Knaff, "Objective estimation of 24-h probability of tropical cyclone formation," *Weather and Forecasting*, vol. 24, pp. 456–471, 2009.

[5] J. H. Cossuth, R. D. Knabb, D. P. Brown, and R. E. Hart, "Tropical cyclone formation guidance using pregenesis dvorak climatology. part i: Operational forecasting and predictive potential," *Weather and Forecasting*, vol. 28, no. 1, pp. 100–118, 2013.

[6] W. Zhang, B. Fu, M. S. Peng, and T. Li, "Discriminating developing versus nondeveloping tropical disturbances in the western north pacific through decision tree analysis," *Weather and Forecasting*, vol. 30, no. 2, pp. 446–454, 2015.

[7] J. D. Hall, A. J. Matthews, and D. J. Karoly, "The modulation of tropical cyclone activity in the australian region by the maddenjulian oscillation," *Monthly Weather Review*, vol. 129, no. 12, pp. 2970–2982, 2001.

[8] D. J. Halperin, H. E. Fuelberg, R. E. Hart, J. H. Cossuth, P. Sura, and R. J. Pasch, "An evaluation of tropical cyclone genesis forecasts from global numerical models," *Weather and Forecasting*, vol. 28, no. 6, pp. 1423–1445, 2013.

[9] D. J. Halperin, R. E. Hart, H. E. Fuelberg, and J. H. Cossuth, "The development and evaluation of a statisticaldynamical tropical cyclone genesis guidance tool," *Weather and Forecasting*, vol. 32, no. 1, pp. 27–46, 2017.

[10] "Teca: A parallel toolkit for extreme climate analysis," *Procedia Computer Science*, vol. 9, pp. 866 – 876, 2012. Proceedings of the International Conference on Computational Science, ICCS 2012.

[11] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, 2016.