1. Climate Informatics

Claire Monteleoni, Department of Computer Science, George Washington University

Gavin A. Schmidt, NASA Goddard Institute for Space Studies

Francis Alexander, Los Alamos National Laboratory

Alexandru Niculescu-Mizil, NEC Laboratories America

Karsten Steinhaeuser, Department of Computer Science & Engineering, University of Minnesota

Michael Tippett, International Research Institute for Climate and Society, Earth Institute, Columbia University

Arindam Banerjee, Department of Computer Science & Engineering and Institute on the Environment, University of Minnesota, Twin Cities

M. Benno Blumenthal, International Research Institute for Climate and Society, Earth Institute, Columbia University

Auroop R. Ganguly, Civil and Environmental Engineering, Northeastern University

Jason E. Smerdon, Lamont-Doherty Earth Observatory of Columbia University

Marco Tedesco, The City College of New York – CUNY and The Graduate Center of the City University of New York

1.1 Introduction

The impact of present and potential future climate change will be one of the most important scientific and societal challenges in the 21st Century. Given observed changes in temperature, sea ice, and sea level, improving our understanding of the climate system is an international priority. This system is characterized by complex phenomena that are imperfectly observed and

even more imperfectly simulated. But with an ever-growing supply of climate data from satellites and environmental sensors, the magnitude of data and climate model output is beginning to overwhelm the relatively simple tools currently used to analyze them. A computational approach will therefore be indispensable for these analysis challenges. This chapter introduces the fledgling research discipline, *Climate Informatics*: collaborations between climate scientists and machine learning researchers in order to bridge this gap between data and understanding. We hope that the study of climate informatics will accelerate discovery in answering pressing questions in climate science.

Machine learning is an active research area at the interface of computer science and statistics, concerned with developing automated techniques, or algorithms, to detect patterns in data. Machine learning (and data mining) algorithms are critical to a range of technologies including web search, recommendation systems, personalized internet advertising, computer vision, and natural language processing. Machine learning has also made significant impacts on the natural sciences, for example Biology; the interdisciplinary field of Bioinformatics has facilitated many discoveries in genomics and proteomics. The impact of machine learning on climate science promises to be similarly profound.

The goal of this chapter is to define Climate Informatics and to propose some grand challenges for this nascent field. Recent progress on Climate Informatics, by the authors as well as by other groups, reveals that collaborations with climate scientists also open interesting new problems for machine learning. There are a myriad of collaborations possible at the intersection of these two fields. This chapter uses both top-down and bottom-up approaches to stimulate research progress on a range of problems in climate informatics, some of which have yet to be proposed. For the former, we present challenge problems posed by climate scientists, and discussed with machine

learning, data mining, and statistics researchers at Climate Informatics 2011, the First International Workshop on Climate Informatics, the inaugural event of a new annual workshop in which all coauthors participated. To spur innovation from the bottom-up, we also describe and discuss some of the types of data available. In addition to summarizing some of the key challenges for climate informatics, this chapter also draws on some of the recent climate informatics research of the coauthors.

The chapter is organized as follows. First we discuss the types of climate data available, and outline some challenges for Climate Informatics, including problems in analyzing climate data. Then we go into further detail on several key climate informatics problems: seasonal climate forecasting, predicting climate extremes, reconstructing past climate, and some problems in polar regions. We then discuss some machine learning and statistical approaches that might prove promising (that were not mentioned in previous sections). Finally we discuss some challenges and opportunities for climate science data and data management. Due to the broad coverage of the chapter, related work discussions are interspersed throughout the sections.

1.2 Machine Learning

Over the past few decades, the field of Machine Learning has matured significantly, drawing ideas from several disciplines including Optimization, Statistics, and Artificial Intelligence [4][34]. Application of Machine Learning has led to important advances in a wide variety of domains ranging from internet applications to scientific problems. Machine Learning methods have been developed for a wide variety of predictive modeling as well as exploratory data analysis problems. In the context of predictive modeling, important advances have been made in linear classification and regression, hierarchical linear models, nonlinear models based on

kernels, as well as ensemble methods that combine outputs from different predictors. In the context of exploratory data analysis, advances have been made in clustering and dimensionality reduction, including nonlinear methods to detect low-dimensional manifold structures in the data. Some of the important themes driving research in modern machine learning are motivated by properties of modern datasets from scientific, societal, and commercial applications. In particular, the datasets are extremely large scale, running into millions or billions of data points, are high-dimensional going up to tens of thousands or more dimensions, and have intricate statistical dependencies which violate the 'independent and identically distributed' assumption made in traditional approaches. Such properties are readily observed in climate datasets, including observations, reanalysis, as well as climate model outputs. These aspects have led to increased emphasis in scalable optimization methods [94], online learning methods [11], and graphical models [47], which can handle large scale data in high dimensions with statistical dependencies.

1.3 Understanding and Using Climate Data

A profuse amount of climate data of various types is available, providing a rich and fertile playground for future data mining and machine learning research. Here we discuss some of the varieties of data available, and provide some suggestions on how they can be used. This discussion itself will open some interesting problems. There are multiple sources of climate data, ranging from single site observations scattered in an unstructured way across the globe, to climate model output that is global and uniformly gridded. Each class of data has particular characteristics that need to be appreciated before it can be successfully used or compared. We provide here a brief introduction to each, with a few examples and references for further information. Common issues that arise in cross-class syntheses are also addressed.

1.3.1 In-situ Observations

In-situ measurements refer to raw (or only minimally processed) measurements of diverse climate system properties that can include temperatures, rainfall, winds, column ozone, cloud cover, radiation etc., taken from specific locations. These locations are often at the surface (for instance, from weather stations), but can also include atmospheric measurements from radiosonde balloons, sub-surface ocean data from floats, data from ships, aircraft, and special intensive observing platforms.

Much of this data is routinely collected and is available in collated form from National Weather Services or special projects such as AEROCOM (for aerosol data), ICOADS (ocean temperature and salinity from ships), Argo (ocean floats), etc. Multivariate data related to single experiments (for instance the Atmospheric Radiation Measurement (ARM) program or the Surface Heat Budget of the Arctic (SHEBA), are a little less well organized, though usually available at specialized websites.

This kind of data is useful for looking at coherent multivariate comparisons, though usually on limited time and space domains, as input to weather model analyses or as the raw material for processed gridded data (see next subsection). The principal problem with this data is their sparseness spatially and in time, inhomogeneities due to differing measurement practices or instruments and overall incompleteness (not all variables are measured at the same time or place) (for instance, see [45][62]).

1.3.2 Gridded/Processed Observations

Given a network of raw in-situ data, the next step is synthesizing those networks into quality-controlled regularly gridded datasets. These have a number of advantages over the raw data in that they are easier to work with, are more comparable to model output (discussed below) and

have less non-climatic artifacts. Gridded products are usually available on 5° latitude by 5° longitude grids or even higher resolution. However, these products use interpolation, gap-filling in space and time, and corrections for known biases all of which affect the structural uncertainty in the product. The resulting error estimates are often dependent upon space and time. Different products targeting the same basic quantity can give some idea of the structural uncertainty in these products and we strongly recommend using multiple versions. For instance, for different estimates of the global mean surface temperature anomalies can be found from the National Climatic Data Center (NCDC), the Hadley Centre, and NASA [6][33][90] that differ in processing and details but show a large amount of agreement at the large scale.

1.3.3 Satellite Retrievals

Since 1979, global and near global observations of the Earth's climate have been made from low-earth orbit and geostationary satellites. These observations are based either on passive radiances (either emitted directly from the Earth, or via reflected solar radiation), or by active scanning via lasers or radars. These satellites, mainly operated by US agencies (NOAA, NASA), the European Space Agency and the Japanese program (JAXA) and data are generally available in near-real time. There are a number of levels of data ranging from raw radiances (Level 1), processed data as a function of time (Level 2), and gridded averaged data at the global scale (Level 3).

Satellite products do have specific and particular views of the climate system, which requires that knowledge of the 'satellite-eye' view be incorporated into any comparison of satellite data with other products. Many satellite products are available for specific instruments on specific platforms, synthesis products across multiple instruments and multiple platforms are possible, but remain rare.

1.3.4 Paleo-climate proxies

In-situ instrumental data only extends on a global basis to the mid-19th Century, though individual records can extend to the 17th or 18th Century. For a longer term perspective, climate information must be extracted from so-called 'proxy' archives, such as ice cores, ocean mud, lake sediments, tree rings, pollen records, caves, or corals, which retain information that is sometimes highly correlated to specific climate variables or events [41].

As with satellite data, appropriate comparisons often require a forward model of the process by which climate information is stored that incorporates the multiple variables that influence any particular proxy (e.g. [75]). However, the often dramatically larger signals that can be found in past climate can overcome the increase in uncertainty due to spatial sparseness and non-climatic noise, especially when combined in a multi-proxy approach [58]. Problems in paleo-climate will be discussed further detail in Section 1.8.

1.3.5 Re-analysis products

Weather forecast models use as much observational data (in-situ, remote sensing etc.) as can be assimilated in producing 6 hour forecasts (the ‘analyses’) which are excellent estimates of the state of the climate at any one time. However, as models have improved over time, the time series of weather forecasts can contain trends related only to the change in model rather than changes in the real world. Thus, many of the weather forecasting groups have undertaken ‘re-analyses’ that use a fixed model to re-process data from the past in order to have a consistent view of the real world (see reanalyses.org for more details). This is somewhat equivalent to a physics-based interpolation of existing data sets and often provides the best estimate of the climate state over the instrumental period (for instance, ERAInterim [16]).

However, not all variables in the re-analyses are equally constrained by observational data. Thus

sea level pressure and winds are well characterized, but precipitation, cloud fields or surface fluxes are far more model dependent and thus are not as reliable. Additionally, there remain unphysical trends in the output as a function of changes in the observing network over time. In particular, the onset of large scale remote sensing in 1979 imparts jumps in many fields that can be confused with real climate trends (e.g. [105]).

1.3.6 Global Climate model (GCM) output

Global climate models are physics-based simulations of the climate system, incorporating (optionally) components for the atmosphere, ocean, sea ice, land surface, vegetation, ice sheets, atmospheric aerosols and chemistry and carbon cycles. Simulations can either be transient in response to changing boundary conditions (such as hindcasts of the 20th Century), or time-slices for periods thought to be relatively stable (such as the mid-Holocene 6000 years ago). Variations in output can depend on initial conditions (because of the chaotic nature of the weather), the model used, variations in the forcing fields (due to uncertainties in the time history, say, of aerosol emissions). A number of coordinated programs, notably the Coupled Model Intercomparison Project (CMIP), have organized coherent model experiments that have been followed by multiple climate modeling groups across the world and which are the dominant source for model output (e.g. [96]).

These models are used to define fingerprints of forced climate change that can be used in the detection and attribution of climate change [39], for hypothesis generation about linkages in the climate system, as testbeds for evaluating proposed real world analyses [24], and, of course, future predictions [61]. Quantifying the structural uncertainty in model parameterizations or the model framework, the impact of known imperfections in the realizations of key processes, and the necessity of compromises at small spatial or temporal scales are all important challenges.

1.3.7 Regional Climate model (RCM) output

Global models necessarily need to compromise on horizontal resolution. In order to incorporate more details at the local level (particularly regional topography), output from the global models or the global reanalyses can be used to drive a higher resolution, regional climate model. The large-scale fields can then be transformed to higher resolution using physical principles embedded in the RCM code. In particular, rainfall patterns that are very sensitive to the detailed topography, are often far better modeled within the RCM than in the global-scale driving model. However, there are many variables to consider in RCMs - from variations in how the boundary field forcing is implemented and in the physics packages - and the utility of using RCMs to improve predictions of change is not yet clear. A coordinated experiment to test these issues is the North American Regional Climate Change Assessment Program (NARCCAP) [60].

1.4 Scientific Problems in Climate Informatics

There are a number of different kinds of problems that climate scientists are working on where machine learning and computer science techniques may make a big impact. This is a brief description of a few examples (with discussion of related work in the literature) that typify these ideas, though any specific implementation mentioned should not be considered the last word. This section provides short descriptions of several challenge problems in Climate Informatics broadly defined. In Section 1.5 we will present problems in climate data analysis. In subsequent sections we will go into more detail on some specific problems in Climate Informatics.

1.4.1 Parameterization Development

Climate models need to deal with physics that occurs at scales smaller than any finite model can resolve. This can involve cloud formation, turbulence in the ocean, land surface heterogeneity,

ice floe interactions, chemistry on dust particle surfaces, etc. This is dealt with by using parameterizations that attempt to capture the phenomenology of a specific process and its sensitivity in terms of the (resolved) large scales. This is an ongoing task, and is currently driven mainly by the scientists' physical intuition and relatively limited calibration data. As observational data become more available, and direct numerical simulation of key processes becomes more tractable, there is an increase in the potential for machine learning and data mining techniques to help define new parameterizations and frameworks. For example, neural net frameworks have been used to develop radiation models [50].

1.4.2 Using Multi-Model Ensembles of Climate Projections

There are multiple climate models that have been developed and are actively being improved at ~25 centers across the globe. Each model shares some basic features with at least some other models, but each has generally been designed an implemented independently and has many unique aspects. In coordinated "Model Intercomparison Projects" (MIPs) (most usefully, the Coupled MIP (CMIP3, CMIP5), the Atmospheric Chemistry and Climate MIP (ACCMIP), the PaleoClimate MIP (PMIP3) etc.), modeling groups have attempted to perform analogous simulations with similar boundary conditions but with multiple models. These 'ensembles' offer the possibility of assessing what is robust across models, what are the roles of internal variability, structural uncertainty, and scenario uncertainty in assessing the different projections at different time and space scales, and multiple opportunities for model-observation comparisons. Do there exist skill metrics for model simulations of the present and past that are informative for future projections? Are there weighting strategies that maximize predictive skill? How would one explore this? These are questions that also come up in weather forecasts, or seasonal forecasts, but are made more difficult for the climate problem because of the long time

scales involved [40][97]. Some recent work has applied machine learning to this problem with encouraging results [63].

1.4.3 Paleo-reconstructions

Understanding how climate varied in the past before the onset of widespread instrumentation is of great interest - not least because the climate changes seen in the paleo-record dwarf those seen in the 20th Century and hence may provide much insight into the significant changes expected this century. Paleo-data is however even sparser than instrumental data, and moreover is not usually directly commensurate with the instrumental record. As mentioned in Section 1.3, paleo-proxies (such as water isotopes, tree rings, pollen counts, etc.) are indicators of climate change but often have non-climatic influences or their behavior, or whose relation to what would be considered more standard variables (such as temperature or precipitation) is perhaps non-stationary or convolved. There is an enormous challenge to bringing together disparate, multi-proxy evidence to produce large scale patterns of climate change [59], or from the other direction build in enough "forward modeling" capability into the models to use the proxies directly as modeling targets [76]. This topic will be discussed in further detail in Section 1.8.

1.4.4 Data Assimilation and Initialized Decadal predictions

The main way in which sparse observational data is used to construct complete fields is through data assimilation. This is a staple of weather forecasts and the various reanalyses in the atmosphere and ocean. In many ways, this is the most sophisticated use of the combination of models and observations, but its use in improving *climate* predictions is still in its infancy. For weather timescales this works well, but for longer term forecasts (seasons to decades) the key variables are in the ocean, not the atmosphere, and initializing a climate model so that the evolution of ocean variability models the real world in useful ways is very much a work in

progress [44][90]. First results have been intriguing, if not convincing, and many more examples are slated to come on line in the new CMIP5 archive [61].

1.4.5 Developing and Understanding Perturbed Physics Ensembles (PPE)

One measure of structural uncertainty in models is the spread among the different models from different modeling groups. But these models cannot be considered to be a random sample from the space of all possible models. Another approach is to take a single model, and within the code vary multiple (uncertain) parameters in order to generate a family of similar models that nonetheless sample a good deal of the intrinsic uncertainty that arises in choosing any specific set of parameter values. These ``Perturbed Physics Ensembles'' (PPEs) have been used successfully in the climateprediction.net and QUMP projects to generate controlled model ensembles that can be compared systematically to observed data and make inferences [46][64]. However, designing such experiments and efficiently analyzing sometimes 1000's of simulations is a challenge, but one which is increasingly going to be attempted.

1.5 Climate Data Analysis: Problems and Approaches

Here we discuss some additional challenge problems in analyzing climate data. The rate of data acquisition via the satellite network and the re-analyses projects is very rapid. Similarly, the amount of model output is equally fast growing. Model-observation comparisons based on processes (i.e., the multivariate changes that occur in a single event (or collection of events), such as a N. Atlantic storm, an ocean eddy, and ice floe melting event, a hurricane, a jet stream excursion, a stratospheric sudden warming etc.) have the potential to provide very useful information on model credibility, physics and new directions for parameterization improvements. However, data services usually deliver data in single variable, spatially fixed, time varying

formats that make it very onerous to apply space and time filters to the collection of data to extract generic instances of the process in question. As a first step, algorithms for clustering data streams will be critical for clustering and detecting the patterns listed. There will also be the need to collaborate with systems and database researchers on the data challenges mentioned here and in Section 1.11. Here we present several other problems to which cutting-edge data analysis and machine learning techniques are poised to contribute.

1.5.1 Abrupt Changes

Earth system processes form a non-linear dynamical system and, as a result, changes in climate patterns can be abrupt at times [74]. Moreover, there is some evidence, particularly in glacial conditions, that climate tends to remain in relatively stable states for some period of time, interrupted by sporadic transitions (perhaps associated with so-called *tipping points*) which delineate different climate regimes. Understanding the causes behind significant abrupt changes in climate patterns can provide a deeper understanding of the complex interactions between earth system processes. The first step towards realizing this goal is to have the ability to detect and identify abrupt changes from climate data.

Machine learning methods for detecting abrupt changes, such as extensive droughts which last for multiple years over a large region, should have the ability to detect changes with spatial and temporal persistence, and should be scalable to large datasets. Such methods should be able to detect well known droughts like the Sahel drought in Africa, the 1930s Dust Bowls in the United States, and droughts with similar characteristics where the climatic conditions were radically changed for a period of time for an extended region [23][37][78][113]. A simple approach for detecting droughts is to apply a suitable threshold to a pertinent climate variable, such as precipitation or soil moisture content, and label low precipitation regions as droughts. While

such an approach will detect major events like the Sahel drought and dust bowls, it will also detect isolated events, such as low precipitation in one month for a single location that is clearly not an abrupt change event. Thus, the number of "false positives" from such a simple approach would be high, making subsequent study of each detected event difficult.

In order to identify drought regions that are spatially and temporally persistent, one can consider a discrete graphical model that ensures spatiotemporal smoothness of identified regions. Consider a discrete Markov Random Field (MRF) with a node corresponding to each location at each time step and a meaningful neighborhood structure which determines the edges in the underlying graph *G = (V,E)* [111]. Each node can be in one of two states: 'normal' or 'drought'. The maximum a posteriori (MAP) inference problem in the MRF can be posed as:

$$x^* = \underset{x \in \{0,1\}^N}{argmax} \left\{ \sum_{u \in V} \theta_u(x_u) + \sum_{(u,v) \in E} \theta_{uv}(x_u, x_v) \right\}$$

where θ_u, θ_{uv} are node-wise and edge-wise potential functions which respectively encourage agreement with actual observations and agreement amongst neighbors, and x_u is the state, i.e., 'normal' or 'drought', at node $u \in V$. The MAP inference problem is an integer programming problem often solved using a suitable linear programming (LP) relaxation [70][111].

Figure **1** shows results on drought detection over the past century based on the MAP inference method. For the analysis, the Climatic Research Unit's (CRU) precipitation dataset was used at 0.5°×0.5° latitude-longitude spatial resolution from 1901-2006. The LP involved around 7 million variables and was solved using efficient optimization techniques. The method detected almost all well-known droughts over the past century. More generally, such a method can be used to detect and study abrupt changes for a variety of settings, including heat waves, droughts, precipitation, and vegetation. The analysis can be performed on observed data, reanalysis data, as

well as model outputs as appropriate.

1.5.2 Climate Networks

Identifying dependencies between various climate variables and climate processes form a key part of understanding the global climate system. Such dependencies can be represented as climate networks [19][20][106][107], where relevant variables or processes are represented as nodes and dependencies are captured as edges between them. Climate networks are a rich representation for the complex processes underlying the global climate system, and can be used to understand and explain observed phenomena [95][108].

A key challenge in the context of climate networks is to construct such networks from observed climate variables. From a statistical machine learning perspective, the climate network should reflect suitable dependencies captured by the joint distribution of the variables involved. Existing methods usually focus on a suitable measure derived from the joint distribution, such as the covariance or the mutual information. From a sample-based estimate of the pairwise covariance or mutual information matrix, one obtains the climate network by suitably thresholding the estimated matrix. Such approaches have already shown great promise, often identifying some key dependencies in the global climate system [43] (Figure **2**).

Going forward, there are a number of other computational and algorithmic challenges that must be addressed to achieve more accurate representations of the global climate system. For instance, current network construction methods do not account for the possibility of time-lagged correlations, yet we know that such relationships exist. Similarly, temporal autocorrelations and signals with varying amplitudes and phases are not explicitly handled. There is also a need for better balancing of the dominating signal of spatial autocorrelation with that of possible teleconnections (long-range dependencies across regions), which are often of high interest. In

addition, there are many other processes that are well-known and documented in the climate science literature, and network representations should be able to incorporate this *a priori* knowledge in a systematic manner. One of the initial motivations and advantages of these network-based approaches is their interpretability, and it will be critical that this property be retained as these various aspects are integrated into increasingly complex models and analysis methods.

1.5.3 Predictive Modeling: Mean Processes and Extremes

Predictive modeling of observed climatic phenomena can help in understanding key factors affecting a certain observed behavior. While the usual goal of predictive modeling is to achieve high accuracy for the response variable, say, the temperature or precipitation at a given location, in the context of climate data analysis, identifying the covariates having the most significant influence on the response is often more important. Thus, in addition to getting high predictive accuracy, feature selection will be a key focus of predictive modeling. Further, one needs to differentiate between mean processes and extremes, which are rather different regimes for the response variable. In practice, different covariates may be influencing the response variable under different regimes and timescales.

In recent literature, important advances have been made in doing feature selection in the context of high-dimensional regression [66][101]. For concreteness, consider the problem of predicting the mean temperature in Brazil based on multiple ocean variables over all ocean locations. While the number of covariates p runs into tens of thousands, the number of samples n based on monthly means over a few decades are a few hundred to a few thousand. Standard regression theory does not extend to this $n \ll p$ scenario. Since the ocean variables at a particular location are naturally grouped, only a few such locations are relevant for the prediction, and only a few

variables in each such location are relevant, one can pose the regression problem as a sparse group lasso problem [25][24]:

$$\min_{\theta \in \mathbb{R}^{Nm}} \left\{ \|y - X\theta\|^2 + \lambda_1 \|\theta\|_1 + \lambda_2 \sum_{g=1}^{N} \|\theta_g\|_2 \right\}$$

where N is the number of ocean locations, m is the number of ocean variables in each location so that $p = Nm$, θ is the weight vector over all covariates to be estimated, θ_g is the set of weights over variables at location g, and λ_1, λ_2 are non-negative constants. The sparse group lasso regularizer ensures that only few locations get non-zero weights, and even among these locations only a few variables are selected. Figure **3** shows the locations and features that were consistently selected for the task of temperature prediction in Brazil.

1.6 Seasonal Climate Forecasting

Seasonal climate forecasts are those beyond the timeframe of standard weather forescasts (say 2 weeks) out to a season or two ahead (up to 6 months). Fundamental questions concern what is (and is not) predictable and exactly how predictable it is. Addressing these questions often gives a good indication of how to make a prediction in practice, too. These are hard questions because much in the climate system is unpredictable, and the observational record is short. Methods from data mining and machine learning applied to observations and data from numerical climate prediction models provide promising approaches. Key issues including finding components of the climate state-space that are predictable, and constructing useful associations between observations and corresponding predictions from numerical models.

1.6.1 What is the basis for seasonal forecasting?

The chaotic nature of the atmosphere and the associated sensitivity of numerical weather

forecasts to their initial conditions is described by the well-known "butterfly effect" – that the flap of a butterfly's wings in Brazil could set off a tornado in Texas. Small errors in the initial state of a numerical weather forecast quickly amplify until the forecast has no value. This sensitive dependence on initial conditions provides an explanation for the limited time horizon (a few days to a week) for which useful weather forecasts can be issued, and the belief until the early 1980s that seasonal forecasting was impossible [81]. This also explains why effort is needed to find the needle of predictability in the haystack of chaos. Given the limited predictability of weather, how is it that quantities such as precipitation and near-surface temperature are skillfully forecast seasons (3 – 6 months) in advance?

First, it should be noted that the format of climate predictions is different from that of weather forecasts. Weather forecasts target the meteorological conditions of a particular day or hour. Climate predictions are made in terms of weather statistics over some time range. For instance, the most common quantities in current climate forecasts are 3-month (seasonal) averages of precipitation and near-surface temperature. Two fundamental facts about the earth system make climate forecasts possible. First, the oceans evolve on time-scales that are generally slower than those of the atmosphere, and some ocean structures are predictable several months in advance. The outstanding predictable ocean structure is associated with the El Niño–Southern Oscillation (ENSO) and is manifest in the form of widespread, persistent departures (anomalies) of equatorial Pacific sea surface temperature (SST) from its seasonally adjusted long-term value. The first ENSO forecasts were made in the late 1980s [10]. The second fact is that some components of the atmosphere respond to persistent SST anomalies. The atmospheric response to SST on any given day tends to be small relative to the usual weather variability. However, since the SST forcing and the associated atmospheric response may persist for months or

seasons, the response of a seasonal average to SST forcing may be significant [82]. For instance, ENSO has impacts on temperature, precipitation, tropical cyclones, human health and perhaps even conflict [31][38][49][72]. Early seasonal forecasts constructed using canonical correlation analysis (CCA) between antecedent SST and climate responses [3] took advantage of this persistence of SST. Such statistical (or empirical, in the sense of not including explicit fundamental physical laws) remain attractive because of their generally low dimensional and cost relative to physical process based models (typically general circulation models; GCMs) with many millions of degrees of freedom.

1.6.2 Data Challenges

Here we introduce some challenges posed by the available data. Data challenges will be further discussed in Section 1.11. Serious constraints come from the dimensions of the available data. Reliable climate observations often do not extend more than 40 or 50 years into the past. This means that, for instance, there may be only 40 or 50 available observations of January-March average precipitation. Moreover the quality and completeness of that data may vary in time and space. Climate forecasts from GCMs often do not even cover this limited period. Many seasonal climate forecast systems start hindcasts in the early 1980s when satellite observations, particularly of SST, became available. In contrast to the sample size, the dimension of the GCM state space may be of the order 10^6 depending on spatial grid resolution. Dimension reduction, (principal component analysis (PCA) is commonly used), is necessary before applying classical methods like canonical correlation analysis (CCA) to find associated features in predictions and observations [5]. There has been some use of more sophisticated dimensionality reduction methods in seasonal climate prediction problems [53]. Methods that can handle large state-spaces and small sample size are needed. An intriguing recent approach that avoids the problem

of small sample size is to estimate statistical models using long climate simulations unconstrained by observations and test the resulting model on observations [18][115]. This approach has the challenge of selecting GCMs whose climate variability is “realistic”, which is a remarkably difficult problem given the observational record.

1.6.3 Identifying predictable quantities

The initial success of climate forecasting has been in the prediction of seasonal averages of quantities such as precipitation and near-surface temperature. In this case, time averaging serves as a filter with which to find predictable signals. A spatial average of SST in a region of the equatorial Pacific is used to define the NINO3.4 index which is used in ENSO forecasts and observational analysis. This spatial average serves to enhance the large-scale predictable ENSO signal by reducing noise. The Madden-Julian Oscillation (MJO) is a sub-seasonal component of climate variability which is detected using time and space filtering. There has been some work on constructing spatial filters that designed to optimize measures of predictability [17] and there are opportunities for new methods that incorporate optimal time and space filtering and that optimize more general measures of predictability.

While predicting the weather of an individual day is not possible in a seasonal forecast, it may be possible to forecast statistics of weather such as the frequency of dry days or the frequency of consecutive dry days. These quantities are often more important to agriculture than seasonal totals. Drought has a complex time-space structure that depends on multiple meteorological variables. Data Mining and Machine Learning (DM/ML) methods can be applied to observations and forecasts to identify drought, as was discussed in Section 1.5.

Identification of previously unknown predictable climate features may benefit from the use of

DM/ML methods. Cluster analysis of tropical cyclone tracks has been used to identify features that are associated with ENSO and MJO variability [9]. Graphical models, the non-homogeneous hidden Markov model in particular, have been used to obtain stochastic daily sequences of rainfall conditioned on GCM seasonal forecasts [32].

The time and space resolution of GCMs forecasts limits the physical phenomena they can resolve. However, they may be able to predict proxies or large-scale antecedents of relevant phenomena. For instance, GCMs that do not resolve tropical cyclones (TCs) completely do form TC-like structures that can be used to make TC seasonal forecasts [8][110]. Identifying and associating GCMs "proxies" with observed phenomena is also a DM/ML problem.

Regression methods are used to connect climate quantities to associated variables that are either unresolved by GCMs or not even climate variables. For instance, Poisson regression is used to related large-scale climate quantities with hurricanes [104], and generalized additive models are used to relate heat waves with increased mortality [68]. Again, the length of the observational record makes this challenging.

1.6.4 Making the best use of GCM data

Data from multiple GCM climate forecasts are routinely available. However, converting that data into a useful forecast product is a nontrivial task. GCMs have systematic errors that can be identified (and potentially corrected) through regression-like procedures with observations. Robust estimates of uncertainty are needed to construct probabilistic forecasts. Since forecasts are available from multiple GCMs, another question is how best to combine information from multiple sources given the relatively short observation records with which to estimate model performance.

1.7 Climate Extremes, Uncertainty and Impacts

1.7.1 The Climate Change Challenge

The Fourth Assessment Report of the Intergovernmental Panel on Climate Change (IPCC, AR4) has resulted in a wider acceptance of global climate change caused by anthropogenic drivers of emission scenarios. However, earth system modelers struggle to develop precise predictions of extreme events (e.g., heat waves, cold spells, extreme rainfall events, droughts, hurricanes and tropical storms) or extreme stresses (e.g., tropical climate in temperate regions or shifting rainfall patterns) at regional and decadal scales. In addition, the most significant knowledge gap relevant for policymakers and stakeholders remains the inability to produce credible estimates of local to regional scale climate extremes and change impacts. Uncertainties in process studies, climate models, and associated spatiotemporal downscaling strategies, may be assessed and reduced by statistical evaluations. But a similar treatment for extreme hydrological and meteorological events may require novel statistical approaches and improved downscaling. Scenario uncertainty for climate change impacts is fundamentally intractable, but other sources of uncertainty may be amenable to reduction. Regional impacts need to account for additional uncertainties in the estimates of anticipatory risks and damages, whether on the environment, infrastructures, economy or society. The cascading uncertainties from scenarios, to models, to downscaling, and finally to impacts, make costly decisions difficult to assess. This problem grows acute if credible attributions need to be made to causal drivers or policy impacts.

1.7.2 The Science of Climate Extremes

One goal is to develop quantifiable insights on the impacts of global climate change on weather or hydrological extreme stresses and extreme events at regional to decadal scales. Precise and local predictions, for example the likelihood of an extreme event on a given day of any given

year a decade later, will never be achievable owing to the chaotic nature of the climate system as well as the limits to precision of measurements and our inability to model all aspects of the process physics. However, probability density functions of the weather and hydrology, for example, likelihoods of intensity-duration-frequency (IDF) of extreme events or of mean change leading to extreme stress, may be achievable targets. The tools of choice range from the two traditional pillars of science: theory (e.g., advances in physical understanding and high-resolution process models of atmospheric or oceanic climate, weather or hydrology) to experimentation (e.g., development of remote and in-situ sensor systems as well as related cyber-infrastructures to monitor the earth and environmental systems). However, perhaps the most significant breakthroughs are expected from the relatively new pillars: computational sciences and informatics. Research in the computational sciences for climate extremes science include the computational data sciences (e.g., high-performance analytics based on extreme value theory and nonlinear data sciences to develop predictive insights based on a combination of observations and climate model simulations) and computational modeling (e.g., regional scale climate models, models of hydrology, improvements in high-resolution processes within general circulation models, as well as feedback to model development based on comparisons of simulations with observations), while the informatics aspects include data management and discovery (e.g., development of methodologies for geographic data integration and management, knowledge discovery from sensor data and geospatial-temporal uncertainty quantification).

1.7.3 The Science of Climate Impacts

The study of climate extremes is inextricably linked to the study of impacts, including risks and damage assessments as well as adaptation and mitigation strategies. Thus, an abnormally hot summer or high occurrence of hurricanes in unpopulated or remote regions of the world, which

do not otherwise affect resources or infrastructures, have little or no climate impact on society. On the other hand, extreme events like the after-effects of hurricane Katrina have extreme impacts owing to complex interactions among multiple effects: a large hurricane hitting an urban area, an already vulnerable levee breaking down because of the flood waters, as well as an impacted society and response systems which are neither robust nor resilient to shocks. In general, climate change mitigation (e.g., emission policies and regulations to possible weather modification and geo-engineering strategies) and adaptation (e.g., hazards and disaster preparedness, early warning and humanitarian assistance or the management of natural water, nutritional and other resources, as well as possible migration and changes in regional population growth or demographics), need to be based on actionable predictive insights which consider the interaction of climate extremes science with the critical infrastructures and key resources, population and society. While the science of impacts can be challenging and relatively difficult to quantify, given recent advances in machine learning, geospatial modeling, data fusion, and Geographic Information Systems (GIS), this is a fertile area for progress on Climate Informatics.

1.8 Reconstructing Past Climate

The most comprehensive observations of Earth's climate span only the last one to two hundred years [105]. This time period includes the establishment of long-term and widespread meteorological stations across the continental landmasses (e.g. ref. [6]), ocean observing networks from ships and buoys (e.g. ref. [114]), and, within the more recent past, remote sensing from satellites (e.g. ref. [109]). Much of our understanding about the climate system and contemporary climate change comes from these and related observations and their fundamental role in evaluating theories and models of the climate system. Despite the valuable collection of

modern observations, however, two factors limit their use as a description of the Earth's climate and its variability: 1) relative to known timescales of climate variability, they span a brief period of time; and 2) much of the modern observational interval is during an emergent and anomalous climate response to anthropogenic emissions of greenhouse gases [36]. Both of these factors limit assessments of climate variability on multi-decadal and longer timescales, or characterizations of climatic mean states under different forcing[1] scenarios (e.g. orbital configurations or greenhouse gas concentrations). Efforts to estimate climate variability and mean states prior to the instrumental period are thus necessary to fully characterize how the climate can change and how it might evolve in the future in response to increasing greenhouse gas emissions.

Paleoclimatology is the study of Earth's climate history and offers estimates of climate variability and change over a range of timescales and climate regimes. Among the many time periods of relevance, the Common Era (CE; the last two millennia) is an important target because the abundance of high-resolution paleoclimatic proxies (e.g. tree rings, ice cores, cave deposits, corals, and lake sediments) over this time interval allows seasonal-to-annual reconstructions on regional-to-global spatial scales (see ref. [40] for a review). The CE also spans the rise and fall of many human civilizations, making paleoclimatic information during this time period important for understanding the complicated relationships between climate and organized societies [7][15]. Given the broad utility and vast number of proxy systems that are involved, the study of CE climate is a wide-ranging and diverse enterprise. The purpose of the following discussion is not

[1] A 'forcing' is a specific driver of climate change, external to the climate models - for instance changes in the composition of well-mixed greenhouse gases (like CO2 or CH4), changes in the land surface due to deforestation or urbanization, changes in air pollution, changes in the sun's input or the impact of large volcanic eruptions. Each forcing can be usefully characterized by the impact it has on the radiative balance at the top the atmosphere - positive forcings increase the energy coming into the climate system and hence warm the planet, while negative forcings cool the climate.

meant to survey this field as a whole, but instead to focus on a relatively recent pursuit in CE paleoclimatology that seeks to reconstruct global or hemispheric temperatures using syntheses of globally distributed multi-proxy networks. This particular problem is one that may lend itself well to new and emerging data analysis techniques, including machine learning and data mining methods. The motivation of the following discussion therefore is to outline the basic reconstruction problem and describe how methods are tested in synthetic experiments.

1.8.1 The Global Temperature Reconstruction Problem

It is common to separate global or hemispheric (large-scale) temperature reconstruction methods into two categories. The first involves index methods that target large-scale indices such as hemispheric mean temperatures [13] [35][51][58]; the second comprises climate field reconstruction (CFR) methods that target large-scale patterns, i.e. global maps of temperature change [21][55][56][59][88]. Although both of these approaches often share common methodological foundations, the following discussion will focus principally on the CFR problem.

Large-scale temperature CFRs rely on two primary data sets. The first is monthly or annual gridded (5° latitude × 5° longitude) temperature products that have near global coverage beginning in the mid-to-late 19th century. These gridded temperature fields have been derived from analyses of land and sea-based surface temperature measurements from meteorological stations, ship and buoy-based observing networks [6][42]. The second dataset comprises collections of multiple climate proxy archives [58], each of which have been independently analyzed to establish their sensitivity to some aspect of local or regional climate variability. These proxy records are distributed heterogeneously about the globe (Figure **4**), span variable periods of time, and each are subject to proxy-specific errors and uncertainties.

The basic premise of CFR techniques is that a relationship can be determined between observed

climate fields and multi-proxy networks during their common interval of overlap. Once defined, this relationship can be used to estimate the climate fields prior to their direct measurement using the multi-proxy network that extends further into the past. Figure **4** represents this concept schematically using a data matrix that casts the CFR formalism as a missing data problem. Note that this missing data approach was originally proposed for CFRs using regularized expectation maximization [77], and has since become a common method for reconstructions targeting the CE [56][57][59]. The time-by-space data matrix in Figure **4** is constructed first from the instrumental data, with rows corresponding to years and columns corresponding to the number of grid cells in the instrumental field. For a typical CFR targeting an annual and global 5°×5° temperature field, the time dimension is several centuries to multiple millennia and the space dimension is on the order of one to two thousand grid cells. The time dimension of the data matrix is determined by the length of the calibration interval during which time the temperature observations are available, plus the reconstruction interval that is determined by the length of available proxy records. The number of spatial locations may be less than the 2592 possible grid cells in a 5° global grid, and depends on the employed surface temperature analysis product. A reconstruction method may seek to infill grid cells that are missing temperature observations [103], or simply leave them missing depending on the number of years that they span [59]. The second part of the composite data matrix is formed from the multi-proxy network, the dimensions of which are determined by the longest proxy records and the total number of proxies (typically on the order of a few hundred to a thousand). The number of records in multi-proxy networks typically decreases back in time, and may reduce to a few tens of records in the earliest period of the reconstruction interval. The temporal resolution of the proxy series may also vary from seasonal to decadal.

Multiple methods have been used for CFRs, including a number of new and emerging techniques within Bayesian frameworks [52][103]. The vast majority of CFRs to date, however, have applied forms of regularized, multivariate linear regression, in which a linear regression operator is estimated during a period of overlap between the temperature and proxy matrices. Such linear regression approaches work best when the time dimension in the calibration interval (Figure **4**) is much larger than the spatial dimension, because the covariance between the temperature field and the proxies is more reliably estimated. The challenge for CFR methods involves the manner in which the linear regression operator is estimated in practical situations when this condition is not met. It is often the case in CFR applications that the number of target variables exceeds the time dimension, yielding a rank-deficient problem. The linear regression formalism therefore requires some form of regularization. Published linear methods for global temperature CFRs vary primarily in their adopted form of regularization (see [88] and [102] for general discussions on the methodological formalism). Matrix factorizations such as Singular Value Decomposition [29] of the temperature and proxy matrices are common first steps. If the squared singular values decrease quickly, as is often the case in climatological data where leading climate patterns dominate over many more weakly expressed local patterns or noise, reduced-rank representations of the temperature and proxy matrices are typically good approximations of the full-rank versions of the matrices. These reduced-rank temperature and proxy matrices therefore are used to estimate a linear regression operator during the calibration interval using various multivariate regression techniques. Depending on the method used, this regression operator may be further regularized based on analyses of the cross-covariance or correlation of the reduced temperature and proxy matrices. Multiple means of selecting rank reductions at each of these steps have been pursued, such as selection rules based on analyses of the singular value (or eigenvalue) spectrum

(e.g. ref [57]) or minimization of cross-validation statistics calculated for the full range of possible rank-reduction combinations (e.g. ref [88]).

1.8.2 Pseudoproxy Experiments

The literature is replete with discussions of the variously applied CFR methods and their performance (see ref. [29] for a cogent summary of many employed methods). Given this large number of proposed approaches, it has become important to establish means of comparing methods using common datasets. An emerging tool for such comparisons is millennium-length, forced transient simulations from coupled General Circulation Models (CGCMs) [1][30]. These model simulations have been used as synthetic climates in which to evaluate the performance of reconstruction methods in tests that have been termed pseudo-proxy experiments (see ref. [85] for a review). The motivation for pseudo-proxies experiments is to adopt a common framework that can be systematically altered and evaluated. They also provide a much longer, albeit synthetic, validation period than can be achieved with real-world data, and thus methodological evaluations can extend to lower frequencies and longer time scales. Although one must always be mindful of how the results translate into real-world implications, these design attributes allow researchers to test reconstruction techniques beyond what was previously possible and to compare multiple methods on common datasets.

The basic approach of a pseudo-proxy experiment is to extract a portion of a spatiotemporally complete CGCM field in a way that mimics the available proxy and instrumental data used in real-world reconstructions. The principal experimental steps proceed as follows: (1) pseudo-instrumental and pseudoproxy data are subsampled from the complete CGCM field from locations and over temporal periods that approximate their real-world data availability; (2) the time series that represent proxy information are added to noise series to simulate the temporal

(and in some cases spatial) noise characteristics that are present in real-world proxy networks; and (3) reconstruction algorithms are applied to the model-sampled pseudo-instrumental data and pseudoproxy network to produce a reconstruction of the climate simulated by the CGCM. The culminating fourth step is to compare the derived reconstruction to the known model target as a means of evaluating the skill of the applied method and the uncertainties expected to accompany a real-world reconstruction product. Multi-method comparisons can also be undertaken from this point.

Multiple datasets are publicly available for pseudoproxy experiments through supplemental websites of published papers [57][87][89][103]. The Paleoclimate Reconstruction Challenge is also a newly established online porthole through the Paleoclimatology Division of the National Oceanographic and Atmospheric Administration that provides additional pseudoproxy datasets[1]. This collection of common datasets is an important resource for researchers wishing to propose new methodological applications for CFRs, and is an excellent starting point for these investigations.

1.8.3 Climate Reconstructions and the Future

More than a decade of research on deriving large-scale temperature reconstructions of the CE has yielded many insights about our past climate and established the utility of such efforts as a guide to the future. Important CFR improvements are nevertheless still necessary and leave open the potential for new analysis methods to have significant impacts on the field. Broad assessments of the multivariate linear regression framework have shown the potential for variance losses and mean biases in reconstructions on hemispheric scales (e.g. [13][51][86]), although some methods have demonstrated significant skill for reconstructions of hemispheric and global indices [57]. The spatial skill of CFRs, however, has been shown in pseudo-proxy experiments to vary widely,

[1] http://www.ncdc.noaa.gov/paleo/pubs/pr-challenge/pr-challenge.html

with some regions showing significant errors [89]. Establishing methods with improved spatial skill is therefore an important target for alternative CFR approaches. It also is critical to establish rigorous uncertainty estimates for derived reconstructions by incorporating a more comprehensive characterization of known errors into the reconstruction problem. Bayesian and ensemble approaches lend themselves well to this task and constitute another open area of pursuit for new methodological applications. Process-based characterizations of the connection between climate and proxy responses also are becoming more widely established [2][22][76][100]. These developments make it possible to incorporate physically-based forward models as constraints on CFR problems and further open the possibility of methodological advancement. Recent Bayesian studies have provided the groundwork for such approaches [52][103], while paleoclimatic assimilation techniques have also shown promise [112].

In the context of machine learning, the problem of reconstructing parts of a missing data matrix has been widely studied as the matrix completion problem (see Figure **4**). A popular example of the problem is encountered in movie recommendation systems, in which each user of a given system rates a few movies out of tens of thousands of available titles. The system subsequently predicts a tentative user rating for all possible movies, and ultimately displays the ones that the user may like. Unlike traditional missing value imputation problems where a few entries in a given data matrix are missing, in the context of matrix completion one works with mostly missing entries, e.g. in movie recommendation systems 99% or more of the matrix is typically missing. Low-rank matrix factorization methods have been shown to be quite successful in such matrix completion problems [48][73]. Further explorations of matrix completion methods for the paleoclimate reconstruction problem therefore are fully warranted. This includes investigations into the applicability of existing methods, such as probabilistic matrix factorization [73] or low-

rank and sparse decompositions [114], as well as explorations of new methods that take into account aspects specific to the paleoclimate reconstruction. Methods that can perform completions along with a confidence score are more desirable because uncertainty quantification is an important desideratum for paleoclimate.

Finally, it is important to return to the fact that extensive methodological work in the field of CE paleoclimatology is aimed, in part, at better constraining natural climate variability on decadal-to-centennial time scales. This timescale of variability, in addition to expected forced changes, will be the other key contribution to observed climate during the 21st century. Whether we are seeking improved decadal predictions (e.g. ref. [93]) or refined projections of 21st-century regional climate impacts (e.g. ref. [28]), these estimates must incorporate estimates of both forced and natural variability. It therefore is imperative that we fully understand how the climate naturally varies across a range of relevant times scales, how it changes when forced, and how these two components of change may couple together. This understanding cannot be achieved from the modern instrumental record alone, and the CE is a strategic paleoclimate target because it provides both reconstructions with high temporal and spatial resolution and an interval over which CGCM simulations are also feasible. Combining these two sources of information to assess model projections of future climate therefore is itself an important future area of discovery. Analyses that incorporate both the uncertainties in paleoclimatic estimates and the ensemble results of multiple model simulations will be essential for these assessments and is likely a key component of climate informatics as the field evolves into the future.

1.9 Applications to Problems in Polar Regions

Another potential application of machine learning concerns the impact of climate change at the poles and the interaction between the poles and climate in general. Because of the difficulty in

collecting data from the polar regions, the relatively expensive costs and logistics, it is important to maximize the potential benefit deriving from the data. The paucity of surface-measured data is complemented by the richness and increasing volume of either satellite/airborne data and model outputs. In this regard, powerful tools are needed, not only to analyze, manipulate and visualize large data sets but also to search and discover new information from different sources, in order to exploit relationships between data and processes that are not evident or captured by physical models.

The number of applications of machine learning to study polar regions is not high though it has been increasing over the past decade. This is especially true in those cases when data collected from space-borne sensors is considered. For example, Tedesco and colleagues [98] [99] use artificial neural networks (ANNs) or genetic algorithms to estimate snow parameters from space-borne microwave observations. Soh and Tsatsoulis [91] use an Automated Sea Ice Segmentation (ASIS) system that automatically segments Synthetic Aperture Radar (SAR) sea ice imagery by integrating image processing, data mining, and machine learning methodologies. The system is further developed in [92], where an intelligent system for satellite sea ice image analysis named Advanced Reasoning using Knowledge for Typing Of Sea ice (ARKTOS) "mimicking the reasoning process of sea ice experts" is presented. Lu and Leen [54] use semi-supervised learning to separate snow and non-snow areas over Greenland using a multispectral approach. Reusch [71] applies tools from the field of ANNs to reconstruct centennial-scale records of West Antarctic sea-ice variability using ice-core datasets from 18 West Antarctic sites and satellite-based records of sea ice. ANNs are used as a non-linear tool to ice-core predictors to sea-ice targets such as sea salt chemistry to sea ice edge. One of the results from this study is that, in general, reconstructions are quite sensitive to predictor used and not all predictors appear to be

useful. Lastly, Gifford [27] shows a detailed study of team learning, collaboration, and decision applied to ice-penetrating radar data collected in Greenland in May 1999 and September 2007 as part of a model-creation effort for subglacial water presence classification.

The abovementioned examples represent a few cases where machine learning tools have been applied to problems focusing on studying the polar regions. Though the number of studies appears to be increasing, likely because of both the increased research focusing on climate change and the poles and the increased computational power allowing machine learning tools to expand in their usage, they are still relatively rare compared to simpler but often less efficient techniques.

Machine learning and data mining can be used to enhance the value of the data by exposing information which would not be apparent from single-data set analyses. For example, identifying the link between diminishing sea ice extent and increasing melting in Greenland can be done through physical models attempting at modeling the connections between the two through the exchange of atmospheric fluxes. However, large scale connections (or others at different temporal and spatial scales) might be revealed through the use of data-driven models or, in a more sophisticated fashion, through the combination of both physical and data-driven models. Such approach would, among other things, overcome the limitation of the physical models that, even if they represent the state of the art in the corresponding fields, are limited by our knowledge and understanding of the physical processes. ANN's can also be used in understanding not only the connections among multiple parameters (through the analysis of the neurons connections) but also to understand potential temporal shifts in the importance of parameters on the overall process (e.g., increase importance of albedo due to the exposure of bare ice and reduced solid precipitation in Greenland over the past few years). Applications are

not limited to a pure scientific analysis but also to the management of information, error analysis, missing linkages between databases, and improving data acquisition procedures.

In synthesis, there are many areas in which machine learning can support studies of the poles within the context of climate and climate change. These include climate model parameterizations and multi-model ensembles of projections for variables such as sea ice extent, melting in Greenland, and sea level rise contribution, in addition to those discussed in previous sections.

1.10 Towards a Climate Informatics Toolbox

Recent additions to the toolbox of modern machine learning have considerable potential to contribute to and greatly improve prediction and inference capability for climate science. Climate prediction has significant challenges including high dimensionality, multiscale behavior, uncertainty, and strong nonlinearity, but also benefits from having historical data and physics-based models. It is imperative that we bring all available, relevant tools to bear on the climate arena. In addition to the methods mentioned in Section 1.2 and in subsequent sections, here we briefly describe several other methods (some proposed recently) that one might consider to apply to problems in climate science.

We begin with CalTech and Los Alamos National Laboratory's recently developed Optimal Uncertainty Quantification (OUQ) formalism [67][79]. OUQ is a rigorous, yet practical, approach to uncertainty quantification which provides optimal bounds on uncertainties for a given, stated set of assumptions. For example, OUQ can provide a guarantee that the probability that a physical variable exceeds a cutoff is less than some value ϵ. This method has been successfully applied to assess the safety of truss structures to seismic activity. In particular OUQ can provide the maximum and minimum values of the probability of failure of a structure as a

function of an earthquake magnitude. These probabilities are calculated by solving an optimization problem that is determined by the assumptions in the problem. As input, OUQ requires a detailed specification of assumptions. One form of assumptions may be (historical) data. The method's potential for practical use resides in a reduction from an infinite-dimensional, nonconvex optimization problem to a finite (typically low) dimensional one. For a given set of assumptions, the OUQ method returns one of three answers. 1) Yes, the structure will withstand the earthquake with probability greater than p. 2) No, it will not withstand it with probability p or 3) given the input one cannot conclude either (i.e., undetermined). In the undetermined case, more/different data/assumptions are then required to say something definite. Climate models are typically infinite-dimensional dynamical systems and a given set of assumptions will reduce this to a finite dimensional problem. The OUQ approach could address such questions as whether (given a potential scenario) the global mean temperature increase will exceed some threshold T, with probability some ϵ.

To improve the performance (e.g., reduce the generalization error) in statistical learning problems, it sometimes helps to encorporate domain knowledge. This approach is particularly beneficial when there is limited data from which to learn, as is often the case in high-dimensional problems (genomics is another example). This general philosophy described in a number of approaches such as learning with side information, Universum Learning [84] and learning from non-examples [83]. Learning with the Universum and learning from non-examples involve augmenting the available data with related examples of from the same problem domain, but not necessarily from the same distribution. Quite often the generalization error for predictions can be shown to be smaller for carefully chosen augmented data, but this is a relatively uncharted field of research and it is not yet know how to use this optimally. One can imagine using an ensemble

of climate models in conjunction with data from model simulations to improve predictive capacity. How to optimally select Universum or non-examples is an open problem.

Domain knowledge in the form of competing models provides the basis of a game-theoretic approach of model selection [11]. This relates to recent work applying algorithms for online learning with experts to combining the predictions of the multi-model ensemble of GCMs [63]. On historical data, this online learning algorithm's average prediction loss nearly matched that of the best performing climate model. Moreover, the performance of the algorithm surpassed that of the average model prediction, which is a common state-of-the-art method in climate science. A major advantage of these approaches, as well as game-theoretic formulations, is their robustness, including the lack of assumptions regarding linearity and noise. However, since future observations are missing, algorithms for unsupervised or semi-supervised learning with experts should be developed and explored.

Conformal prediction is a recently developed framework for learning based on the theory of algorithmic randomness. The strength of conformal prediction is that it allows one to quantify the confidence in a prediction [80]. Moreover, the reliability of the prediction is never overestimated. This is of course very important in climate prediction. To apply the suite of tools from conformal prediction, however, one needs to have iid (independent, identically distributed) or exchangeable data. While this is a serious restriction, one can imagine using iid computer simulations and checking for robustness. Conformal Prediction is fairly easy to use and can be implemented as a simple wrapper to existing classifiers or regression algorithms. Conformal prediction has been applied successfully in genomics and medical diagnoses. It is likely worthwhile to apply conformal prediction to other complex problems in computational science.

Statistical Relational Learning [26] offers a natural framework for inference in climate. Included

within this set of methods are graphical models [47], a flexible and powerful formalism with which to carry out inference for large, highly complex systems (some of which were discussed in Sections 1.5 and 1.6). At one extreme, graphical models can be derived solely from data. At the other extreme, graphical models provide a generalization of Kalman filters or smoothers, where data is integrated with a model. This general approach is quite powerful but requires efficient computation of conditional probabilities. As a result, one might explore how to adapt or extend the current suite of belief propagation methods to climate-specific problems.

Finally, for all of the above methods, it would be helpful if the learning algorithm could automatically determine which information or data it would be useful to get next. The 'optimal learning' formalism addresses this question [69]. This gradient learning approach can be applied to a whole host of problems for learning where one has limited resources to allocate for information gathering. Optimal learning has been applied successfully to experiment design, in particular in the pharmaceutical industry, where it has the potential to reduce the cost (financial, time, etc.) of the drug discovery process. Optimal learning might be applied to climate science, in order to guide the next sets of observations and/or the next simulations.

To conclude, there is a suite of recently developed machine learning methods whose applicability usefulness in climate science should be explored. At this point, we have only begun to scratch the surface. If these methods prove successful in climate studies, we would expect them to apply elsewhere where one has a model of the physical system and can access data.

1.11 Data Challenges and Opportunities in Climate Informatics

Here we discuss additional challenges and important issues in analyzing climate data.

1.11.1 Issues with Cross-Class Comparisons

There is often a need to compare across different classes of data, whether to provide ground truth

for a satellite retrieval or to evaluate a climate model prediction or to calibrate a proxy measurement. But because of the different characteristics of the data, comparing 'apples to apples' can be difficult.

One of the recurring issues is the difference between internal variability (or weather) and climate responses tied to a specific external forcing. The internal variability is a function of the chaotic dynamics in the atmosphere, and can't be predicted over time periods longer than 10 days or so (see Section 1.6). This variability, which can exist on all time scales, exists also in climate models, but because of the sensitive dependence on initial conditions, any unique simulation will have a different realization of the internal variability. *Climate* changes are then effectively defined as the ensemble mean response (i.e. after averaging out any internal variability). Thus any single realization (such as the real world record) must be thought of as a forced signal (driven by external drivers) combined with a stochastic weather component.

The internal variability increases in relative magnitude as a function of decreasing time or spatial scale. Thus comparisons of the specific time evolution of the climate system need to either take the variability into account, or use specific techniques to minimize the difference from the real world. For instance, 'nudged' simulations use observed winds from the reanalyses to keep the weather in the model loosely tied to the observations. Simulations using the observed ocean temperatures as a boundary condition can do a good job at synchronizing the impacts of variability in the ocean on the atmospheric fields. Another way to minimize the impact of internal variability, is to look for property-to-property correlations to focus on specific processes which, though they may happen at different points in time or space, can nonetheless be compared

across models and observations.

Another issue is that model output does not necessarily represent exact topography or conditions related to an in-situ observation. The average height of a specific grid box might not correspond to the height of a mountain-based observing platform, or the resolved shape of the coastline might make a difference of a 200 km or so in the distance of a station to the shore. These issues can be alleviated to some extent if comparisons are focused on large-scale gridded data. Another technique is to 'downscale' the model output to specific locations, either statistically (based on observed correlations of a local record to larger scale features of the circulation), or dynamically (using an embedded RCM). These methods have the potential to correct for biases in the large-scale model, but many practical issues remain in assessing by how much.

Finally, observations are a function of a specific observing methodology, which encompasses technology, practice and opportunity. These factors can impart a bias or skewness to the observation relative to what the real world may nominally be doing. Examples in satellite remote sensing are common - a low cloud record from a satellite will only be able to see low clouds when there are no high clouds for instance. Similarly, a satellite record of 'mid-tropospheric' temperatures might actually be a weighted integral of temperatures from the surface to the stratosphere. A paleo-climate record may be of a quantity that while related to temperature or precipitation, may be a complex function of both, weighted towards a specific season. In all these cases, it is often advisable to create a 'forward model' of the observational process itself to post-process the raw simulation output to create more commensurate diagnostics.

1.11.2 Climate System Complexity

A further issue arises in creating statistical models of the climate system because both the real world and dynamical models have a large number of different physical variables.

Even simplified models can have hundreds of variables, and while not all of them are essential to determining the state of the system, one variable is frequently not sufficient. Land, atmosphere, and ocean processes all have different dominant time scales, and thus different components are essential at different scales. Some physical understanding is thus necessary to make the proper variable/data choices, even with analysis schemes that extract structure from large datasets. Furthermore, these systems are chaotic, i.e. initial conditions that are practically indistinguishable from each other in any given observing system will diverge greatly from each other on some short timescale. Thus extracting useful predictions requires more than creating more accurate models – one needs to determine what aspects are predictable and which are not.

1.11.3 Challenge: Cloud-computing-based Reproducible Climate Data Analysis

The study of science requires reproducible results: science is a body of work where the community strives to ensure that results are not from the unique abilities and circumstances of one particular person or group. Traditionally this has been done in large part by publishing papers, but the scale of modern climate modeling and data analysis efforts has far outstripped the ability of a journal article to convey enough information to allow reproducibility. This is an issue both of size and of complexity: model results are much larger than can be conveyed in a few pages, and both models and analysis procedures are too complex to be adequately described in a few pages.

The sheer size of GCM and satellite datasets are also outstripping our traditional data storage and

distribution methods: frequently only a few variables from a model's output are saved and distributed at high resolution, and the remaining model output is heavily averaged to generate datasets that are sufficiently small.

One promising approach to addressing these problems is cloud-computing-based reproducible climate data analysis. Having both the data and the analyses resident in the computational cloud allows the details of the computation to be hidden from the user, so, for example, data-intensive portions of the computation could be executed close to where the data resides. But these analyses must be reproducible, which brings not only technical challenges of archiving and finding, describing, and publishing analysis procedures, but also institutional challenges of ensuring that the large datasets that form the basis of these analyses remain accessible.

Data Scale. The size of datasets is rapidly outstripping the ability to store and serve the data. We have difficulty storing even a single copy of the complete archive of the CMIP3 model results, and making complete copies of those results and distributing them for analysis becomes both a large undertaking and limits the analysis to the few places that have data storage facilities of that scale. Analysis done by the host prior to distribution, such as averaging, reduces the size to something more manageable, but currently those reductions are chosen far in advance, and there are many other useful analyses that are not currently being done.

A cloud-based analysis framework would allow such reductions to be chosen and still executed on machines with fast access to the data.

Reproducibility and Provenance Graphs. A cloud-based analysis framework would have to generate reproducible documented results, i.e. we would not only need the ability to rerun a calculation and know that it would generate the same results, but also know precisely what analysis had been done. This could be achieved in part by having standardized analysis schemes,

so that one could be sure precisely what was calculated in a given data filter, but also important is systematically tracking the full provenance of the calculation. This *provenance graph*, showing the full network of data filters and initial, intermediate, and final results, would provide the basis of both reproducibility and communication of results: the provenance graphs provide the information necessary to rerun a calculation and get the same results; they also provide the basis of the full documentation of the results. This full network would need to have layers of abstraction so that the user could start with an overall picture and then proceed to more detailed versions as needed.

1.12 Conclusion

The goal of this chapter is to inspire future work in the nascent field of Climate Informatics. We hope to encourage work not only on some of the challenge problems proposed here but also on new problems. A profuse amount of climate data of various types is available, providing a rich and fertile playground for future machine learning and data mining research. Even exploratory data analysis could prove useful for accelerating discovery. To that end, we have prepared a Climate Informatics wiki as a result of the First International Workshop on Climate Informatics, which includes climate data links with descriptions, challenge problems, and tutorials on machine learning techniques [14]. We are confident that there are myriad collaborations possible at the intersection of climate science and machine learning, data mining, and statistics. We hope our work will encourage progress on a range of emerging problems in Climate Informatics.

Acknowledgements

The First International Workshop on Climate Informatics (2011) served as an inspiration for this chapter, and some of these topics were discussed there. The workshop sponsors were: LDEO/GISS Climate Center, Columbia University; Information Science and Technology Center, Los Alamos National Laboratory; NEC Laboratories America, Department of Statistics, Columbia University; Yahoo! Labs; The New York Academy of Sciences.

KS was supported in part by NSF Grant 1029711.

MKT and MBB are supported by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NOAA. NA05OAR4311004). The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its sub-agencies.

AB was supported in part by NSF grants IIS-1029711, IIS-0916750, and IIS-0812183, and NSF CAREER award IIS-0953274.

ARG's research reported here has been financially supported by the Oak Ridge National Laboratory and Northeastern University grants, as well as the National Science Foundation award 1029166, in addition to funding from the US Department of Energy and the Department of Science and Technology of the Government of India.

The work of JES was supported in part by NSF grant ATM0902436 and by NOAA grants NA07OAR4310060 and NA10OAR4320137.

MT would like to acknowledge the NSF grant ARC 0909388. GAS is supported by the NASA Modeling and Analysis Program.

References

[1] C. M. Ammann, F. Joos, D. S. Schimel, B. L. Otto-Bliesner, and R. A. Tomas. Solar influence on climate during the past millennium: Results from transient simulations with the NCAR Climate System Model. Proc. U. S. Natl. Acad. Sci., 104(10):3713-3718, 2007.

[2] K. J. Anchukaitis, M. N. Evans, A. Kaplan, E. A. Vaganov, M. K. Hughes, H. D. Grissino-Mayer. Forward modeling of regional scale tree-ring patterns in the southeastern United States and the recent influence of summer drought, Geophys. Res. Lett., 33, L04705, doi:10.1029/2005GL025050.

[3] A. G. Barnston and T. M. Smith. Specification and prediction of global surface temperature and precipitation from global SST using CCA. J. Climate, 9:2660–2697, 1996.

[4] C. M. Bishop. *Machine Learning and Pattern Recognition*. Springer, 2007.

[5] Christopher S. Bretherton, Catherine Smith, and John M. Wallace. An intercomparison of methods for finding coupled patterns in climate data. J. Climate, 5:541–560, 1992.

[6] Brohan, P., J.J. Kennedy, I., Harris, S.F.B. Tett, and P.D. Jones. Uncertainty estimates in regional and global observed temperature changes: A new dataset from 1850. *J. Geophys. Res.* 111, D12106, 2006.

[7] Buckley, B.M., K.J. Anchukaitis, D. Penny, et al. Climate as a contributing factor in the demise of Angkor, Cambodia. *Proc. Nat. Acad. Sci. USA* 107, 6748-6752, 2010.

[8] S. J. Camargo and A. G. Barnston. Experimental seasonal dynamical forecasts of tropical cyclone activity at IRI. Wea. Forecasting, 24:472–491, 2009.

[9] S. J. Camargo, A. W. Robertson, A. G. Barnston, and M. Ghil. Clustering of eastern

North Pacific tropical cyclone tracks: ENSO and MJO effects. Geochem. Geophys. and Geosys., 9:Q06V05, 2008. doi:10.1029/2007GC001861.

[10] M.A. Cane, S.E. Zebiak, and S.C. Dolan. Experimental forecasts of El Niño. Nature, 321:827–832, 1986.

[11] Cesa-Bianchi, N. and G. Lugosi. *Prediction, Learning and Games.* Cambridge University Press, 2006.

[12] V. Chandrasekaran, S. Sanghavi, P. Parril, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. SIAM Journal of Optimization, 21(2), 2011.

[13] B. Christiansen, T. Schmith, and P. Thejll. A surrogate ensemble study of climate reconstruction methods: Stochasticity and robustness. J. Climate, 22(4):951-976, 2009.

[14] Climate Informatics wiki: http://sites.google.com/site/1stclimateinformatics/

[15] Cook, E.R., R. Seager, M.A. Cane, and D.W. Stahle. North American drought: Reconstructions, causes, and consequences. *Earth Science Reviews* 81, 93-134, 2007.

[16] Dee, D.P., S.M. Uppala, A.J. Simmons, et al. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Quart. J. Roy. Meteorol. Soc.* 137, 553-597, 2011.

[17] T. DelSole and M. K. Tippett. Average Predictability Time: Part I. Theory. J. Atmos. Sci., 66:1188-1204, 2009.

[18] T. DelSole, M. K. Tippett, and J. Shukla. A significant component of unforced multidecadal variability in the recent acceleration of global warming. J. Climate, 24:909-926, 2011.

[19] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. The backbone of the climate network. *European Physics Letters*, 87(4):48007, 2007.

[20] R. Donner, S. Barbosa, J. Kurths, and N. Marwan. Understanding the Earth as a Complex System – recent advances in data analysis and modeling in Earth sciences. European Physics Journal Special Topics, 174:1-9, 2009.

[21] M. N. Evans, A. Kaplan, and M. A. Cane. Pacific sea surface temperature field reconstruction from coral δ18O data using reduced space objective analysis. Paleoceanography, 17, 2002.

[22] M. N. Evans, B. K. Reichert, A. Kaplan, K. J. Anchukaitis, E. A. Vaganov, M. K. Hughes, and M. A. Cane. A forward modeling approach to paleoclimatic interpretation of tree-ring data. J. Geophys. Res., 111(G3), 2006.

[23] J. A. Foley, M. T. Coe, M. Scheffer, and G. Wang. Regime Shifts in the Sahara and Sahel: Interactions between Ecological and Climatic Systems in Northern Africa. *Ecosystems*, 6:524-532, 2003.

[24] Foster, G., J.D. Annan, G.A. Schmidt, and M.E. Mann. Comment on ”Heat capacity, time constant, and sensitivity of Earth’s climate system” by S.E. Schwartz. *J. Geophys. Res.* 113, D15102, 2008.

[25] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. Preprint, 2010.

[26] Getoor, L. and B. Tasker (eds). *Introduction to Statistical Relational Learning.* MIT Press, 2007.

[27] Gifford, C.M. *Collective Machine Learning: Team Learning and Classification in Multi-Agent Systems*. PhD Dissertation, University of Kansas, 2009.

[28] F. Giorgi and N. Diffenbaugh. Developing regional climate change scenarios for use in assessment of effects on human health and disease. Clim. Res., 36:141-151, 2008.

[29] G. H. Golub and C. F. Van Loan. Matrix Computations. The Johns Hopkins University Press, Washington D.C., third edition, 1996.

[30] J F González-Rouco, H. Beltrami, E. Zorita, and H. Von Storch. Simulation and inversion of borehole temperature profiles in surrogate climates: Spatial distribution and surface coupling. Geophys. Res. Lett., 33(1), 2006.

[31] W.M. Gray. Atlantic seasonal hurricane frequency. PartI: El-Nin˜o and 30-MB quasi-biennial oscillation influences. Mon. Wea. Rev, 112:1649–1688, 1984.

[32] Arthur M. Greene, Andrew W. Robertson, Padhraic Smyth, and Scott Triglia. Downscaling forecasts of Indian monsoon rainfall using a nonhomogeneous hidden Markov model. Quart. J. Royal Meteor. Soc., 137:347–359, 2011.

[33] Hansen, J., R. Ruedy, M. Sato, and K. Lo. Global surface temperature change. *Rev. Geophys*.48, RG4004, 2010.

[34] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.

[35] Hegerl, G.C., T.J. Crowley, M. Allen, et al. Detection of human influence on a new, validated 1500-year temperature reconstruction. *J. Climate* 20, 650-666, 2007.

[36] Hegerl, G.C., F.W. Zwiers, P. Braconnot, et al. Understanding and attributing climate change. *Climate Change 2007: The Physical Science Basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, S. Solomon, et al. (eds), Cambridge University Press, 2007.

[37] M. Hoerling, J. Hurrell, J. Eischeid, and A. Phillips. Detection and Attribution of Twentieth-Century Northern and Southern African Rainfall Change. *Journal of Climate*, 19(16):3989-4008, August 2006.

[38] Solomon M. Hsiang, Kyle C. Meng, and Mark A. Cane. Civil conflicts are associated with the global climate. Nature, 476:438–441, 2011.

[39] IDAG (International ad hoc Detection and Attribution Group. Detecting and attributing external influences on the climate system: A review of recent advances. *J. Clim.* 18, 1291-1314, 2005.

[40] IPCC (Intergovernmental Panel on Climate Change). Expert Meeting on Assessing and Combining Multi Model Climate Projections: Good Practice Guidance Paper on Assessing and Combining Multi Model Climate Projections, R. Knutti, et al., 2010.

[41] Jones, P.D., K.R. Briffa, T.J. Osborn, et al. High-resolution palaeoclimatology of the last millennium: a review of current status and future prospects. *The Holocene* 19, 3-49, 2009.

[42] Kaplan A., M.A. Cane, and Y. Kushnir. Reduced space approach to the optimal analysis interpolation of historical marine observations: Accomplishments, difficulties, and prospects. In Advances in the Applications of Marine Climatology: The Dynamic Part of the WMO Guide to the Applications of Marine Climatology, pages 199-216, Geneva, Switzerland, 2003. World Meteorological Organization.

[43] J. Kawale, S. Liess, A. Kumar, et al. Data guided discovery of dynamic dipoles. In *Proceedings of the NASA Conference on Intelligent Data Understanding*, 2011.

[44] Keenlyside, N.S., M. Latif, J. Jungclaus, L. Kornblueh, and E. Roeckner. Advancing decadal-scale climate prediction in the North Atlantic Sector. *Nature* 453, 84-88, 2008.

[45] Kennedy, J.J., N.A. Rayner, R.O. Smith, D.E. Parker, and M. Saunby. Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement sampling uncertainties. *J. Geophys. Res.* 116, D14103, 2011.

[46] Knutti, R., G.A. Meehl, M.R. Allen, and D.A. Stainforth. Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Clim.* 19, 4224-4233, 2006.

[47] Koller, D. and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques.* MIT Press, 2009.

[48] Y. Koren, R. M. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. IEEE Computer, 42(8):30-37, 2009.

[49] R Sari Kovats, Menno J Bouma, Shakoor Hajat, Eve Worrall, and Andy Haines. El Niño and health. The Lancet, 362:1481–1489, 2003.

[50] V. M. Krasnopolsky and M. S. Fox-Rabinovitz. Complex hybrid models combining deterministic and machine learning components for numerical climate modeling and weather prediction. Neural Networks,19(2):122–134, 2006.

[51] Lee, T.C.,K., F.W. Zwiers, and M. Tsao. Evaluation of proxy-based millennial reconstruction methods. *Climate Dyn*. 31, 263-281, 2008.

[52] B. Li, D.W. Nychka, and C.M. Ammann. The value of multiproxy reconstruction of past climate. J. Am. Stat. Assoc., 105:883–895, 2010.

[53] Carlos H. R. Lima, Upmanu Lall, Tony Jebara, and Anthony G. Barnston. Statistical prediction of ENSO from subsurface sea temperature using a nonlinear dimensionality reduction. J. Climate, 22:4501–4519, 2009.

[54] Lu, Z. and T.K. Leen. Semi-supervised Learning with Penalized Probabilistic Clustering. In *Advances of Neural Information Processing System*, MIT Press, 2005.

[55] M. E. Mann, R. S. Bradley, and M. K. Hughes. Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations. Geophys. Res. Lett., 26:759-762, 1999.

[56] M. E. Mann, S. Rutherford, E. Wahl, and C. Ammann. Testing the fidelity of methods used in proxy-based reconstructions of past climate. J. Climate, 18:4097-4107, 2005.

[57] M. E. Mann, S. Rutherford, E. Wahl, and C. Ammann. Robustness of proxy-based climate field reconstruction methods. J. Geophys. Res., 112(D12109), 2007.

[58] Mann, M.E., Z. Zhang, M.K. Hughes, et al. Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia. *Proc. Nat. Acad. Sci. USA 105,* 13252-13257, 2008.

[59] Mann, M.E., Z. Zhang, S. Rutherford, et al. Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly. *Science* 326, 1256-1260, 2009.

[60] Mearns, L.O., W.J. Gutowski, R. Jones, et al. A regional climate change assessment program for North America. *EOS* 90, 311-312, 2009.

[61] Meehl, G.A., T.F. Stocker, W.D. Collins, et al. Global climate projections. *Climate Change 2007: The Physical Science Basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, S. Solomon, et al. (eds), Cambridge University Press, 2007.

[62] Menne, M.J., C.N. Williams Jr., and M.A. Palecki. On the reliability of the U.S. surface temperature record. *J. Geophys. Res.* 115, D11108, 2010.

[63] Monteleoni, C., G.A. Schmidt, S. Saroha, and E. Asplund. Tracking climate models. *Statistical Analysis and Data Mining* 4, 372-392, 2011.

[64] Murphy, J.M., B.B. Booth, M. Collins, et al. A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Phil. Trans. Roy. Soc. A* 365, 2053-2075, 2007.

[65] G. T. Narisma, J. A. Foley, R. Licker, and N. Ramankutty. Abrupt changes in rainfall

during the twentieth century. *Geophysical Research Letters*, 34:L06710, March 2007.

[66] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. Arxiv, 2010. http://arxiv.org/abs/1010.2731v1.

[67] Owhadi, H., J.C. Scovel, T. Sullivan, M. McKems, and M. Ortiz. Optimal Uncertainty Quantification, *SIAM Review*, 2011 (submitted).

[68] Roger D. Peng, Jennifer F. Bobb, Claudia Tebaldi, Larry McDaniel, Michelle L. Bell, and Francesca Dominici. Toward a quantitative estimate of future heat wave mortality under global climate change. Environ Health Perspect, 119, 2010.

[69] Powell, W.B., and P. Frazier. Optimal Learning. In *Tutorials in Operations Research: State-of-the-art decision making tools in the Information Age.* Hanover, MD, 2008.

[70] P. Ravikumar, A. Agarwal, and M. J. Wainwright. Message-passing for graph-structured linear programs: proximal projections, convergence and rounding schemes. *Journal of Machine Learning Research*, 11:1043-1080, 2010.

[71] Reusch, D.B. Ice-core Reconstructions of West Antarctic Sea-Ice Variability: A Neural Network Perspective. *Fall Meeting of the American Geophysical Union*, 2010.

[72] C.F Ropelewski and M.S. Halpert. Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. Mon. Wea. Rev., 115:1606–1626, 1987.

[73] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems 20*, 2008.

[74] M. Scheffer, S. Carpenter, J. A. Foley, C. Folke, and B. Walker. Catastrophic shifts in ecosystems. *Nature*, 413(6856):591-596, October 2001.

[75] Schmidt, G.A. Error analysis of paleosalinity calculations. *Paleoceanography* 14, 422-429, 1999.

[76] Schmidt, G.A., A. LeGrande, and G. Hoffmann. Water isotope expressions of intrinsic and forced variability in a coupled ocean-atmosphere model. *J. Geophys. Res.* 112, D10103, 2007.

[77] T. Schneider. Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. J. Climate, 14:853-871, 2001.

[78] S. D. Schubert, M. J. Suarez, P. J. Pegion, R. D. Koster, and J. T. Bacmeister. On the cause of the 1930s dust bowl. *Science*, 303:1855-1859, 2004.

[79] Scovel, C. and Steinwart, I. Hypothesis testing for validation and certification. *J. Complexity*, 2010 (submitted).

[80] Shafer, G. and V. Vovk. A tutorial on conformal prediction. *J. Mach. Learn. Res.* 9, 371-421, 2008.

[81] J.Shukla. Dynamical predictability of monthly means. Mon. Wea. Rev.,38:2547–2572,1981.

[82] J. Shukla. Predictability in the midst of chaos: A scientific basis for climate forecasting. Science, 282:728–731, 1998.

[83] Sinz. F.H. *A priori Knowledge from Non-Examples*. Diplomarbeit (Thesis), Universität Tübingen, Germany, 2007.

[84] Sinz, F.H., O. Chapelle, A. Agrawal and B. Schölkopf. An analysis of inference with the universum. In *Advances in Neural Information Processing Systems 20*, 2008.

[85] J. E. Smerdon. Climate models as a test bed for climate reconstruction methods: pseudoproxy experiments. Wiley Interdisciplinary Reviews Climate Change, in revision,

2011.

[86] J. E. Smerdon and A. Kaplan. Comment on "Testing the Fidelity of Methods Used in Proxy-Based Reconstructions of Past Climate": The Role of the Standardization Interval. J. Climate, 20(22):5666-5670, 2007.

[87] J. E. Smerdon, A. Kaplan, and D. E. Amrhein. Erroneous model field representations in multiple pseudoproxy studies: Corrections and implications. J. Climate, 23:5548–5554, 2010.

[88] 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. J. Climate, 24:1284-1309, 2011.

[89] J. E. Smerdon, A. Kaplan, E. Zorita, J. F. González-Rouco, and M. N. Evans. Spatial performance of four climate field reconstruction methods targeting the Common Era. Geophys. Res. Lett., 38, 2011.

[90] Smith, D.M., S. Cusack, A.W. Colman, et al. Improved surface temperature prediction for the coming decade from a global climate model. *Science*317, 769-799, 2007.

[91] Soh, L.-K. and C. Tsatsoulis. Unsupervised segmentation of ERS and Radarsat sea ice images using multiresolution peak detection and aggregated population equalization. *Int. J. Remote S.* 20, 3087-3109, 1999.

[92] Soh, L.-K.. C. Tsatsoulis, D. Gineris, and C. Bertoia. ARKTOS: an intelligent system for SAR sea ice image classification. *IEEE T. Geosci. Remote S.*, 42, 229-248, 2004.

[93] A. Solomon, L. Goddard, A. Kumar, J. Carton, C. Deser, I. Fukumori, A. Greene, G. Hegerl, B. Kirtman, Y. Kushnir, M. Newman, D. Smith, D. Vimont, T. Delworth, J. Meehl, and T. Stockdale. Distinguishing the roles of natural and anthropogenically forced

decadal climate variability: Implications for prediction. Bull. Amer. Meteor. Soc., 92:141-156, 2010.

[94] S. Sra, S. Nowozin, and S. Wright. *Optimization for Machine Learning*. MIT Press, 2011.

[95] K. Steinhaeuser, A. R. Ganguly, and N. V. Chawla. Multivariate and Multiscale Dependence in the Global Climate System Revealed Through Complex Networks. *Climate Dynamics*, doi:10.1007/s00382-011-1135-9, in press, 2011.

[96] Taylor, K.E., R. Stouffer, and G. Meehl. The CMIP5 experimental design. *Bull Amer. Meteorol. Soc.*, 2011 (submitted).

[97] Tebaldi, C. and R. Knutti. The use of the multi-model ensemble in probabilistic climate projections in probabilistic climate projections. *Phil. Trans. Roy. Soc. A* 365, 2053-2075, 2007.

[98] Tedesco, M. and E.J. Kim. A study on the retrieval of dry snow parameters from radiometric data using a Dense Medium model and Genetic Algorithms. *IEEE T. Geosci. Remote S.* 44, 2143-2151, 2006.

[99] Tedesco, M., J. Pulliainen, P. Pampaloni, and M. Hallikainen. Artificial neural network based techniques for the retrieval of SWE and snow depth from SSM/I data. *Remote Sens. Environ.* 90, 76-85, 2004.

[100] D.M. Thompson, T.R. Ault, M.N. Evans, J.E. Cole, and J. Emile-Geay. Comparison of observed and simulated tropical climate trends using a forward model of coral δ18O. Geophys. Res. Lett., in review, 2011.

[101] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*, 58:267-288, 1996.

[102] Martin P. Tingley, Peter F. Craigmile, Murali Haran, Bo Li, Elizabeth Mannshardt-

Shamseldin, and Bala Rajaratnam. Piecing together the past: Statistical insights into paleoclimatic reconstructions. Technical Report 2010-09, Department of Statistics, Stanford University, 2010.

[103] M. P. Tingley and P. Huybers. A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part I: Development and Applications to Paleoclimate Reconstruction Problems. J. Climate, 23(10):2759-2781, 2010.

[104] M. K. Tippett, S. J. Camargo, and A. H. Sobel. A Poisson regression index for tropical cyclone genesis and the role of large-scale vorticity in genesis. J. Climate, 24:2335–2357, 2011.

[105] Trenberth, K.E. P.D. Jones, P. Ambenje, et al. Observations: Surface and atmospheric climate change. *Climate Change 2007: The Physical Science Basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, S. Solomon, et al. (eds), Cambridge University Press, 2007.

[106] A. A. Tsonis, K. L. Swanson and P. J. Roebber. What Do Networks Have To Do With Climate? *Bulletin of the American Meteorological Society*, 87(5):585-595, 2006.

[107] A. A. Tsonis and P. J. Roebber. The architecture of the climate network. *Physica A*, 333:497-504, 2004.

[108] A. A. Tsonis and K. L. Swanson. Topology and Predictability of El Niño and La Niña Networks. *Physical Review Letters*, 100(22):228502, 2008.

[109] Vinnikov, K.Y., N.C. Grody, A. Robok, et al. Temperature trends at the surface and in the troposphere. *J. Geophys. Res.* 111, D03106, 2006.

[110] F. D. Vitart and T. N. Stockdale. Seasonal forecasting of tropical storms using coupled GCM integrations. Mon. Wea. Rev., 129:2521–2537, 2001.

[111] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1-305, 2008.

[112] M. Widmann, H. Goosse, G. van der Schrier, R. Schnur, and J. Barkmeijer. Using data assimilation to study extratropical northern hemisphere climate over the last millennium. Clim. Past, 6:627–644, 2010.

[113] C. A. Woodhouse and J. T. Overpeck. 2000 years of drought variability in the central United States. *Bulletin of the American Meteorological Society*, 79:2693-2714, 1998.

[114] Woodruff, S.D., S.J. Worley, S.J. Lubker, et al. ICOADS Release 2.5: Extensions and enhancements to the surface marine meteorological archive. *J. Geophys. Res.* 31, 951-967, 2011.

[115] Qiaoyan Wu and Dake Chen. Ensemble forecast of Indo-Pacific SST based on IPCC twentieth-century climate simulations. Geophys. Res. Lett., 37, 2010.

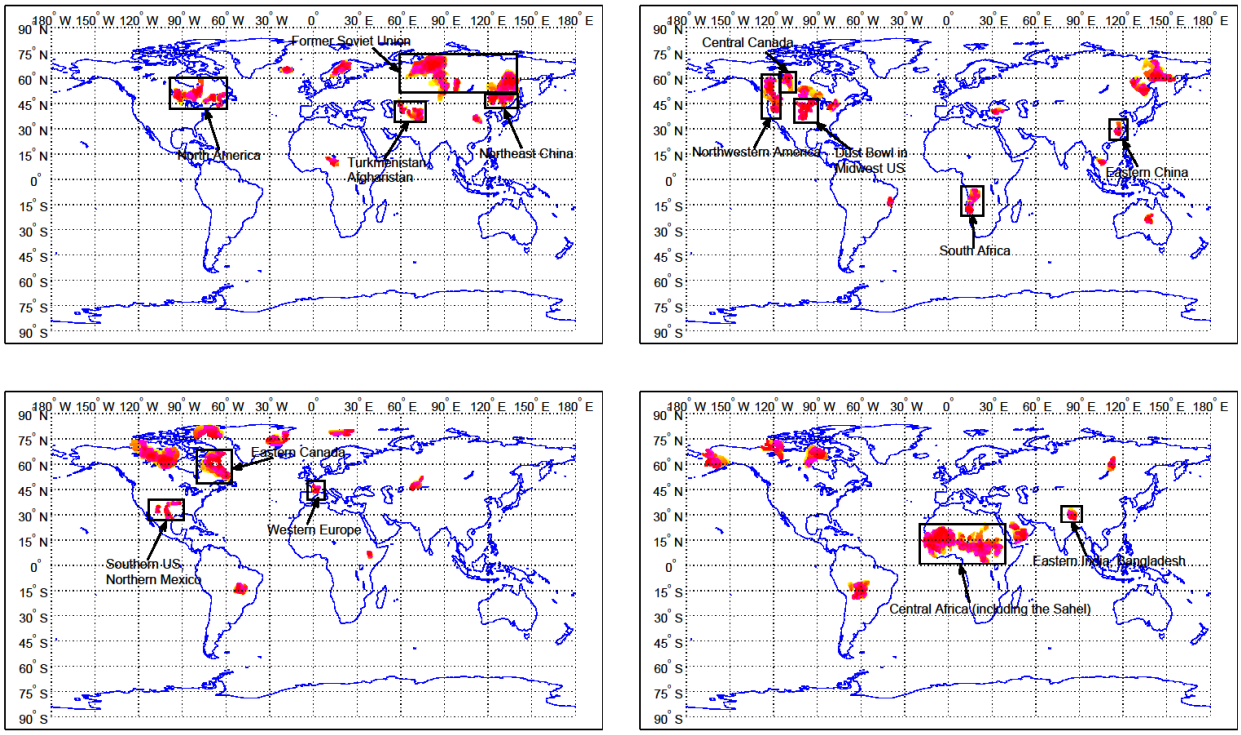

Figure 1. The drought regions detected by our algorithm. Each panel shows the drought starting from a particular decade: 1905-1920 (top left), 1921-1930 (top right), 1941-1950 (bottom left), and 1961-1970 (bottom right). The regions in black rectangles indicate the common droughts found by [63].

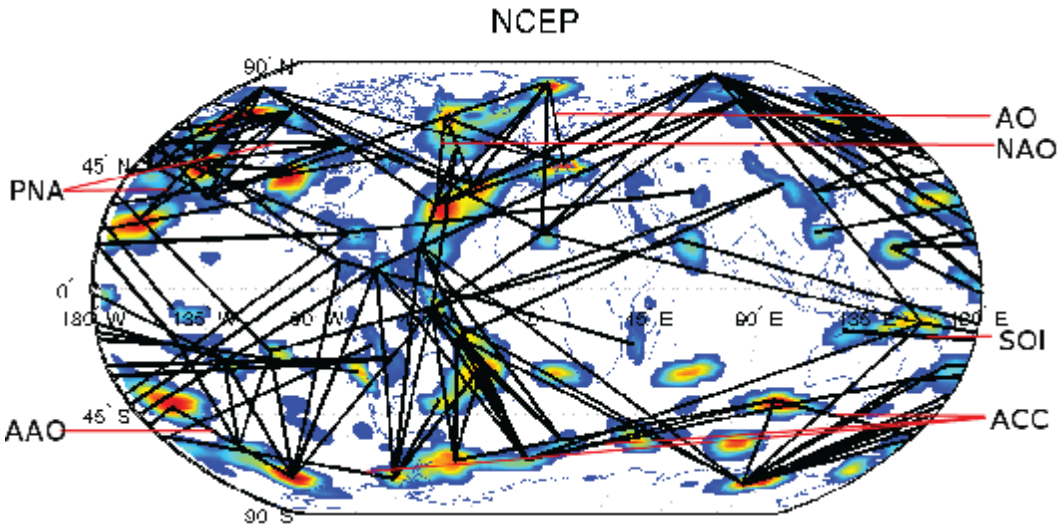

Figure 2. Climate dipoles discovered from sea level pressure (reanalysis) data using graph-based analysis methods (see [42] for details).

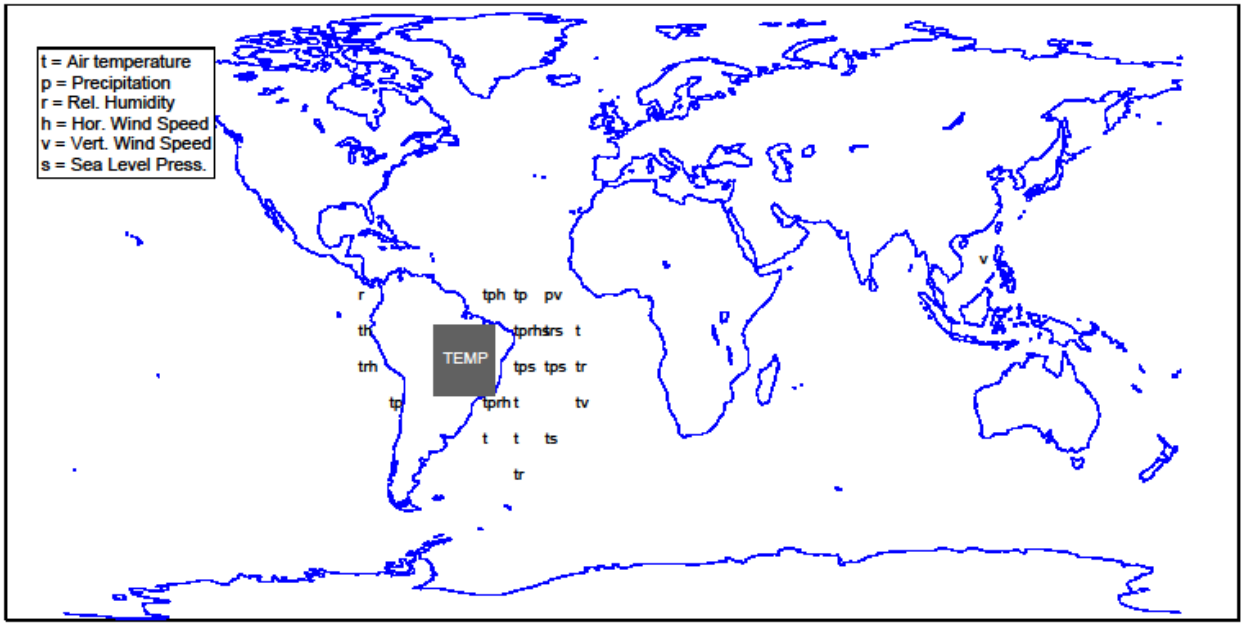

Figure 3. Temperature prediction in Brazil: Variables chosen through cross-validation.

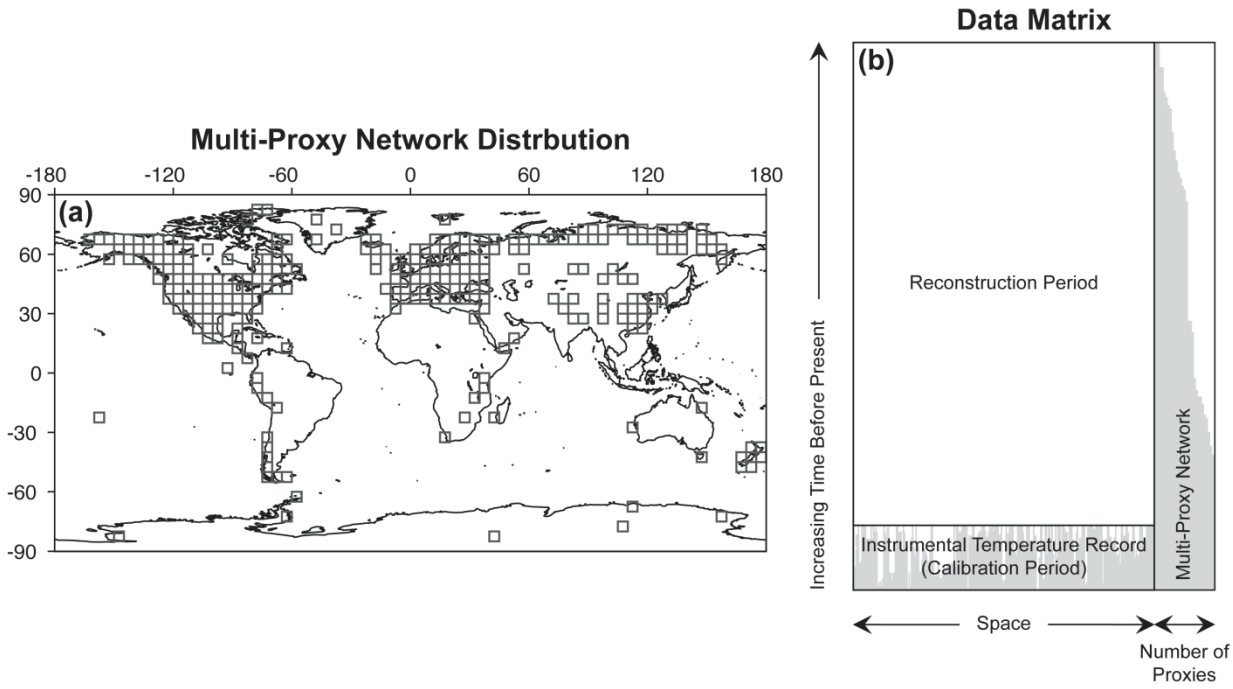

Figure 4. Representation of the global distribution of the most up-to-date global multi-proxy network used in ref. [58]. Grey squares indicate the 5° grid cells that contain at least one proxy in the unscreened network from ref. [58]. (b) Schematic of the data matrix for temperature field reconstructions spanning all or part of the CE. Grey regions in the data matrix are schematic representations of data availability in the instrumental temperature field and the multi-proxy matrix. White regions indicate missing data in the various sections of the data matrix.