Nonlin. Processes Geophys. Discuss., 1, 51–96, 2014
www.nonlin-processes-geophys-discuss.net/1/51/2014/
doi:10.5194/npgd-1-51-2014
© Author(s) 2014. CC Attribution 3.0 License.

Nonlinear Processes in Geophysics
Discussions
Open Access

This discussion paper is/has been under review for the journal Nonlinear Processes in Geophysics (NPG). Please refer to the corresponding final paper in NPG if available.

Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques

A. R. Ganguly[1,*], E. A. Kodra[1,*], A. Banerjee[2], S. Boriah[2], S. Chatterjee[3], S. Chatterjee[2], A. Choudhary[4], D. Das[1], J. Faghmous[2], P. Ganguli[1], S. Ghosh[5], K. Hayhoe[6], C. Hays[7], W. Hendrix[4], Q. Fu[2], J. Kawale[2], D. Kumar[1], V. Kumar[2], S. Liess[8], R. Mawalagedara[1], V. Mithal[2], R. Oglesby[7], K. Salvi[5], P. K. Snyder[8], K. Steinhaeuser[2], D. Wang[1], and D. Wuebbles[9]

[1]Sustainability and Data Sciences Laboratory, Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA
[2]Department of Computer Science and Engineering, University of Minnesota, Twin Cities, MN, USA
[3]School of Statistics, University of Minnesota, Twin Cities, MN, USA
[4]Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL, USA
[5]Department of Civil Engineering, Indian Institute of Technology Bombay, Mumbai, Maharashtra, India
[6]Texas Tech University, Lubbock, TX, USA
[7]Department of Earth and Atmospheric Science, University of Nebraska, Lincoln, NE, USA

[8]Department of Soil, Water, and Climate, University of Minnesota, Twin Cities, MN, USA
[9]Department of Atmospheric Sciences, University of Illinois, Urbana-Champaign, IL, USA
[*]These authors contributed equally to this work.

Received: 28 January 2014 – Accepted: 2 February 2014 – Published: 14 February 2014

Correspondence to: A. R. Ganguly (a.ganguly@neu.edu)

Published by Copernicus Publications on behalf of the European Geosciences Union & American Geophysical Union.

Abstract

Extreme events such as heat waves, cold spells, floods, droughts, tropical cyclones, and tornadoes have potentially devastating impacts on natural and engineered systems, and human communities, worldwide. Stakeholder decisions about critical infrastructures, natural resources, emergency preparedness and humanitarian aid typically need to be made at local to regional scales over seasonal to decadal planning horizons. However, credible climate change attribution and reliable projections at more localized and shorter time scales remain grand challenges. Long-standing gaps include inadequate understanding of processes such as cloud physics and ocean-land-atmosphere interactions, limitations of physics-based computer models, and the importance of intrinsic climate system variability at decadal horizons. Meanwhile, the growing size and complexity of climate data from model simulations and remote sensors increases opportunities to address these scientific gaps. This perspectives article explores the possibility that physically cognizant mining of massive climate data may lead to significant advances in generating credible predictive insights about climate extremes and in turn translating them to actionable metrics and information for adaptation and policy. Specifically, we propose that data mining techniques geared towards extremes can help tackle the grand challenges in the development of interpretable climate projections, predictability, and uncertainty assessments. To be successful, scalable methods will need to handle what has been called “Big Data” to tease out elusive but robust statistics of extremes and change from what is ultimately small data. Physically-based relationships (where available) and conceptual understanding (where appropriate) are needed to guide methods development and interpretation of results. Such approaches may be especially relevant in situations where computer models may not be able to fully encapsulate current process understanding, yet the wealth of data may offer additional insights. Large-scale interdisciplinary team efforts, involving domain experts and individual researchers who span disciplines, will be necessary to address the challenge.

NPGD
1, 51–96, 2014

1 Introduction

Observed and projected changes in the frequency and severity of heat waves and heavy precipitation events have been explicitly linked to human-induced climate change by the recent literature, as summarized in the IPCC Special Report on Managing the Risk of Extreme Events and Disasters to Advance Climate Change Adaptation, also known as the IPCC-SREX (Field et al., 2012) and in a perspective article in Nature Climate Change (Coumou and Rahmstorf, 2012). However, the impact of the changing climate on other types of extremes such as severe weather and hydrological events, including floods, droughts, storms, hurricanes, cyclones, and tornadoes, remains unclear. Mitigation policy requires quantifying the benefits of reducing emissions in terms of impacts avoided. Adaptation to natural hazards and constrained natural resources requires credible projections of extremes, along with their uncertainties, at local to regional scales. Delineating possible links between changes in weather extremes with changes in climate or land use are therefore directly relevant to both mitigation and adaptation planning.

High-resolution global climate models, in conjunction with downscaling based on statistical approaches or regional climate models, may bridge the gap. Unfortunately, the recent literature and our analyses suggest that physics-based modeling alone may not be able to keep pace with the urgency of stakeholder requirements. Each generation of climate models brings new advances, such as the recent expansion of more traditional atmosphere-ocean general circulation models into fully coupled earth system modelling systems in the Coupled Model Intercomparison Project version 5 (CMIP; Taylor et al., 2012). Coupling new models brings its own issues, however, and evaluation studies suggest that, despite noticeable improvements, regional-scale biases persist in the latest generation of climate models, despite enhanced resolutions and the incorporation of additional physical and biogeochemical processes (e.g. Ryu and Hayhoe, 2013; Kumar et al., 2014).

NPGD
1, 51–96, 2014

Climate extremes continue to represent a major challenge. Consider the example of droughts: according to the IPCC-AR5 (IPCC, 2013) report on the "physical science basis" of climate change, scientific confidence in the ability to characterize and project droughts may have reduced over the last several years (see also Table SPM.1 in the IPCC-AR5 summary for policymakers). Two recent papers on droughts, one published in Nature Climate Change (Dai, 2012) and another in Nature (Sheffield et al., 2012), offered diametrically opposite insights. While Dai (2012) concluded that droughts globally have shown an increasing trend in the past and will worsen in the future; Sheffield et al. (2012) found a lack of trends in global drought over the past 60 yr. The differing insights are summarized by Trenberth et al. (2014) in a perspectives article in Nature Climate Change.

Climate-related data are rapidly increasing in size and complexity (Overpeck et al., 2011; Taylor et al., 2012). This begs the question whether data science, which has already transformed disparate data-rich fields from biological sciences to social media to information retrieval, may also offer fresh insights to address fundamental knowledge-gaps related to climate extremes. Going back to the droughts example, Trenberth et al. (2014) suggest that the reasons for the diverging insights were the different choices of underlying data and metrics.

While confidence in scientific understanding and attribution of observed trends to human-induced climate change continues to increase, the IPCC-SREX highlights key gaps in present scientific understanding of climate extremes. Previous and current research on climate extremes typically focuses on one of three areas: the physical science basis, statistics of extremes, or adaptation and potential impacts. Physical science-based analyses tend to emphasize mechanistic understanding and attribution; statistical analyses generally develop data-driven techniques for descriptive and predictive analyses (for example, recent applications of extreme value theory, change detection and sparse regression to climate extremes); and impact studies tend to focus on exposure, vulnerability and consequence assessments. Despite significant progress in all three areas, our ability to establish credible links between climate variability, climate

NPGD
1, 51–96, 2014

change, and climate extremes is still insufficient to facilitate confident and risk-informed decision-making, particularly at regional and decadal scales.

Reliable projections need to generate interpretable predictive insights while accounting for the knowledge-gaps and intrinsic system variability. The wealth of data continues to increase, as does our conceptual understanding of processes that may generate extremes, such as the influence of oceans and climate oscillators, and local or regional terrestrial drivers. The lack of significant improvement in the latest generation of computer models may suggest that the enhanced understanding may not yet translate to improved projections. Data-driven methods by themselves may not be adequate for long-lead time projections of a nonlinear dynamical system such as climate. Data assimilation methods have limited ability to contribute in the future when projection lead times are large. However, dependence characterization and data-driven predictive modeling may be conditioned on the results of physics-based models, and further based on physical or process understanding, that in turn may be difficult to capture within the current set of model parameterizations. In such cases, pure data-driven methods may lead to spurious correlations or predictions, but physical constraints in the design and interpretation of such methods may guard against the possibility. Thus, ocean or atmospheric temperatures from climate models may generate better characterizations and projections of precipitation extremes statistics with uncertainties (e.g., Kao and Ganguly, 2011; Steinhaeuser et al., 2012).

2 The climate science question: interdisciplinary perspectives

This article focuses on what may be viewed as three inter-related grand challenges in climate change studies: (1) characterization of climate extremes, (2) comprehensive assessment of uncertainties, and (3) enhanced predictive understanding, with a goal of improving projections. Climate and earth sciences have grown from data–poor to data–rich sciences over the last couple of decades, and are likely to be at the forefront of societal challenges pertaining to Big Data in this century (Overpeck et al., 2011).

NPGD
1, 51–96, 2014

Can the rapid and recent increases in computational power and analysis capabilities, as well as steady progress on foundational theories in statistics, nonlinear physics, information theory, signal processing, computer science, and econometrics, enable fundamental advances in climate science through computational data sciences? Can the data science methods be carefully designed to avoid spurious generalizations, and to extract physically-based patterns that can be interpreted by climate scientists?

Solutions for massive data volume and complexity have already made their mark in scientific and engineering disciplines as diverse as biology, astrophysics, and Internet phenomena such as Google or Facebook (Berriman et al., 2010; Langmead et al., 2010; Yang et al., 2011) and spawned new fields of research such as sensor networks (Ganguly et al., 2009a). Climate problems increasingly demand data-driven solutions, but the relevant approaches need to consider relatively unique challenges not present, or not as predominant, in fields where data sciences have proved enormously successful thus far. Thus, carefully-designed parallel and distributed algorithms may be required to ensure that sophisticated methods designed for nonlinear processes and complex long-memory or long-range associations can scale and remain resilient to spurious "discoveries".

3 Big data challenges in climate science

Over the last few decades, climate data has expanded rapidly in both size and complexity. While weather station records remain small and relatively manageable, the advent of the satellite era and remote sensors in general, and the evolution of high-resolution weather and climate models, both of which divide the planet up into ever-decreasing grid-sizes, are the primary factors driving data increases. Ensembles of archived climate model outputs have grown from a few hundred terabytes after the last IPCC assessment cycle (AR4) to the petabyte scale (AR5). The global archive of climate data is projected to grow to about 50 PB around 2015, exceed 100 PB by 2020 and reach up to 350 PB by 2030, mainly from model simulations and remote sensing observations,

NPGD
1, 51–96, 2014

but also from in-situ observations (Overpeck et al., 2011; Taylor et al., 2012). The pace of data growth appears to suggest that even these projections may represent lower bounds.

Disk space and processing speeds are perennial challenges. Today, however, the major technical barriers for mining massive data lie in scalable data–intensive analysis capabilities, where fast storage and scalable input/output are major concerns (Schadt et al., 2010; Trelles et al., 2011), along with mathematical and algorithmic capabilities. Data–driven methods are not new in climate, meteorology or geophysics; the novelty is in the scalability challenges for massive data as well as the opportunities to infer novel process understanding and new predictive insights.

Recent developments in data science have sometimes focused almost exclusively on scalability to massive data rather than data complexity (Armbrust et al., 2010; Dean and Ghemawat, 2008). New methods need to consider several crucial aspects that are rather unique to climate science and related disciplines: climate data exhibit complex space-time dependence; the data-generation processes are highly nonlinear and may be extremely sensitive to initial conditions; variability may occur over long time frames and thus may be difficult to evaluate with limited historical data; spatial dependencies may be based on proximity as well as long range teleconnections with time lags or leads, which makes the discovery of associations a combinatorial challenge; and extremes, unusual patterns or anomalies are of interest, particularly at higher resolutions. The dominance of nonlinear and non-stationary processes, combined with the need for projections (e.g., for extreme values) over long-lead time precludes data-driven projections alone.

Predictability studies (e.g., Karamperidou et al., 2014) leading to characterization of irreducible uncertainties is a major challenge in climate science that may be relatively unique among the urgent Big Data challenge areas. Sterk et al. (2012) measured predictability of extremes with relatively simple geophysical models using finite-time Lyapunov exponent. Delsole and Tippett (2009a, b) proposed a measure based on average predictability considering all lead times without time averaging. Koster et al. (2000)

NPGD
1, 51–96, 2014

studied predictability of precipitation in the context of climate variability. Branstator and Teng (2010) and Branstator et al. (2012) studied decadal-scale predictability from an ensemble of multiple initial condition runs using relative entropy. Giannakis and Majda et al. (2012) have used data driven methods for dynamical systems (with applications in climate atmosphere ocean science) to quantify predictability and extract spatiotemporal patterns. The approaches are relatively new to climate but have tremendous implications for stakeholders and decision makers. The implications of adapting these methods to Big Data have not been studied in detail.

Big Data has its own unique problems. A major challenge related to working with large datasets is avoiding false positives, especially when looking for patterns in the data that are rare. The problem arises from the fact that when a large amount of data is considered, the probability of encountering random occurrence of the target pattern in the data is also high. From the view-point of statistical tests, the p value has little relevance for a sample size big enough to be called really Big Data. Thus, virtually any hypothesis will be accepted if the sample size is large enough, since the p value of the null hypothesis will always be almost zero. Bonferroni correction, a theorem of statistics that gives a statistically naive way to avoid these false positive responses to a search through the data, has been used widely in the past with large datasets. However, avoiding false discoveries is still an active research area and several new methods have been proposed in last two decades (Benjamini and Hochberg, 1995; Bogdan et al., 2008; Dudoit et al., 2003; Efron, 2007) that improve upon the Bonferroni theorem both by new methodological and theoretical developments. Another problem with big data arises if one tries to identify the distribution a variable follows based solely on p values. The goodness-of-fit tests become extremely sensitive to small, inconsequential changes when the sample size is large. The issue of false positives with Big Data has been discussed in the context of a commonly used statistical approach for climate extremes. Resampling techniques have been used to study properties of climate extremes (e.g., Kharin and Zwiers, 2005; Kharin et al., 2007), where the authors also list caveats and challenges for such usage. A recent proposal for bootstrap in big data

NPGD
1, 51–96, 2014

(Kleiner et al., 2012) and other alternatives require further study, in order to understand how to use resampling for extremes of climate variables from large datasets.

Due to this ever-present risk of coming up with spurious discoveries and insights with Big Data, the importance of physics-guided data mining needs to be emphasized further. We can either use physical constraints to validate the data-driven knowledge discoveries or incorporate the physical constraints in the knowledge discovery process by mapping them either as statistical constraints or in the selection of variables and distributions.

4 Societal urgency and state of the science

The types of extreme events discussed here have the potential to cause significant devastation; as shown in Fig. 1, the largest number of deaths results from droughts, tropical cyclones and floods, and the most significant economic loss from hurricanes/cyclones and floods. Mortality and economic losses from tornadoes and severe thunderstorms has been of significant concern in the United States, given the devastating losses in 2011 (Simmons et al., 2012). The size depicting each type of hazard provides a measure of our uncertainty under climate change; unfortunately, we find that the level of uncertainty is generally high for the most destructive hazards (Bouwer, 2011). Even for the relatively better understood temperature extremes, such as heat waves and cold snaps, large uncertainties remain, especially at regional scales (Ganguly et al., 2009b). Hazards are expected to be more severe for poorer and more vulnerable regions; developed economies, however, are not immune to loss either, as demonstrated by Paris and Chicago heatwave mortality (Hayhoe et al., 2010) and United States Gulf Coast hurricane impacts (Burby, 2006). Recent studies have advanced our understanding of observed trends in heavy precipitation or flooding and attributions to global warming (Min et al., 2011; Pall et al., 2011). However, uncertainties remain in interpreting observed extremes (Ghosh et al., 2011; Goswami et al., 2006) and in reliable projections of extremes' intensity-duration-frequency at regional scales (Kao and Ganguly,

NPGD
1, 51–96, 2014

2011; Kharin et al., 2007) that are crucial for water and flood management. Floods in particular are less well understood owing to cascading uncertainty from projections of heavy rainfall to consequences for surface hydrology and impacts on water management (Schneider and Kuntz-Duriseti, 2002). Nevertheless, generating credible projections of climate variables at regional or even local scales remains an important step for reliable assessments of hazards and their consequences.

Can improvements in physics and higher-resolution models, increase otherwise inadequate precision and enhance the accuracy, of projections for climate-related extremes? Projections from global models tend to grow more uncertain with increased spatial and temporal resolution, especially for precipitation, particularly so over the tropics (Kao and Ganguly, 2011). The possibility that the current-generation and higher resolution CMIP5 models will improve projections compared to the previous-generation phase 3 (CMIP3) models remains to be tested at appropriate scales.

While comparing the performance of GCMs (or the newer generation of earth system models), it is important to carefully distinguish between model evaluation versus translating model outputs into information relevant for impacts, adaptation, and vulnerability (IAV) studies. GCMs are designed to model large-scale atmospheric dynamics, and from that perspective, recent results suggest general improvement of the ensemble of CMIP5 models compared to CMIP3 (e.g. Ryu and Hayhoe, 2013). However, any improvement in the internal physics or dynamical behavior of models may not be immediately manifested in, for example, model ability to reproduce absolute values of temperature or precipitation at regional and seasonal scales, or in their extremes. Nonetheless, IAV studies may occasionally rely on GCM simulations of temperature and precipitation for future assessments, either directly or indirectly after statistical or dynamical downscaling. One of the primary functions of downscaling, particularly statistical, is to remove GCM-simulated biases in absolute values for IAV applications that require absolute values to assess impacts. The importance of this step is illustrated in Fig. 2, which compares a 7-member CMIP5 versus CMIP3 ensemble with National Center for Environmental Prediction (NCEP-I and NCEP-II) reanalysis temperature and

NPGD
1, 51–96, 2014

Global Precipitation Climatology Project (GPCP) precipitation. Based on a straightforward comparison, no improvements are apparent either in terms of the multimodel median projections or in terms of the uncertainty bounds. In fact, CMIP5 almost consistently predicts higher temperatures and precipitation compared to the CMIP3 multimodel median, but these higher values do not necessarily agree better with the observations. These preliminary results (further details in Kumar et al., 2014) may appear to provide further support to arguments (Hulme et al., 2009) that model improvements alone may not provide immediate answers to stakeholder questions or adaptation needs and additional analyses are clearly required in order to extract information from GCM simulations directly relevant to and able to be used by IAV assessments.

This is precisely where Big Data solutions (and in the case of extremes, Big Data solutions that are ultimately geared towards rare events and small data, or elusive indicators thereof) may provide value. Improvements in internal physics and large-scale dynamics of GCMs may not directly improve the variables of most immediate interest to IAV studies. However, data-driven methods may still be able to leverage the improvements in the larger-scale or internal model variables and yield improved projections for the variables of interest to IAV. For the data-driven projections to be interpretable and useful, they need to be guided by physical understanding, where physics may not be directly captured by GCMs, perhaps even after downscaling.

5 Characterization of climate extremes

Climate extremes often refer to well-defined weather or climate events that are quantified using measurable physical quantities such as temperature, precipitation, or wind speed and that are rare (i.e., occurring at the tails of the distribution) relative to current climate states (Zwiers et al., 2013). The definition of climate extremes, in general, varies with the nature of the phenomena and may be based on their impacts. Extremes such as hurricanes, tornadoes and floods cause immediate and widespread devastation, while droughts tend to unfold slowly, are spatially extensive, non-structural and

NPGD
1, 51–96, 2014

have longer-lasting impacts. While phenomena like heat waves under climate change are better understood than most other climate-related extremes (Coumou and Rahmstorf, 2012; Field et al., 2012), their very definitions may depend on the impact sector of interest (Ebi and Meehl, 2007). Figure 2 (top) shows that different definitions of heat waves can significantly impact the final insights; however each definition remains useful for its specific context, such as energy demand (Christenson et al., 2006) or public health (Kovats and Kristie, 2006).

As model-simulated and observational databases, and the importance of informing adaptation or mitigation policy, continue to grow, descriptive analysis of multiple definitions of model-projected and observed extremes will at once become a larger and more complex task. Surprising insights about cold temperature extremes (Kaspi and Schneider, 2011; Kodra et al., 2011) are still being discovered from observed and model-simulated data. Thus, while decreasing frequency of cold extremes has been reported (Coumou and Rahmstorf, 2012), there is still a need for better characterization and improved mechanistic understanding of their potential persistence in a warming world.

Recent advances in attribution of heavy rainfall do not directly translate to improved information for adaptation (Min et al., 2011; Pall et al., 2011). Thus, intensification of precipitation extremes under warming, which is partially explained through our conceptual process understanding (O'Gorman and Schneider, 2009; Sugiyama et al., 2010), is projected relatively credibly in the extra-tropics and at continental to global average scales (Kao and Ganguly, 2011; Kharin et al., 2007). However, large uncertainties remain in estimating the precise degree of change and for specific regions such as the tropics (Kharin et al., 2007), where diverging insights (Ghosh et al., 2011; Goswami et al., 2006) have been recently reported owing to differing characterizations of extremes. Extreme value theory (EVT) has been used in hydrology (Towler et al., 2010) or climate (Ghosh et al., 2011; Kao and Ganguly, 2011; Kharin et al., 2007; Min et al., 2011) to characterize rainfall extremes. Moreover, hydrological extremes are described by several mutually correlated characteristics; such as peak flow, volume and duration

NPGD
1, 51–96, 2014

(Zhang and Singh, 2007) for floods and severity, duration, intensity and spatial extent for droughts (Reddy and Ganguli, 2013; Song and Singh, 2010).

Univariate frequency analyses cannot provide accurate assessment of the probability of occurrence of extremes if the underlying event is characterized by mutually correlated random variables and may lead to over or under estimation of associated risk (Chebana and Ouarda, 2011). Hence, multivariate statistical approaches are often necessary in order to completely assess risk of hydrological extremes. Further developments in the statistical theories related to multivariate extremes are needed for advancing our ability to quantify the complex dependencies of climate extremes more completely, and with greater certainty (Kuhn et al., 2007; Marty and Blanchet, 2011; Mastrandrea et al., 2011; Turkman et al., 2009; Wadsworth and Tawn, 2012). Descriptions of rainfall extremes, whether based on EVT or fixed/dynamic thresholds, need to characterize changing statistics of storm events (Kao and Ganguly, 2011), droughts (van Huijgevoort et al., 2012) and be relevant to multiple sectors, including hydraulic infrastructure design, flood and drought management policy. A recent study of probable maximum precipitation (PMP) and climate change (Kunkel et al., 2013) may offer new ways to blend physics and data-driven insights for precipitation extremes.

Can data-driven methods provide new insights for understanding and characterizing these extremes? Figure 3 (bottom) presents fully automated and computationally efficient spatio-temporal characterization of long-term droughts using a Markov random field-based approach (Fu et al., 2012). The algorithm was able to detect some of the major global droughts and proved to be efficient in detecting droughts as compared to fixed percentile-based approaches for drought detection. The method has been applied to detect all persistent droughts over the past century (1901–2006). Negative precipitation anomalies of at least 5 yr are considered as significant (hydrologic) droughts and shown here for data from 1970 to 1998. The Sahel drought is clearly detected, as are several others. While this analysis uses a single variable, specifically CRU precipitation observations, the method is capable of handling multiple variables that contribute to the characterization of droughts, such as precipitation, soil moisture, and geopotential

NPGD
1, 51–96, 2014

height. In fact, this MRF-based approach, once generalized to multiple variables, may be viewed as a methodological improvement to the wavelet-based method (Narisma et al., 2007) for abrupt drought detection in the literature. One of the advantages is the ability to fully automate the drought detection procedure with a lesser number of predefined parameters, which may be useful for the detection of megadroughts from paleoclimate data or plausible megadroughts from model projections. The value-addition of the MRF-based approach, beyond proof-of-concept detection of known droughts, would be demonstrated when the methods are generalized for multiple variables, and subsequently used for the evaluation of historical multi-model ensembles as well as for the generation of future projections with uncertainty from model projections in forecast mode. On a completely different scale, our recent research (Ganguli and Ganguly, 2013) explores severity-duration-frequency curves for observed meteorological droughts over the continental US during the last few decades through copula-based approaches.

6 Computational challenges in downscaling

As long as the spatiotemporal scales relevant to stakeholders and policymakers are inadequately resolved by global climate models, downscaling will continue to remain highly relevant to impact analyses. Driven by global climate model outputs, downscaling inherits many of their problems and generates (often massive volumes of) additional data, thus amplifying the Big Data challenge in terms of both data size and complexity. Statistical downscaling (Bürger et al., 2012; Mannshardt-Shamseldin et al., 2010; Robertson et al., 2004) model outputs are relatively computationally inexpensive to generate, but criticisms (Eden et al., 2012; Schmith, 2008) have focused on model complexity and the lack of clarity on whether statistical models will perform well far into the future or on disparate regions. Dynamical downscaling (Pierce et al., 2012; Trapp et al., 2010), based on regional climate models, is much more resource-intensive and is not independent of stationarity assumptions in sub-grid scale parameterizations,

NPGD
1, 51–96, 2014

either. The primary advantages over statistical downscaling are explicit incorporation of topography and higher-resolution process-models, which are critical given the possible importance of finer-scale processes (Jung et al., 2012; Diffenbaugh et al., 2005). However, regional climate models parameterize such processes, often leading to significant inter-model disagreement, e.g., on precipitation (Palmer et al., 2004).

Figure 4 illustrates the ability of both statistical (Ghosh, 2010) and dynamical (Heikkilä et al., 2010) downscaling to provide precise insights compared to the original global model results. Dynamical downscaling over the island-nation of Sri Lanka (Fig. 4, top) suggests, upon visual inspection, that the approach may be able to better capture the expected influence of topography on heat waves beyond global models, particularly since successive resolution-enhancements reveal distinct orographic patterns. On the other hand, the statistical hypothesis test does not necessarily indicate significant improvement, which suggests the importance of multi-metric explorations and rigorous evaluation of downscaling results. However, while the value of dynamical downscaling as a tool for hypothesis testing cannot be denied (despite news articles such as Kerr, 2013), the propagation of uncertainty (Sain et al., 2011) remains a challenge for projections. Over India, while global models suggest a uniform increase in rainfall extremes trends, the results from statistical downscaling (Fig. 4, bottom) show evidence for considerable geographical heterogeneity, which in turn agree with the latest findings on spatial variability of extremes (Ghosh et al., 2011).

7 Complexity of uncertainty assessments

A thorough and comprehensive characterization and quantification of uncertainty, which may result from imprecise observational data, inadequate models, or intrinsic climate system variability, is invaluable for stakeholders and policy-makers but difficult and often even impossible to achieve. Best-estimate projections and corresponding uncertainty bounds under climate change are sometimes thought to be better captured with multimodel ensembles. It is important to evaluate the ability of models to simulate

NPGD
1, 51–96, 2014

historical climate patterns (Pierce et al., 2009), but that alone may not be sufficient for climate models in view of non-stationarity and long lead time projections. Multimodel agreement in the future becomes an important metric, with the notion that consensus implies higher certainty (Overpeck et al., 2011; Weigel et al., 2010). Empirical studies suggest that averages of output from multiple models outperform individual models, this insight being insensitive to which specific models are averaged (Pierce et al., 2009). However, the value of multimodel averages has been questioned (Knutti, 2010), particularly for regional assessments (Knutti et al., 2010; Kodra et al., 2012). Recent attempts at regional assessments include the development of statistical methods that consider both model performance relative to historical observed data and model ensemble agreement (Smith et al., 2009; Ganguly et al., 2013).

One way to improve the uncertainty assessment approaches may be to consider physical and correlative relations in combination with historical model skills and future multimodel agreement. For example, in Fig. 5, observations and model simulations may exhibit regional differences in their adherence to known physical relations. Evaluating the extent to which observed rainfall extremes follow physical relationships like the Clausius-Clapeyron (CC) may help identify systematic patterns in extreme rainfall behavior that could be encapsulated in multimodel uncertainty quantification methodology. We are not aware of any existing statistical strategy (e.g., along the lines of Smith et al., 2009) that attempts to explicitly utilize theoretical physical processes in addition to historical skills and multimodel agreement. In the top panel of Fig. 5, the observations and multiple ESMs are compared to the theoretical CC (scaled to compare with the other curves) over the Eastern United States. An analogous plot is shown for the southwestern United States in (b); the use of different regions makes apparent the degree to which data (observed and modeled) adheres to conceptual physical relations (in this example, the CC). Each point represents a 20 yr mean temperature (1980–1999, x axis) and an estimated 30 yr return rainfall value (calculated from 1980–1999, y axis (Kharin and Zwiers, 2007)) for a land-based grid cell. Polynomial spline regression, a nonparametric smoothing regression approach, is used to fit the rainfall

NPGD
1, 51–96, 2014

return values on mean temperature. This is performed for all models and for NCEP2 reanalysis. Regression model fits are depicted by the colored lines. The theoretical CC relation is depicted by the red line; a manually calibrated multiplication scaling factor of 0.00023 (0.00027 for the southwest) was applied for visual purposes (to line the CC up in the same space as the data) that should not affect the results significantly. Note that the level of the CC line has no real meaning beyond this scaling; only the exponential pattern does. Uncertainty bounds for the multimodel ensemble are created with a resampling scheme combined with the same spline regression.

Besides model-to-model uncertainty, internal model variability due to different choices of parameters is also a major source of uncertainty but is more difficult to quantify due to computational constraints. Generating model simulations with multiple sets of parameters generates (Stainforth et al., 2005) a large number of simulations from a single global climate model but requires enormous computational resources (Stainforth et al., 2002, 2005). Such an approach may generate substantial single model insights (Stainforth et al., 2005) but is not yet feasible on a massive, multimodel scale. Evaluation of multiple models remains an important step in comprehensive uncertainty assessments, even though structural differences may make inter-model comparisons difficult and at times even infeasible.

Requirements to provide uncertainty estimates almost invariably magnify the data challenge, both by generating more model-simulated data (Stainforth et al., 2005) and/or by requiring more data-intensive approaches. Even relatively easily-parallelizable approaches like the bootstrap method, which has been used with EVT (Ghosh et al., 2011; Kao and Ganguly, 2011; Kharin et al., 2007) to characterize uncertainties in return level estimates of climate variables, can benefit significantly from parallel processing. The recently developed method of "bag of little bootstraps" (Kleiner et al., 2012), claims to significantly improved the time-complexity of bootstrap method for large datasets with theoretical guarantees of correctness of uncertainty estimates. The adaptation of these techniques in space and time for observed and model-simulated

NPGD
1, 51–96, 2014

climate data across multimodel and multiple initial condition runs and with different statistical estimation approached may represent major challenges.

8 Enhanced understanding and predictions

Climate extremes, such as heavy rainfall or tropical cyclones, are known to depend on other climate variables (including mean states, local or regional variables, as well as large-scale effects such as oceanic indices) that may be better simulated by models, such as land and sea surface temperatures. Developments in correlative analysis (Reshef et al., 2011; Khan et al., 2007), extended to handle correlated data at multiple spatial and temporal scales, may help quantify conceptual understanding and possibly even discover new dependencies (Khan et al., 2006). Challenges in analyses of historical extreme events such as tornado and hurricane data involve attributing spatial and temporal scales of their behavior to climate change versus natural variability (Emanuel et al., 2008; Webster et al., 2005), as well as to data collection issues for tornadoes and cyclones (Brooks and Doswell, 2001; Emanuel, 2005) and discontinuity of operational definitions for tornadoes (Doswell et al., 2009). Innovative data-driven approaches that consider these complexities are needed to build understanding of the physical behavior and drivers of tornadoes and hurricanes because physics-based modeling for these types of processes is still in early stages (Emanuel et al., 2008; Trapp et al., 2010).

New process understanding or novel insights from mining climate data may help enhance projections and ultimately reduce uncertainty. Although relatively coarse-resolution global climate models are not able to directly simulate tropical cyclones, they have been used to develop aggregate statistics of hurricanes (Emanuel et al., 2008) under climate change. In the same manner, temperature and updraft velocity profiles have been used to constrain or enhance multimodel projections of precipitation extremes (Knutson et al., 2010; Wilhite and Glantz, 1985). These approaches point to the information content in auxiliary variables relevant for climate extremes, and with appropriate adaptations, may lead to a virtuous cycle where data-driven insights

NPGD
1, 51–96, 2014

and process understanding mutually inform, complement, and improve each other. Recently, even tornado occurrences have been associated with monthly environmental parameters (Tippett et al., 2012), though not necessarily in a climate change context.

Linear dimensionality reduction has been used (Mishra et al., 2012) for advancing understanding of climate processes like monsoons, which are known to be important for hydrometeorological extremes. The relationships among large and high dimensional climate data can improve understanding of dominant processes and lead to enhanced projections through predictive modeling. The IPCC-SREX indicates that crucial processes that may influence climate extremes, such as El Niño or other climate oscillators and monsoons, are not well understood. Inferences from surrogate data may yield new insights on extremes processes: the use of ocean salinity data to understand the intensification of climate extremes (Durack et al., 2012) provides an example using a proxy data set for precipitation. Figure 6 (top) provides an example where new data mining methods (Kawale et al., 2011, 2013) for dipole discovery were used to extract information about climate oscillators that may be useful for model evaluation.

An intelligent combination of process understanding with data mining methods may yield new explainable predictive insights beyond statistical downscaling. In fact, the premise of statistical downscaling (discussed earlier), where one overall approach is linear dimensionality reduction followed by nonlinear regression (Ghosh, 2010), is that lower-resolution model outputs have information content about higher-resolution variables. We propose taking this one step further. Variables that are more reliably projected by climate models may be used not only to improve our process understanding, but also to enhance projections of the climate extremes of interest. For enhanced climate projections, especially given the importance of spatiotemporal neighborhoods, prevailing winds, intra-decadal to multi-decadal climate oscillators, and teleconnections, the number of potential explanatory variables may far exceed the number of observations available, which creates problems for classic regression.

Popular dimensionality reduction approaches like empirical orthogonal functions (Hannachi et al., 2007) summarize complex data succinctly but may not necessarily

NPGD
1, 51–96, 2014

do so in a way that maximizes information useful for predicting a specific variable. Sparse regression (Negahban and Wainwright, 2011; Negahban et al., 2012) represents promising alternatives under these situations. Sparse regressions based on constraining the L1-norm of the regression coefficients became popular due to their ability to handle high dimensional data unlike the regular regressions, which suffer from overfitting and model identifiability issues especially when sample size is small. They are often the method of choice in many fields of science and engineering for simultaneously selecting covariates and fitting parsimonious linear models that are better generalizable and easily interpretable. Sparse regularization methods have just begun to be applied to statistical downscaling (Ebtehaj et al., 2012; Phatak et al., 2011). However, this method can also be applied for improved understanding of the complex dependence structure between climate variables, especially in a high-dimensional setting (Chatterjee et al., 2012; Das et al., 2012, 2013). High-performance computational challenges related to this general approach represent an active area of research.

Networks that connect nodes defined as spatial grid-cells (Steinhaeuser et al., 2011b; Donges et al., 2013) or climate oscillators (Donges et al., 2009a), often known as "climate networks", may be useful to represent climate dependencies and develop process understanding (Donges et al., 2009a, b). Figure 6 (bottom) provides an example of new data-driven predictive approaches (Chatterjee et al., 2012) that appear well-suited for high-dimensional and geographically-distributed climate data with complex dependence structures. Network-based graphical models have been used to discover causality among different modes of climate variability (Ebert-Uphoff and Deng, 2012). New methods in nonlinear data sciences, from complex networks (Steinhaeuser et al., 2011a) to multifractals (García-Marín et al., 2013; Muzy et al., 2006), have demonstrated initial promise for better description and predictive insights on climate-related extremes, such as extreme monsoonal rainfall over South Asia (Malik et al., 2011). Certain methods may eventually be applicable in a climate change detection context, potentially making similar innovations useful for not only long horizon prediction and

NPGD
1, 51–96, 2014

uncertainty reduction but also for relatively abrupt change and disturbance analysis or even for early warning systems.

9 Summary

One of the largest scientific gaps in climate change studies is the inability to develop credible projections of extremes with the degree of precision required for adaptation decisions and policy. The dire consequences of climate-related extremes, even in developed economies (Gall et al., 2011), may call for a range of well-informed adaptation strategies from low-regret (Wilby and Keenan, 2012) to transformative (Kates et al., 2012). Improving regional projections (e.g., through variable selection or statistical downscaling) and characterizing natural variability (e.g., irreducible uncertainty at decadal scales: Deser et al., 2012) are necessary for informing adaptation at stakeholder-relevant scales and planning horizons. As climate-related data approaches the scale of hundreds of petabytes (Overpeck et al., 2011), and climate data mining research continues to improve (Smyth et al., 1999; Robertson et al., 2004, 2006; Khan et al., 2006; Camargo et al., 2007a, b; Gaffney et al., 2007), new opportunities will emerge (e.g., Monteleoni et al., 2013; Ganguly et al., 2013). Data-driven methods are complementary to, and indeed conditioned on, physics-based models; however, they need to be tailored to the complexity of climate data and processes. The methods may be motivated from often disparate data-science disciplines such as statistics and econometrics, machine learning and data mining in computer science, nonlinear dynamics in physics and signal processing in engineering. The blend of physics and data-driven insights has conceptual similarities with data assimilation methods (e.g., Gerber and Joos, 2013). However, data assimilation methods are ultimately constrained by the physics encoded within climate models, and updates to parameters or state variables cannot be made in the future where no observations exist. The physics-guided data mining discussed here refers, for example, physics-motivated decomposition into component processes (e.g., Ganguly and Bras, 2003, offers an example in weather

NPGD
1, 51–96, 2014

forecasting), physically motivated variable selection in statistical downscaling (e.g., certain analog methods: Zorita and Von Storch, 1998), or physics-based model selection (Fasullo and Trenberth, 2012) and physically-guided climate networks (Donges et al., 2009b; Steinhaeuser and Tsonis, 2013). The climate extremes exemplars discussed here are a collection of outstanding challenges where data mining already does or can play an innovative new role; various scientific communities will have to decide which specific directions to pursue guided by a combination of stakeholder priorities and which problems they are best positioned to address. Once developed and refined, physics-guided data mining methods are well positioned to produce new scientific understanding and credible projections of climate extremes leading to more informed adaptation and policy.

Acknowledgements. The work was primarily funded by the United States (US) National Science Foundation (NSF) Expeditions in Computing Grant #1029711. Partial funding was provided by the US Nuclear Regulatory Commission, the Planetary Skin Institute, the US NSF Grant IIS-0905581, and the US Department of Energy (DOE) BES and BER Offices. We thank Debadrita Das, David J. Erickson III, Joseph Kanney, Vimal Mishra and Habib Najm for helpful discussions. We are grateful to the Survey of India of the Government of India and the United Nations Refugee Agency (UNHCR) for the maps of India and Sri Lanka respectively used in Fig. 4.

References

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M.: A view of cloud computing, Commun. ACM, 53, 50–58, doi:10.1145/1721654.1721672, 2010.

Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B, 57, 289–300, 1995.

Berriman, G. B., Juve, G., Deelman, E., Regelson, M., and Plavchan, P.: The Application of Cloud Computing to Astronomy: A Study of Cost and Performance, in 2010 Sixth IEEE International Conference on e-Science Workshops, 1–7, IEEE, 2010.

NPGD
1, 51–96, 2014

Bogdan, M., Ghosh, J., and Tokdar, S.: A comparison of the Benjamini-Hochberg procedure with some Bayesian rules for multiple testing, Honor Prof. Pranab K., available at: http://projecteuclid.org/euclid.imsc/1207058275 (last access: 17 January 2014), 2008.

Bouwer, L. M.: Have Disaster Losses Increased Due to Anthropogenic Climate Change?, Bull. Am. Meteorol. Soc., 92, 39–46, doi:10.1175/2010BAMS3092.1, 2011.

Branstator, G. and Teng, H.: Two limits of initial-value decadal predictability in a CGCM, J. Climate, 23, 6292–6311, 2010.

Branstator, G., Teng, H., Meehl, G. A., Kimoto, M., Knight, J. R., Latif, M., and Rosati, A.: Systematic estimates of initial-value decadal predictability for six AOGCMs, J. Climate, 25, 1827–1846, 2012.

Brooks, H. and Doswell, C. A.: Some aspects of the international climatology of tornadoes by damage classification, Atmos. Res., 56, 191–201, doi:10.1016/S0169-8095(00)00098-3, 2001.

Burby, R. J.: Hurricane Katrina and the Paradoxes of Government Disaster Policy: Bringing About Wise Governmental Decisions for Hazardous Areas, Ann. Am. Acad. Pol. Soc. Sci., 604, 171–191, doi:10.1177/0002716205284676, 2006.

Bürger, G., Murdock, T. Q., Werner, A. T., Sobie, S. R., and Cannon, a. J.: Downscaling Extremes – An Intercomparison of Multiple Statistical Methods for Present Climate, J. Climate, 25, 4366–4388, doi:10.1175/JCLI-D-11-00408.1, 2012.

Camargo, S. J., Robertson, A. W., Gaffney, S. J., Smyth, P., and Ghil, M.: Cluster analysis of typhoon tracks. Part I: General properties, J. Climate, 20, 3635–3653, 2007a.

Camargo, S. J., Robertson, A. W., Gaffney, S. J., Smyth, P., and Ghil, M.: Cluster analysis of typhoon tracks. Part II: Large-scale circulation and ENSO, J. Climate, 20, 3654–3676, 2007b.

Chatterjee, S., Steinhaeuser, K., Banerjee, A., Chatterjee, S., and Ganguly, A. R.: Sparse Group Lasso: Consistency and Climate Applications., in SDM, 47–58, SIAM., 2012.

Chebana, F. and Ouarda, T. B. M. J.: Multivariate quantiles in hydrological frequency analysis, Environmetrics, 22, 63–78, doi:10.1002/env.1027, 2011.

Christenson, M., Manz, H., and Gyalistras, D.: Climate warming impact on degree-days and building energy demand in Switzerland, Energy Convers. Manag., 47, 671–686, doi:10.1016/j.enconman.2005.06.009, 2006.

Coumou, D. and Rahmstorf, S.: A decade of weather extremes, Nat. Clim. Change, 2, 491–496, doi:10.1038/nclimate1452, 2012.

NPGD
1, 51–96, 2014

Dai, A.: Increasing drought under global warming in observations and models, Nat. Clim. Change, 3, 52–58, 2012.
Das, D., Ganguly, A., Banerjee, A., and Obradovic, Z.: Towards understanding dominant processes in complex dynamical systems, in Proceedings of the Sixth International Workshop on Knowledge Discovery from Sensor Data – SensorKDD '12, 16–24, ACM Press, New York, New York, USA, 2012.
Das, D., Ganguly, A. R., and Obradovic, Z.: A Sparse Bayesian Model for Dependence Analysis of Extremes: Climate Applications, in the International Conference on Machine Learning (ICML) workshop on Inferning: Interactions between Inference and Learning, 2013.
Dean, J. and Ghemawat, S.: MapReduce: simplified data processing on large clusters, Commun. ACM, 51, 107–113, doi:10.1145/1327452.1327492, 2008.
DelSole, T. and Tippett, M. K.: Average predictability time. part I: Theory, J. Atmos. Sci., 66, 1172–1187, 2009a.
DelSole, T. and Tippett, M. K.: Average predictability time. Part II: Seamless diagnoses of predictability on multiple time scales, J. Atmos. Sci., 66, 1188–1204, 2009b.
Deser, C., Phillips, A., Bourdette, V., and Teng, H.: Uncertainty in climate change projections: the role of internal variability, Clim. Dynam., 38, 527–546, 2012.
Diffenbaugh, N. S., Pal, J. S., Trapp, R. J., and Giorgi, F.: Fine-scale processes regulate the response of extreme events to global climate change, Proc. Natl. Acad. Sci. USA, 102, 15774–15778, 2005.
Donges, J. F., Zou, Y., Marwan, N., and Kurths, J.: Complex networks in climate dynamics, Eur. Phys. J. Spec. Top., 174, 157–179, doi:10.1140/epjst/e2009-01098-2, 2009a.
Donges, J. F., Zou, Y., Marwan, N., and Kurths, J.: The backbone of the climate network, Eur. Phys. Lett., 84, 48007, doi:10.1209/0295-5075/87/48007, 2009b.
Donges, J. F., Petrova, I., Loew, A., Marwan, N., and Kurths, J.: Relationships between eigen and complex network techniques for the statistical analysis of climate data. arXiv preprint arXiv:1305.6634, 2013.
Doswell, C. A., Brooks, H. E., and Dotzek, N.: On the implementation of the enhanced Fujita scale in the USA, Atmos. Res., 93, 554–563, doi:10.1016/j.atmosres.2008.11.003, 2009.
Dudoit, S., Shaffer, J., and Boldrick, J.: Multiple hypothesis testing in microarray experiments, Stat. Sci., available at: http://www.jstor.org/stable/10.2307/3182872 (last access: 17 January 2014), 2003.

NPGD
1, 51–96, 2014

Durack, P. J., Wijffels, S. E., and Matear, R. J.: Ocean salinities reveal strong global water cycle intensification during 1950 to 2000, Science, 336, 455–458, doi:10.1126/science.1212222, 2012.
Ebert-Uphoff, I. and Deng, Y.: Causal Discovery for Climate Research Using Graphical Models, J. Climate, 25, 5648–5665, doi:10.1175/JCLI-D-11-00387.1, 2012.
Ebi, K. L. and Meehl, G. A.: The heat is on: climate change and heatwaves in the Midwest, in: Regional Impacts of Climate Change: Four Case Studies in the United States, Pew Center on Global Climate Change, Arlington, Virginia, 2007.
Ebtehaj, A. M., Foufoula-Georgiou, E., and Lerman, G.: Sparse regularization for precipitation downscaling, J. Geophys. Res., 117, D08107, doi:10.1029/2011JD017057, 2012.
Eden, J. M., Widmann, M., Grawe, D., and Rast, S.: Skill, Correction, and Downscaling of GCM-Simulated Precipitation, J. Climate, 25, 3970–3984, doi:10.1175/JCLI-D-11-00254.1, 2012.
Efron, B.: Size, power and false discovery rates, Ann. Stat., available at: http://projecteuclid.org/euclid.aos/1188405614 (last access: 17 January 2014), 2007.
Emanuel, K.: Increasing destructiveness of tropical cyclones over the past 30 years, Nature, 436, 686–688, doi:10.1038/nature03906, 2005.
Emanuel, K., Sundararajan, R., and Williams, J.: Hurricanes and Global Warming: Results from Downscaling IPCC AR4 Simulations, Bull. Am. Meteorol. Soc., 89, 347–367, doi:10.1175/BAMS-89-3-347, 2008.
Fasullo, J. T. and Trenberth, K. E.: A less cloudy future: The role of subtropical subsidence in climate sensitivity, Science, 338, 792–794, 2012.
Field, C. B., Barros, V., Stocker, T. F., and Qin, D.: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation, edited by: Field, C. B., Barros, V., Stocker, T. F., and Dahe, Q., Cambridge University Press, Cambridge, 2012.
Fu, Q., Banerjee, A., Liess, S., and Snyder, P.: Drought Detection of the Last Century: An MRF-based Approach, in: Proceedings of the 2012 SIAM International Conference on Data Mining, p. 11., 2012.
Gaffney, S. J., Robertson, A. W., Smyth, P., Camargo, S. J., and Ghil, M.: Probabilistic clustering of extratropical cyclones using regression mixture models, Clim. Dynam., 29, 423–440, 2007.
Gall, M., Borden, K. A., Emrich, C. T., and Cutter, S. L.: The Unsustainable Trend of Natural Hazard Losses in the United States, Sustainability, 3, 2157–2181, doi:10.3390/su3112157, 2011.

NPGD
1, 51–96, 2014

Ganguly, A. R. and Bras, R. L.: Distributed quantitative precipitation forecasting using information from radar and numerical weather prediction models, J. Hydrometeorol., 4, 1168–1180, 2003.

Ganguly, A. R., Gama, J., Omitaomu, O. A., Gaber, M., and Vatsavai, R. R. (Eds.): Knowledge discovery from sensor data, CRC Press, 215 pp., 2009a.

Ganguly, A. R., Steinhaeuser, K., Erickson, D. J., Branstetter, M., Parish, E. S., Singh, N., Drake, J. B., and Buja, L.: Higher trends but larger uncertainty and geographic variability in 21st century temperature and heat waves, Proc. Natl. Acad. Sci. USA, 106, 15555–15559, doi:10.1073/pnas.0904495106, 2009b.

Ganguly, A. R., Kodra, E., Chatterjee, S., Banerjee, A., and Najm, H. N.: Computational Data Sciences for Actionable Insights on Climate Extremes and Uncertainty, Computational Intelligent Data Analysis for Sustainable Development, Chap. 5, 1127–1156, 2013.

Ganguli, P. and Ganguly, A. R.: Severity-Duration-Frequency curves of meteorological droughts over continental United States, Abstract No. H44C-04, presented at 2013 Fall Meeting, AGU, San Francisco, Calif., 9–13 December, 2013.

García-Marín, A. P., Ayuso-Muñoz, J. L., Jiménez-Hornero, F. J., and Estévez, J.: Selecting the best IDF model by using the multifractal approach, Hydrol. Process., 27, 433–443, doi:10.1002/hyp.9272, 2013.

Gerber, M. and Joos, F.: An Ensemble Kalman Filter multi-tracer assimilation: Determining uncertain ocean model parameters for improved climate-carbon cycle projections, Ocean Model., 64, 29–45, 2013.

Ghosh, S.: SVM-PGSL coupled approach for statistical downscaling to predict rainfall from GCM output, J. Geophys. Res., 115, D22102, doi:10.1029/2009JD013548, 2010.

Ghosh, S., Das, D., Kao, S.-C., and Ganguly, A. R.: Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes, Nat. Clim. Chang., 2, 86–91, doi:10.1038/nclimate1327, 2011.

Giannakis, D. and Majda, A. J.: Quantifying the predictive skill in long-range forecasting. Part I: Coarse-grained predictions in a simple ocean model, J. Climate, 25, 1793–1813, 2012.

Goswami, B. N., Venugopal, V., Sengupta, D., Madhusoodanan, M. S., and Xavier, P. K.: Increasing trend of extreme rain events over India in a warming environment, Science, 314, 1442–1445, doi:10.1126/science.1132027, 2006.

NPGD
1, 51–96, 2014

Hannachi, A., Jolliffe, I. T., and Stephenson, D. B.: Empirical orthogonal functions and related techniques in atmospheric science: A review, Int. J. Climatol., 27, 1119–1152, doi:10.1002/joc.1499, 2007.

Hayhoe, K., Sheridan, S., Kalkstein, L., and Greene, S.: Climate change, heat waves, and mortality projections for Chicago, J. Great Lakes Res., 36, 65–73, doi:10.1016/j.jglr.2009.12.009, 2010.

Heikkilä, U., Sandvik, A., and Sorteberg, A.: Dynamical downscaling of ERA-40 in complex terrain using the WRF regional climate model, Clim. Dynam., 37, 1551–1564, doi:10.1007/s00382-010-0928-6, 2010.

Hulme, M., Pielke, R., and Dessai., S.: Keeping prediction in perspective, Nat. Reports Clim. Chang., 0911, 126–127, doi:10.1038/climate.2009.110, 2009.

Jung, T., Miller, M. J., Palmer, T. N., Towers, P., Wedi, N., Achuthavarier, D., Adams, J. M., Altshuler, E. L., Cash, B. A., Kinter, J. L., Marx, L., Stan, C., and Hodges, K. I.: High-Resolution Global Climate Simulations with the ECMWF Model in Project Athena: Experimental Design, Model Climate, and Seasonal Forecast Skill, J. Climate, 25, 3155–3172, doi:10.1175/JCLI-D-11-00265.1, 2012.

Kao, S.-C. and Ganguly, A. R.: Intensity, duration, and frequency of precipitation extremes under 21st-century warming scenarios, J. Geophys. Res., 116, D16119, doi:10.1029/2010JD015529, 2011.

Karamperidou, C., Cane, M. A., Lall, U., and Wittenberg, A. T.: Intrinsic modulation of ENSO predictability viewed through a local Lyapunov lens, Clim. Dynam., 42, 253–270, doi:10.1007/s00382-013-1759-z, 2014.

Kaspi, Y. and Schneider, T.: Winter cold of eastern continental boundaries induced by warm ocean waters., Nature, 471, 621–624, doi:10.1038/nature09924, 2011.

Kates, R. W., Travis, W. R., and Wilbanks, T. J.: Transformational adaptation when incremental adaptations to climate change are insufficient., Proc. Natl. Acad. Sci. USA, 109, 7156–7161, doi:10.1073/pnas.1115521109, 2012.

Kawale, J., Steinbach, M., and Kumar, V.: Discovering Dynamic Dipoles in Climate Data., in SDM, 107–118, available at: http://siam.omnibooksonline.com/2011datamining/data/papers/321.pdf, 2011.

Kawale, J., Liess, S., Kumar, A., Steinbach, M., Snyder, P., Kumar, V., Ganguly, A. R., Samatova, N. F., and Semazzi, F.: A graph-based approach to find teleconnections in climate data, Stat. Anal. Data Min., 6, 158–179, doi:10.1002/sam.11181, 2013.

NPGD
1, 51–96, 2014

Kerr, R. A.: Forecasting regional climate change flunks its first test, Science, 339, 638 pp., 2013.
Khan, S., Ganguly, A. R., Bandyopadhyay, S., Saigal, S., Erickson III, D. J., Protopopescu, V., and Ostrouchov, G.: Nonlinear statistics reveals stronger ties between ENSO and the tropical hydrological cycle, Geophys. Res. Lett., 33, L24402, doi:10.1029/2006GL027941, 2006.
Khan, S., Bandyopadhyay, S., Ganguly, A. R., Saigal, S., Erickson III, D. J., Protopopescu, V., and Ostrouchov, G.: Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data, Phys. Rev. E, 76, 026209, doi:10.1103/PhysRevE.76.026209, 2007.
Kharin, V. V. and Zwiers, F. W.: Estimating extremes in transient climate change simulations, J. Climate, 18, 1156–1173, 2005.
Kharin, V. and Zwiers, F.: Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations, J. Climate, 20, 1419–1444, doi:10.1175/JCLI4066.1, 2007.
Kharin, V. V., Zwiers, F. W., Zhang, X., and Hegerl, G. C.: Changes in Temperature and Precipitation Extremes in the IPCC Ensemble of Global Coupled Model Simulations, J. Climate, 20, 1419–1444, doi:10.1175/JCLI4066.1, 2007.
Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M.: The big data bootstrap, ArXiv Prepr. ArXiv12066415, available at: http://arxiv.org/abs/1206.6415, 2012.
Knutson, T. R., McBride, J. L., Chan, J., Emanuel, K., Holland, G., Landsea, C., Held, I., Kossin, J. P., Srivastava, A. K., and Sugi, M.: Tropical cyclones and climate change, Nat. Geosci., 3, 157–163, doi:10.1038/ngeo779, 2010.
Knutti, R.: The end of model democracy?, Clim. Change, 102, 395–404, doi:10.1007/s10584-010-9800-2, 2010.
Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G.A .: Challenges in Combining Projections from Multiple Climate Models, J. Climate, 23, 2739–2758, doi:10.1175/2009JCLI3361.1, 2010.
Kodra, E., Steinhaeuser, K., and Ganguly, A. R.: Persisting cold extremes under 21st-century warming scenarios, Geophys. Res. Lett., 38, L08705, doi:10.1029/2011GL047103, 2011.
Kodra, E., Ghosh, S., and Ganguly, A. R.: Evaluation of global climate models for Indian monsoon climatology, Environ. Res. Lett., 7, 014012, doi:10.1088/1748-9326/7/1/014012, 2012.
Koster, R. D., Suarez, M. J., and Heiser, M.: Variance and predictability of precipitation at seasonal-to-interannual timescales, J. Hydrometeorol., 1, 26–46, 2000.

NPGD
1, 51–96, 2014

Kovats, R. S. and Kristie, L. E.: Heatwaves and public health in Europe., Eur. J. Public Health, 16, 592–599, doi:10.1093/eurpub/ckl049, 2006.
Kuhn, G., Khan, S., Ganguly, A. R., and Branstetter, M. L.: Geospatial-temporal dependence among weekly precipitation extremes with applications to observations and climate model simulations in South America, Adv. Water Resour., 30, 2401–2423, doi:10.1016/j.advwatres.2007.05.006, 2007.
Kumar, D., Kodra, E., and Ganguly, A. R.: Regional and seasonal intercomparison of CMIP3 and CMIP5 climate model ensembles for temperature and precipitation, Clim. Dynam., doi:10.1007/s00382-014-2070-3, in press, 2014.
Kunkel, K. E., Karl, T. R., Easterling, D. R., Redmond, K., Young, J., Yin, X., and Hennon, P.: Probable maximum precipitation and climate change, Geophys. Res. Lett., 40, 1402–1408, doi:10.1002/grl.50334, 2013.
Langmead, B., Hansen, K. D., and Leek, J. T.: Cloud-scale RNA-sequencing differential expression analysis with Myrna., Genome Biol., 11, R83, doi:10.1186/gb-2010-11-8-r83, 2010.
Malik, N., Bookhagen, B., Marwan, N., and Kurths, J.: Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks, Clim. Dynam., 39, 971–987, doi:10.1007/s00382-011-1156-4, 2011.
Mannshardt-Shamseldin, E. C., Smith, R. L., Sain, S. R., Mearns, L. O., and Cooley, D.: Downscaling extremes: A comparison of extreme value distributions in point-source and gridded precipitation data, Ann. Appl. Stat., 4, 484–502, doi:10.1214/09-AOAS287, 2010.
Marty, C. and Blanchet, J.: Long-term changes in annual maximum snow depth and snowfall in Switzerland based on extreme value statistics, Clim. Change, 111, 705–721, doi:10.1007/s10584-011-0159-9, 2011.
Mastrandrea, M. D., Tebaldi, C., Snyder, C. W., and Schneider, S. H.: Current and future impacts of extreme events in California, Clim. Change, 109, 43–70, doi:10.1007/s10584-011-0311-6, 2011.
Min, S.-K., Zhang, X., Zwiers, F. W., and Hegerl, G. C.: Human contribution to more-intense precipitation extremes, Nature, 470, 378–381, doi:10.1038/nature09763, 2011.
Mishra, V., Smoliak, B. V, Lettenmaier, D. P., and Wallace, J. M.: A prominent pattern of year-to-year variability in Indian Summer Monsoon Rainfall., Proc. Natl. Acad. Sci. USA, 109, 7213–7217, doi:10.1073/pnas.1119150109, 2012.
Monteleoni, C., Schmidt, G. A., and McQuade, S.: Climate informatics: accelerating discovering in climate science with machine learning, Comput. Sci. Eng., 15, 32–40, 2013.

NPGD
1, 51–96, 2014

Muzy, J., Bacry, E., and Kozhemyak, A.: Extreme values and fat tails of multifractal fluctuations, Phys. Rev. E, 73, 66114, doi:10.1103/PhysRevE.73.066114, 2006.
Narisma, G. T., Foley, J. A., Licker, R., and Ramankutty, N.: Abrupt changes in rainfall during the twentieth century, Geophys. Res. Lett., 34, L06710, doi:10.1029/2006GL028628, 2007.
Negahban, S. N. and Wainwright, M. J.: Simultaneous Support Recovery in High Dimensions: Benefits and Perils of Block $\ell 1/\ell \infty$-Regularization, IEEE Trans. Inf. Theory, 57, 3841–3863, doi:10.1109/TIT.2011.2144150, 2011.
Negahban, S. N., Ravikumar, P., Wainwright, M. J., and Yu, B.: A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers, Stat. Sci., 27, 538–557, doi:10.1214/12-STS400, 2012.
O'Gorman, P. A. and Schneider, T.: The physical basis for increases in precipitation extremes in simulations of 21st-century climate change, Proc. Natl. Acad. Sci. USA, 106, 14773–14777, doi:10.1073/pnas.0907610106, 2009.
Overpeck, J. T., Meehl, G. A., Bony, S., and Easterling, D. R.: Climate data challenges in the 21st century, Science, 331, 700–702, doi:10.1126/science.1197869, 2011.
Pall, P., Aina, T., Stone, D. A., Stott, P. A., Nozawa, T., Hilberts, A. G. J., Lohmann, D., and Allen, M. R.: Anthropogenic greenhouse gas contribution to flood risk in England and Wales in autumn 2000, Nature, 470, 382–385, doi:10.1038/nature09762, 2011.
Palmer, T. N., Doblas-Reyes, F. J., Hagedorn, R., Alessandri, A., Gualdi, S., Andersen, U., Feddersen, H., Cantelaube, P., Terres, J.-M., Davey, M., Graham, R., Délécluse, P., Lazar, A., Déqué„ M., Guérémy, J.-F., Díez, E., Orfila, B., Hoshen, M., Morse, A. P., Keenlyside, N., Latif, M., Maisonnave, E., Rogel, P., Marletto, V., and Thomson, M. C.: Development of a European multimodel ensemble system for seasonal-to-interannual prediction, Bull. Am. Meteorol. Soc., 85, 853–872, doi:10.1175/BAMS-85-6-853, 2004.
Phatak, A., Bates, B. C., and Charles, S. P.: Statistical downscaling of rainfall data using sparse variable selection methods, Environ. Model. Softw., 26, 1363–1371, doi:10.1016/j.envsoft.2011.05.007, 2011.
Pierce, D. W., Barnett, T. P., Santer, B. D., and Gleckler, P. J.: Selecting global climate models for regional climate change studies, Proc. Natl. Acad. Sci. USA, 106, 8441–8446, doi:10.1073/pnas.0900094106, 2009.

NPGD
1, 51–96, 2014

Pierce, D. W., Das, T., Cayan, D. R., Maurer, E. P., Miller, N. L., Bao, Y., Kanamitsu, M., Yoshimura, K., Snyder, M. A., Sloan, L. C., Franco, G., and Tyree, M.: Probabilistic estimates of future changes in California temperature and precipitation using statistical and dynamical downscaling, Clim. Dynam., 40, 839–856, doi:10.1007/s00382-012-1337-9, 2012.

Reddy, M. J. and Ganguli, P.: Spatio-temporal analysis and derivation of copula-based intensity–area–frequency curves for droughts in western Rajasthan (India), Stoch. Environ. Res. Risk Assess., 27, 1975–1989, doi:10.1007/s00477-013-0732-z, 2013.

Reshef, D. N., Reshef, Y. A, Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C.: Detecting novel associations in large data sets, Science, 334, 1518–24, doi:10.1126/science.1205438, 2011.

Robertson, A., Kirshner, S., and Smyth, P.: Downscaling of daily rainfall occurrence over north-east Brazil using a hidden Markov model, J. Climate, 17, 4407–4424, 2004.

Robertson, A. W., Kirshner, S., Smyth, P., Charles, S. P., and Bates, B. C.: Subseasonal-to-interdecadal variability of the Australian monsoon over North Queensland, Q. J. Roy. Meteor. Soc., 132, 519–542, 2006.

Ryu, J. and Hayhoe, K.: Understanding the sources of Caribbean precipitation biases in CMIP3 and CMIP5 simulations, Clim. Dynam., doi:10.1007/s00382-013-1801-1, 2013.

Sain, S. R., Furrer, R., and Cressie, N.: A spatial analysis of multivariate output from regional climate models, Ann. Appl. Stat., 5, 150–175, doi:10.1214/10-AOAS369, 2011.

Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., and Nolan, G. P.: Computational solutions to large-scale data management and analysis, Nat. Rev. Genet., 11, 647–657, doi:10.1038/nrg2857, 2010.

Sheffield, J., Wood, E. F., and Roderick, M. L.: Little change in global drought over the past 60 years, Nature, 491, 435–438, 2012.

Schmith, T.: Stationarity of Regression Relationships: Application to Empirical Downscaling, J. Climate, 21, 4529–4537, doi:10.1175/2008JCLI1910.1, 2008.

Schneider, S. H. and Kuntz-Duriseti, K.: Uncertainty and climate change policy, in: Climate change policy: a survey, edited by: Schneider, S. H., Rosencranz, A., and Niles, J. O., p. 368, Island Press., 2002.

Simmons, K., Sutter, D., and Pielke, R. A.: Blown away: monetary and human impacts of the 2011 US tornadoes, in: Extreme Events and Insurance: 2011 Annus Horribilis, Vol. 5, edited by: Courbage, C. and Stahel, W. R., p. 147, The Geneva Reports: Risk and Insurance Research, 2012.

NPGD
1, 51–96, 2014

Smith, R. L., Tebaldi, C., Nychka, D., and Mearns, L. O.: Bayesian Modeling of Uncertainty in Ensembles of Climate Models, J. Am. Stat. Assoc., 104, 97–116, doi:10.1198/jasa.2009.0007, 2009.
Smyth, P., Ide, K., and Ghil, M.: Multiple Regimes in Northern Hemisphere Height Fields via MixtureModel Clustering, J. Atmos. Sci., 56, 3704–3723, 1999.
Song, S. and Singh, V. P.: Meta-elliptical copulas for drought frequency analysis of periodic hydrologic data, Stoch. Environ. Res. Risk Assess., 24, 425–444, doi:10.1007/s00477-009-0331-1, 2010.
Stainforth, D., Kettleborough, J., Allen, M., Collins, M., Heaps, A., and Murphy, J.: Distributed computing for public-interest climate modeling research, Comput. Sci. Eng., 4, 82–89, doi:10.1109/5992.998644, 2002.
Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., Kettleborough, J. A., Knight, S., Martin, A., Murphy, J. M., Piani, C., Sexton, D., Smith, L. A., Spicer, R. A., Thorpe, A. J., and Allen, M. R.: Uncertainty in predictions of the climate response to rising levels of greenhouse gases., Nature, 433, 403–406, doi:10.1038/nature03301, 2005.
Steinhaeuser, K. and Tsonis, A. A.: A climate model intercomparison at the dynamics level, Clim. Dynam., 1–6, doi:10.1007/s00382-013-1761-5, 2013.
Steinhaeuser, K., Chawla, N. V., and Ganguly, A. R.: Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science, Stat. Anal. Data Min., 4, 497–511, doi:10.1002/sam.10100, 2011a.
Steinhaeuser, K., Ganguly, A. R., and Chawla, N. V.: Multivariate and multiscale dependence in the global climate system revealed through complex networks, Clim. Dynam., 39, 889–895, doi:10.1007/s00382-011-1135-9, 2011b.
Steinhaeuser, K., Ganguly, A. R., and Chawla, N. V.: Multivariate and multiscale dependence in the global climate system revealed through complex networks, Clim. Dynam., 39, 889–895, 2012.
Sterk, A. E., Holland, M. P., Rabassa, P., Broer, H. W., and Vitolo, R.: Predictability of extreme values in geophysical models, Nonlin. Processes Geophys., 19, 529–539, doi:10.5194/npg-19-529-2012, 2012.
Sugiyama, M., Shiogama, H., and Emori, S.: Precipitation extreme changes exceeding moisture content increases in MIROC and IPCC climate models, Proc. Natl. Acad. Sci. USA, 107, 571–575, doi:10.1073/pnas.0903186107, 2010.

NPGD
1, 51–96, 2014

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, Bull. Am. Meteorol. Soc., 93, 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012.
Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective, J. R. Stat. Soc. Ser. B, 73, 273–282, doi:10.1111/j.1467-9868.2011.00771.x, 2011.
Tippett, M. K., Sobel, A. H., and Camargo, S. J.: Association of U.S. tornado occurrence with monthly environmental parameters, Geophys. Res. Lett., 39, L02801, doi:10.1029/2011GL050368, 2012.
Towler, E., Rajagopalan, B., Gilleland, E., Summers, R. S., Yates, D., and Katz, R. W.: Modeling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory, Water Resour. Res., 46, W11504, doi:10.1029/2009WR008876, 2010.
Trapp, R. J., Robinson, E. D., Baldwin, M. E., Diffenbaugh, N. S., and Schwedler, B. R. J.: Regional climate of hazardous convective weather through high-resolution dynamical downscaling, Clim. Dynam., 37, 677–688, doi:10.1007/s00382-010-0826-y, 2010.
Trelles, O., Prins, P., Snir, M., and Jansen, R. C.: Big data, but are we ready?, Nat. Rev. Genet., 12, 224 pp., doi:10.1038/nrg2857-c1, 2011.
Trenberth, K. E., Dai, A., van der Schrier, G., Jones, P. D., Barichivich, J., Briffa, K. R., and Sheffield, J.: Global warming and changes in drought, Nat. Clim. Change, 4, 17–22, 2014.
Turkman, K. F., Amaral Turkman, M. A., and Pereira, J. M.: Asymptotic models and inference for extremes of spatio-temporal data, Extremes, 13, 375–397, doi:10.1007/s10687-009-0092-8, 2009.
Van Huijgevoort, M. H. J., Hazenberg, P., van Lanen, H. A. J., and Uijlenhoet, R.: A generic method for hydrological drought identification across different climate regions, Hydrol. Earth Syst. Sci. Discuss., 9, 2033–2070, doi:10.5194/hessd-9-2033-2012, 2012.
Wadsworth, J. L. and Tawn, J. A.: Dependence modelling for spatial extremes, Biometrika, 99, 253–272, doi:10.1093/biomet/asr080, 2012.
Webster, P. J., Holland, G. J., Curry, J. A., and Chang, H.-R.: Changes in tropical cyclone number, duration, and intensity in a warming environment, Science, 309, 1844–1846, doi:10.1126/science.1116448, 2005.
Weigel, A. P., Knutti, R., Liniger, M. A., and Appenzeller, C.: Risks of Model Weighting in Multimodel Climate Projections, J. Climate, 23, 4175–4191, doi:10.1175/2010JCLI3594.1, 2010.
Wilby, R. L. and Keenan, R.: Adapting to flood risk under climate change, Prog. Phys. Geogr., 36, 348–378, doi:10.1177/0309133312438908, 2012.

NPGD
1, 51–96, 2014

Wilhite, D. A. and Glantz, M. H.: Understanding: the Drought Phenomenon: The Role of Definitions, Water Int., 10, 111–120, doi:10.1080/02508068508686328, 1985.

Yang, B.-W., Tsai, W.-C., Chen, A.-P., and Ramandeep, S.: Cloud Computing Architecture for Social Computing – A Comparison Study of Facebook and Google, in 2011 International Conference on Advances in Social Networks Analysis and Mining, IEEE, 741–745, 2011.

Zhang, L. and Singh, V. P.: Gumbel–Hougaard Copula for Trivariate Rainfall Frequency Analysis, J. Hydrol. Eng., 12, 409–419, doi:10.1061/(ASCE)1084-0699(2007)12:4(409), 2007.

Zorita, E. and Von Storch, H.: The analog method as a simple statistical downscaling technique: comparison with more complicated methods, J. Climate, 12, 2474–2489, 1999.

Zwiers, F. W., Alexander, L. V., Hegerl, G. C., Knutson, T. R., Kossin, J. P., Naveau, P., Nicholls, N., Christoph, S., Seneviratne, S. I., and Zhang, X.: Climate Extremes: Challenges in Estimating and Understanding Recent Changes in the Frequency and Intensity of Extreme Climate and Weather Events, in: Climate Science for Serving Society, edited by: Asrar, G. R. and Hurrell, J. W., Springer Netherlands, Dordrecht, 2013.

NPGD
1, 51–96, 2014

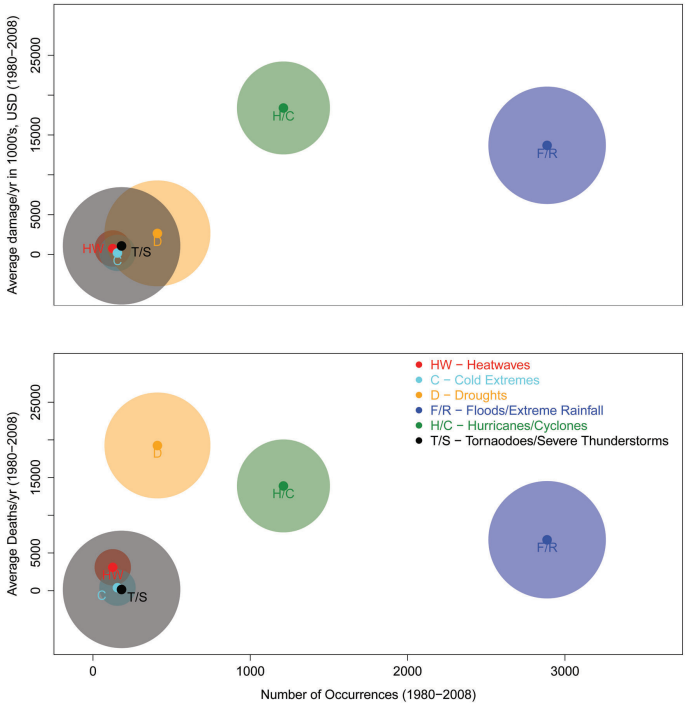


Fig. 1. Frequency and severity of hydrometeorological hazards, indicated as damages in US dollars (top) and annual fatalities (bottom), along with their uncertainties in our current understanding in a climate change context. Damage and fatality data are taken from the UN Office for Disaster Risk Reduction (UNISDR) at PreventionWeb (http://www.preventionweb.net/english/hazards/) averaged over 1980–2008. The uncertainties are represented by the size of the bubbles with larger circles indicative of greater uncertainty and are derived from the IPCC confidence and likelihood estimates (Field et al., 2012).

NPGD
1, 51–96, 2014

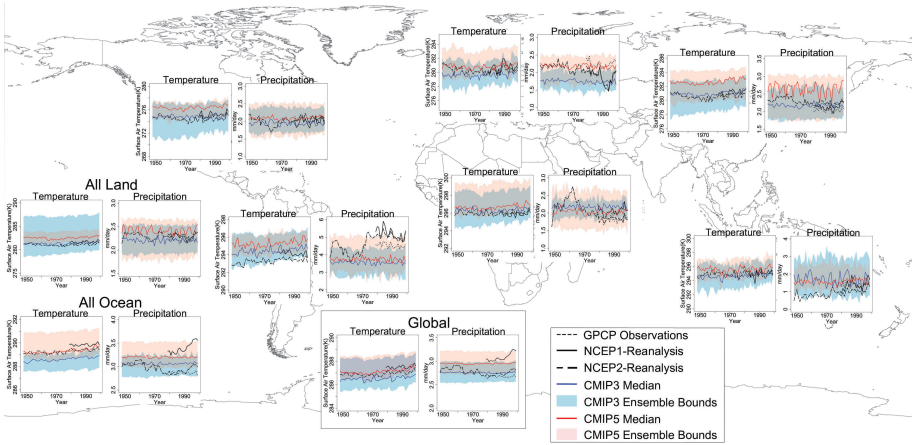
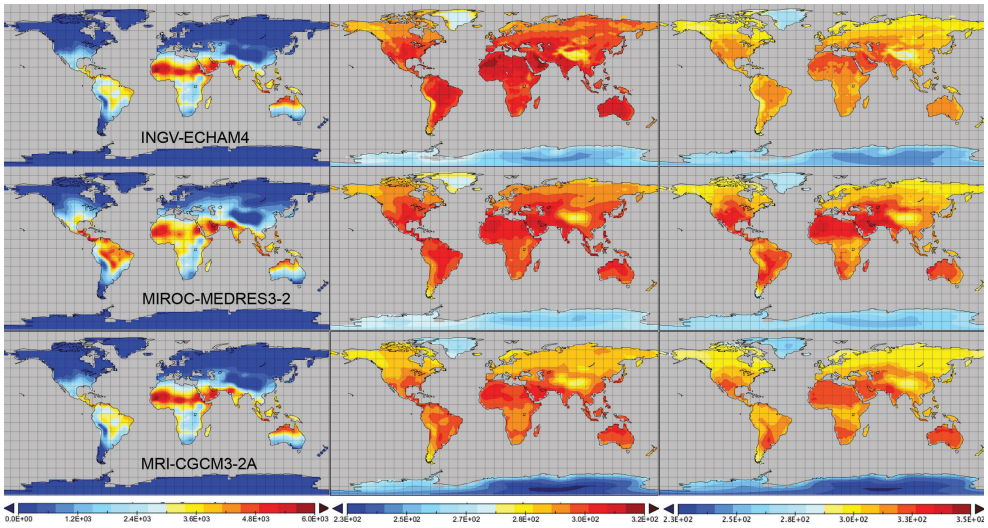


Fig. 2. Global climate models are designed to simulate the large-scale circulation of the atmosphere and its response to external forcing, and needs to be evaluated from that perspective. However, adaptation, impacts and vulnerability (IAV) studies occasionally use GCM model simulations directly, or after statistical and dynamical downscaling. Model-based assessments in the future need to ultimately rely on GCM projections. Nevertheless, relatively naïve utilization of GCM projections for IAV studies may yield non-informative or even misleading conclusions. This is illustrated though comparisons between the CMIP3 and CMIP5 climate model simulations at continental and global scales in terms of average temperature and precipitation. The two GCM-ensembles are evaluated against the observation-based NCEP/NCAR (NCEP-1) and NCEP/DOE (NCEP-2) reanalysis data. For precipitation, the Global Precipitation Climatology Project (GPCP) observational data is used in addition. Aggregate comparisons do not appear to suggest significant improvements of CMIP5 over CMIP3. While this does not necessarily imply a lack of improvement in CMIP5 over CMIP3 in terms of large-scale dynamics, this does suggest the need for caution when GCMs (with or without downcaling) need to be used for IAV studies.

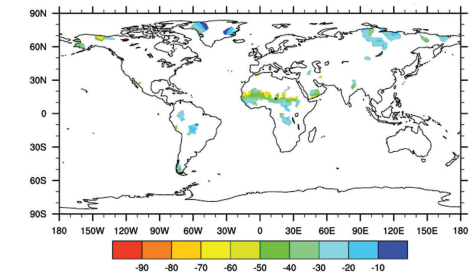
NPGD
1, 51–96, 2014



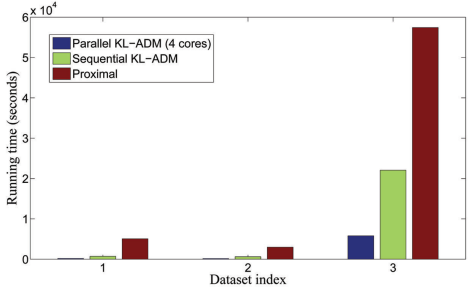
Average Annual Cooling Degree Days, 2081-2100

Average Heatwave Intensity (K), 2081-2100

30-year Maximum Temperature Event (K), 2081-2100



Precipitation 1970-1998: Anomaly in Percent



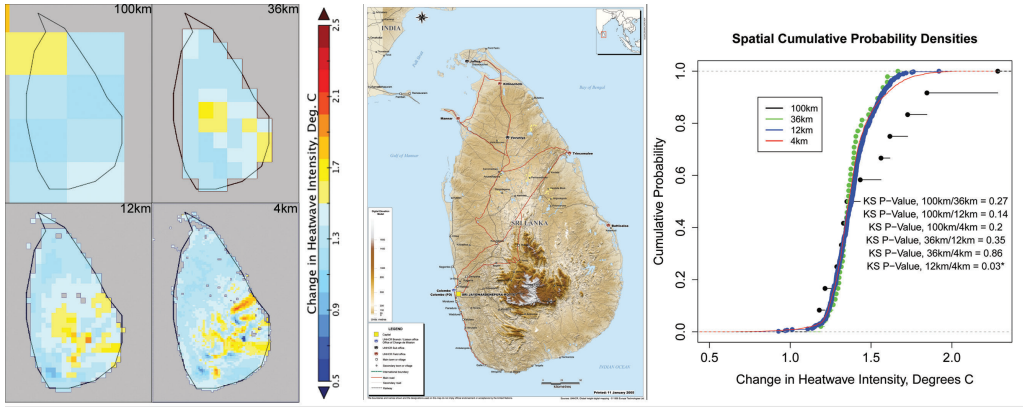
NPGD
1, 51–96, 2014

Physics-guided data mining techniques

A. R. Ganguly et al.

Fig. 3. Characterization of extremes from big climate data. (Top) Characterization of extremes requires a summarization of the statistical properties of climate-related observations and model-simulations. A thorough assessment of extremes from massive climate data may be especially challenging because the definitions of extremes definitions and indices may depend on stakeholder needs. Here we present three different choices: an energy-consumption related metric called Cooling Degree Days or CDD (left); a heatwave intensity index (Ganguly et al., 2009b) thought to be relevant for human mortality defined as consecutive nighttime minima events (middle); an index grounded in the statistical theory of extreme values (Kharin et al., 2007). The substantial regional differences suggest the differences in the nature of the insights. (Bottom) Novel data-driven approaches can help detect climate-related extremes, particularly ones like droughts that are especially difficult to characterize[43–44]. Our analysis (bottom left) suggests that Markov Random Field (MRF) based approaches may improve the detection process but traditional implementations may not scale to large data. We have developed a new, computationally efficient optimization solver to implement the MRF (Fu et al., 2012). As a proof-of-concept, here we show how the new method detects persistent and significant droughts over space and time. We used three popular methods (bottom right) to solve the same MRF inference. Our algorithm for characterizing droughts, 'KL-ADM', is approximately one order of magnitude faster than an existing popular routine called 'Proximal' (dark red) and much faster than any commercially available software (e.g., IBM ILOG CPLEX Optimizer, http://www-01.ibm.com/software/integration/optimization/cplex-optimizer/). The first (second) dataset (x axis) is a simulated dataset with 100 000 (200 000) variables and 293 500 (586 000) two-way relationships among them, where each variable can take on 3 (4) possible values. The third dataset is the Climate Research Unit precipitation dataset, which has 7 146 520 variables (i.e., points in space) and each can take on 2 possible values (drought or no drought). This example clearly shows a significant speedup in computation using KL-ADM, especially with four parallel processors.

NPGD
1, 51–96, 2014



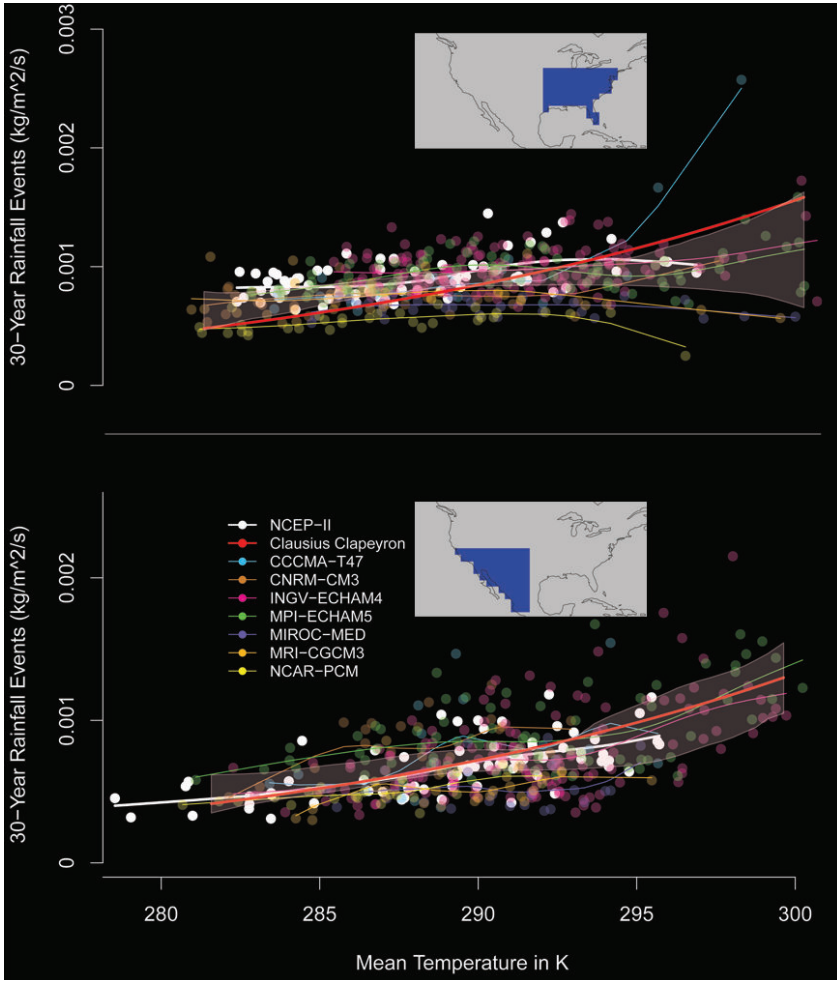
NPGD
1, 51–96, 2014

Physics-guided data mining techniques

A. R. Ganguly et al.

Fig. 4. Precise projections with downscaling generate bigger data. (Top) An example of dynamical downscaling that shows changes in heatwave intensity over Sri Lanka from 2006–2015 to 2056–2065 projected by the Community Climate System Model Version 4 (CCSM4) at 100 km spatial resolution and dynamically downscaled by the Weather Research and Forecasting (WRF) model at three successively enhanced spatial resolutions: 36, 12, and 4 km. Dynamical downscaling yields significant computational challenges. A goodness-of-fit comparison, via the Kolmogorov-Smirnov (KS) tests, does not yield substantial evidence for differences in spatial distributions of model runs, which is probably owing to small sample sizes for the 100 and 36 km resolution data. However, the effects of topography in the mid-southern Sri Lanka appear more prominent at higher resolutions. The sheer size of the newly generated dynamically downscaled simulations, as well as the problem complexity, further intensifies the need for Big Data solutions. (Bottom) Statistical downscaling is complementary to dynamical downscaling and usually requires significantly less computational resources. Here we perform statistical downscaling by relating fine-resolution rainfall observations with a large set of climate model-simulated variables and using the relation first to validate on unseen data and then for precise projections in the future. The results show how an ensemble of five runs of the CCCMA CGCM3.1 T47 global climate model is validated for 20th century simulations (left), and then applied to the SRES A2 scenario for 2070–2099 (right). Geographical heterogeneity in the trends of rainfall extremes over India, shown in a recent observation-based study, is suggested after downscaling but not directly from the global model runs (Ghosh et al., 2011).

NPGD
1, 51–96, 2014



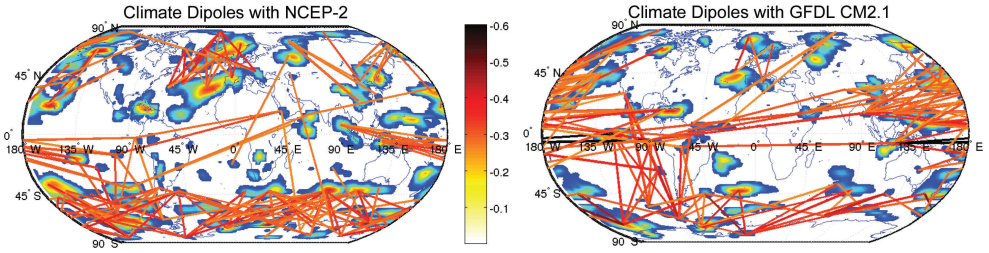
NPGD
1, 51–96, 2014

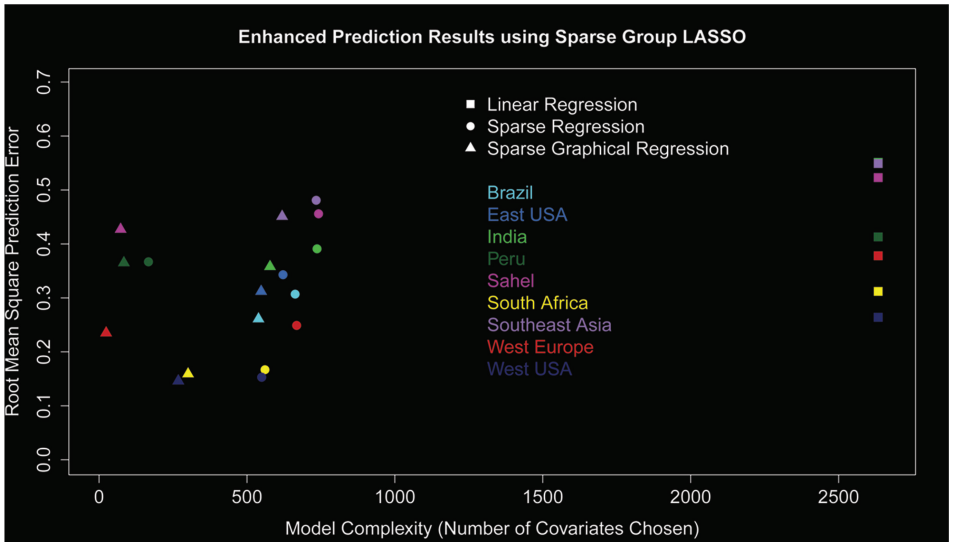
Physics-guided data mining techniques

A. R. Ganguly et al.

Fig. 5. Uncertainty quantification adds to the Big Data challenge. Multimodel ensembles have been used to quantify uncertainty in the structural representation of climate physics; their performance has been evaluated by investigating skills in reproducing historical behavior (skills) and multimodel agreement (convergence) in the future. Here we investigate the uncertainty in precipitation extremes and explore whether physically based relations, like the temperature-dependence of precipitation extremes through the saturation vapor pressure (known as the Clausius-Clapeyron, or CC, relation), may help further inform uncertainty assessments and skill-based model selection. For the southwestern and southeastern United States, a 7-member CMIP3 model ensemble is used for the analysis, with NCEP2 used as a baseline model and the theoretical CC curve shown for comparison. Every point from each model represents a 20 yr mean temperature (1980–1999) on the x axis and a 30 yr rainfall (1980–1999, y axis) with nonlinear regressions fit to each dataset and uncertainty bounds computed using a bootstrap-based resampling procedure. The value of using the multivariate physically based CC relation in uncertainty quantification is suggested, particularly for extremes (specifically, heavy rainfall) where covariate relations (specifically, temperature-dependence) are known from process physics (e.g., Clausius-Clapeyron).

NPGD
1, 51–96, 2014





NPGD
1, 51–96, 2014

Physics-guided data mining techniques

A. R. Ganguly et al.

Fig. 6. Predictive insights from observations and model simulations. (Top) Dipoles were detected using sea level pressure (SLP) from NCEP2 reanalysis and the GFDL CM2.1 global climate model from CMIP3 (from left to right, respectively) from the year 1979–2000. Dipoles are a class of teleconnections, or long-range dependence in space, that represent a persistent and large-scale temporal negative correlation in a given climate variable between two neighboring or distant geographical locations. The dipoles shown here are generated using the shared reciprocal nearest neighbors (SRNN) algorithm graphical approach (Kawale et al., 2013). The edges of the graph, shown in the figure, represent dipole connections between two regions, while the color in the background shows the SRNN density, where darker colors signify regions of higher connectivity. This class of methods may be useful for systematically detecting, refining existing, or even identifying new, climate teleconnections or oscillations, as well as for model evaluation. (Bottom) While critical for statistical downscaling and relating ocean-based indices to regional land climatology, regression problems in climate may be particularly difficult to solve reliably, owing to issues like high-dimensionality (large input variables compared to the number of calibration or training data), proximity-based spatial correlations, and teleconnections. Here we use multiple ocean variables (as predictors or covariates) to predict changes in land precipitation for multiple regions using NCEP1 reanalysis. The results indicate that a new approach, called the Sparse Group Lasso (SGL: Chatterjee et al., 2012), outperforms ordinary least squares and LASSO regressions (Tibshirani, 2011) as per both error-based predictive accuracy and model parsimony (i.e., the number of covariates selected for prediction). Model parsimony refers to simpler models with lesser number of parameters, which in turn tend to generalize better than more complex models, especially if predictive accuracy on training data remains identical or also gets lower. Where climate extremes of interest (e.g., hurricanes or rainfall extremes) are projected less reliably but relate to variables (i.e., potential covariates) that are better projected (e.g., oceanic or land temperature), methods such as the SGL and future innovations may enhance projections beyond model-simulations alone. We note that the ordinary least squares (OLS) approach serves only as a very naïve baseline for this analysis. In the OLS approach shown here, the number of covariates selected is simply all covariates considered; OLS intrinsically assigns non-zero coefficients to all covariates. With so many covariates, almost certainly the OLS model will have nonsensical non-zero parameters due to issues like multi-collinearity. We acknowledge the fact that procedures like stepwise least squares may improve on the naïve OLS reduce shown here by reducing the dimensionality of the problem. However, forward versus backward versus mixed stepwise procedures have their own set of

NPGD
1, 51–96, 2014

problems related to multi-collinearity and changes in coefficients with addition or removal of covariates, among others. Still, we present the naïve all-covariate OLS purely as a baseline for comparison without implying that it is or should be used in high-dimensional problems of this nature.

NPGD
1, 51–96, 2014