Towards understanding dominant processes in complex dynamical systems: Case of precipitation extremes

Debasish Das[1,2], Auroop Ganguly[1], Arindam Banerjee[3], Zoran Obradovic[2]
[1]Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA
[2]Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA, USA
[3]Computer Science and Engineering, University of Minnesota, Minneapolis, MN, USA
d.das@neu.edu, a.ganguly@neu.edu, banerjee@cs.umn.edu, zoran.obradovic@temple.edu

ABSTRACT

Complex dynamical systems like precipitation extremes under climate variability or change are typically governed by multiple processes at multiple scales. The processes themselves may be manifested at multiple scales and would need to be captured through key indicator variables, which in turn may be better projected by physical models than the variables of interest. We posit that hybrid approaches based on physically-motivated approaches and data-driven methods, which in turn are conditioned on both observations and simulations from large-scale physics-based models, may offer novel and quantifiable insights. The data-driven approaches may need to extend and adapt methods developed and tested in statistics, data mining and machine learning to the concept of dominant processes. In this paper, we performed some exploratory data analysis to characterize the effect of dominant processes on precipitation extremes, annually and seasonally, and from global and century scale to regional and decadal scale, and found some interesting insights that pointed towards need of improved understanding. We identified the gaps in understanding the regional drivers where data-driven methods can make useful improvements and eventually lead to a predictive model for precipitation extremes. Although we do not propose any specific method for solving the problem, we realize that any successful data mining solution should include all or a subset of tools like dimensionality reduction, grouped variable selection, non-linear regression and graphical models. The concepts of dominant processes proposed here would likely generalize broadly to climate extremes while the solution frameworks themselves may generalize beyond climate.

Keywords

Precipitation Extremes, Dominant Processes, Predictive Models, Graphical Models

1. INTRODUCTION

Improved understanding as well as precise and accurate prediction of precipitation and other climate extremes is essential for informing preparedness and policy, but progress is often limited by long-standing gaps in the science, non-stationary behavior, complex dependence and nonlinear dynamical generating processes [16,18,21,26,28]. The continued availability of massive data in recent years from remote or in-situ sensor-based observations as well as large-scale physics-driven computational models offer new challenges and opportunities. Nonlinear dynamical systems may exhibit sensitivity to initial conditions as well as short decorrelation space-time scales, thresholds, and intermittences, which limit the applicability of data-driven methods based solely on observations. Computational models which rely on physical mechanisms may not be able to adequately capture processes at sub-grid scales or mechanistic behavior that are manifested in the aggregate owing to multiple confounding factors [21,23]. The non-stationary nature of climate variability of change, combined with the need for long lead times projections to inform preparedness and policy, impose further restrictions on the ability to generalize data-driven or physics-model based projections.

On the other hand, variables of interest such as precipitation extremes exhibit mechanistic dependence on auxiliary variables such as atmospheric and oceanic temperatures, which in turn may be better projected by models. The temperature of an atmospheric column is directly related to the saturation vapor pressure through a physical relation referred as the Clausius-Clapeyron [27], which in turn dictates the amount of water vapor that can evaporate into the column and which in turn may ultimately drain out as rain. Temperature profiles and updraft velocities in the column relate to atmospheric instabilities and the propensity of convection processes, which govern precipitation extremes [25,27]. Ultimately, the availability of water in the atmosphere and the surrounding areas dictates the total evaporation. These dominant processes impact one another, and are in turn impacted by less well measured or understood variables and processes ranging from properties and volume of aerosols to detailed cloud physics. An entire atmospheric column is subject to translation based on prevailing motions, which in turn may be guided in a climate context by wind vectors and topography.

We examine the extent to which data-driven and physics-based methods may be able to delineate the dominant processes of precipitation extremes as well as quantify their inter-relations and help construct a predictive model. The predictive model would use the physics embedded within the large-scale computational models by ensuring the mechanisms are conditioned on relatively better-predicted model-simulated variables, the physical or conceptual insights at multiple scales that may not be well-captured in the models, as well as an overall data-driven methodology that would attempt to understand the contributions and interactions of the dominant processes besides statistically

SensorKDD '12 Beijing, China
Copyright 2012 ACM 978-1-4503-1554-8/12/08 ...$10.00.

modeling the influence of the less-understood processes. We present preliminary ideas, problems definitions, data analysis results, and solution frameworks for such a dominant process model which brings together known physics with data-driven methods.

The ability to extend statistical or machine learning concepts to dominant processes may be necessary to advance our understanding, eventually leading to predictive models. We hypothesize that such hybrid approaches and the concept of dominant processes may be crucial to help bridge gaps in the science of climate extremes, leading to new paths forward for informed adaptations or mitigation decision. The solution framework proposed here may generalize not only to other climate extremes but beyond climate to areas where dominant but interacting processes operating at multiple scales govern the evolution of variables in complex dynamical systems.

2. MOTIVATION

One of the largest gaps in climate science relevant for informing stakeholders and policy makers is the inability to develop credible projections for extreme events and regional change at spatiotemporal scales that matter to decision-makers. Precipitation and their extremes are particularly relevant in this context because of the deep science challenges as well as the widespread impacts on flood hazards and water resources decisions. The data mining literature in climate applications have tended to focus on teleconnections (long-range spatial dependence), especially on oceanic influence over regional land climatology [18, 19, 20]. While teleconnections are important, local and regional atmospheric conditions typically tend to dominate in the context of climate-related extremes. However, precipitation and regional climate prediction have been highlighted as two of four real holes in climate science [5], which makes the prediction of precipitation or their extremes at local to regional scales a major challenge.

Large-scale climate models solve a system of partial differential equations (PDEs) based on first principles, but also contain parameterizations for processes that are not so well understood. Unfortunately, processes pertaining to precipitation are among the least well understood and precipitation is not a state variable in the PDEs. In addition, precipitation is known to be extremely variable in space and time and the underlying processes are subject to thresholds and intermittences. However, as pointed out in the literature [1,2,3,22,23,24], precipitation extremes tend to have a dependence on atmospheric variables ranging from temperature, humidity and precipitable water, to updraft velocity and horizontal wind components. These atmospheric variables, which can be thought of as potential covariates for precipitation extremes, are often better predicted than precipitation itself.

One promising recent approach has been the development of physically-based approaches [1,2,3] which attempt to exploit the above relations to relate the atmospheric covariates with precipitation extremes through what could be viewed as hypothesis-guided approaches. While these approaches have demonstrated significant promise, they may not be able to leverage the full information content in atmospheric covariates and translate these to predictive insights, primarily because they have to rely on known physics-based hypotheses. Additionally, these approaches are only concerned about explaining the precipitation extremes at a global level and over a multi-decadal scale and therefore still lack relevance towards informing the regional water resources policy decisions. In order for these approaches to be successful at the regional level they need to acknowledge the regional and seasonal variability in dependence of precipitation extremes on the dominant processes. For example, at the global level temperature influences the probability of occurrence of a precipitation extreme since temperature controls both convection and capacity of air to hold water (quantified by saturation vapor pressure). However, at regional level the dependence may be different on two different processes. In a region where overall water availability is less, a condition that can induce high amount of convection is not enough to cause an extreme rainfall. Similarly, a high amount of water availability in air may not necessarily result in an extreme if the condition for convection is non-existent.

A large number of prior works related to precipitation extremes involves finding global relation of precipitation extremes or it's change with ancillary climate variables [1,2,3,15,16,19]. Statistical trend analysis in precipitation extremes and it's attribution has been another focal point of related works [17,18,19,21]. The dominant processes concept has been introduced in hydrology domain recently [20]. Climate has emerging as a new field of application for data mining methods. Complex networks [10,11] and sparse regression [5,6,7,12] were shown to be two useful tools for estimating dependence structures between different climate variables and indices.

2.1 Scope of Work and Main Contributions

The main goals of this paper is to show, with a few motivating data analysis, (a) the gaps that remain in understanding the seasonal and regional variability in effects of dominant processes on the precipitation extremes and (b) how these problems can be mapped into data mining challenges and motivate new innovations over state-of-the-art in data mining methodologies. With some exploratory data analysis at different spatial and temporal aggregation level, we made following useful observations.

A. Existing knowledge about processes driving the increase of precipitation extremes is confirmed at a global level.

B. At the regional level, the existing knowledge cannot explain the increase in precipitation extremes, which seemed surprising initially.

C. Further data analysis at seasonal level revealed a completely different factor driving the increase.

D. This new understanding does not completely explain the pattern of precipitation increase in a different region.

With these interesting observations, we proceeded to make a case for the need of improved understanding of the processes that drives precipitation extremes. We did not attempt to provide any specific solutions; rather our goal is to make the data mining community aware of this important climate problem. In the next section, we will describe the datasets we used to perform the experiments we described in this paper. In the fourth section, we described the data pre-processing steps for estimating indicator variables from available climate data based on climate physics. In the fifth section, we described the gaps in understanding the relation between precipitation extremes and dominant processes with a few motivating examples whereas in sixth section we attempted to cast the climate problem as a data mining task.

3. DATASETS

The sheer volume of climate data available at different sources is quite staggering. However, it is often required to integrate these diverse datasets from different sources which is a challenging task due to wide variations in their quality, format and spatial and temporal resolutions. The climate datasets can be divided into three broad categories based on how they are generated. In the first category there are observations from the ground-based and satellite-based sensors. Observations are not available for the entire globe and their quality and resolution varies widely over different sources, regions and time-periods. In some cases the data is not even available publicly. However, in most cases observations are the closest to the ground-truth. In our experiments we used the precipitation observations from Climate Prediction Center (CPC) which are available at 0.25° x 0.25° grids [22] over entire US (see Table 1). The second category of climate datasets is simulated data from physics models generated independent of the observations. These are available over the entire globe and are mostly of uniform quality. There are a number of such models available that generate multiple climate variables having different spatial resolution. The third category of datasets is called "re-analysis" and they are generated from physics models that are forced to match with available observations. Reanalysis is therefore considered to be closer to the observations. Again there are different sources of reanalysis data among which we used the dataset generated from NCEP-NCAP (National Center for Environmental Physics - National Center for Atmospheric Research) reanalysis project. This particular dataset is available over entire globe at 2.5° x 2.5° spatial resolution from 1948 to present (see Table 1).

Among the above three categories of datasets, only the physics models generate future projections which are not available for the other two due to obvious reason. Also we needed to interpolate the CPC observation to coarser 2.5° x 2.5° spatial resolution to match the reanalysis datasets. On the other hand, we could only use the reanalysis over US since the CPC observations are available over US only.

Table 1: Description of the datasets used in the experiments

	Temporal Resolution	Spatial Resolution	Availa-bility	Variables Used
Observa-tion	Daily	.25° x .25°	US	Precipitation
Physics Models (MIROC)	Daily	1.125° x 1.125°	Global	Precipitation, Temperature
Reanalysis	Daily	2.5° x 2.5°	Global	Precipitation, Temperature

4. PHYSICS GUIDED PREPROCESSING

We describe, in detail, the dominant processes that dictate the occurrence of precipitation extremes at regional level and how their indicators are computed. Not all of them are equally important in influencing extremes and their importance changes drastically over different region and season. Primary climate variables like temperature, relative humidity, horizontal wind velocities etc. that are generated directly from climate models, are available at multiple vertical pressure levels for each grid-point and time instant thereby increasing the dimensionality of the prediction problem mentioned before. The process of generating the indicator variables for the dominant processes may be seen as one form of dimensionality reduction guided by known climate physics. However, we may still need further data-driven dimensionality reduction and feature selection techniques to develop the final predictive model.

4.1 Clausius-Clapeyron (CC) Relationship

CC relation [1,2,23] governs the change of water holding capacity of a parcel of air with temperature. Higher temperature causes the air to expand making room for additional water vapor. Precisely, under a constant relative humidity condition, the rate of change of saturation vapor pressure [23], a measure of water holding capacity of air, as a function of temperature is dictated by the CC relation, which is estimated from temperature using the following equation.

$$e_s(T) = 6.1094exp\left(\frac{17.625T}{T+243.04}\right) \qquad (1)$$

where e_s is the saturation vapor pressure in *hPa* (hecto-Pascal) and T is the temperature in *celsius* scale. So, we computed this indicator from temperature only which is available from the reanalysis dataset that we used. Since T is available at multiple pressure levels at any given grid-point, we will get as many values of saturation vapor pressure for each location. We averaged these values over pressure level to come up with a single aggregate measure of SVP for each location.

Now, given that enough water is available in the surface for evaporation, the rate of increase of precipitable water in the air will follow the rate of increase of saturation vapor pressure, which for typical lower tropospheric temperatures, is estimated to be around 7% [1,2]. Intuitively, the amount of precipitable water in the air should directly influence the amount of rainfall in a region. However, it has been shown in the climate literature [1,2] that at high percentiles, extreme rainfalls far exceed the prediction by CC relation. The value of the threshold changes based upon region (e.g. over different latitudes) and time. This clearly points towards changing dynamics of dominant processes for the precipitation extremes over precipitation mean process.

4.2 Convection

Most of the evaporated moistures are generally concentrated at the lower levels of atmosphere to be eventually carried up as a result of convective processes. Radiation from the sun heats up the surface and the adjacent lower layers of the air causing the air to be expanded and lighter. This condition creates an imbalance in potential energy between the layers within a column of air, and potentially initiates a convection process within the column [25]. Convection can also be initiated by a parcel of air saturated with lighter than air water vapor at the lower layers of the atmosphere. Convection is an essential condition for precipitation and convective rainfalls often lead to precipitation extremes in certain region. Here we describe two indicators of convection process that can be estimated using standard equations.

4.2.1 Convective Available Potential Energy

Although both temperature and water vapor content of a column of air can be an indicator of existence of convection, a more informative measure is Convective Available Potential Energy (CAPE) [24, 25] which is given by the following equation.

$$\mathrm{CAPE} = \sum_{LFC}^{EL}\left[\frac{T_{vp}-T_{ve}}{T_{ve}}\vec{g}\right]dz \qquad (2)$$

Where LFC is the layer of free convection and EL is the equilibrium layer [24], T_{vp} and T_{ve} respectively are the *virtual temperatures* [24, 25] of the parcel of the air under consideration and its neighboring layers at different pressure levels [24], g is the acceleration due to gravity and z is the height of the pressure levels [24] which is more commonly known as the *geopotential height*. Following are short descriptions of some of the terms used above. The virtual temperatures can be computed from temperature and relative humidity.

LFC: This layer of air below which the convective upward motion of air is inhibited by negatively buoyant energy which is generally a result of the parcel of air being cooler than the layers surrounding it. Beyond this layer the convection occurs freely.

EL: This is layer where the equilibrium is reached between the parcel of air and its surrounding and no convection occurs beyond this layer.

Virtual Temperature: It is the temperature of a moist parcel of air at which a theoretical dry air parcel would have a total pressure and density equal to the moist parcel of air.

We computed CAPE within a column of air over a grid-point from available temperature data at pressure levels, relative humidity, geo-potential height, elevation at that grid-point and surface pressure [24].

4.2.2 Updraft Velocity

A second indicator of convection is updraft velocity of a parcel of air which is often referred to as "*omega*" in climate literature. There is no direct way to measure this variable and it is not a primary state variable in the climate models. Therefore it is generally computed from east-ward (U-wind) and south-ward (V-wind) horizontal winds by integrating the continuity equation over vertical axis (equation 3).

$$\omega(\mathrm{p}) = \omega(\mathrm{p_s}) - \int_{\mathrm{p_s}}^{\mathrm{p}}\left(\frac{\partial \mathrm{u}}{\partial \mathrm{x}}+\frac{\partial \mathrm{v}}{\partial \mathrm{y}}\right).\mathrm{dp} \qquad (3)$$

Omega is also available as a separate variable in the reanalysis dataset; however, reanalysis does not provide future projections. Omega is available or computed at vertical pressure levels and therefore need to be aggregated over pressure levels for each location. Either of the two indicators (CAPE and Omega) described above gives a good estimate of the convective process. However, equation (2) and (3) only give an approximate estimate of the corresponding quantities and therefore cannot be relied upon as the sole estimate of convection. So, it will be wise to use as many independent indicators of the process as possible (two in this case).

Both CC relation and convection influence the precipitation extremes within a column of air and they have significant inter-dependence. So, when we consider the processes that influence the precipitation extremes, we should also consider the interaction between CC relation and convection as a separate influencing process since neither of the processes can separately captures this dependence. The interaction can be represented by the product of indicator variables for each of the processes.

4.3 Water Availability

High water holding capacity of air and convection cannot guarantee rainfall if there is not enough water available on the surface that can be evaporated into the atmosphere. The water may evaporate from sources like large water bodies on land, vegetation and soil moisture. Unlike CC relation and convection, influence of water availability is not limited to a column of air over a single grid-point. Surface water from neighboring grid-points can also evaporate into the air and come down as rain. It is hard to measure water availability on the surface and it is not a quantity that is made available by any of the climate models. Therefore the indicator variable that is used to represent this quantity is the precipitable water in an entire column of water which is estimated by integrating relative humidity over all pressure levels in a vertical column. Precipitable water is available as a separate variable within the reanalysis dataset. However, the choice of a neighborhood within which the total amount of precipitable water can come down as rain at a grid-point is difficult one since the neighborhood can be different for different grid-points.

4.4 Translation

All the processes we discussed above are generally confined within a column of air and influence the rainfall at a specific grid-point or within a limited neighborhood. However the process of translation can carry an entire column of air horizontally over to a completely different location. So if there is a development of atmospheric conditions conducive to rainfall within a column of air over a certain location, it may not still rain at that location, if the entire column of water is carried away by a translation process. It may cause a rainfall extreme in the new location instead. The process of translation may be controlled by the horizontal wind velocity (U-wind and V-wind) and therefore we use them as indicator variables for the process of translation. However, the process of translation can be quite complex due to presence of multiple small scale movement within the larger scale transport mechanisms.

Note that, although we specified indicator variables for each of the processes mentioned above, they cannot be taken as the proxy for the corresponding processes. The processes can be regarded as unobserved latent variables which have some dependence on the available indicator variables. Indicator variables themselves in most cases are non-linear functions of primary atmospheric variables that are generated by the climate models with a certain degree of accuracy.

4.5 Interaction between Processes

There can be interaction between these processes which may influence the precipitation extremes as well. One way to compute the indicator for these interactions is to simply taking the product of the indicator variables of the interacting processes at pressure levels and integrating them over pressure levels. For example, saturation vapor pressure and convection may have some dependence between then and their interaction term can be defined as

$$I = \int_p \omega(p).\frac{\partial q_s}{\partial p} \qquad (4)$$

Where q_s is saturation specific humidity, a quantity closely related to saturation vapor pressure and the integral is taken over entire troposphere. For this quantity to be meaningful, some constraints need to be satisfied which we omitted here. Please refer to [2] and [3] for more details.

5. A MOTIVATING CASE STUDY: TEMPERATURE DEPENDENCE

The Clausius-Clapeyron relationship mandates that the saturation vapor pressure (a measure of the water holding capacity) of the air increases by approximately 7.5% with a unit Kelvin increase in temperature. Although at a global level, the intensity of precipitation follow the same rate of increase at the relatively lower percentiles of the rainfall distribution, at the higher (> 90) percentiles, the rate of increase far exceeds 7.5% [1,2]. This means, at the lower percentiles, CC relation may dominate rainfall events but a combination of events might dominate the extreme precipitations. We performed some preliminary experiments on the model-simulated climate datasets (detailed description of the dataset provided below) and the results show that when we go at the regional level, a combination of processes influence a rainfall extreme rather than a single process. Moreover, the combination may change over seasons even for a single location.

In the first experiment, we attempted to reproduce the results from [1] which shows the fractional increase in precipitation percentiles for unit increase in average temperature over future projections of a climate model. In order to generate this plot, we collected daily precipitation and temperature projections between 1981-2000 and 2081-2100 from the MIROC model [1]. Let us denote the average temperature over a target region (we used global, tropical and extra-tropics) for these 20-yr periods as T_{20} and T_{21} respectively and let ΔT be their difference, i.e. $\Delta T = T_{21}- T_{20}$. Now we computed the values of higher percentiles (90-99.999) of precipitation for each of these intervals. Let us denote the *i*-th percentile of precipitation for each of these intervals as $P_{20}^{(i)}$ and $P_{21}^{(i)}$ respectively. So the fractional increase in *i*-th percentile per unit increase in average temperature over the 100 year interval is then given by $[(P_{21}^{(i)}- P_{20}^{(i)})/ P_{20}^{(i)}]/ \Delta T$. Figure 1 shows these fractional increases for different values of percentiles *i* for entire globe, tropical and extratropical regions.

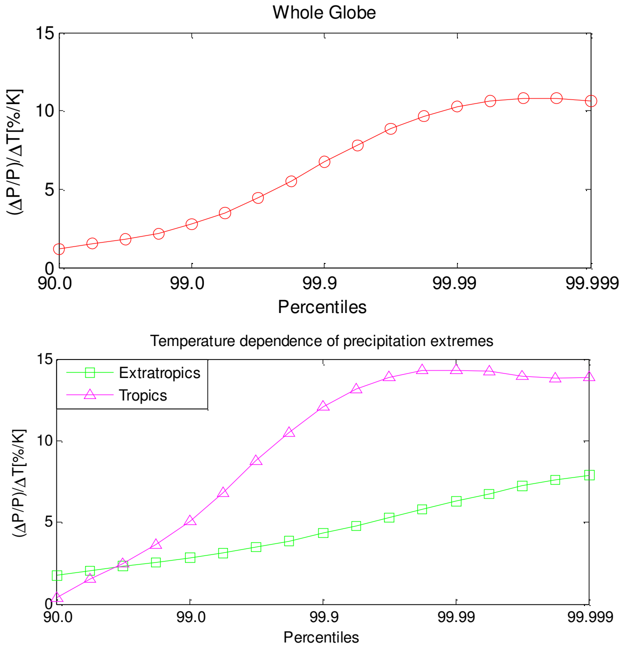


Figure 1. Fractional increase in precipitation extremes percentiles per unit increase in average temperature between 1981-2000 and 2081-2100

There are at least two interesting information that can be inferred from these global plots. Firstly, even at the global scale, higher percentiles of extreme precipitations increase at a faster rate with increasing temperature and in the tropical region the rate of increase near higher percentiles far exceeds the rate estimated from the CC relation. This hints at the presence of significant amount of convective rainfall which tends to cause extremes more than they cause average precipitation. Secondly, we can also see significant differences exist between the rates of increase in tropical and sub-tropical regions. In tropical regions, extremes increase at a higher rate than the extra-tropics.

In Figure 2, we showed the same metric for temperature dependence of precipitation extremes over latitudinal regions, this time for 90th and 99th percentile of rainfalls. They show wide variations over latitudes and higher variance for higher percentile.

With this motivating example we would like to emphasize the importance of regional level knowledge about which processes dominates the rainfall at different regions and at different times of the year. Just by looking at the global distribution of precipitation extremes it is hard for one to guess the possible difference between tropical and extra-tropics unless one is made aware of the fact that convective processes dominate rainfalls in the tropics. Similarly, the knowledge of rate of increase in precipitation over a hundred years and over entire tropical region (or extra-tropical region) is not enough for the stakeholders. In order to obtain more useful information for policymakers and resource managers, we need to be able to understand which processes are dominating the precipitation extremes at a much higher spatial and temporal resolution.

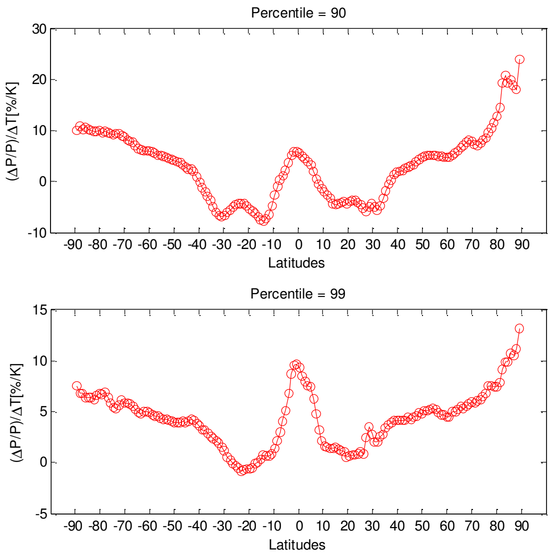


Figure 2. Latitudinal variation of fractional increase in precipitation extremes percentiles per unit increase in average temperature between 1981-2000 and 2081-2100

To begin with, we focused at a much smaller region of north-west US. First we generated the same plot showing increase in extremes per unit increase in average temperature only for NW US using the model data. Next, we attempted to understand the dependence of precipitation extremes on increase of temperature at the level of individual precipitation events. For this purpose we used the daily precipitation and temperature between 1948-2006 from the reanalysis dataset since they are closer to the actual observations than the model simulations. Here we followed a more direct method to estimate the dependence of precipitation extremes on temperature. We selected all the precipitation events over a certain percentile threshold P_t and the corresponding temperatures. Now for each selected event, we computed the difference ΔP_i between corresponding value of precipitation and the threshold percentile value. We also computed the difference ΔT_i between corresponding temperature and temperature at the precipitation threshold. Now we considered the gradient of the line fitted on values $\{\Delta P_i/P_t, \Delta T_i\}$ to be an estimator of CC-relation-driven increase in precipitation extremes per unit increase in temperature. In Figure 3, we show scatter plots of $\{\Delta P_i/P_t, \Delta T_i\}$ for percentile values 90 and 99.99. Although the precipitation extremes increase with temperature initially, after a certain temperature it seems to decrease with increasing temperature. At a first glance this may seem counter-intuitive; however, a closer look at the weather patterns of a major city in NW US (Seattle) explains this anomaly.

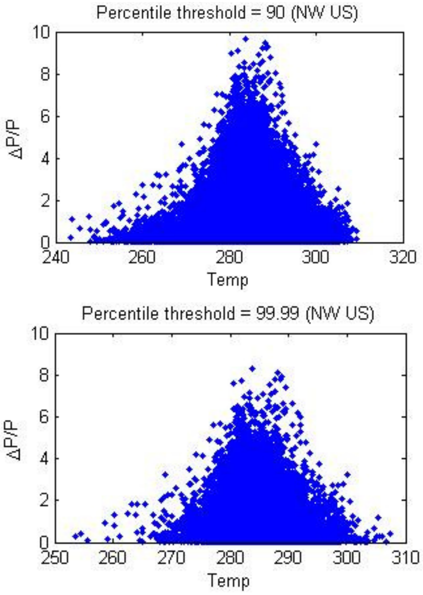


Figure 3. Scatter plots of fractional increase in precipitation extremes vs increase in temperature for two percentile thresholds 90 and 99.9

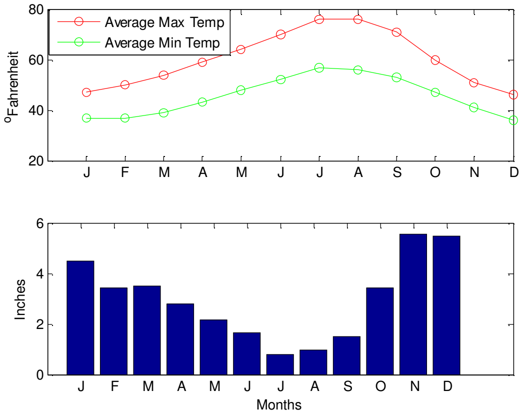


Figure 4. Average monthly precipitation and temperature in Seattle, WA. (Data obtained from weather.com)

We can see that Seattle has a hot and dry summer. Therefore, even though temperature is high in summer allowing for more evaporation, there is not enough water that can be evaporated into the atmosphere. So, the water availability dominates the occurrence of extreme rainfalls in summer over CC-relation. To distinguish between the effects of temperature and water availability on precipitation extremes, we attempted to investigate the rainy season separately from dry season since during the dry seasons (e.g. for Seattle it is the summer months) there is not enough water available in the atmosphere. However, a closer look at the precipitation patterns at individual grid-points revealed two different rainfall patterns in the region. One set of grid-points experience dry summer and the other set of grid-points receive more rainfall in summer than in other months.

So, we separated these two types of region and examined them separately. We followed a simple algorithm for separating two types of behavior which is listed in Table 2 and the grid-points classified according to the rainfall pattern shown in Figure 5 over a plot of topography in the region.

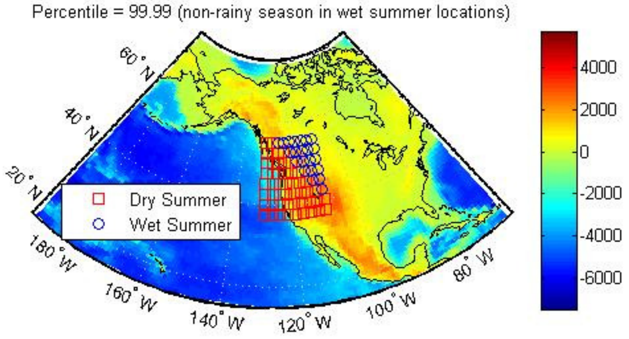


Figure 5. Distribution of grid-points having dry summer and wet summer

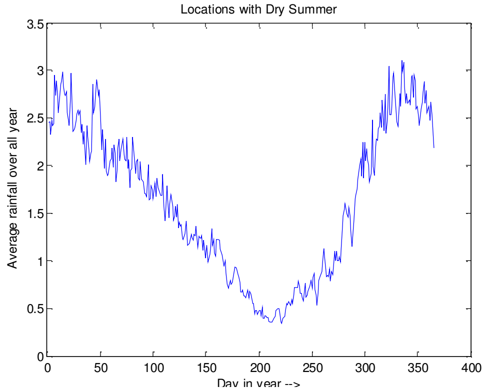


Figure 6: Distribution of average daily rainfalls over grid-points with dry summer.

The grid-points with similar precipitation patterns seem to be spatially cohesive and the demarcation line between these two regions seems to be influenced by the topography. This hints towards topography being one more regional driver of precipitation patterns.

Now we focus on each of two regions and investigate the rainy and non-rainy regions separately. In Figure 6, we show the distribution of average daily rainfalls for region with dry summer (similar to Seattle) whereas in Figure 7, we show the scatterplots of $\{\Delta P_i/P_t, \Delta T_i\}$ with percentile values 99.99 for the dry summer region separately.

Figure 7 clearly shows that during rainy seasons, majority of extreme precipitation events follows an increasing trend with increasing temperature whereas during dry season (summer) majority of extreme events follow a decreasing trend with increasing temperature since days with higher temperatures happens to be drier leading to less precipitation extremes. We also performed similar analysis for the southeast US, which did not reveal any surprising trend (Figure 8) which can be attributed to the fact that this region experiences major rainfall during the summer months.

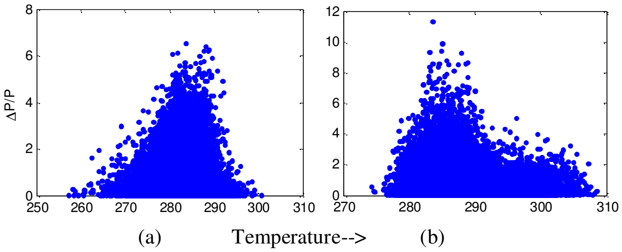


Figure 7. Scatter plot showing temperature dependence of precipitation extremes for (a) rainy and (b) non-rainy seasons in dry summer locations in NW US (percentile =99.99)

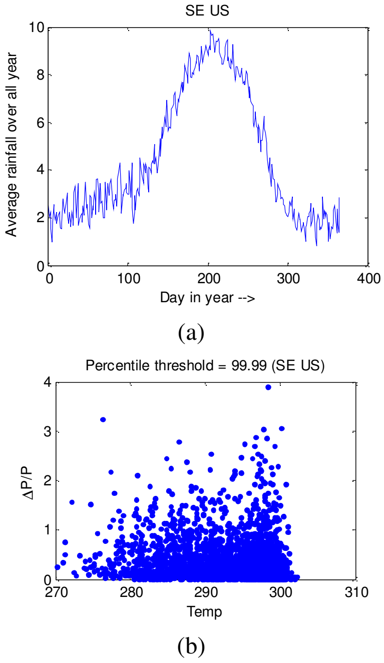
6. Data Mining Challenges and Future Research Directions

From the motivating example we described in the previous section, it is obvious that in order to predict precipitation extremes with a certain level of precision and accuracy that is useful to the policymakers; we need an improved understanding of precipitation extremes at the regional level.

The climate system is nonlinear dynamical (even chaotic) and non-stationary, while projections are sought for long lead-times. Thus, physics-based climate models, which in turn have become very computationally expensive, are essential. Purely data-guided methods are not best suited for multi-step ahead prediction under these circumstances, thus the problem is not cast here as a standard data-mining prediction problem. On the other hand, climate models by themselves may not be adequate, especially for critical challenges like precipitation extremes. Now based on the hypothesis that some variable that are well predicted by the physics models carry information about the precipitation extremes, our final goal is to develop a predictive model that will predict the probability of future occurrences and/or intensity of a precipitation extreme given the future projections of the variables that are well-predicted by the physics model. Since the physics models generate projections for these variables for hundreds of years in the future, the expectation is that the predictive model would also predict precipitation extremes for the distant future as well.

Climate is a highly non-linear and complex dynamical system and therefore multi-step ahead predictions are extremely difficult without any knowledge of the non-stationarity involved in the underlying physical processes. We assume that the inherent non-stationarity and multi-decadal low-frequency oscillations involved in the climate systems are handled best by the informed physics models and therefore are accounted for in the future projections of well-predicted variables. So, to that end, we need to better understand the relation between these variable with precipitation extremes based on past climate data, both simulated and observed, and be able to exploit the information content within these variables about precipitation extremes. This is exactly where the use of data-driven methods could be appropriate.

Now, even the relation between these variables and precipitation extremes are highly non-linear and exhibits high amount of spatial and temporal variability. Furthermore, these variables do not influence the extremes directly. Rather, they influence the dominant processes described above which in turn creates the conditions conducive to the occurrence of extremes. We cannot directly observe the actual strength of these processes over a region during a certain time-period and therefore they can be considered as latent variables. However we can compute one or more indicator variable(s) for each of these processes using the equations introduced in section 4 from the known physics. These equations already accounts for some of the nonlinearities present in the target predictive model. As a next step, we either need to adopt existing data mining methods or need to develop entirely novel data mining algorithms for improving our understanding of the spatial and temporal variability in dependence between precipitation extremes and the dominant processes. In the next subsections we will attempt to delineate the solution framework for this problem.



(a)

(b)

Figure 8. (a) Average daily rainfall and (b) Scatter plot showing temperature dependence of precipitation extremes for all seasons in SE US

6.1 Process Understanding

The relation between the indicator variables and precipitation extremes is highly nonlinear and they are related through the unobserved dominant processes. In order to improve our understanding of the dominant processes driving the precipitation extremes, we need to learn the above relation as accurately as possible. However, this problem is not trivial and may require application of sophisticated non-linear dimensionality reduction and multi-manifold embedding techniques on the input space of indicator variables to be able to visualize any structure. A simple 3-D plot (Figure 9) of precipitation extremes with respect to three different indicator variables aggregated over all grid-points in NW US region does not show any perceptible relation whatsoever.

One should also remember that above relation varies over regions and seasons and therefore it may be wise to consider the indicator variables at the smallest possible grid-points and over different season as separate features, which will make the input space really high-dimensional. Consequently, we may need sparse regression techniques and graphical models to estimate the underlying graphical structure among these features. On the other hand, we may have to resort to some suitable aggregation on these features at different levels depending on at what level the large scale dominant processes operate.

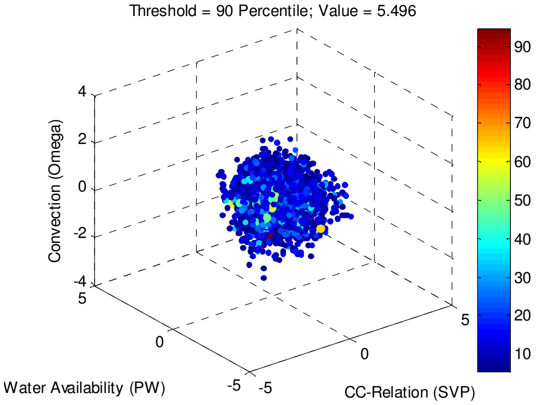


Figure 9: Plot of precipitation extremes (events over 90th percentile of wet days) with respect to indicator CC-relation, convection and water availability. Colors and size of dots are proportional to the strength of extremes

6.2 The Prediction Problem

Let X_1, X_2, …,X_M be the variables that are well predicted by the physics models; let $\mathbf{I}_1$,…. $\mathbf{I}_R$ be the indicator variables for different dominant processes, where each process may be affected by one or more of the X_i variables; and let Z_1, Z_2, … Z_R be the latent variables representing the strengths of different dominant processes affecting the observed rainfall y. One possible dependency relation between these variables is shown using a graphical model in Figure 10 to demonstrate the complexity of the problem. Now depending on what assumptions are made about the scales of the dominant processes and their mutual dependence, this structure may change.

Here the shaded variables are observed and the transparent variables are unobserved. The variable S is a switch variable that determines whether a rainfall will be considered extreme or not. S=1 will mean an extreme has occurred and S=0 will mean otherwise. This variable is required since the nature of dependence between precipitation and dominant processes changes depending on whether the rainfall event is an extreme or not. The rectangular plate indicates the repeatability of the entire structure over space and time [29]. Also, we have some knowledge about the non-linear dependence between the first two layers from physics which we described in section 4, but they are still not accurate. None of the observed variables are free of noise. Few challenges that need to be addressed are listed below:

1) At what spatio-temporal resolution should the prediction model operate? Maximum possible resolution is limited by the resolution of the observed and simulated data. But predictive models operating at the highest resolution might be unrealistic since sometimes the dominant processes operate at a much larger scale.

2) How can we handle the spatio-temporal variability in the predictive models?

3) Is it possible to cluster the grid-points over regions and seasons so that these clusters show uniform behavior in terms of their dependence on the dominant processes?

4) What is the best way to handle the state variable S that determines whether a precipitation event is an extreme? This is important because there is no universal threshold to distinguish between extremes and non-extremes. This threshold varies over regions and seasons and may even change over time. Furthermore this threshold separates two sets of precipitation events which may well be generated by different data generating processes. Assigning a prior over the threshold values may be a better idea than assuming a firm deterministic threshold.

5) Should we assume continuous (strength of process) or a binary (presence of process) latent variable to describe the dominant processes?

Note that, although our goal is to be able to predict precipitation extremes in future, one may need to make some assumption and approximations to simplify the problem to start with.

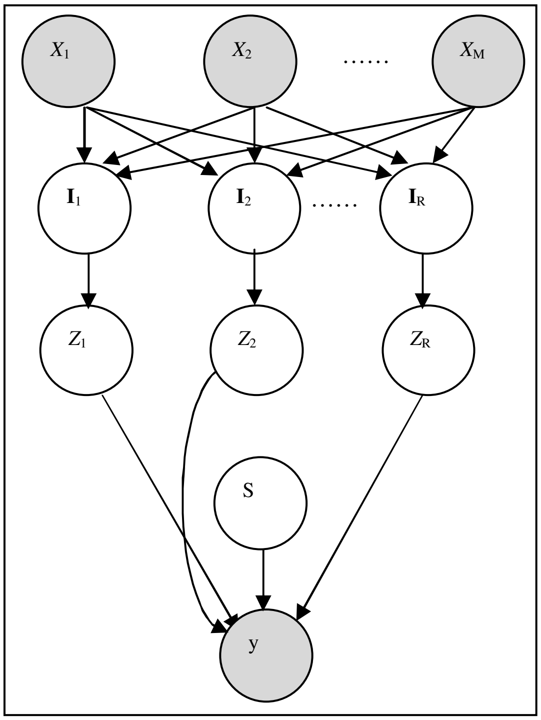


Figure 10: Graphical model showing the dependency between different variables.

7. Conclusion

In this paper we presented an important climate problem of better understanding and predicting precipitation extremes from future projections of some other related climate variables. To this end, we introduced the concepts of dominant processes that are primary drivers of the precipitation extremes and they are indicated by non-linear functions of primary climate variables. We performed some exploratory data analysis that confirmed the existing knowledge about relation of precipitation extremes with increasing temperature [1,2] at a global scale. However, we also discovered, with some exemplary case studies focused on specific regions, that this relationship does not solely explain the increase in precipitation extremes at regional level. We performed further data analysis at the seasonal level, which revealed interesting insights about completely different driver of extremes.

With the above motivating examples, we argued the need for improved understanding of the regional drivers of the precipitation extremes and tried to point out where data mining methods might be useful in mitigating the gaps in understanding. We also tried to cast the relevant parts of the problem as a data mining task. However, there can be other alternative ways in which this problem can be cast. Furthermore, we tried to demonstrate the fact that any useful data mining solutions for prediction of precipitation extremes, a result of complex dynamical processes, should adopt physics-based domain knowledge in order to remain interpretable and justifiable in terms of the processes driving the extremes.

Acknowledgements

This research is partially supported by NSF grant IIS-1029711, as well as a grant to ARG from the Nuclear Regulatory Commission and ARG's start-up funds from Northeastern University. We would like to thank Dr. Paul A. O'Gorman for helpful discussions.

8. References

[1] M. Sugiyama, H. Shiogama, and S. Emori. Precipitation extreme changes exceeding moisture content increases in MIROC and IPCC climate models. *PNAS.* 2010.

[2] P. A. O'Gorman and T. Schneider. The physical basis for increases in precipitation extremes in simulations of 21st-century climate change. *PNAS* 2009.

[3] P. A. O'Gorman and T. Schneider. Scaling of precipitation extremes over a wide range of climates simulated with an idealized GCM. *Journal of Climate* 22, 5676-5685

[4] Q. Schiermeier. The real holes in climate science. *Nature* (2010).

[5] X. Chen, Y. Liu, H. Liu, J. G. Carbonell. Learning Spatial-Temporal Varying Graphs with Applications to Climate Data Analysis. *AAAI.* 2010.

[6] A. C. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, C. Perlich, J. Hosking and N. Abe, Spatial-temporal causal modeling for climate change attribution. In *ACM SIGKDD*, 2009.

[7] Y. Liu, A. Niculescu-Mizil, A.C. Lozano and Y. Lu Learning Temporal Causal Graphs for Relational Time-Series Analysis, in *ICML.* 2010.

[8] Kalnay et al., The NCEP/NCAR 40-year reanalysis project, *Bull. Amer. Meteor. Soc.*, 1996.

[9] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, C. Potter, Discovery of climate indices using clustering, in *KDD* 2003,

[10] K. Steinhaeuser, N. V. Chawla and A. R. Ganguly, Complex Networks as a Unified Framework for Descriptive Analysis and Predictive Modeling in Climate Science. *Statistical Analysis and Data Mining*, 2010.

[11] J. Kawale et. al. Data Guided Discovery of Dynamic Climate Dipoles *NASA Conference on Intelligent Data Understanding (CIDU)*, 2011.

[12] S. Chatterjee, K. Steinhaeuser, A Banerjee, S. Chatterjee, A. Ganguly. Sparse Group Lasso: Consistency and Climate Applications. *SDM.* 2012

[13] P. C. D. Milly. Stationarity Is Dead: Whither Water Management?. *Science*, 2008.

[14] C. J. Muller and P. A. O'Gorman. An energetic perspective on the regional response of precipitation to climate change. *Nature Climate Change*, 2011.

[15] S. C. Liu et al. Temperature dependence of global precipitation extremes. *Geophysical Research Letters.* 2009.

[16] R. P. Allan and B. J. Soden. Atmospheric Warming and the Amplification of Precipitation Extremes. *Science*, 2008.

[17] V. V. Kharin et. al. Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. *Journal of Climate*, 2007.

[18] S. C. Kao, and A. R. Ganguly, Intensity, duration, and frequency of precipitation extremes under 21st-century warming scenarios, *Journal of Geophysical Research - Atmospheres*, 2011.

[19] R. E. Benestad, D. Nychka and L. O. Mearns, Spatially and temporally consistent prediction of heavy precipitation from mean values. *Nature Climate Change*, 2012

[20] B. Sivakumar, Dominant processes concept, model simplification and classification framework in catchment hydrology, *Stochastic Environmental Research and Risk Assessment,* 2008

[21] S. Ghosh, D. Das, S.-C. Kao, and A.R. Ganguly, (2012): Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes. *Nature Climate Change* 2012.

[22] CPC US Unified Precipitation data *provided by* NOAA/OAR/ESRL PSD, Boulder, Colorado, USA. website: http://www.esrl.noaa.gov/psd/

[23] P. Pall, M. R.. Allen and D. A. Stone. Testing the Clausius–Clapeyron constraint on changes in extreme precipitation under CO_2 warming. *Climate Dynamics*, 2007.

[24] K. Riemann-Campe, K. Fraedrich, F. Lunkeit, Global climatology of Convective Available Potential Energy (CAPE) and Convective Inhibition (CIN) in ERA-40 reanalysis. *Atmospheric Research.* 2009.

[25] K. A. Emanuel, Atmospheric Convection. Oxford University Press. 2005

[26] D. Coumou, S. Rahmstorf. A decade of weather extremes. *Nature Climate Change* 2012.

[27] C.F. Bohren, and B. Albrecht . *Atmospheric Thermodynamics*. Oxford University Press. 1998

[28] Field, C.B., Et Al., IPCC, Summary for Policymakers. In: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Special Report of the Intergovernmental Panel on Climate Change, 2012, Cambridge University Press, pp. 1-19.

[29] D. Koller, N. Friedman. Probabilistic Graphical Models. The MIT Press 2009.