Data Mining for Climate Change and Impacts

Auroop R Ganguly and Karsten Steinhaeuser
Geographic Information Science & Technology Group, CSE Division
Oak Ridge National Laboratory, Oak Ridge, TN, USA
{gangulyar, steinhaueskj} @ ornl.gov

Abstract

Knowledge discovery from temporal, spatial and spatiotemporal data is critical for climate change science and climate impacts. Climate statistics is a mature area. However, recent growth in observations and model outputs, combined with the increased availability of geographical data, presents new opportunities for data miners. This paper maps climate requirements to solutions available in temporal, spatial and spatiotemporal data mining. The challenges result from long-range, long-memory and possibly nonlinear dependence, nonlinear dynamical behavior, presence of thresholds, importance of extreme events or extreme regional stresses caused by global climate change, uncertainty quantification, and the interaction of climate change with the natural and built environments. This paper makes a case for the development of novel algorithms to address these issues, discusses the recent literature, and proposes new directions. An illustrative case study presented here suggests that even relatively simple data mining approaches can provide new scientific insights with high societal impacts.

1. Introduction

The Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC) [1] clearly points to anthropogenic greenhouse gas emissions as the cause of global warming. This has been possible by the analysis of massive volumes of observations from sensors as well as precise outputs from global-scale climate models.

Climate related observations from remote sensors like satellites and weather radars or from in situ sensors and sensor networks, as well as outputs of climate or earth system models from large-scale computational platforms, yield terabytes of temporal, spatial and spatiotemporal data. In addition, the rapid growth of geographical information systems implies availability of multi-source data to inform climate impacts analysis. However, the rate of data generation and storage far exceeds the rate of data analyses. This represents lost opportunities in terms of scientific insights not gained and impacts or adaptation strategies not adequately informed. While there is a mature literature in climate statistics and scattered applications of data mining, systematic efforts in climate data mining are lacking. The time is ripe for the spatial and spatiotemporal data mining (SSTDM) community to take a lead in this area. SSTDM deals with dependence of learning samples and auto- or cross-correlations. Climate data are geographical and hence inherit the spatial or temporal correlation properties. Additional challenges stem from nonlinear dependence, long memory processes in time, and long-range dependence or teleconnections in space. Post-AR4, the emphasis in climate research has shifted from global change at century scales to regional change and impacts at decadal scales. In particular, the need to develop anticipatory insights about extreme weather and hydrological events, as well as extreme hydro-meteorological stresses caused by regional change, has been recognized. The analysis results need to inform regional impacts assessments, which in turn use geographic data about the environment, land use, infrastructures and population. A major challenge in the analysis of extremes, regional change, and corresponding impacts, is the characterization of uncertainty for risk-informed decision making.

2. Motivation

According to [1], "*it is very likely that hot extremes, heat waves, and heavy precipitation events will continue to become more frequent"*. In addition to probable increase in intensity-duration-frequency (IDF) of extreme events and consequent exacerbation of natural hazards, [1] also mentions that regional

climate change is expected to cause stresses to the environment and society owing to increased temperatures and regional changes in precipitation patterns. Increase in global population, especially in the vulnerable regions of the world, may result in loss of human lives and reduction of living conditions, caused by acute scarcity of natural resources, greater damage from natural disasters, as well as large-scale migration. Climate change is expected to be a major contributor and/or exacerbate an already worsening situation in developing countries. Developed countries may have to face the brunt of the migration and may be called upon to provide disaster and humanitarian relief. The economic damage from weather or hydrologic extremes may actually be higher in developed nations because of greater exposed assets. In a climate change war game organized by the Center for a New American Security (CNAS), these very issues were debated through role playing. The war game was covered in a *Nature* news article [2], which also mentioned that the climate change scenarios [3] were provided by the Oak Ridge National Laboratory (ORNL). As a source of climate change assessments to multiple agencies, ORNL recognizes that perhaps one of the most significant challenges is the generation of credible information at local to regional scales for resource managers and policy makers. The core needs are the generation of predictive insights, risk management, and uncertainty characterization, with the ultimate aim of informing adaptation and mitigation decisions.

One of the primary climate models used by the IPCC is the Community Climate System Model version 3 (CCSM3) developed jointly by the National Center for Atmospheric Research (NCAR) and ORNL. Generating a set of outputs from a model like CCSM3 is a non-trivial computational exercise (e.g., [5]). The models are run from 1870 till now in "hindcast" mode and from 2000 to 2100 in projection mode. Future climate simulated by models depend on greenhouse gas emissions, which in turn depend on future socio-economic factors. These are captured through the IPCC Special Reports Emissions Scenarios (SRES) [1]: there are a total of 40 SRES scenarios. Ensemble runs are generated corresponding to each scenario, and take into account the facts that climate systems are potentially chaotic and hence sensitive to initial conditions or that model parameters are uncertain. Ensembles of multiple models are also available. The number of ensembles per scenario typically varies from a few to a few tens. The CCSM3 model generates over 100 output variables at spatial resolutions of 1.4 lat-long degree grids and 6-hour time resolutions for the surface and for several atmospheric layers, with global coverage. Thus, each run may correspond to a few terabytes of data and the total number of runs may run into several hundreds. The generation of risk profiles and uncertainty assessments at regional scales for the entire globe requires a detailed analysis of multiple model runs. The model outputs (hindcasts and prior projections) need to be compared with observations, which in turn may be obtained from satellites and remote sensors or historical archives of station measurements and in situ sensors. Remote sensors (e.g., NASA's Earth Observing Satellites) and environmental sensor networks generate massive volumes of data at rates higher than they are analyzed.

The massive volumes of climate related observations and climate model outputs are dimensioned by space and time; hence theoretical principles of SSTDM [6-7] remain valid. However, mining or analysis of climate data presents unique challenges. SSTDM methodologies may need to be significantly adapted or new approaches may have to be developed by drawing from multiple disciplinary areas, for example, statistics, mathematics, computer science, nonlinear dynamics and operations research. Finally, risk and uncertainty management requires analysis of impacts on lives, economy and the environment, and hence integration of geographic data.

This paper is an attempt to bridge the gap between climate scientists and the SSTDM community. A comprehensive overview of SSTDM [7] is used as a basis for formulating key climate data mining challenges. Adaptations of SSTDM concepts as well as unique challenges in climate are described. A case study illustrates how a simple data mining application led to new insights in climate science beyond published results in the literature (e.g., *Science* magazine: [8-10]).

3. Contributions

The primary contributions of this paper are three-fold: (a) introduction of climate challenges to the data mining (specifically the SSTDM) community to motivate future research (Sections 1-2), (b) defining the climate data mining problem by comparing and contrasting with SSTDM (Section 4), and (c) presenting a case study to demonstrate how even simple data mining applications can lead to novel insights in climate science.

We select heat waves for the illustrative case study because (a) mean and extremes of temperature are better predicted from climate models compared to other variables [1]; (b) Europe is known to be particularly vulnerable to heat waves, and (c) a recent Science paper [8] discussed these topics. The analysis

of Spain is an attempt to quantify one aspect of the "Africanization of Spain" issue which has been getting significant attention in the international media (e.g., [11]), among Spanish policy makers, as well as world bodies. We discuss uncertainty and population impacts [12] because the climate science community has barely begun to address some of these issues in depth, hence even a simple case like the one presented here clearly demonstrates the value-add from SSTDM.

4. Climate Data Mining

4.1 Knowledge Discovery Requirements

There is a scale disparity between atmospheric, hydrological and land process models on the one hand and water, energy, environmental or infrastructural impacts assessments on the other. One purpose of knowledge discovery in climate is to build a bridge between these disparate scales. Thus, climate model outputs at lower resolutions need to be downscaled to higher resolution impacts assessments, not just to provide guidance on mean values at local to regional scales but also for extreme events and extreme stresses caused by regional change. The need to provide credible information from climate models all the way to impacts assessments implicitly touches upon uncertainty reduction, risk formulations and uncertainty characterization.

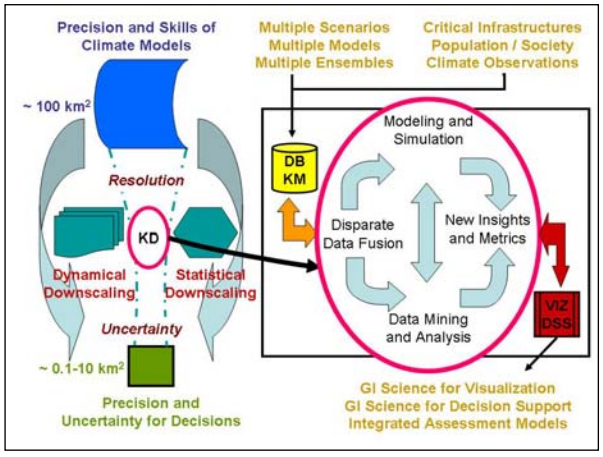

Figure 1. Knowledge Discovery in Climate

The Knowledge Discovery (KD) challenge for climate is depicted schematically in Fig. 1. The spatial scales are currently about 100 km^2 for climate models and anywhere from 10's of meters to 10's of kilometers for impacts assessments and decisions. Climate model analyses are credible at decadal trends (e.g., behavior of monthly averages or extremes in one future decade compared with the current decade), while the temporal horizons for decision making, infrastructural impacts assessments or mitigation policy decisions may range from a few days to decades. The downscaling issue can be solved in variety of ways, but the key requirement is one of uncertainty reduction and assessment.

Two major KD areas are (a) data analysis and mining, which extracts patterns from massive volumes of climate related observations and model outputs and (b) data-guided modeling and simulation (e.g., models of water and energy or other assessments of impacts) which take downscaled outputs as the inputs. Data fusion is broadly construed to describe a set of capabilities which can deal with multiple ensembles (e.g., based on clustering approaches) and scenarios (e.g., based on co-occurrence and persistence of space-time features), handle cascading uncertainty from models to downscaled outputs to impacts, and develop actionable predictive insights based on geographical information about impacted population and infrastructures.

Knowledge Management (KM) systems and data repositories serve as enablers of and feeders to the KD process while scalable visualizations and decision support systems present the KD results in an intuitive manner for dissemination to scientists, modelers, resource managers, decision makers and policy makers.

A preliminary illustration of KD in climate impacts is shown in Fig. 2. Daily precipitation observations at 2.5^o spatial grids from 1940 to 2004 (top left) in South America are used to extract information about extreme values based on a Generalized Pareto Distribution (GPD). An extremes volatility index (ratio of the 50-year and the 200-year return levels subtracted from unity) is calculated (bottom left). The extremes volatility ratio relates to the shape parameter of the GPD and provides an intuitive measure of "surprise" (e.g., the stress on an infrastructure which is designed to last a 50-year storm when a 200-year storm occurs), which can be roughly interpreted as a likelihood measure. The middle panels show population counts from the LandScan Global database [4], which is assumed in this preliminary study to scale linearly with exposed assets (lives and economy). A product of the hazard likelihood and the cost (exposure scaled by coping ability or resilience, roughly approximated here by the GDP of a nation) yields anticipated risks, which in turn can be used to optimize resources for disaster preparedness and humanitarian aid.

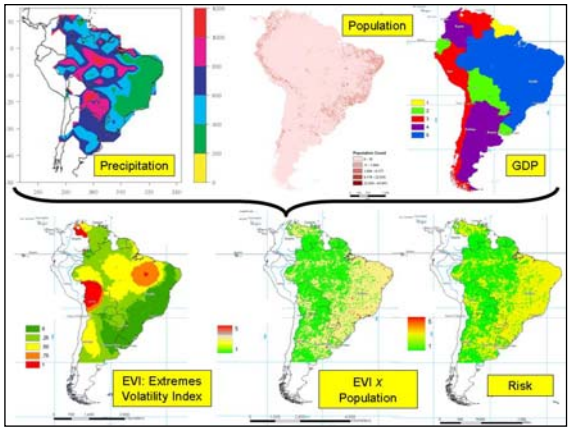

Figure 2. End-to-End Knowledge Discovery [35]

4.2 Correlations and Extremes

The theoretical foundations of SSTDM rely on the premise that learning samples are not independent and hence traditional data mining methods are inadequate. Correlations and seasonal effects, which are typically ignored by traditional data mining, must be considered in time series and spatial statistics, as well as SSTDM.

The entire field of time series analysis [13] and forecasting relies on the autocorrelation function (ACF) and the Fourier transform of the ACF as a fundamental concept. Spatial statistics and SSTDM owe their origins to these concepts, which were originally developed in time series. However, in time series, there are at least a couple of conditions under which the ACF is inadequate or inappropriate: (a) non-stationary data and (b) nonlinear behavior. Certain non-stationary or nonlinear behavior can be approximated by ACF-based analysis, often after simple stabilization transformations. However, not all types of non-stationarity can be handled in a simplistic manner, while for certain kinds of nonlinear behavior, a whole other set of tools may be appropriate [14]. In all cases, the dependence among time series, whether linear or nonlinear, remains important. One form of dependence, long-memory processes [15], is in principle captured by ACF but may be difficult to handle.

The uniqueness of spatial [16-17] and spatiotemporal statistics [18] arises to a great extent from auto- or cross-correlations among variables in multiple directions. SSTDM draws upon the insights developed in these areas when it attempts to extend traditional data mining by relaxing the explicit or implicit independence assumptions. Climate data mining inherits these challenges from SSTDM since climate data are geographical. However, the ability to develop scalable computational solutions for challenges unique to climate data analysis [19] is important. In the context of correlation and dependence, these include long-memory temporal processes, long-range spatial dependence (e.g., "teleconnections") and nonlinear dependence. Thus, the theoretical foundations of climate data mining need to inherit from both SSTDM and nonlinear, non-stationary time (and space-time) series analyses.

A second major challenge in climate data mining is the importance of extremes, both extreme hydro-meteorological events as well as extreme hydro-meteorological stresses caused by regional change. The fact that understanding and modeling deviations from the norm are typically more important than the ability to model the norm makes statistical estimation difficult and represents an important departure from many traditional data mining approaches. In the context of climate change, extreme events could refer to intensity-duration-frequency (IDF) of heat waves or large precipitation events and droughts, while extreme stresses could be caused by a region getting hotter and dryer like the Western United States [10]. The uncertainty and validation problems become acute when extremes are of interest. Thus, historical and recent observations need to be investigated in depth and compared with climate model hindcasts to estimate the ability of the model to capture extremes, develop uncertainty bounds and potentially help narrow these bounds, as well as develop projections of future extremes by statistical analysis of model outputs and uncertainty. In addition, extremes are not just high or low values or large stresses in the natural systems, but their interactions with built and social environments. Thus, hurricane Katrina is remembered till date not just because that was a Category 5 but also because a levee broke causing flooding in New Orleans, which in turn resulted in loss of lives and property at a scale unprecedented in recent years for a developed nation.

A major generic challenge in climate data mining results from the nature of historical observations. In recent years, climate model outputs and remote or in situ sensor observations have grown rapidly. However, for climate and geophysics, historical data may still be noisy and incomplete, with uncertainty and incompleteness typically increasing deeper into the past. Therefore, in climate data mining the need to develop scalable solutions for massive geographical data co-exist with the need to develop solutions for noisy and incomplete data.

These issues complicate the study of correlations and extremes, especially when the dependence is nonlinear or when extremes / anomalies are correlated.

4.3 Relation with State-of-the-Art SSTDM

This sub-section compares and contrasts climate data mining with SSTDM challenges described in [7].

1. Spatial Data: The so called first law of geography regarding proximity based relations is valid for climate. In addition, the presence of long-range spatial dependence (teleconnections), long memory processes and nonlinear dependence are important. The need to go deep into time and far out in space for the extraction of correlations makes automation of knowledge discovery a challenging computational task. The mutual information (MI) provides a way to measure nonlinear correlations. Recipes for MI computations include kernel density estimators, k-nearest neighbors, Edgeworth partitioning of differential entropy and adaptive partitioning of the XY plane [20]. The base algorithms for these approaches are typically amenable to efficient computational implementations.

2. Prediction and Classification: Input output pairs denoted by (x_i, y_i) are classified (i.e., "the input vectors x_i are assigned to a few discrete number of classes y_i") or regressed (i.e., approximated by a function of the form $y \equiv f(x)$ and used for prediction). In climate applications, statistical predictions based on observations are valid in a steady climate, but not under climate change scenarios. Therefore, for climate change, large scale physically-based climate models like CCSM3 are used, where a set of partial differential equations describe the evolution of the vector of state variables according to physical laws.

Predictive insights based on statistical models are useful for short leads or spatial downscaling and for natural climatic oscillations (e.g., the El Nino Southern Oscillation). Consider the Spatial Autoregressive (SAR) model [7]: $y = \rho Wy + X\beta + \varepsilon$. In climate applications like downscaling, the downscaled value on any one high-resolution grid may depend on its own neighbors as well as the low resolution variables. Thus, surrogates for the "convective available potential energy" from climate model outputs, or topographical information from remote sensing, may have information content about convective precipitation over and above what is already contained in the projected precipitation from the climate model. In this sense SAR appears well-suited. However, linear and stationary approaches cannot be necessarily assumed for climate applications. Thus, rather than using linear regression and linear AR coefficients, a nonlinear form may need to be prescribed. Support vector machines may be one way to handle nonlinear regressions in this context. The approach can be compared with [21], where a neural network based version of the ARMA was used to downscale numerical weather prediction model outputs. The equation becomes: $y = f_1(y) + f_2(X) + \varepsilon$, where the two functional forms were modeled independently with relatively simple neural networks (multilayer perceptrons) in an ensemble mode. In general, climate predictions are based on nonlinear dynamical tools (e.g., [21-22]).

Classification is useful in situations where regions need to be grouped into known categories based on climate conditions. Thus, if the classes y_i are the known climate categories (e.g., tundra) and the input vectors are climate observation or projections (e.g., mean and extremes of temperature and precipitation), then the probability of a particular grid or region belonging to a class is given as $p(y_i|x)$, which by Bayes' theorem becomes $[p(x|y_i) * p(y_i) / p(y_i|x)]$ leading to the discriminant function $g_i(x) = \ln p(x|y_i) + \ln p(y_i)$ as described in [7]. One other example of classification is when sea surface temperature anomalies need to be processed through dimensionality reduction algorithms (e.g., empirical orthogonal functions) and classified into categories (e.g., an "El Nino" year). Climate applications may require nonlinear dimensionality reduction, for example, manifold learning like LLE or ISOMAP [23-25].

3. Outlier Detection: A spatial outlier has been informally defined as "a local instability (in values of non spatial attributes) of a spatially referenced object whose non spatial attributes are extreme relative to its neighbors, even though the attributes may not be significantly different from the entire population" [7]. The tests for detecting spatial outliers include finding outlying points in variogram cloud and Moran scatterplots, while for normally distributed spatial data S(x), a normalized spatial statistic is often used [7]: $Z_{S(x)} = |\{(S(x)-\mu_s) / \sigma_s\}| > \theta$. Extensions to spatiotemporal data and/or for multiple attributes are conceptually straightforward. In climate data, outliers such as these may arise because of, for example, measurement errors or anthropogenic influence (e.g., urban heat island effect or sudden deforestation).

An intriguing problem in climate data mining is the need to distinguish between a measurement error (alternatively, any outlier generated by non-repeatable conditions) versus recurring but low probability patterns (e.g., 100-year extremes) and nonlinear effects (e.g., chaotic dynamics and transition behavior). An approach to distinguish chaos versus measurement error was presented by [26], which used short-term predictability. However, the possibility of distinguishing recurring extreme values or sudden or small sustained change from noise or measurement errors needs to be investigated in more depth.

Extreme value theory [27] is a statistical theory that develops parametric approaches to infer low probability high or low values based on analysis of values that are above certain thresholds but not necessarily low probability extremes. The development and use of extreme value theory for space-time climate applications was motivated by the IPCC [28]. Case studies with precipitation extremes are presented in [29-30], in the context of univariate extremes based on the Generalized Pareto Distribution (GPD) and copula based multivariate extremes dependence. While statistical definitions and theories are necessary for a rigorous study of extremes, user definitions based on impacts may be more meaningful to non-scientists and decision makers. In either case, a systematic exploration of the statistical properties may help distinguish recurrent and non-repeating patterns.

The presence of small sustained change is relatively easier to distinguish from outliers. The method proposed by [31] for network data and refined by [32] in the context of remote sensing change detection offers a first step, but may need to be refined for climate applications [36].

4. Co-Location and Clustering: Clustering, or the process of categorization, is used in many SSTDM applications. Multivariate clusters have been used to categorize climate regimes [33] and extract climate indices [37]. According to [7], "one important application of clustering is hot spot detection". However, in general, supervised, or at least semi-supervised, approaches which utilize both the data and any available domain knowledge may be better suited for climate applications.

Co-location or associations among variables has not been used as much in the climate change literature but may become very useful when impacts of climate on infrastructures are quantified in a rigorous manner, especially infrastructural grids like water or energy (electric) grids are considered.

5. Uncertainty and Risks: The characterization of unknown probabilities (uncertainty) and management of known probabilities (risks) uses the results of data mining to develop value-added decision support solutions. The need to deal with ensembles of initial conditions (to manage chaotic dynamics), model parameters (e.g., Monte Carlo type analysis), model physics (multi-model ensembles) and future conditions (scenarios of emissions and impacts) make the tasks computationally challenging. The integration of disparate data for the comprehensive characterization of uncertainty and risk management becomes a major challenge.

4.4 Computational Challenges

The computational challenges for SSTDM are presented in [7]. The unique considerations for climate data mining have been discussed in brief earlier. Here we focus on one specific example, which is of immediate concern to our ongoing research and the case study.

The CCSM3 global climate models output data for a single ensemble run of the A1FI (fossil fuel intensive) scenario are available at daily intervals in $1.4^{o} \times 1.4^{o}$ grids (or cells $\sim 100 km^2$), from 2000 to 2099. The data size makes even simple analyses (e.g. range, mean, and exceedance computations) a non-trivial task from a computational perspective. Thus, we are currently working which has a dataset with 100 years of daily data for the IPCC A1FI scenario with 100+ variables (the atmosphere models have 26 layers). The data has a size of approximately 850GB on disk and rough calculations show that loading the entire dataset at the surface level only (single layer) would require 480GB of memory. Preliminary tests on a desktop computer (3Ghz Pentium 4, 1GB RAM, Windows XP Pro.) indicate that computing mean, standard deviation, and exceedances in a sliding window for a subset of eight variables on a desktop machine would take nearly a week. High-performance computing can mitigate this problem to some extent but straightforward programming implementations are only possible when the task is "embarrassingly parallel", as in the example here. In this case, the data can be divided into disjoint subsets and distributed to many processors in a large-scale system performing independent computations (although, data management and distribution issues must be considered). However, more powerful analysis methods require simultaneous access to data from multiple geographic locations and time intervals, demanding more sophisticated algorithmic solutions for efficient parallel implementation. Specifically, we are interested in studying climate extremes such as heat waves, floods, or droughts, and underlying the analysis of such phenomena is a hierarchy of increasingly complex statistical techniques, ranging from (for example) grid counting and dimensionality reduction to space-time linear or nonlinear correlations, geographically weighted regressions and computation of the intensity-duration-frequency of extreme events or extreme stresses as well as the trends thereof.

One step toward achieving these goals would be a comprehensive scientific data mining toolkit designed specifically for large-scale climate data analysis tasks, e.g., an extension of [38].

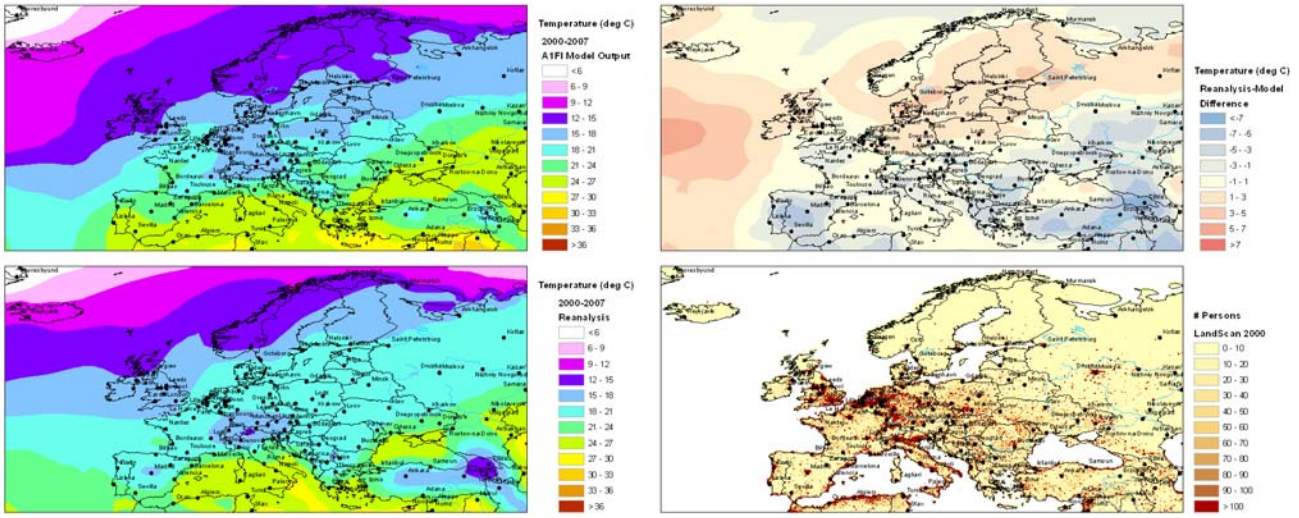

Figure 3. Model vs. Observed Extreme Heat Events and Population for the Present Time (2000-2007) in Europe

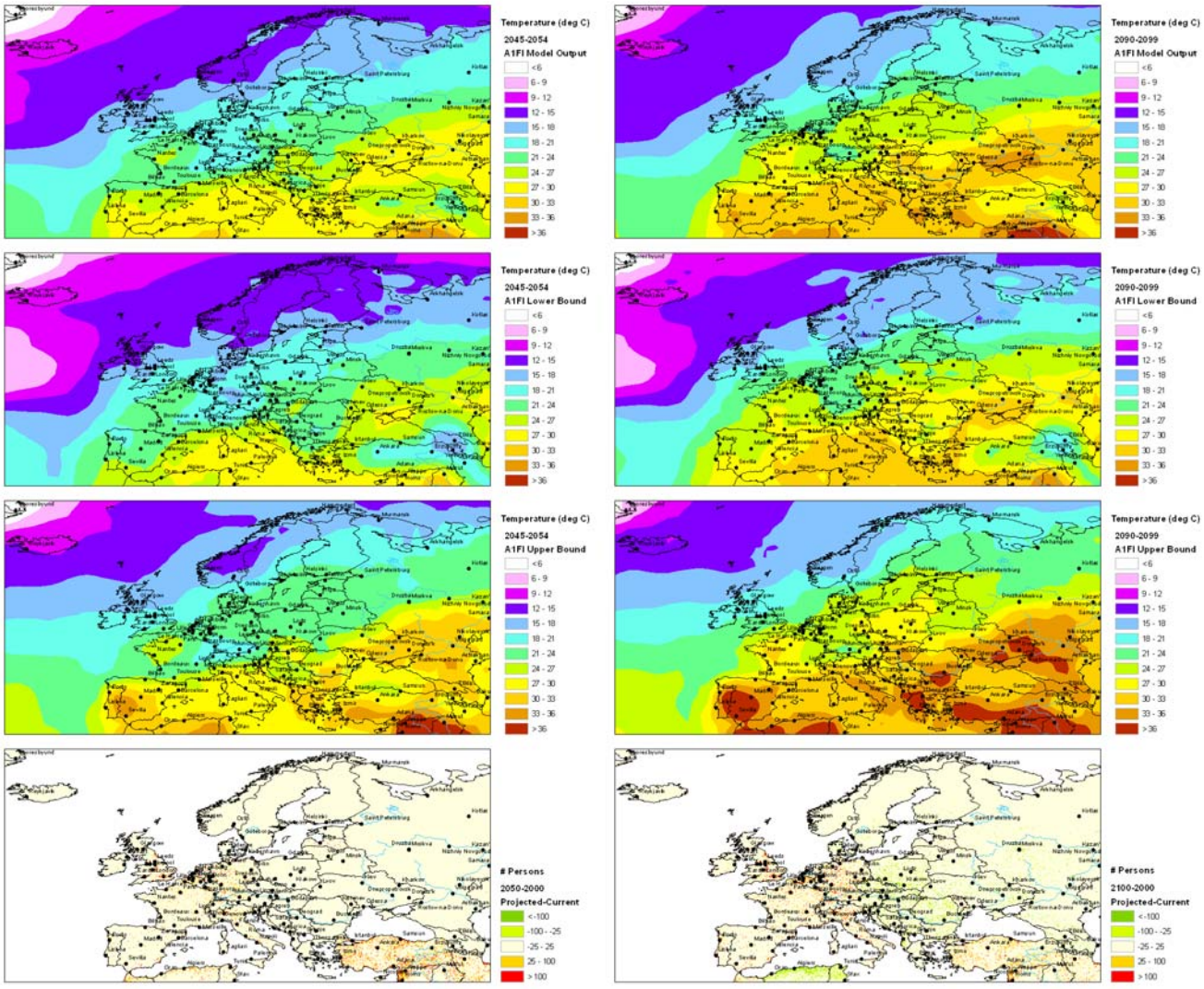

Figure 4. Projected Extreme Heat Events and Population for the Middle and End of the 21st Century in Europe

5. Case Study

In this section, we provide an illustrative example of mining spatiotemporal climate data and demonstrate that even an application of simple analysis techniques can lead to novel insights for climate change. Specifically, we describe a case study on temperature extremes (heat waves) consisting of three parts: first, we compare model projections for the present time to observed data and extract uncertainty bounds; second, we consider future projections and examine the potential effects of uncertainty on predictions; finally, we focus on one particular region and discuss the possible impacts of projected climate change.

5.1 Data Sources

This study used data from multiple disparate sources. Climate projections were based on IPCC SRES A1FI, the worst-case fossil fuel intensive scenario which nonetheless has started to look credible in recent years with increased trends in temperature observations. The model simulations were performed at ORNL and NCAR and output data are available through the Earth System Grid (ESG). The observation data was compiled by the National Centers for Environmental Prediction (NCEP) and are available for download on the NOAA Earth System Research Laboratory (ESRL) website. Population projections for Europe are based on the IPCC SRES A1FI population, followed by downscaling to country-levels by the Center for International Earth Science Information Network (CIESIN) at Columbia University, while the grid-based allocations were done based on current LandScan data from ORNL [4]. The grid-based allocations are preliminary, since future allocations are assumed to remain identical to the current.

5.2 Methodology: Extremes and Uncertainty

A heat wave can be defined in many different ways, for example as the exceedance of temperature over a threshold or as a period of sustained high temperatures. Here, we choose a definition from a couple of prior studies of the 1995 Chicago heat wave [8, 34], which focuses on an annual event marked by several consecutive nights with persistent high temperatures. Specifically, this annual "worst heat event" is defined as the mean over the three-day period with the highest average nighttime low temperature. All temperature measurements are taken at reference height of 2m above the earth's surface.

As a first order estimate of uncertainty in the model, we use the difference (absolute value) between the outputs and observations from the same time period. The upper/lower bounds for model projections are then created by adding/subtracting the difference to/from the model outputs. Note that this definition provides only a rough and possibly conservative estimate of uncertainty as we might expect the bounds to expand as we project further into the future.

In this study we consider three periods at the beginning, middle, and end of the 21st century: the present time (2000-2007 only to allow for comparison to observations), 2045-2054, and 2090-2099. For each period we take the average of annual worst heat events and visualize these values using a (commercial) GIS.

Figure 3 shows the worst three-day heat events from the model as well as the observed values (left panels), and by visual inspection we can glean a fair amount of agreement between the two. The top right panel shows the difference between the observed and the model values, where blue indicates areas where the model suggested higher temperatures than observed while red shows areas with lower predicted temperatures than observed. We find that the model tends to overpredict at lower latitudes (e.g., Greece, the Balkans, Italy, Spain) and underpredict at higher latitudes (e.g., Germany, Poland, UK, Scandinavia). With the exception of the region in the far Northwest, the model performs better over bodies of water than it does over land. The bottom right panels shows the current distribution of population in Europe.

Next, we examine the climate model projections for 2045-2054 (left) and 2090-2099 (right) in Figure 4. The top row corresponds to the actual model outputs; the next two rows show lower and upper bounds given the uncertainty estimates, respectively; the bottom row shows population projections (difference). These data provide several interesting insights. Most notably, we see a trend of increasing intensity in extreme events over time. Even the best case (lower bound), while still comparable to current levels at 2050, shows a significant increase in extreme temperatures across all of Europe by the end of the century. Moreover, the average and especially the worst case (upper bound) paint a grim picture, with sustained nighttime lows above 30ºC around the Mediterranean Sea and well into the 20s over much of the continent. Note that the A1FI population is expected to grow in the West but decrease in Eastern Europe by the end of the century.

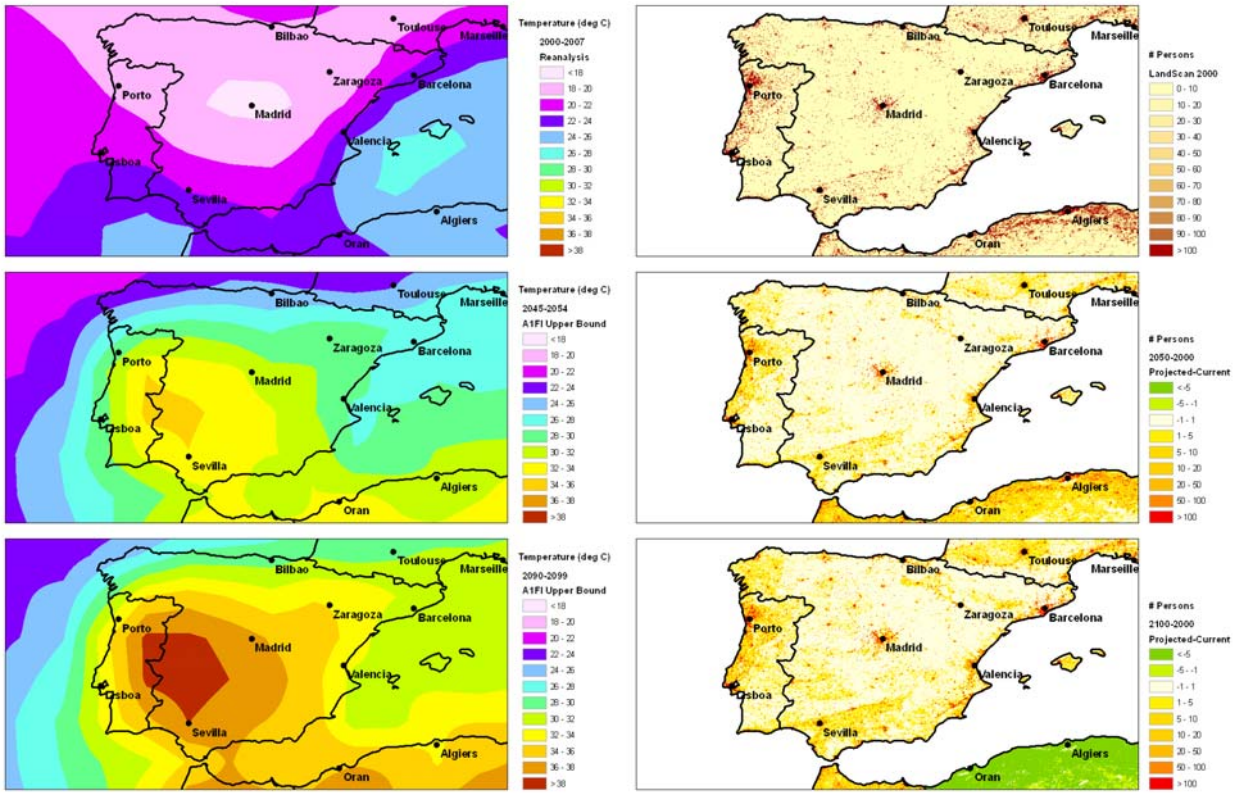

Figure 5. Observed and Projected Extreme Heat Events (Upper Bound Only) and Population in Spain

5.3 Local Study: Africanization of Spain

Spain has recently received press (e.g., see [11]) for its threat of "Africanization", a combination of increasing temperatures and decreasing precipitation. We concentrate on Spain for our local case study.

Figure 5 shows extreme heat events (left) and population (right) for each of the three time periods. Note that the top left panel represent observations for the present time, while the bottom two panels in that column depict the upper bound for the middle and end of the century, respectively. From the top left panel we find that the observed events are limited to moderate temperatures ranging from 17ºC around Madrid to 24ºC along the Mediterranean coast. In contrast, the middle left panel shows a drastic increase in intensity, with sustained nighttime temperatures reaching 35ºC. The increasing trend is projected to continue through the end of the century, when nearly half of the country can expect nightly heat waves exceeding 35ºC.

Intense heat waves may have direct impacts on human lives and well-being as well as indirect effects on scarcity of water and agriculture. The top right panel shows the current population of Spain, and the next two panels the projected difference for the future periods. Note that significant growth is expected, especially around the existing urban centers (including Porto and Lisboa in Portugal, which are similarly affected by these changes). Focusing again on Madrid, the model projects a worst-case scenario of nearly 20ºC higher nighttime temperature extremes between now and the end of the century, along with moderate to strong population growth in and around the city.

References

[1] Intergovernmental Panel on Climate Change. "Climate Change 2007: Fourth Assessment Report (AR4)" (2007).

[2] Tollefson, J. "Climate war games", *Nature*, Published online 5 August 2008, doi: 10.1038/454673a.

[3] Ganguly, A.R., E.S. Parish, N. Singh, K. Steinhaeuser, D.J. Erickson, M.L. Branstetter, A.W. Wayne, and E.J. Middleton. "Regional and decadal analysis of climate change induced extreme hydro-meteorological stresses informs adaptation and mitigation policies", 21st Conf. on Climate Variability and Change, Submitted (2009). [For information prior to acceptance/publication, please see http://www.ornl.gov/knowledgediscovery/WarGaming]

[4] Bhaduri, B., E. Bright, P. Coleman, J. Dobson. "LandScan: Locating People is What Matters", *Geoinformatics*, 4(2), 34-37 (2002).

[5] Drake, J.B., P.W. Jones, and G.R. Carr, Jr. "Overview of the Software Design of the Community Climate System Model", *Int. J. High. Perform. C.*, 19(3), 177-186 (2005).

[6] Roddick, J.F., K. Hornsby, and M. Spiliopoulou. "An updated bibliography of temporal, spatial and spatio-temporal data mining research", in *Temporal, Spatial, and Spatio-Temporal Data Mining*, Springer (2001).
[7] Shekhar, S., R.R. Vatsavai, and M. Celik. "Spatial and Spatiotemporal Data Mining: Recent Advances", in *Data Mining: Next Generation Challenges and Future Directions*, AAAI Press (2008).
[8] Meehl, G.A., and C. Tebaldi. "More intense, more frequent, and longer lasting heat waves in the 21st century", *Science*, 305(5686), 994-997 (2004).
[9] Goswami, B.N., V. Venugopal, D. Sengupta, M.S. Madhusoodanan, and P.K. Xavier. "Increasing Trend of Extreme Rain Events over India in a Warming Environment", *Science*, 314(5804), 1442-1445 (2006).
[10] Barnett, T.P., Et Al.. "Human-induced changes in the hydrology of the Western United States", *Science*, 319(5866), 1080-1083 (2008).
[11] Rosenthal, E. "In Spain, Water is a New Battleground", *The New York Times*, 3rd June, 2008.
[12] Vorosmarty, C.J., P. Green, J. Salisbury, and R.B. Lammers. "Global Water Resources: Vulnerability from Climate Change and Population Growth", *Science*, 289(5477), 284-288 (2000).
[13] Brockwell, P.J. and R.A. Davis. *Introduction to Time Series and Forecasting, 2nd ed.*, Springer (1996).
[14] Kantz, H., and T. Schreber. *Nonlinear Time Series Analysis*, Cambridge University Press (2004).
[15] Palma, W. *Long-Memory Time Series: Theory and Methods*, John Wiley and Sons (2007).
[16] Ripley, B.D. *Spatial Statistics*, Wiley (2004).
[17] Noel, A.C. *Statistics for Spatial Data*, *Revised Edition*, John Wiley (1993).
[18] Finkenstadt, B., L. Held, and V. Isham. *Statistical Methods for Spatio-Temporal Systems*, CRC Press (2006).
[19] von Storch, H., and A. Navarra. *Analysis of climate variability*, Springer (1999).
[20] Ganguly, A.R., and R.L. Bras. "Distributed quantitative precipitation forecasting combining information from radar and numerical weather prediction model outputs", *J. Hydrometeorol.* 4(6), 1168-1180 (2003).
[21] Khan, S., A.R. Ganguly, and S. Saigal. "Detection and predictive modeling of chaos in finite hydrological time series", *Nonlinear Proc. Geoph.,* 12, 41-53 (2005).
[22] Khan, S., A.R. Ganguly, and S. Saigal. "Detection and predictive modeling of chaos in finite hydrological time series", *Nonlinear Proc. Geoph.*, 12, 41-53 (2005).
[23] Roweis, S.T., and L.K. Saul. "Nonlinear dimensionality reduction by locally linear embedding", *Science*, 290(5500), 2323-2326 (2000).
[24] Tenenbaum, J.B., V. de Silva, and J.C. Langford. "A global geometric framework for nonlinear dimensionality reduction", *Science*, 290(5500), 2319-2323 (2000).
[25] Jenkins, O.C., and M.J. Mataric. "A spatio-temporal extension to Isomap nonlinear dimension reduction", 21st Int'l Conf. on Machine Learning (2004).
[26] Sugihara, G., and R.M. May. "Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series", *Nature*, 344(6268), 734-741 (1990).
[27] Coles, S., *An introduction to statistical modeling of extreme values*, Springer (2001).
[28] Intergovernmental Panel on Climate Change, Workshop on Changes in Extreme Weather and Climate Events (2002).
[29] Khan, S., G. Kuhn, A.R. Ganguly, D.J. Erickson, and G. Ostrouchov. "Spatio-temporal variability of daily and weekly precipitation extremes in South America", *Water Resour. Res.,* 43, W11424 (2007).
[30] Kuhn, G., S. Khan, A.R. Ganguly, and M. Branstetter. "Geospatial-temporal dependence among weekly precipitation extremes with applications to observations and climate model simulations in South America", *Adv. Water Resour.*30(12), 2401-2423 (2007).
[31] Lambert, D., and C. Liu, "Adaptive thresholds: Monitoring streams of network counts online", *J. Am. Stat. Assoc.* 101, 78-89 (2006).
[32] Y. Fang, A.R. Ganguly, N. Singh, V. Vijayaraj, N. Feierabend, D.T. Potere. "Online change detection: Monitoring land cover from remotely sensed data," 6th Int'l Conf. on Data Mining – Workshops, 626-631 (2006).
[33] Hoffman, F.M., W.W. Hargrove, and D.J. Erickson. "Using Clustered Climate Regimes to Analyze and Compare Predictions from Fully Coupled General Circulation Models", *Earth Interactions*, 9(10), 1-27 (2005).
[34] Karl, T.R., and R.W. Wright. "The 1995 Chicago heat wave: How likely is a Recurrence?" *B. Am. Metereol. Soc.,* 78, 1107-1119 (1997).
[35] Fuller, C.T, A. Sabesan, S. Khan, G. Kuhn, A.R. Ganguly, D.J. Erickson, G. Ostrcouchov. "Quantification and visualization of the human impacts of anticipated precipitation extremes in South America", *Eos Trans. AGU* 87(52), Fall Meet. Suppl., Abstract GC44A-03.
[36] Boriah, S., V. Kumar, M. Steinbach, C. Potter. S. Klooster. "Land Cover Change Detection: A Case Study", *ACM SIGKDD Conf. on KDD,* Las Vegas, NV (2008).
[37] Steinbach, M., P.-N. Tan, V. Kumar, S. Klooster, C. Potter. "Discovery of climate indices using clustering", *ACM SIGKDD Conf. on KDD,* Washington, DC (2003).
[38] Rushing, J. R. Ramachandran, U. Nair, S. Graves, R. Welch, H. Lin. "ADaM: A data mining toolkit for scientists and engineers", *Comput. Geosci.,* 31(5), 607-618 (2005).

Acknowledgments

We thank D. Erickson and M. Branstetter of ORNL for the A1FI data and N. Singh of ORNL for the population projections. Reanalysis data are available from the NOAA website. A major portion of the funding for this research was provided by the Institute for a Secure and Sustainable Environment (ISSE) at the University of Tennessee, Knoxville, through a research grant to ORNL. This research was funded by the Oak Ridge National Laboratory (ORNL), managed by UT Battelle, LLC, for the U.S. Department of Energy under Contract DE-AC05-00OR22725. The United States Government retains, and the publisher by accepting the article for publication, acknowledges that the United Stated Government retains, a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.