A Semantic Framework for Climate Metadata Interoperability

M. Benno Blumenthal, John del Corral, and Haibo Liu

International Research Institute for Climate and Society, Columbia University

Dan Holloway and Nathan Potter

OPeNDAP, Inc

Abstract

The Semantic Web provides a single framework that allows describing datasets according to multiple standards, creating a more complete description than any single standard provides. Going beyond standards, it can explicitly describe the data models implicit in programs that display and manipulate data. Writing Models, Crosswalks , and Objects all within RDF/SemanticWeb means these data models and metadata standards can be interrelated in a single framework, leading to interoperability.

Crosswalking between different standards can be as simple as two different names for the same quantity, but sooner or later the mapping gets more complicated. Frequently, different objects are related conceptually but are very different structurally. Our framework thus has both structure and conceptual models: structure models that describe how dataset metadata is written (e.g. cf-att which describes the attributes of a CF convention netcdf file), and conceptual models which describe the conceptual objects represented in the convention, e.g. cf-obj which describes the more abstract objects (like geo-located data) that are being described in the CF convention. XML Schema is a common way to represent structure models for XML files, and we have a translation of XML Schema to RDF/OWL which allows us to create conforming XML files from RDF information. We have applied this to the WCS Schema, for example, to extract the needed information for an OPeNDAP WCS service based on RDF extracted from CF/netcdf files. We also have included controlled vocabularies such as CF standard names or GCMD scientific parameters. Controlled vocabularies are a common way to structure classifications, and important for us to build a faceted search that works across diverse datasets.

Our working example is composed of datasets and metadata in the IRI/LDEO Climate Data Library (http://iridl.ldeo.columbia.edu). These

data services enable access and analysis by providing data in a framework which allows format translation, rendering, and application of a variety of analysis functions, including sampling, averaging, regridding, EOFs, and statistical operators. Datasets are both local and remote, allowing a federation of data servers to appear in a uniform space of data access and functionality.

Describing the library's contents requires concepts like datasets, units, dependent variables, and independent variables. These datasets have been provided under diverse frameworks that have varied levels of associated metadata. We have created an RDF expression of a taxonomy that forms the basis of a dynamic earth data search interface. The concepts include location, time, quantity, realm, author, and institution. We have also started cross-walking these metadata into various existing metadata schema, so that our data can be found in the corresponding systems.

A persistence framework incorporating inference and crawling is used to ingest the metadata information for a specified starting point as well as infer the connections between the diverse data-oriented concepts of the data library and the conceptual framework of the data search. This persistence framework includes inferred crawling and rule construction, OWL/SWRL as well as custom SeRQL construct rule inferencing, and XSL Transform on ingest (both GRDDL/RDFa based, and inferred from RDF information).

Spatial survey of elementary data mining for drought

Matthew Collier

Department of Geography, University of Oklahoma

Amy McGovern

School of Computer Science, University of Oklahoma

While long term teleconnections, such as the El Niño-Southern Oscillation, have been shown to modulate the spatiotemporal extent of drought on land, local hydroclimatic and surface conditions may influence the overall severity of drought. In this study we seek to exploit this dependence on local, neighborhood drought index values in the prediction of next month's value. Our methodology consists of constructing appropriate attributes from raw spatiotemporal drought data, and searching predictive hypothesis spaces with appropriate data mining schemes. Input data consist of the monthly Palmer Drought Severity Index values on a 2.5 degree geographical grid spacing. To this data we add the following monthly teleconnection data: the Atlantic Multi-decadal Oscillation, the Bivariate ENSO Time Series, the Northern Hemisphere land temperature, and the Pacific Decadal Oscillation. New attributes are derived to include month, season, random drought values for overfitting assessment, and first order differences of PDSI. Finally, a class attribute is constructed to investigate the short term predictability of drought one month into the future. Each of the attributes is then discretized for ingestion into data mining schemes. Separate data sets are created with artificially injected noise at various ratios. They are used to quantify the uncertainty in the usefulness of the data mining schemes at drought class prediction in the presence of real world data that may contain noise. Results include abstract maps of the Southern Great Plains of the United States that present performance characteristics of three elementary mining schemes applied to the PDSI across the grid cells in the geographic region. Briefly, the three schemes include: 1) "ZeroR" - a scheme that returns the mode of the class to be predicted. This serves as a baseline to judge improvements obtained from more complex data mining schemes and is well grounded in traditional statistics. 2) "OneR" - A simple algorithm that has been shown to have remarkably low classification error rates on many canonical data mining data sets. It, too, serves as a baseline for more complex mining schemes. However, OneR also begins to reveal potentially predictive structure in the spatiotemporal data. And, 3) AODE - or, "Averaged One-Dependence Estimators". This data mining scheme consists of the average prediction from an ensemble of elementary estimators whose models are naïve-Bayes-like. This is the most complex of the schemes applied in this poster presentation. This work is part of an extended effort that is also looking at the 3-month Standardized Precipitation Index drought indicator, and at using more complex data mining schemes such as C4.5, release 8, decision trees. All analyses are done in the open source WEKA data mining software system, and on publicly available datasets.

Data-Guided Discovery of Dependence Structure between Precipitation Extremes and Local Covariates

Debasish Das[1,2], Snigdhansu Chatterjee[3], Auroop Ganguly[4], Vipin Kumar[1], Zoran Obradovic[2]
[1]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455
[2]Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, PA 19122
[3]School of Statistics, University of Minnesota, MN 55455
[4]Oak Ridge National Laboratory, Oak Ridge, TN

Abstract: **Discovery of dependence structure between precipitation extremes and other climate variables (covariates) within a smaller spatial and temporal neighborhood is an important step in better understanding the drivers of this complex phenomenon as well as short-term prediction of extremes occurrence. Apart from the inherent spatio-temporal variability of the dependence, it is further complicated by the availability of the covariates at different vertical levels. The above problem can be split into three different sub-problems. Firstly, a spatio-temporal neighborhood of influence has to be discovered, which can be different for different locations. Secondly, the dependence structure between the precipitation extremes and the covariates has to be discovered within this neighborhood and thirdly, it has to be investigated whether this dependence structure can be exploited for any predictive power. Climate scientists have already discovered some physics-based relations between some of the covariates (e.g. temperature, relative humidity, precipitable water etc.) and precipitation extremes. We are exploring data-dependent alternatives for these problems and any possibility of incorporating the physics-based relations into the resulting data model. In particular, we used elastic net-based sparse optimization technique which solves all three problems of neighborhood discovery, covariate dependence discovery and predictive modeling and at the same time maintains the interpretability of the resulting model. Preliminary results look promising and show potential for some interesting knowledge discovery. We are currently exploring non-linear correlations and the alternatives to combine the physics-based relationships into the data model.**

Atlantic Cyclogenesis: A Data Driven Approach

James Faghmous[1], Stefan Liess[1,2], and Vipin Kumar[1]

[1] Department of Computer Science, The University of Minnesota - Twin Cities

[2] Department of Soil, Water, and Climate, The University of Minnesota - Twin Cities

July 8, 2011

Abstract

Global climate change and its effect on Atlantic tropical cyclone (TC) activity has become one of the most contested issues in climate science. The difficulty of attributing a change in TC frequency to global climate change stems from the lack of reliable historical data as well as the large amplitude fluctuations in present-day storms. Understanding future TC activity is crucial, especially in light of TC's potential role in the ocean's poleward heat transport [5, 1], impact on marine ecosystems [4, 3], and increasing destructiveness [2, 6].

Currently, a theory of TC formation (cyclogenesis) is still lacking [1], which makes predicting future TC frequency highly uncertain. Existing high-resolution climate models fail to consistently predict an increase or decrease in the total number of TCs in a warming environment. Globally, the majority of global circulation models (GCMs) forecast a decrease in the total number of TCs as the atmosphere continues to warm. At the individual basin level, however, regional circulation models (RCMs) have been significantly more uncertain with projected changes of up to $+/-50\%$.

In this work we attempt to gain a better understanding into Atlantic cyclogenesis by leveraging the recently available climate and TC data. First, we show that not all regions in the Atlantic are equally important when it comes to cyclogenesis. Previous work monitoring TC trends employed basin-wide averaging to study the cyclogenesis-TC relationship. Our work shows that by focusing our attention on smaller and more meaningful regions in the Atlantic we are able to better capture sea surface temperature's (SST) relationship to Atlantic TCs. Our results suggest that the warming of the Atlantic off the West African coast near 20°- 30°N prior to the TC season, as well as the warming westward of $10° - 20°$N between 18° and 60°W during the TC season have a pronounced effect on TC formation. Furthermore, we propose that, unlike other basins, the recent increase in TC activity can be linked to Atlantic SST.

The second part of our work, leverages the insights gained from the previous section to build a nonparametric probabilistic model to learn the climatology of Atlantic cyclogenesis. Previous work has focused on season TC frequency predictions. Instead of predicting seasonal counts, we are interested in learning the climate factors that trigger (individual) cyclogenesis. To this end, we monitor every TC that has been recorded and attempt to learn the factors that influenced its formation. The advantage of using some of the emerging nonparametric models is to relax common assumptions of independence on the data. Furthermore, we are interested in learning how the relationship between Atlantic cyclogenesis and various climate variable evolves over time. Finally, using such (soft) models would ensure that we can adapt as more data are observed to account for the potential changing relationship between cyclogenesis and climate factors, especially as the atmosphere continues to warm.

References

[1] K. Emanuel. The hurricane-climate connection. *Bulletin of the American Meteorological Society*, 89(5), 2008.

[2] Kerry Emanuel. Increasing destructiveness of tropical cyclones over the past 30 years. *Nature*, 436(7051):686–688, 2005.

[3] C. Hansen, E. Kvaleberg, and A. Samuelsen. Anticyclonic eddies in the norwegian sea; their generation, evolution and impact on primary production. *Deep Sea Research Part I: Oceanographic Research Papers*, 57(9):1079–1091, September 2010.

[4] I. Lin, W.T. Liu, C.C. Wu, G.T.F. Wong, C. Hu, Z. Chen, W.D. Liang, Y. Yang, and K.K. Liu. New evidence for enhanced ocean primary production triggered by tropical cyclone. *Geophys. Res. Lett*, 30(13):1718, 2003.

[5] R.L. Sriver and M. Huber. Observational evidence for an ocean heat pump induced by tropical cyclones. *Nature*, 447(7144):577–580, 2007.

[6] P. J. Webster, G. J. Holland, J. A. Curry, and H.-R. Chang. Changes in tropical cyclone number, duration, and intensity in a warming environment. *Science*, 309(5742):1844 –1846, 2005.

Uncertainty in Calibration with Multiple Targets

James R. Gattiker

Statistical Sciences, Los Alamos National Laboratory

Background is provided on fully-Bayesian calibration of computer models using Gaussian process emulators. This approach allows the quantification of uncertainty in parameters and predictions by comparing complex computer models to observations, including multivariate response fields. An open question in analysis is how to combine the results from multiple target observations. If one believes the models are true, the posterior probabilities should be multiplied. If one believes they are merely indicative, the posteriors should be subsumed. We demonstrate a hierarchical model as a balance between these alternatives. These approaches are demonstrated on calibration of simulations of an intermediate complexity climate model.

Meteorological Satellite Image Retrieval & Indexing

Mohamed Gebril, Abdollah Homaifar , and Ruben Buaba

NOAA-ISET Center
Autonomous Control and Information Technology Center
Electrical and Computer Engineering
North Carolina A&T State University
Greensboro, NC 27411
mmgebril,rbuaba, homaifar@ncat.edu

Eric Kihn
NOAA /NGDC
325 Broadway
Boulder, CO 80305
303-497-6346
Eric.A.Kihn@noaa.gov

Vision is arguably our most significant sense, giving rise to efforts to empower computers to represent, process, understand, and act on visual imagery. As a result, images are being generated from a variety of sources. Terabytes of data are being generated in the form of satellite imagery, surveillance images, fingerprints, trademarks and logos, graphic illustrations, engineering line drawings, documents, manuals, medical images, images from sports events, documentation of environmental resources in the form of images, and entertainment industry photos and videos.

A mixture of image classification techniques and a Locality Sensitive Hashing (LSH) based searching algorithm to search for similarity based on Learning Mertics and to classify satellite imagery is presented. The problem of recognizing classes of objects in images is important for annotation and indexing of Satellite image databases.

Texture and Shape descriptors have been used frequently as features to characterize an image for classification and image retrieval tasks. The problem of recognizing classes of objects in images is important for annotation and indexing of Satellite image databases. In this paper, a comparison between shape and texture features for classification is presented. The classification is based on Support Vector Machine (SVM) learning. A SVM classifier can be learned from training data of relevance images and irrelevance images marked by users. Using the classifier, the system can retrieve more images relevant to the query in the database efficiently.

Active learning has been shown as a key technique for improving content-based image retrieval (CBIR) performance. Among various methods, SVM active learning is popular for its application to relevance feedback in CBIR. However, the regular SVM active learning has two main drawbacks when used for relevance feedback. In this paper, we propose a a new scheme that exploits both semi-supervised kernel learning and batch mode active learning for relevance feedback in CBIR. In particular, a kernel function is first learned from a mixture of labeled and unlabeled examples. The kernel will then be used to effectively identify the informative and diverse examples for active learning via a min-max framework. In the past, several approaches to fuse the advantages of generative and discriminative approaches were presented, often leading to improved robustness and recognition accuracy. SVMs are a well known discriminative classification framework but, similar to other discriminative approaches, suffer from a lack of robustness with respect to noise and overfitting. Gaussian mixtures, on the contrary, are a widely used generative technique. We present a method to directly fuse both approaches, effectively allowing for fully exploit the advantages of both.

The goal is to build an accurate and fast query-by-example using content based image retrieval based on the information extracted from satellite image data. Large scale image search has recently attracted considerable attention due to easy availability of huge amounts of data. Several hashing methods have been proposed to al- low approximate but highly efficient search. Unsupervised hashing methods show good performance with metric distances but, in image search, semantic similarity is usually given in terms of labeled pairs of images. There exist supervised hashing methods that can handle such semantic similarity but they are prone to overfitting when labeled data is small or noisy. Moreover, these methods are usually very slow to train. In this work, we propose a semi-supervised hashing method that is formulated as minimizing empirical error on the labeled data while maximizing variance and independence of hash bits over the labeled and unlabeled data. The proposed method can handle both metric as well as semantic similarity.

We have investigated and described various feature extraction methods relevant to our work in this paper. The results demonstrate satisfactory classification accuracy based on the proposed model. The results show the effectiveness of our approach. .

Authors: David S. Grass and Mark A. Cane

Affiliation: Lamont Doherty Earth Observatory of Columbia University
Department of Earth and Environmental Science
61 Route 9W
Palisades, NY 10964

Title:
Assessing the Impacts of Climate Change on High-Risk Synoptic Weather Patterns: Implications of Changes in Regional Circulation Patterns for Air Pollution and Mortality Levels in Santiago, Chile

Abstract:

The principal objective of this study is to determine how changes in the climate of the 21st century will affect the frequency, persistence, and character of high-risk winter weather classes in the airshed surrounding Santiago, Chile. We apply a synoptic weather typing procedure, based on principal components analysis and cluster analysis, to assign daily NCEP/NCAR Reanalysis II data from 1981-2000 to classes with homogenous meteorological conditions. Observed daily respiratory and cardiovascular mortality counts and pollution levels are used to identify weather classes associated with elevated or depressed mortality and/or air pollution levels. A linear discriminant function is used to assign NCEP/NCAR data from 2001-2005 to the weather classes defined during the 1981-2000 period on the basis of the mean sea level pressure, and the air temperature, specific humidity, and northward and eastward wind components at three standard air pressure levels: 1000 hPa, 850 hPa, and 500 hPa, in 4 cell by 4 cell domain centered over Santiago. We conduct two validation experiments to test the skill of the linear discriminant function in identifying days associated with elevated or depressed mortality and/or air pollution levels during the 2001-2005 period. We apply the validated linear discriminant function to daily output from seven high-resolution model runs from the CMIP3 multi-model dataset archive for the 1981-2000 period. We select the Max Planck Institute ECHAM5 model for all subsequent analyses based on its ability to best reproduce the frequency of occurrence of each weather class observed in the NCEP/NCAR Reanalysis data for the same period. Each winter day in ECHAM5 output for the years 1961-1980, 1981-2000, 2046-2065 and 2081-2100 is assigned to one of the predefined weather patterns. The four 20-year periods are compared to determine how the frequency of occurrence, the mean residence time (i.e. persistence), and the mean conditions (i.e. character) of high-risk weather classes change over time. Excess mortality risk resulting from these changes is estimated. Changes in weather class character, frequency, and persistence are also compared under the B1, A1B, and A2 emissions scenarios during the 2081 to 2100 period. Finally, we test the regional climate prediction that the frequency of extra-tropical cyclones in the study domain will decrease as a result of a poleward shift in the storm track and an intensification of the Pacific Subtropical Anticyclone, and examine the implications of these changes on weather classes historically associated with elevated air pollution concentrations.

Ad-hoc Event-based Satellite Data Query and Retrieval for Climate Informatics

Shen-Shyang Ho HOSHENSH@UMD.EDU
Center for Automated Research
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742

Wenqing Tang WENQING.TANG@JPL.NASA.GOV
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91107

Kwo-Sen Kuo KWO-SEN.KUO@NASA.GOV
Goddard Earth Science Technology and Research Center - Caelum Research Corporation
NASA Goddard Space Flight Center
Greenbelt, MD 20771

Current research in climate informatics focuses mainly on the development of novel (machine learning, data mining, or statistical) techniques to analyze climate data (e.g. model, in-situ, or satellite) or to make prediction based on these climate data. One important component missing from this workflow is data management that allows efficient and flexible data retrieval, (ease of) reproducibility, and the (ease of) techniques reuse on user-defined data subsets or other data. As a result, the utility of the developed techniques may be limited.

In this poster, we describe such a database management system that supports satellite data-driven climate science research by providing scientists with previously unavailable accurate and efficient ad-hoc event query and Earth science satellite sensor data retrieval based on user-defined criteria (Schneider et al., 2010, 2011) to study and analyze atmospheric events such as tropical cyclones and mesoscale convective systems (MCS). The system is based on moving objects database technology (Güting and Schneider, 2005) to store the dynamic atmospheric events and the satellite data retrieval solution is formulated as a spatiotemporal join query to identify the spatiotemporal location where moving sensor objects and dynamic atmospheric event objects intersect, either precisely or within a user-defined proximity (Ho et al., 2010). Based on the query output, subsetted satellite data is retrieved from NASA data centers (e.g., JPL PODAAC or GES DISC) and then analyze by scientists.

From published scientific journal papers, one observes that such a query and retrieval capability is extremely important to scientists who retrieve specific sensor data of specific atmospheric events for statistical analysis. Two query examples derived from these published scientific papers that require search, retrieval, and analysis of satellite data containing cyclone features, are listed below:

. **Acknowledgement:** This work was funded by the National Aeronautics and Space Adminstration (NASA) Advanced Information Systems Technology (AIST) Program under grant number AIST-08-0081.

1. Retrieve TRMM precipitation data for tropical cyclones that attained tropical storm intensity or higher over western North Pacific and the South China Sea between longitudes 100^oE and 180^o. 138 sensor datasets for 61 tropical cyclones retrieved (Kodama and Yamada, 2005).

2. Retrieve TRMM precipitation data for tropical cyclones from December 1997 to December 2003. 3703 sensor datasets for 563 tropical cyclones retrieved (Yokoyama and Takayabu, 2008).

Moreover, such a capability supports scientists in their investigations where large amount of problem-specific, event-specific data are needed. An example of tropical storm characteristics investigation using QuikSCAT ocean surface wind data subset is as follows. Wind structure is one of the important factors controlling the intensity change (intensification or weakening) of tropical storms (Tang and Liu, 2009). One can retrieve ocean surface wind fields measured by QuikSCAT associated to tropical storms in the Atlantic ocean between 2000 and 2009, with the ability to further specify the search criteria to find the wind measurements associated with the group of cases satisfying characteristics such as "storm paths when the translation speed is greater than $5m/s$" or "ocean surface vector wind for hurricanes reaching categories 4 or 5", where the storm translation speed or categories is determined from the tropical cyclone trajectories.

While the current system prototype only has the tropical cyclone events, work is underway to automatically identify and track events such as MCS and tornadoes to be ingested into the system. In future, users can contribute and share their event data to enable a highly flexible query and retrieval system. Climate data other than satellite data can also be retrieved.

References

R.H. Güting and M. Schneider. *Moving Objects Databases*. Morgan Kaufmann Publishers, 2005.

S.-S. Ho, W. Tang, W. T. Liu, and M. Schneider. A Framework for Moving Sensor Data Query and Retrieval of Dynamic Atmospheric Events. In *22nd Int. Conf. on Scientific and Statistical Database Management (SSDBM)*, volume 6187 of *Lecture Notes in Computer Science*, pages 96–113. Springer, 2010.

Y. M. Kodama and T. Yamada. Detectability and configuration of tropical cyclone eyes over the western north pacific in trmm pr and ir observations. *Monthly Atmospheric Review*, 133:2213–2226, 2005.

M. Schneider, S.-S. Ho, T. Chen, A. Khan, G Viswanathan, W. Tang, and W. T. Liu. Moving Objects Database Technology for Ad-Hoc Querying and Satellite Data Retrieval of Dynamic Atmospheric Events. In *Earth Science Technology Forum*, 2010.

M. Schneider, S.-S. Ho, M. Agrawal, T. Chen, H. Liu, and G Viswanathan. A Moving Objects Database Infrastructure for Hurricane Research: Data Integration and Complex Object Management. In *Earth Science Technology Forum*, 2011.

W. Tang and W. T. Liu. Dependence of hurricane asymmetry and intensification on translation speed revealed by a decade of quikscat measurements. In *NASA Ocean Vector Wind Science Team Meeting, Boulder, Colorado*, 2009.

C. Yokoyama and Y. N. Takayabu. A statistical study on rain characteristics of tropical cyclones using trmm satellite data. *Monthly Atmospheric Review*, 136:3848–3862, 2008.

Dealing with intermodel bias in downscaling using analogues followed by Random Forests: a case study in the Iberian Peninsula and implications for climate studies.

G. Ibarra-Berastegi (*), J. Saénz, A. Ezcurra, J.Diaz Argandoña , I. Errasti

[1] Dept. of N.E. & Fluid Mechanics. University of the Basque Country, Spain
[2] Dept. of Applied Physics II. University of the Basque Country, Spain
[3] Dept. of Chemical & Environmental Engineering. University of the Basque Country, Spain
[4] Dept. of Applied Physics I. University of the Basque Country, Spain

Corresponding author: email: gabriel.ibarra@ehu.es)

Abstract

In this paper, reanalysis fields from ECMWF reanalysis projects have been statistically downscaled to predict from large-scale atmospheric fields daily horizontal moisture flux at surface level (q_{10} vector) at two observatories located in the Ebro Valley (Spain) during the 1961-2001 period.

Three types of downscaling models have been built: i) analogues, ii) analogues followed by random forests (RF) and iii) analogues followed by multiple linear regression (MLR). At the model building stage, the inputs consist of data (predictor fields) taken from the ERA-40 reanalysis. The predicted fields is zonal and meridional surface moisture flux as measured at Zaragoza and Tortosa observatories according to data from European Climate Assessment (ECA) dataset http://eca.knmi.nl/

Zonal and meridional moisture fluxes (q_x and q_y expressed as kilograms of water vapour per square meter per second) were calculated as follows:

$$q_x = q\ \rho\ U_s \qquad (1)$$

$$q_y = q\ \rho\ V_s \qquad (2)$$

where ρ represents the density of air as a function of temperature, U_s and V_s are the zonal and meridional wind speed while q is the specific humidity (kg water vapour/ kg air) at the surface level. The specific humidity q is calculated using the Clausius-Clapeyron equation from surface pressure, temperature, relative humidity and/or dew point temperature.

With the aim to reduce the dimensionality of the problem, the ERA-40 fields have been decomposed using empirical orthogonal functions. Available daily data has been divided into two parts: a training period (1961-1996) used to find a group of about 300 analogues to build the downscaling model and a test period (1997-2001), where models' performance has been assessed using independent data.

ERA-40 data covers the period from mid-1957 to mid-2002 and since 1989 onwards ERA-Interim is used. Due to the different assimilation algorithms used and the different periods covered, there is not a continuity between ERA-40 and ERA-Interim and a bias exists between both reanalyses.

In this work, the sensitivity of the downscaling techniques to the use of ERA-40 and ERA-Interim has been tested. While all the models have been fitted exclusively on ERA-40 historical records (1961-1996), the models obtained have been fed with i)ERA-40 data belonging to the test period ii) ERA-Interim data from the same period.

The results indicate that models fitted with ERA-40, when they are fed with ERA-Interim data, experiment a deterioration in performance which is notoriously higher in analogues followed by linear regression. In the case of analogues+RF, the degradation

is much smaller being a reasonable option for the future fitting RF models and use them with the more recently available ERA-Interim (from 1989 onwards).
The reason for this might be that MLR models heavily rely on the coefficients of an equation and the differences between ERA-40 and ERA-Interim input data tend to be amplified by the values of the coefficientes. However, the nature of the regression with RF is different and is based on regression trees where homogeneous areas in the final leafs are sought as the trees split at the different stages. In this sense, it can be expected that similar homogeneous areas can be described either using ERA-40 or ERA-Interim EOF as inputs, since leading EOF from both reanalysis are likely to be describing the same physical effects on the studied area.
These results also have implications for climate studies since downscaling transfer functions can be fitted using analogues, RF and historical records of NCEP or ERA reanalyses. At a next stage, these transfer functions could be fed with climatic models projected onto the same grid as NCEP or ERA. Although strong bias between reanalyses and climatic models (higher than between ERA-40 and ERA-Interim) can be expected, using analogues followed by RF could be a valuable tool to obtain local future projections of certain variables of interest.
After two recent publications [1][2], further research is currently being carried out by our group in this line.

[1] G. Ibarra-Berastegi, J. Sáenz, A. Ezcurra, A. Elías, J. Díaz de Argandoña, and I. Errasti(2011). *Downscaling of surface moisture flux and precipitation in the Ebro Valley (Spain) using analogues and analogues followed by random forests and multiple linear regression,* Hydrology and Earth System Sciences, *In Press*, hess-2011-37

[2] I. Errasti, A. Ezcurra, J. Sáenz, G. Ibarra-Berastegi (2011) *Validation of IPCC AR4 models over the Iberian Peninsula.* Theoretical and Applied Climatology, 103(1):61-79. DOI: 10.1007/s00704-010-0282-y

Scalable and Automated Workflow in Mining Large-Scale Severe-Storm Simulations

Lei Jiang[1,3], Gabrielle Allen[1,3] *, and Qin Chen[2,3]

[1] Department of Computer Science, Louisiana State University
[2] Department of Civil and Environmental Engineering, Louisiana State University
[3] Center for Computation and Technology, Louisiana State University
{ljiang,gallen}@cct.lsu.edu, qchen@lsu.edu

Abstract. The simulation of large-scale complex systems, such as modeling the effects of hurricanes or storms in coastal environments, typically requires a large amount of computing resources in addition to data storage capacity. To make an efficient prediction of the potential storm surge height for an incoming hurricane, surrogate models, which are computationally cheap and can reach a comparable level of accuracy with simulations, are desired. In this paper, we present a scalable and automated workflow for surrogate modeling with hurricane-related simulation data.

Keywords: automated workflow, scalability, severe-storm simulation, data mining, surrogate model

As today the cyberinfrastructure development keeps progressing in many research organizations across the world, the abundance of high-performance computing resources brings new possibilities to domain scientists and engineers in terms of large-scale parallel simulations for complex phenomena. A typical application scenario with high significance comes from severe storms such as hurricanes. A 6-day storm surge simulation can take more than 2,000 CPU hours, which means a 35-hour run on a cluster with 64 cores. Deterministic physics-based simulations are the primary way as a guide for decision makers but become time-consuming and inflexible for real-time predictions. Simulation data, especially for a set of simulations with designed input parameter space, can be analyzed and knowledge can be extracted from them. As an effort to simulation data mining [1], we focus on surrogate modeling from storm simulation data, which leads to lightweight models mimicking the behavior of simulations on points of interest (POI).

In the paper, we first describe the surrogate modeling approaches with large-scale simulation data, where each simulation generates values on multiple POIs. In this way, to make predictions on a target variable at a point or location in the simulation, we use functional data analysis for storm simulations with designed parameter space in two circumstances: the prediction of maximum storm surge height as scalar response and the time-series prediction of surge profile

* Corresponding author: Gabrielle Allen (216 Johnston Hall, Baton Rouge, LA 70803)

as functional response. It is needed to train a different model for each point of interest. Also, for complex systems with dynamics on the response surface at a single point, it is desired to dig out more information. Spatio-temporal causal links [2] can be constructed for the same simulation output variable between locations or across variables. As links are contingent upon the simulation input, like hurricane tracks in our scenario, a granger causality test is performed ahead of regression in order to detect the links that tend to be invariant. Such a modeling framework can thereby supply confidence level and measurements of uncertainty to time-critical predictions.

A scalable and automated workflow is then important to facilitate the data mining process. Our workflow combines parallel simulation, distributed data archive, and high-performance data mining in the same framework. We exploit task-level parallelism to achieve scalability. Two modes in data processing are involved: *i)* a piece of data from each simulation is needed for modeling and those from multiple simulations are assembled for a modeling task, namely *Task Assembling*; and *ii)* the processing can be separately performed for each simulation and finally the results are to be reduced for generalization (Map-Reduce [3]). Then, in the implementation, several components are included: simulation data archive [4], parameter space, pattern space, data mining job pool and model base. Besides scalability, it also ensures that modeling process is automatically handled with continuously increasing data in the archive.

In the demonstration at the SSDBM 2011 conference, we show the workflow performance as well as some surrogate modeling results. Future directions of the work include creating a generic workflow for more data sources and the surrogate models themselves can also be further elaborated.

Acknowledgments

The authors acknowledge the contributions of Kelin Hu (Louisiana State University). This work is funded by NSF/EPSCoR EPS-0701491 (CyberTools) and NSF/EPSCoR EPS-1010640 (Coastal Hazards Collaboratory). Computing resources were provided by the Louisiana Optical Network Initiative (LONI).

References

1. T.F. Brady and E. Yellig, "Simulation data mining: a new form of computer simulation output," in *Proc. of the Winter Simulation Conference*, pp. 285-289, 2005.
2. A. Lozano, H. Li and et al., "Spatial-temporal causal modeling for climate change attribution," in *Proc. of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'09)*, 2009.
3. J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," in *Proc. of USENIX Symposium on Operation Systems Design and Implementation (OSDI'04)*, 2004.
4. H. Bhagawaty, L. Jiang and et al., "Design, Implementation and Use of a Simulation Data Archive for Coastal Science, in *Proc. of Int'l ACM Symposium on High-Perf. Parallel and Distributed Computing (HPDC'10) Workshops*, pp. 651-657, 2010.

Copula-based Approaches to Characterization of Droughts

Sergey Kirshner, Xin Lu Tan
Department of Statistics, Purdue University
250 N University St
West Lafayette, IN 47907-2066
{skirshne,tanx}@purdue.edu

Rao S. Govindaraju
School of Civil Engineering, Purdue University
550 Stadium Mall Drive
West Lafayette, IN 47907-2051
govind@purdue.edu

Abstract

In this work we apply machine learning and statistics techniques to analyze and characterize *droughts* on a regional scale. Broadly speaking, drought is a deficit of water, whether in the form of precipitation (*meteorological drought*), soil moisture (*agricultural drought*), or water supply (e.g., streamflow, groundwater, reservoir; such drought is referred to as *hydrological drought*) [1]. Drought, being a recurring phenomenon with diverse geographical and temporal distribution, is one of the least understood and most expensive natural disasters. The actual causes of droughts are too complex to be fully identified, and a precise scientific definition of droughts does not exist.

Drought severity is measured by drought indices that reflect deviation of hydrologic variables from their long-term averages. Among most commonly employed drought indices is the Standardized Index (SI) [3] in which historical observations are used to compute the probability distribution of the monthly and seasonal (2 months, 3 months, etc., up to 48 months) precipitation totals. The fitted probability distributions are then normalized using the standard inverse Gaussian function to calculate SI. A negative value of SI indicates precipitation less than median rainfall, and the magnitude of departure from zero represents the severity of drought. [3] also suggested a classification scale in which (i) extreme drought occurs when SI value is less than -2.0, (ii) severe drought when SI value is between -2.0 and -1.5, and (iii) moderate drought when SI value is between -1.5 and -1.0. These classifications correspond to the drought categories in the Drought Monitor[1] that currently serves as the standard reference for drought status.

However, such approach provides no mechanism to aggregate the indices over multiple locations or over different temporal scales. To enable regional quantification of droughts, we propose a *regional* drought index combining the observations at multiple locations with the help of *copulas*, joint distribution function over ranks of random variables. More formally, for a vector of d continuous random variables $(x_1, \ldots, x_d)$ with a joint CDF H and marginal CDFs $u_j = F_j(x_j)$, $j = 1, ..., d$, a copula is defined as $C(u_1, \ldots, u_d) = H(x_1, \ldots, x_d)$, and for absolutely continuous H such copula is unique [5]. In the context of drought indices, as SI is also based on the ranks of the water accumulation variables, a copula for observations at multiple precipitation locations can be viewed as a joint distribution of functions of SIs at these locations. Hence, copulas provide a particularly convenient framework for constructing a joint index which can be a statistic of the copula governing the joint behavior of the corresponding precipitation observations. One such possible index is the Kendall distribution function $K_C(t) = P(C(u_1, ..., u_d) \leq t)$,

[1] http://drought.unl.edu/dm

the distribution function of the joint cumulative density function of the copula [4] which can also be viewed ranking function for multi-dimensional objects.

One of the major challenges in our study is a relatively small number of observations (between 50 and 100 years for our region of interest). As the shortest temporal scale of droughts consists of months, and since seasons typically exhibit very different joint distribution of precipitation amounts, fairly few observations are available for estimation of d-dimensional copulas (where d may be in the 15-30 range). To reduce the effective dimension of the underlying process, we impose conditional independence assumptions on the water accumulation variables using tree-structured copulas [2], a graphical model for joint distributions of real-valued (possibly, non-normal) random variables. An introduction of tree structure into the distribution enables the joint density $c(u_1, ..., u_d)$ to factorize as a product of $c(u_i, u_j)$ for each (u_i, u_j) belonging to the set of edges in the tree. Then, the estimation of a tree-structured copula reduces to the estimation of bivariate copulas for its edge pairs of variables.

References

[1] R. R. Heim Jr. A review of twentieth-century drought indices used in the United States. *Bullentin of the American Meteorological Society*, 83:1149–1165, August 2002.

[2] S. Kirshner. Learning with tree-averaged densities and distributions. In J. C. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 761–768, Cambridge, MA, 2008. MIT Press.

[3] T. McKee, N. Doesken, and J. Kleist. The relationship of drought frequency and duration to time scales. In *Proceedings of the 8th Conference on Applied Climatology*, volume 184, 1993.

[4] R. Nelsen. Kendall distribution functions. *Statistics & Probability Letters*, 65(3):263–268, Nov. 2003.

[5] A. Sklar. Fonctions de repartition a n dimensions et leures marges. *Publications de l'Institut de Statistique de L'Universite de Paris*, 8:229–231, 1959.

Self-organized mapping of climate model complexity for performance testing

Christian D. Klose, Think Geohazards, New York, NY

Beate G. Liepert, NWRA, Seattle, WA

Climate models build up complex multi-dimensional data in time and space. The field of **machine intelligence** provides a variety of statistical methods to analyze and benchmark multi-dimensional and complex data. One method that we like to present here is called Self-Organizing Mapping (SOM) as an advancement of vector quantization (VQ).[1] This technique is very often used for data compression and clustering problems. SOM has not only the advantage of clustering multi-dimensional climate model data (climate quantities) into homogeneous sub-groups (climate states) but is also able to present resulting outputs in 1-D graphs or 2-D images in a meaningful way. SOM can transform climate quantities into simplified generally 2-D discrete maps (Kohonen layer). Thus, n-dimensional feature vectors, which are built up by n climate quantities, can be visualized in these maps. Clusters of these feature vectors, which represent uniform climate states (groups of similar features), can be also visualized in these maps.[2] A wide spectrum of other machine intelligence methods exist that can be applied to interpreting climate model data. SOM, however, is well tested and applied reliably to several disciplines showing convincing results.[2-4]

Here we present preliminary results where SOM was utilized to delineate the performance of the **GFDL20** and **GFDL21** climate models as part of the Climate Model Intercomparison Project 3 (CMIP3). The analysis was performed on climate variable mean values in a period 1979-1999 and limited for the northern hemisphere. Moreover, eight climate quantities that represent the atmospheric state (wind components, cloud cover, precipitation, temperature, radiation, pressure) were drawn from all climate models and reanalysis data. Performances of GFDL20 and GDFL21 were compared to the CMIP3 climate model ensemble means and the NCEP-NCAR and ERA40 reanalysis.[5,6] Model performances were determined by the difference of the 8-dimensional climate variable vector x={u850, v850, pr, clt, t200, u200, rsut, psi} and NCEP-NCAR and ERA40 reanalysis data x'={u850', v850', pr', clt', t200', u200', rsut', psi'}. The applied SOM algorithm delineates the performance of all 24 climate models[5,6] including both GFDL models. Each of the eight climate quantities shows varying differences in space and time. SOM allows the visualization of these model performances and enables quantitative analysis of the benchmarks of the models.

Finally, we conclude our results show the capability of SOM/VQ for analyzing the complexity of climate models, which are, in general, of higher dimension than shown in this case study. And hence enables an objective and reliable benchmarking of model performances.

References

[1] Kohonen, T. (2001) Self-Organizing Maps. 3rd edition, Springer, Berlin.

[2] Klose, CD (2006) Self-Organising Maps for Geoscientific Data Analysis: Geological Interpretation of Multi-dimensional Geophysical Data, *Comput. Geosciences* 10(3), 265-277.

[3] Klose, CD, Klose, AD, Netz, U. et al. (2008) Multiparameter classifications of optical tomographic images, *Journal of Biomedical Optics*, 13, 050503, DOI:10.1117/1.2981806

[4] Klose, CD, Seo, S, Obermayer K (2005) A new clustering approach for partitioning directional data. *IJRMMS* 42, 315-321.

[5] Reichler, T. and Kim J. (2008) How Well Do Coupled Models Simulate Today's Climate? *BAMS*, 303-311.

[6] Gleckler, P.J., Taylor, K.E. Dotriaux, C. (2008) Performance metrics for climate models. *JGR* 113, D06104.

For: The First International Workshop on Climate Informatics; NYC; August 2011

Title: Toward improvement of Bayesian combination of multiple climate models for regional uncertainty assessment

Evan Kodra[1, 2], Snigdhansu Chatterjee[3], Auroop R. Ganguly[1, 2, *]

[1]University of Tennessee, Knoxville, TN
[2]Oak Ridge National Laboratory, Oak Ridge, TN
[3]University of Minnesota, Minneapolis, MN
*Corresponding Author: gangulyar@ornl.gov/auroop@gmail.com

Abstract

Regional climate prediction is one of the major gaps in climate science; yet, many resource and infrastructural decisions will be made at regional or local levels. This motivates uncertainty assessment relevant for this scale of decision-making. Bayesian approaches attempt to construct uncertainty distributions for regional mean temperature change by weighting global climate models (GCM) simulations via hindcast bias derived from comparison to observations as well as multimodel agreement in the future. Here, we reveal potential problems with the latest Bayesian model and propose several changes which seek to improve data handling and the physical interpretability of GCM weights, thus possibly enhancing the quality of uncertainty assessments. We also highlight alternative approaches to regional climate prediction and uncertainty quantification developed recently in statistics and machine learning.

as functional response. It is needed to train a different model for each point of interest. Also, for complex systems with dynamics on the response surface at a single point, it is desired to dig out more information. Spatio-temporal causal links [2] can be constructed for the same simulation output variable between locations or across variables. As links are contingent upon the simulation input, like hurricane tracks in our scenario, a granger causality test is performed ahead of regression in order to detect the links that tend to be invariant. Such a modeling framework can thereby supply confidence level and measurements of uncertainty to time-critical predictions.

A scalable and automated workflow is then important to facilitate the data mining process. Our workflow combines parallel simulation, distributed data archive, and high-performance data mining in the same framework. We exploit task-level parallelism to achieve scalability. Two modes in data processing are involved: *i)* a piece of data from each simulation is needed for modeling and those from multiple simulations are assembled for a modeling task, namely *Task Assembling*; and *ii)* the processing can be separately performed for each simulation and finally the results are to be reduced for generalization (Map-Reduce [3]). Then, in the implementation, several components are included: simulation data archive [4], parameter space, pattern space, data mining job pool and model base. Besides scalability, it also ensures that modeling process is automatically handled with continuously increasing data in the archive.

In the demonstration at the SSDBM 2011 conference, we show the workflow performance as well as some surrogate modeling results. Future directions of the work include creating a generic workflow for more data sources and the surrogate models themselves can also be further elaborated.

Acknowledgments

The authors acknowledge the contributions of Kelin Hu (Louisiana State University). This work is funded by NSF/EPSCoR EPS-0701491 (CyberTools) and NSF/EPSCoR EPS-1010640 (Coastal Hazards Collaboratory). Computing resources were provided by the Louisiana Optical Network Initiative (LONI).

References

1. T.F. Brady and E. Yellig, "Simulation data mining: a new form of computer simulation output," in *Proc. of the Winter Simulation Conference*, pp. 285-289, 2005.
2. A. Lozano, H. Li and et al., "Spatial-temporal causal modeling for climate change attribution," in *Proc. of the 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'09)*, 2009.
3. J. Dean and S. Ghemawat, "MapReduce: simplified data processing on large clusters," in *Proc. of USENIX Symposium on Operation Systems Design and Implementation (OSDI'04)*, 2004.
4. H. Bhagawaty, L. Jiang and et al., "Design, Implementation and Use of a Simulation Data Archive for Coastal Science, in *Proc. of Int'l ACM Symposium on High-Perf. Parallel and Distributed Computing (HPDC'10) Workshops*, pp. 651-657, 2010.

Modeling Winter Rainfall over Northwest India Using a Non-homogeneous Hidden Markov Model

Indrani Pal[1], Andrew Robertson[1], Upmanu Lall[2], Mark A Cane[3]

[1] *International Research Institute for Climate and Society, The Earth Institute at Columbia University, Palisades, NY 10964*

[2] *Department of Earth and Environmental Engineering, Columbia University, New York, NY 10027*

[3] *Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964*

A multiscale-modeling framework for daily rainfall is considered and diagnostic results are presented for an application to the winter season in Northwest India. The daily rainfall process is considered to follow a Hidden Markov Model, with the hidden states conditioned on slowly varying climate indices that are presumed to modulate the winter jet stream and moisture transport dynamics. The data used are from 11 stations over Satluj River basin in northwest India in winter (Dec-Jan-Feb; 1977/78-2005/06). The HMM identifies four discrete weather states, which were useful to describe daily rainfall variability in winter over the study region. Two states were depicted as very wet (state 2) and very dry (state 1) and another two states were fairly wet (state 3) and fairly dry (state 4) conditions. Interannual changes in daily occurrence frequencies of states 2–4 were found to be negatively correlated with the state 1 (very dry), with a maximum negative correlation with state 4; whereas mutual correlations between state 2, 3 and 4 were low. Each HMM state was found to be associated with a distinct atmospheric circulation pattern in winter. Most notably, state 1 and 2 (also 3) were found to be associated with opposite atmospheric conditions. The occurrence of very wet conditions (state 2) is strongly conditioned by the El-Nino phenomena in winter, which was also captured in the large-scale correlation maps with the sea level pressure and sea surface temperature. This suggests that there is a tendency of higher frequency of wet days in winter in the El-Nino years. On the other hand, the Arctic Oscillation in Nov and the Atlantic Multidecadal Oscillation in Sept also showed significant correlations indicating their effects on the

occurrence of various winter rainfall states. Winter rainfall over northwestern India is brought about by the mid-latitude storms embedded in the westerlies in winter, which is still poorly understood. This study helps to recognize the sequence of Northern Hemisphere mid-latitude storms bringing winter rainfall over the study region using HMM and their association with low frequency modes that may be sources of predictability on seasonal time scales and longer. The generalization of the model presented using a multilevel hierarchical Bayesian Network is also considered. Algorithms for learning such a network are explored.

Draft abstract for The First International Workshop on Climate Informatics, August 26th 2011, The New York Academy of Sciences

Hidden Markov models for stochastic simulation and prediction of daily rainfall across scales

Andrew W. Robertson[1], Arthur M. Greene[1], Padhraic J. Smyth[2] and Scott Triglia[2]

[1] International Research Institute for Climate and Society (IRI), Earth Institute, Columbia University, Palisades, NY
[2] Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA

The statistics of daily sequences of rainy and dry days characterize hydroclimate variability at the local spatial scale. A better understanding how these statistics are modulated by low-frequency modes of climate variability and anthropogenically-forced climate change is urgently needed for societal application of climate science, such as through seasonal forecasting for water management or flood risk.

We have developed nonhomogeneous hidden Markov models (NHMMs) that allow daily rainfall sequences recorded on a network of raingauges to be modeled as a function of (1) GCM seasonal climate forecasts, and (2) GCM climate change projections. An attractive feature of the NHMM is the potential to use the hidden rainfall meta states as a diagnostic linkage between the large-scale exogenous GCM forcing and the station-scale daily rainfall sequences. In the seasonal forecasting context, the NHMMs have been applied to downscale GCM seasonal forecasts over several tropical countries; an Indonesian example is highlighted in which the onset date of the monsoon is well captured by the simulations. In the climate change context we explore the non-stationarity of the rainfall states over India from IPCC 4th Assessment Report GCM runs, and their implications for likely changes in daily rainfall statistics.

Study of the Climate System Bifurcations: Analytical and Numerical Approach

I.A. Sudakov[1,2], S.A. Vakulenko[2]

[1]*St. Petersburg State University*

[2]*St. Petersburg State University of Technology and Design*

Problems of climate warming are under great attention last decades. In particular, it is interesting to estimate effects connected with methane emission and propagation in atmosphere since methane is a very dangerous greenhouse gas. A great attention is given to the problem on methane emission from Siberian tundra. In this region there are a number of small lakes producing methane. The methane emission is connected here, in particular, with existence of such lake system; these lakes slowly evolve in a complicated way.

We propose a new combined model, which can describe methane propagation in the atmosphere. The model is a development of the Goody model based on Rayleigh-Benard convection that serves as a paradigm of fluid mechanical stability and turbulence. We extend the Goody model taking into account effects connected with methane production, a chemical degradation of methane molecules, methane diffusion and convection (here we consider methane as a passive scalar). Although the model proposed is a slightly artificial mathematical idealization, such model can be useful as a first approximation for estimation of the methane emission influence, since this model is close to classical models for Rayleigh-Benard convection, where many linear and nonlinear stability results have been established.

The main results can be described as follows. First, using this model we obtain a general relation for a change of temperature induced by methane emission by tundra lakes. Second, we find, by fairly general variational relations, how the methane emission influences the bifurcation parameter values for the Goody model (earlier bifurcations have been investigated by V. Larson). Third, at the bifurcation point we obtain a system of ordinary differential equations describing a slow time evolution of amplitudes of main modes.

For: The First International Workshop on Climate Informatics; NYC; August 2011

Title: Soil carbon cycle insights from network analysis

Joshua Tolen[1, 2], Evan Kodra[1, 2], Karsten Steinhaeuser[3], Auroop R. Ganguly[1, 2, *], Stan Wullschleger[2], Charles Garten, Jr.[2], Robin Graham[2]

[1]University of Tennessee, Knoxville, TN
[2]Oak Ridge National Laboratory, Oak Ridge, TN
[3]University of Minnesota, Minneapolis, MN
*Corresponding Author: gangulyar@ornl.gov/auroop@gmail.com

Understanding the soil carbon cycle is becoming more important as simultaneous needs for bioenergy, greenhouse gas mitigation, and food security become increasingly evident. Here, an analysis of high dimensional experimental data culminates in the construction of several complex networks, providing analytic representations and visual depictions of relationships that are difficult to capture coherently with the use of traditional statistical methods. Analysis of these networks may reveal interesting and novel insights previously unknown in the soil science domain. Findings need to be interpreted based on current understanding of soil science and examined for new hypothesis generation, further expanding useful knowledge and conclusions within the domain.