Understanding Dominant Factors for Precipitation over the Great Lakes Region

Soumyadeep Chatterjee[1], **Stefan Liess**[2], **Arindam Banerjee**[1] and **Vipin Kumar**[1]
[1]Department of Computer Science and Engineering., [2]Department of Soil, Water, and Climate
University of Minnesota, Twin Cities
Minneapolis, MN 55455
{chat0129, liess}@umn.edu, {banerjee,kumar}@cs.umn.edu

Abstract

Statistical modeling of local precipitation involves understanding local, regional and global factors informative of precipitation variability in a region. Modern machine learning methods for feature selection can potentially be explored for identifying statistically significant features from pool of potential predictors of precipitation. In this work, we consider sparse regression, which simultaneously performs feature selection and regression, followed by random permutation tests for selecting dominant factors. We consider average winter precipitation over Great Lakes Region in order to identify its dominant influencing factors. Experiments show that global climate indices, computed at different temporal lags, offer predictive information for winter precipitation. Further, among the dominant factors identified using randomized permutation tests, multiple climate indices indicate the influence of geopotential height patterns on winter precipitation. Using composite analysis, we illustrate that certain patterns are indeed typical in high and low precipitation years, and offer plausible scientific reasons for variations in precipitation. Thus, feature selection methods can be useful in identifying influential climate processes and variables, and thereby provide useful hypotheses over physical mechanisms affecting local precipitation.

1 Introduction

Understanding climate change and its impacts on policy and infrastructure involves prediction of state of earth's climate under different forcing scenarios (Moss et al. 2010). One of the most important variables of interest in modeling climate is precipitation, particularly at regional or local scales. Earth System Models (ESM) (Randall and others 2007) that model the physics and dynamics of climate, are known to have deficiencies in modeling local precipitation (Kang and others 2002; Gao and others 2008; Piani, Haerter, and Coppola 2010). This shortcoming is mainly due to the spatial resolution of the models, which is often too coarse to accurately model local and regional precipitation (Gao and others 2008). Therefore there exists a gap in understanding of the factors affecting precipitation over small scale regions on the globe.

Increasingly, statistical models are being considered to inform climate science research on factors which may affect precipitation (Das 2015). The goal is to discover statistical dependencies between precipitation and covariates of interest, and then try to gain a mechanistic physical understanding of how the covariates affect precipitation. The covariates or predictors are often multi-scale climate variables and processes, which may manifest their effect with some temporal lags, or in conjunction with each other (Steinhaeuser, Chawla, and Ganguly 2011). For a given region of interest, there is a plethora of possible influencing factors for precipitation, such as ocean oscillations, atmospheric variables, and long-term ocean-atmosphere coupled processes (Cunderlik and Simonovic 2007). Therefore, it is of interest to the climate research and modeling community to understand the most influential factors in this pool of predictors, and derive climatological insights from such a discovery process.

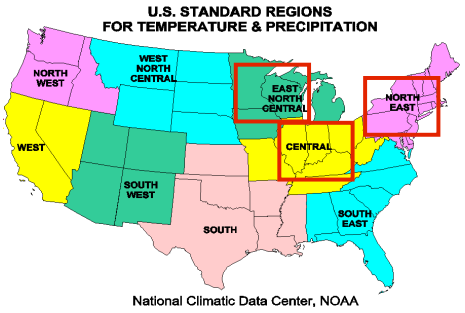
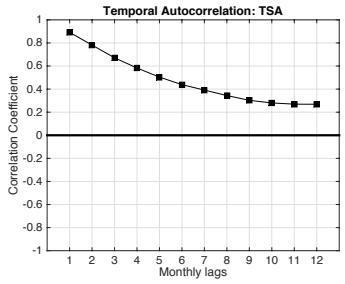
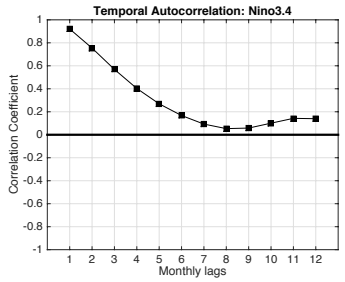


Figure 1: U.S. Standard Climatological Regions (Karl and Koss 1984). The Great Lakes consist of the three marked regions.

In this work, we consider prediction of precipitation over the Great Lakes region of the US (Fig. 1), using predictor variables at multiple spatial scales with temporal lags. The predictors include atmospheric variables at local and regional scales, as well as multiple global climate indices (Stenseth and others 2003) that capture climate processes and oscillations. The global climate indices are time series, derived from oceanic and atmospheric fields over different regions on the earth, that are known to capture certain periodic climate events such as oscillations (Allan and others 1996), and/or correlate with variations in atmospheric fields controlling global climate (Chambers, Tapley,

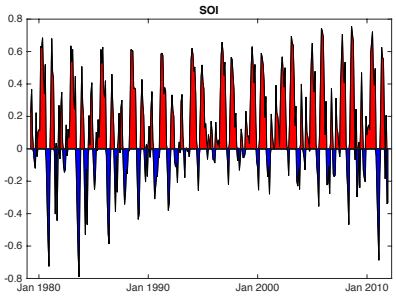


(a) TSA Index

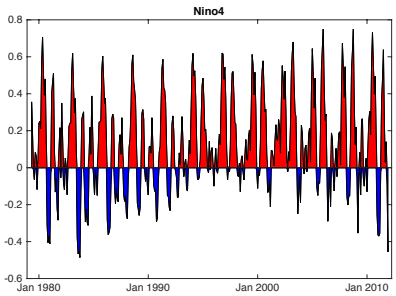


(b) Nino3.4 Index

Figure 2: Monthly temporal autocorrelations in climate indices computed over 1979-2011. Some indices, such as Tropical/Southern Atlantic Index (TSA) have significant correlations for upto 11 months.



(a) Monthly SOI index



(b) Monthly Nino4 index

Figure 3: Climate Indices over Pacific which capture the El-Nino Southern Oscillation (ENSO)

and Stewart 1999; Stenseth and others 2003). We consider the December-January-February (DJF) or winter mean precipitation at a given weather station in the region as the response. The goal, therefore, is to understand the dominant factors for precipitation from among the large pool of possible predictors.

Sparse regression methods, such as LASSO (Tibshirani 1996), are useful in this scenario. Such methods allow simultaneous feature selection and regression, and are often supported by theoretical guarantees (Negahban et al. 2012; Banerjee et al. 2014). LASSO has been found to perform well empirically in multiple other domains, and also provides fast solvers for efficient implementation (Liu and Ye 2010). However, often predictors have temporal autocorrelations (Fig. 2), and since stations located in a region have geographical proximity, the data samples are also spatially correlated. Further, different climate indices related to the same climate phenomenon may be mutually correlated (Figs. 3(b) and 3(a)). In presence of such correlations, the set of features selected by LASSO may exhibit instability, and may include spurious predictors. In order to address this issue, we consider significance testing of selected set of predictors, to obtain stable and statistically significant covariates as dominant predictors. We use a random permutation test (Pendse et al. 2012), to test the significance of each selected predictor, followed by composite analysis to gain a physical understanding of the effect of covariates on precipitation.

The rest of the paper is arranged as follows. We briefly review related work in Section 2. In Section 3, we overview the sparse regression methodology, and the random permutation testing framework. We describe the dataset used and pre-processing techniques in Section 4. In Section 5, we present experimental results, and discussions. Finally, we conclude in Section 6

2 Related Work

In recent years statistical modeling is receiving attention from the climate science community for improving predictive performance of traditional physical models (Peña and van den Dool 2008), as well as for statistical downscaling (Hessami et al. 2008). Ridge regression, particularly, has been widely used for multimodel ensemble forecasting with Earth System Models (ESM) (DelSole 2007), and for modeling transformation functions for computing surface temperature from satellite data (McMillin, Crone, and Crosby 1989). However, regression with dimensionality reduction or feature selection has often been used in the context of statistical downscaling (Corte-Real, Zhang, and Wang). Most commonly, regression methodologies involve application of principal component analysis (PCA) to covariates to reduce dimensionality, followed by multivariate linear or non-parametric regression models on the principal component scores (Ghosh and Mujumdar 2008). Such methods are unsuitable for hypothesis generation since they do not allow feature selection. Recently, (Das 2015) has applied sparse regression for understanding factors for annual precipitation extremes over continental U.S. However, since precipitation

mechanisms vary widely over seasons, and individual climatological regions within the U.S., such an approach cannot capture seasonal factors for precipitation.

3 Sparse Regression for Feature Selection

Sparse regression allows one to simultaneously conduct feature selection and regression, thus enabling selection of the most predictive set of features. We consider a linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}\,, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n, \mathbf{X} \in \mathbb{R}^{n\times p}$ are samples and $\boldsymbol{\beta}^*$ is a p-dimensional coefficient vector. The LASSO (Tibshirani 1996) method estimates a sparse $\hat{\boldsymbol{\beta}}$, by solving the following estimation problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\,, \tag{2}$$

where $\lambda > 0$ is a regularization parameter. In the context of discovering the dominant factors, often there is no prior bias on the sparsity imposed on the coefficients, although some of the covariates considered in the model may have strong correlations among each other, and temporal autocorrelation within itself over monthly or seasonal values. For example, consider the Nino4 index (Fig. 3(b)), which is computed from sea surface temperature, and the SOI index (Fig. 3(a)), which is derived from sea level pressure. Both indices carry information regarding the El-Nino Southern Oscillation (ENSO) (Allan and others 1996), albeit from different climate variables. Hence they exhibit a high negative correlation (about -0.6). Some of the indices also show high temporal autocorrelations (Fig 2).

In presence of such correlations in covariates and samples, the set of features selected by LASSO often have instability (Meinshausen and Bühlmann 2010). Further, for finite samples, there is a non-zero probability that for a given training set and a chosen penalty parameter LASSO selects a non-zero coefficient for a non-informative predictor by random chance. Therefore, we require a significance testing method to test each non-zero coefficient estimated by LASSO on training data, and compute a p-value for significance of each feature.

Testing significance of covariates has been considered in various problems of applied statistics, and the most commonly used testing methodology is random permutation test (Nichols and Holmes 2002; Manly 2006; Ojala and Garriga 2010). Such a test is a nonparametric hypothesis testing framework, which measures the significance of every non-zero coefficient value by constructing a random distribution over the coefficient using random permutations of the data. We adopted a variation of the methodology developed in (Pendse et al. 2012) that we discuss next.

Permutation Test

We fix λ at a particular value. On the training data, we first compute the LASSO estimate $\hat{\boldsymbol{\beta}}$ by solving (2). Next, keeping $\mathbf{X}$ constant, we randomly permute the response $\mathbf{y}$ to obtain a vector $\tilde{\mathbf{y}}$. The random permutation of the response destroys any statistical relationship existing between the covariates in $\mathbf{X}$ and the response $\tilde{\mathbf{y}}$. Thereafter, we run LASSO with $\mathbf{X}$ and $\tilde{\mathbf{y}}$ in order to obtain a random coefficient vector $\tilde{\boldsymbol{\beta}}$, which represents random causal relationships between the covariates and the response. Executing this strategy multiple ($\nu \geq 1000$) times, for the i-th non-zero coefficient in $\hat{\boldsymbol{\beta}}$, we compute the probability that a random value $|\tilde{\boldsymbol{\beta}}_i|$ exceeds the estimated value $|\hat{\boldsymbol{\beta}}_i|$ given by

$$p_i = \frac{count(|\tilde{\boldsymbol{\beta}}_i| \geq |\hat{\boldsymbol{\beta}}_i|)}{\nu + 1}\,. \tag{3}$$

It represents the p-value associated with the corresponding coefficient $\hat{\boldsymbol{\beta}}_i$.

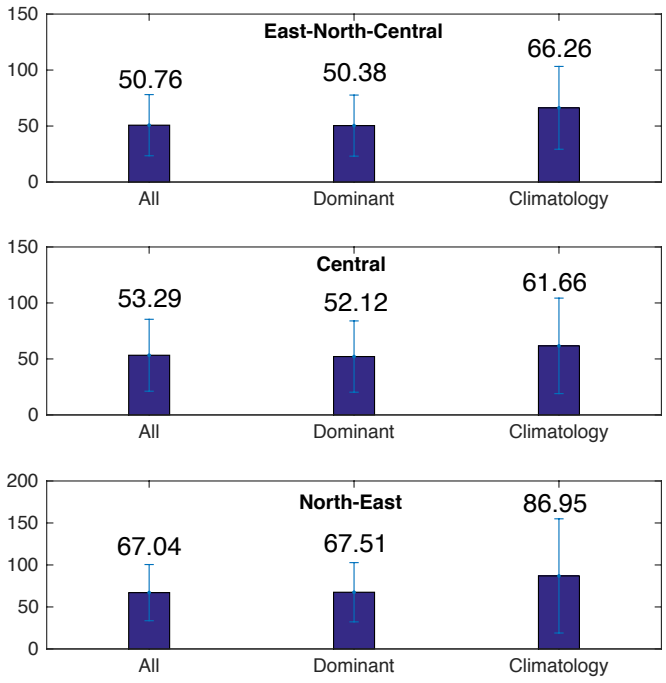


Figure 4: Root Mean Square Error (RMSE) on precipitation prediction (in hundredths of an inch) of Ordinary Least Squares regression using only dominant factors and using all covariates. Prediction errors from long-term climatology is also plotted. The error bars denote one standard deviation.

4 Dataset

We compiled datasets from two sources: (1) United States Historical Climatological Network (USHCN) (Menne, Williams Jr, and Vose 2010), and (2) North American Regional Reanalysis (NARR) (Mesinger, DiMego, and others 2006). We considered the three climate regions that surround the Great Lakes, as shown in Fig. 1. Precipitation data was obtained from station records in USHCN in the above states that lie near ($< 200km$) of the lakes. These stations are located in one of the following 8 states of U.S.:(i) Minnesota (MN), (ii) Wisconsin (WI), (iii) Illinois (IL), (iv) Indiana (IN), (v) Michigan (MI), (vi) Ohio (OH), (vii) Pennsylvania (PA) and (viii) New York (NY). For each station,

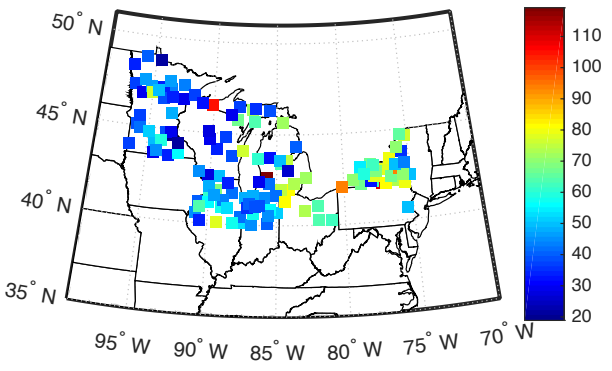


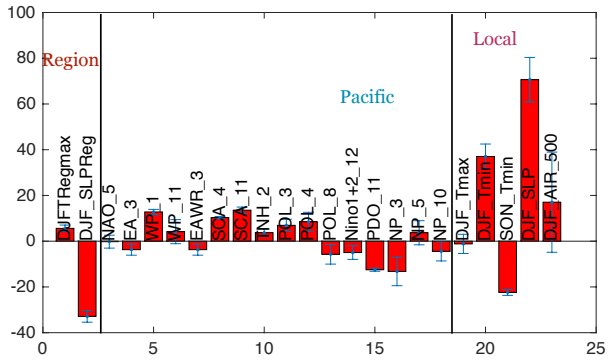
Figure 5: Geographical Spread of RMSE (in hundredths of an inch) over Great Lakes. The North-East has higher errors than other regions.

data for daily maximum/minimum temperature, and precipitation are directly available. We considered the average winter (DJF) precipitation for each station as a response, where the average is over 3 months' daily data. Therefore, for every region, we had winter precipitation data for stations for 1979-2011.

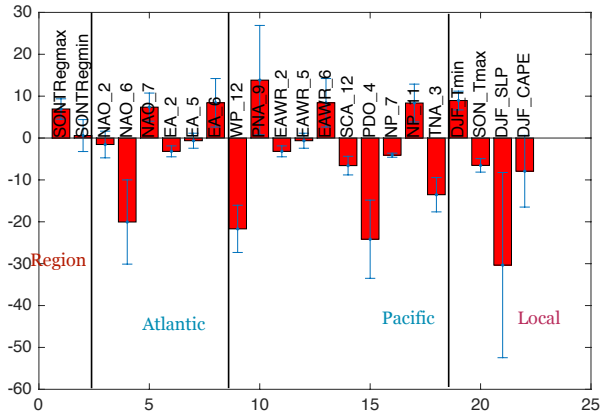
The covariates consisted of local, regional and global climate variables (listed in Table 1 in (Chatterjee et al.)). The surface temperature data was obtained by taking the seasonal maximum and minimum from the daily USHCN data. Further, we obtained seasonal average pressure and convective available potential energy (CAPE) over winter (DJF) and autumn (SON) as covariates by interpolating the NARR data to the local stations available from USHCN. The regional average for local covariate listed above was obtained by computing the area weighted average from the local station values. For each global climate index, we considered all 12 preceding values (from Jan to Dec. of a year) as covariates. We discarded the lower and upper one percentile of the data since these correspond to very low and very high precipitation, and therefore are "extreme events" (Das 2015), which often have very different mechanisms than normal precipitation (Liu et al. 2009). In total, the dataset had $\sim$ 2200 samples over 32 years, where we discarded samples which contained missing values. We divided the data into two sets. The first, comprising of 22 years' data, was used for finding dominant factors. The second set, with the remaining 10 years' data, was used to test predictive performance. Finally, we standardized the covariates in the two datasets by computing the mean and variance of each predictor in the training set of 22 years described above, and using them for the standardization procedure.

5 Results and Discussion

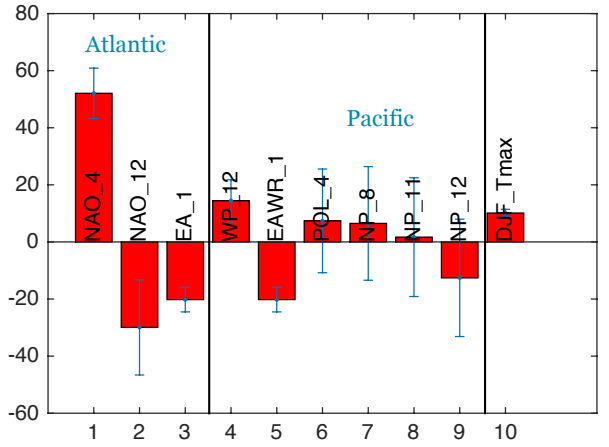
We used a randomly selected 22 years' data for obtaining the dominant features. Further, we conducted leave-one-out cross-validation on the remaining 10 years' data to test the predictive performance of the dominant predictors.



(a) East-North-Central



(b) Central



(c) North-East

Figure 6: Dominant factors for precipitation in each region. The standard abbreviation for each index has been used, along with the month represented as a number. Influences from Atlantic and Pacific are evident in all three regions, mainly from tropical and east pacific, and north atlantic. Multiple summer index values are deemed significant. Further local atmospheric influences are deemed more predictive for regions further inland, while oceanic indices are the sole dominant factors in the North-East.

5.1 Predictive Performance

It is important to assess the predictive performance of the dominant factors found by the proposed method against the climatology of each region. The climatology denotes the long-term average of precipitation over the region. Predictive covariates need to show improvement upon the prediction from long-term climatology, in order to be considered for further hypothesis generation on the mechanism of precipitation.

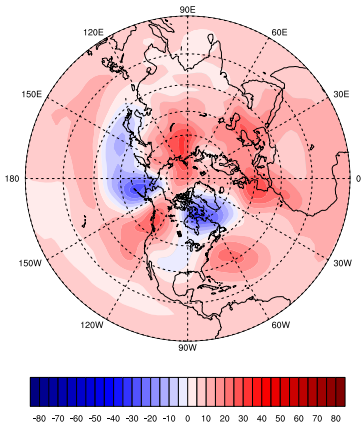
We conducted leave-one-year-out cross-validation on held out test set described earlier. Fig. 4 shows the root mean square error (RMSE) from ordinary least squares regression using dominant factors (less than 25 factors in each region) vs. the entire pool of 232 predictors. The performance is identical (2-sample t-test p-value more than 0.8 on all three cases), and much better than simply predicting the climatological mean. This illustrates that the dominant predictors carry almost all predictive information available in the set of covariates. Note that the dominant factors are discovered using a statistical estimation procedure from data, and are not guided by any physical constraints.

Further, for each station, we computed the MSE in the test set during crossvalidation. In Fig. 5, we have plotted MSE at each geographic location of the stations. MSE in the inland locations (Central and East-North-Central region) are lower than in the North-East region. Higher MSE in the North-East is understandable due to the complex processes which affect variation of precipitation in this region. The north pacific jet stream and the Lake Effect often causes large variation, along with influences of winds from Atlantic, since the area is near the coast.

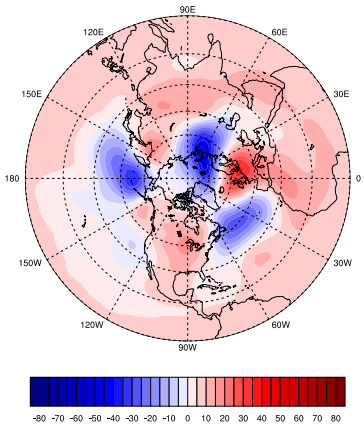
5.2 Dominant Factors

For each climatological region, we obtained a subset of features as the dominant factors, which are plotted in Fig. 6. For choosing the regularization parameter λ, we selected 2% of the training set as a validation set and selected λ that provides the smallest prediction mean square error (MSE) on this validation test. In Fig. 6, for each selected factor, we also plot the mean and standard deviation (as error bars) of the coefficient obtained during the leave-one-out cross-validation. Some interesting patterns emerge from these figures. Surface air temperature in winter plays a prominent role in precipitation during the winter season. It is well known that high snowfall years typically experience lower than normal minimum temperature. Moreover this effect is more pronounced in inland regions (East-North-Central and Central). However, note that since heavy precipitation may itself lower the surface temperature, the relationship may depict correlation rather than causation.

Sea level pressure (SLP), which is a dominant factor in the ENC and Central regions, has influence on the surface level winds that carry moisture from the Pacific across the continent to the Great Lakes. Lower SLP over the region is often associated with higher moisture flow and thus higher precipitation. However, since variations of SLP is a surface phenomenon, it is more noisy as a predictor in seasonal scales than higher atmospheric variables. Therefore, we obtain higher variance in the weights for SLP over crossvalidation runs.



(a) High precipitation



(b) Low precipitation

Figure 7: Average 700mb Geopotential height anomalies in December over (a) 10 highest precipitation years, and (b) 10 lowest precipitation years. There exists strong negative anomaly (low pressure) over mainland U.S. in high precipitation years thus increasing moisture flow from Pacific.

The North-East region behaves differently in that only a single local atmospheric variable is selected as dominant factor. This may be indicative of the fact that the North-East region, due to its proximity to the ocean, is influenced heavily by oceanic effects. As noted earlier, it is known that there are multiple factors for variation of precipitation over this region. For example, the Lake Effect (Niziol, Snyder, and Waldstreicher 1995) has substantial impact on snowfall during winter, which is influenced by the Pacific jet stream. Due to this phenomenon, often the region experiences very heavy snowfall over only a few days or hours.

Atlantic and Pacific influences are prominent across the entire Great Lakes region. Moreover, comparison of the three panels of Fig. 6 shows that Atlantic influences be-

come more prominent on the eastern part of the Great Lakes, while most stable indices in the ENC region are computed over Pacific. Particularly, consider the dominant factors of precipitation over the ENC region. The dominant climate indices are mainly EA (East Atlantic Pattern), WP (West Pacific Pattern), SCA (Scandinavian Pattern), TNH (Tropical/Northern Hemisphere Pattern), POL (Polar/Eurasia Pattern), PDO (Pacific Decadal Oscillation) and NP (Northern Pacific Oscillation). All of these indices are computed from or have high correlation with 700mb – 500 mb geopotential height anomalies. Therefore, we construct composites for geopotential height anomalies in order to further investigate the processes leading to precip variations across the ENC region.

5.3 Composites over Geopotential Height Anomalies

In Fig. 7, we plot average December 700mb geopotential height anomalies over the northern hemisphere, where the average is taken over the 10 highest and 10 lowest precipitation years. Fig. 7(a) shows a strong negative anomaly over Canada and north-central U.S. denoting existence of a low pressure system. The strong low pressure system is conducive for increased wind flow from northern Pacific, which picks up moisture from the Pacific ocean and thus favors higher moisture content in the air. In the presence of colder temperatures over much of the region, this may lead to increased precipitation.

In stark contrast, Fig. 7(b) illustrates that a *positive* anomaly exists over the entire U.S. for seasons with low precipitation. Such anomalies are associated with higher than average pressure system over the region, and may adversely affect precipitation in two ways. First, since the Pacific has negative anomalies (Fig. 7(b)), the system is not conducive for wind flowing into the continent from the Pacific. Thus it leads to less moisture flow into the region. Second, the positive anomalies at higher levels (700mb) may also lead to down drafts from the upper atmosphere, thus decreasing convective precipitation.

The two panels in Fig. 7 represent *typical* patterns for geopotential height anomalies over the northern hemisphere for high and low precipitation seasons. Although such influences are known in climate science, it is reassuring that the statistical estimation procedure is able to discover such influences in a purely data driven manner. The typical patterns seem to be captured by climate indices, and have predictive information about local precipitation. Otherwise, in Fig. 4, the performance of the predictive model would not be better than the climatology of the region. Further such patterns are often persistent over months leading to winter. Fig. 8 illustrates the geopotential height anomalies averaged over 10 highest precipitation years over ENC region. The low pressure region moves East from over Pacific in September to over U.S. in December, which is consistent with the movement of the Westerlies.

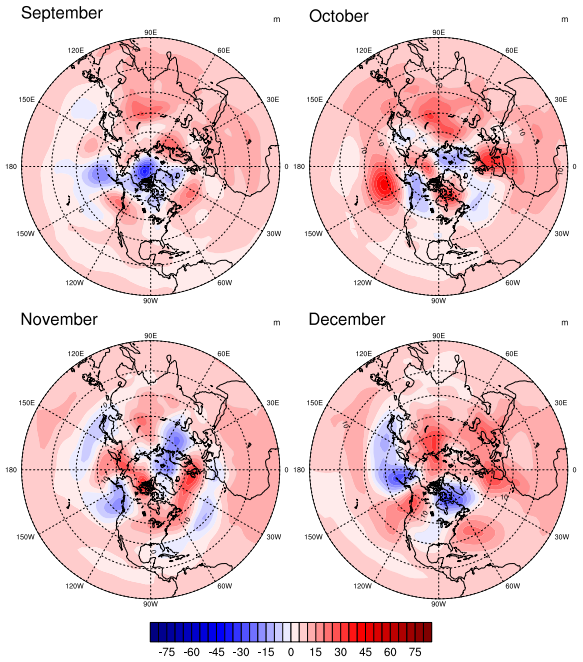


Figure 8: Geopotential height anomalies averaged over 10 highest precipitation years over ENC region in months leading to winter. The low pressure region shifts from Pacific to over the U.S. over the Fall months along the westerlies.

6 Conclusions

In this paper, we proposed a method for discovery of dominant factors for precipitation over the Great Lakes region using a sparse regression method, in conjunction with permutation test for significance. Dominant factors discovered through this process showed high predictive power and produced lower error than obtained from climatology. Further, composite analysis of some of the discovered factors shows that certain seasonal atmospheric patterns may affect precipitation over the region, and is consistent with understanding from climate science. Thus, the proposed method may be useful for deriving hypotheses over how stable atmospheric patterns, such as variations in geopotential heights, may produce scenarios which influence wind and moisture flow, and thus precipitation. In general, the method will be useful for constructing such hypothesis in various statistical modeling scenarios in climate, which can then be further investigated for statistical and physical significance.

Acknowledgments

We would like to thank Debasish Das for help with the climate data. This research was supported in part by NSF Grants IIS-1029711, IIS-0916750, SES-0851705, IIS-0812183, and NSF CAREER Grant IIS-0953274. We are grateful for technical support from University of Minnesota Supercomputing Institute (MSI).

References

Allan, R., et al. 1996. El nino: Southern oscillation and climatic variability.

Banerjee, A.; Chen, S.; Fazayeli, F.; and Sivakumar, V. 2014. Estimation with Norm Regularization. In *Advances in Neural Information Processing Systems 27*. 1556–1564.

Chambers, D. P.; Tapley, B. D.; and Stewart, R. H. 1999. Anomalous warming in the indian ocean coincident with el niño. *J. of Geoph. Res.* 104:3035–3047.

Chatterjee, S.; Liess, S.; Banerjee, A.; and Kumar, V. Supplementary material. `http://www-users.cs.umn.edu/~chatter/papers/15/supplement.pdf`.

Corte-Real, J.; Zhang, X.; and Wang, X. Downscaling gcm information to regional scales: a non-parametric multivariate regression approach. *Climate Dynamics* 11(7):413–424.

Cunderlik, J., and Simonovic, S. 2007. Inverse flood risk modelling under changing climatic conditions. *Hydrological processes* 21(5):563–577.

Das, D. 2015. *Bayesian sparse regression with application to data-driven understanding of climate*. Ph.D. Dissertation, Temple University.

DelSole, T. 2007. A bayesian framework for multimodel regression. *Journal of climate* 20(12):2810–2826.

Gao, X., et al. 2008. Reduction of future monsoon precipitation over china: Comparison between a high resolution rcm simulation and the driving gcm. *Meteorology and Atmospheric Physics* 100(1-4):73–86.

Ghosh, S., and Mujumdar, P. 2008. Statistical downscaling of gcm simulations to streamflow using relevance vector machine. *Advances in Water Resources* 31(1):132–146.

Hessami, M.; Gachon, P.; Ouarda, T. B.; and St-Hilaire, A. 2008. Automated regression-based statistical downscaling tool. *Environmental Modelling & Software* 23(6):813–834.

Kang, I.-S., et al. 2002. Intercomparison of the climatological variations of asian summer monsoon precipitation simulated by 10 gcms. *Climate Dynamics* 19(5-6):383–395.

Karl, T., and Koss, W. J. 1984. *Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983*. National Climatic Data Center.

Liu, J., and Ye, J. 2010. Moreau-Yosida regularization for grouped tree structure learning. In *NIPS*.

Liu, S. C.; Fu, C.; Shiu, C.; Chen, J.; and Wu, F. 2009. Temperature dependence of global precipitation extremes. *Geophysical Research Letters* 36(17).

Manly, B. F. 2006. *Randomization, bootstrap and Monte Carlo methods in biology*, volume 70. CRC Press.

McMillin, L. M.; Crone, L. J.; and Crosby, D. S. 1989. Adjusting satellite radiances by regression with an orthogonal transformation to a prior estimate. *Journal of Applied Meteorology* 28(9):969–975.

Meinshausen, N., and Bühlmann, P. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72(4):417–473.

Menne, M.; Williams Jr, C.; and Vose, R. 2010. United states historical climatology network (ushcn) serial monthly dataset. Technical report, Oak Ridge National Laboratory,.

Mesinger, F.; DiMego, G.; et al. 2006. North american regional reanalysis. *Bulletin of the American Meteorological Society* 87(3):343–360.

Moss, R. H.; Edmonds, J. A.; Hibbard, K. A.; Manning, M. R.; Rose, S. K.; Van Vuuren, D. P.; Carter, T. R.; Emori, S.; Kainuma, M.; Kram, T.; et al. 2010. The next generation of scenarios for climate change research and assessment. *Nature* 463(7282):747–756.

Negahban, S. N.; Ravikumar, P.; Wainwright, M. J.; Yu, B.; et al. 2012. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science* 27(4):538–557.

Nichols, T. E., and Holmes, A. P. 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human brain mapping* 15(1):1–25.

Niziol, T. A.; Snyder, W. R.; and Waldstreicher, J. S. 1995. Winter weather forecasting throughout the eastern united states. part iv: Lake effect snow. *Weather and Forecasting* 10(1):61–77.

Ojala, M., and Garriga, G. C. 2010. Permutation tests for studying classifier performance. *The Journal of Machine Learning Research* 11:1833–1863.

Peña, M., and van den Dool, H. 2008. Consolidation of multimodel forecasts by ridge regression: Application to pacific sea surface temperature. *Journal of Climate* 21(24):6521–6538.

Pendse, S. V.; Tetteh, I. K.; Semazzi, F. H.; Kumar, V.; and Samatova, N. F. 2012. Toward data-driven, semi-automatic inference of phenomenological physical models: Application to eastern sahel rainfall. In *SDM*, 35–46.

Piani, C.; Haerter, J.; and Coppola, E. 2010. Statistical bias correction for daily precipitation in regional climate models over europe. *Theoretical and Applied Climatology* 99(1-2):187–192.

Randall, D. A., et al. 2007. Climate models and their evaluation. In *Climate Change 2007: The physical science basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC (FAR)*. Cambridge University Press. 589–662.

Steinhaeuser, K.; Chawla, N.; and Ganguly, A. 2011. Comparing predictive power in climate data: Clustering matters. In *Advances in Spatial and Temporal Databases*, volume 6849, 39–55.

Stenseth, N. C., et al. 2003. Studying climate effects on ecology through the use of climate indices: the north atlantic oscillation, el nino southern oscillation and beyond. *Proceedings of the Royal Society of London B: Biological Sciences* 270(1529):2087–2096.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.