Spatio-temporal analysis of precipitation data via a sufficient dimension reduction in parallel

Sai K. Popuri*[1], Ross Flieger-Allison[2], Lois Miller[3], Danielle Sykes[1], Pablo Valle[4], Nagaraj K. Neerchal[1], Kofi P. Adragni[1], Amita Mehta[5], and Matthias K. Gobbert[1]

[1]Department of Mathematics and Statistics, University of Maryland, Baltimore County
[2]Department of Computer Science and Department of Statistics, Williams College
[3]Department of Mathematics, DePauw University
[4]Department of Mathematical Sciences, Kean University
[5]Joint Center for Earth Systems Technology, 5523 Research Park Dr, Baltimore, MD 21228

Abstract
Prediction of precipitation using simulations on various climate variables provided by Global Climate Models (GCM) as covariates is often required for regional hydrological assessment studies. In this paper, we use a sufficient dimension reduction method to analyze monthly precipitation data over the Missouri River Basin (MRB). At each location, effective reduced sets of monthly historical simulated data from a neighborhood provided by MIROC5, a Global Climate Model, are first obtained via a semi-continuous adaptation of the Sliced Inverse Regression, a sufficient dimension reduction approach. These reduced sets are used subsequently in a modified Nadaraya-Watson method for prediction. We implement the method on a computing cluster, and demonstrate that it is scalable. We observe a signficant speedup in the runtime when implemented in parallel. This is an attractive alternative to the traditional spatio-temporal analysis of the entire region given the large number of locations and temporal instances.

Key Words: Sufficient Dimension Reduction, Spatio-temporal, MIROC5, Precipitation, Parallel Computing

1. Introduction

Daily precipitation data is often required as an input to hydrological modeling tools (e.g.:Gassman et al. (2007)) to assess the impact of decadal climate changes on crop and water yields at the regional scale. One of the methods to predict precipitation is to use simulated data provided by Global Climate Models (GCMs) as covariates in a regression model with the observed precipitation as the response (Wood et al. (2004)). Under such models, a common approach is to predict the monthly precipitation, and simulate daily precipitation in a manner consistent with the monthly forecasts (e.g.:Gassman et al. (2007)).

In this paper we discuss a method to forecast precipitation at the monthly level over the Missouri River Basin (MRB) using simulated data on several climate variables provided by MIROC5 (Model of Interdisciplinary Research on Climate) (Nozawa et al. (2007))) as predictors. The MRB is the largest river basin in the United States covering more than $510,000$ square miles. It is home to 12% of all U.S. farms and 28% of all land used for farming (of Agriculture Natural Resources Conservation Service (2012)), making it a significant agricultural region. It also accounts for approximately 44% of the nation's wheat, 22% of grain corn, and 34% of cattle. Since approximately 90% of the basin is not irrigated, the region is heavily dependent on rainfall. As a result, it is important to assess the impact of future changes in the climate on the availability

*saiku1@umbc.edu

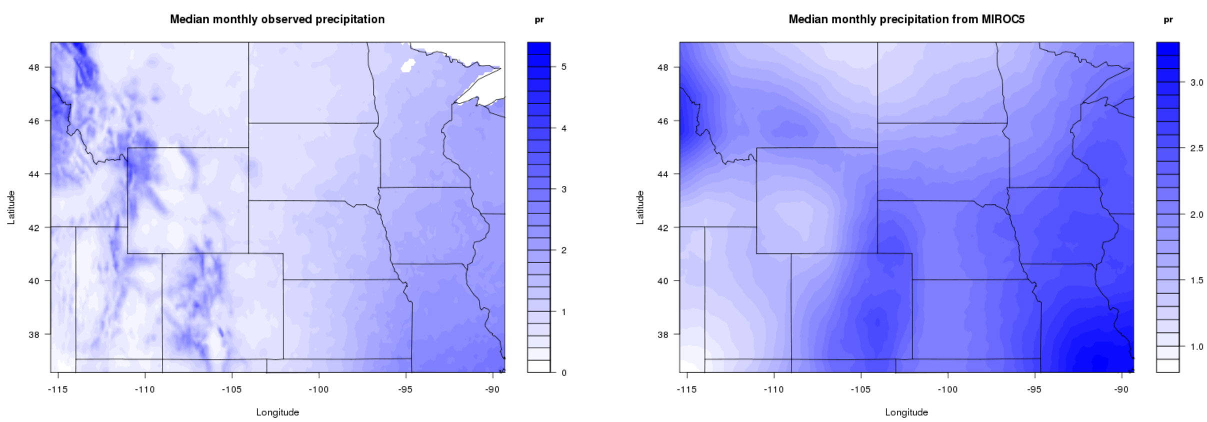
of water in the region (Mehta et al. (2013)).

Daily precipitation data at a location from the MRB region using MIROC5 data as covariate was recently analyzed by Popuri et al. (2015a), and Popuri et al. (2015b) using a Tobit, and a Bayesian AR(1) state space model respectively. In both cases, prediction accuracy was found to suffer because of heavy model dependency. Here we try to address that concern by recognizing the complicated, and possibly non-linear, relationship between the observed, and simulated data from MIROC5, and adopting a non-parametric approach. We extend the analysis from a grid location to the entire MRB region. We circumvent the challenge of fitting traditional spatio-temporal models to data from a large number of locations ($21,000$) as in the MRB region, by reducing the dimension in the set of a large of number of covariates consisting of spatially, and temporally separated values of several climate variables at each location using semi-continuous (point mass at 0) adaptations of the Sliced Inverse Regression (SIR;Li et al. (2006)), and Nadaraya-Watson estimator for prediction. Since the data can be fit at each location, the method is 'embarassingly parallel', which offers signficant computational advantage when implemented on a computing cluster in parallel.

Rest of the paper is organized as follows. In section 2 we describe the data. Section 3 discusses the proposed prediction methodology. Section 4 outlines the results obtained. We conclude with some discussion in secion 5.

2. Study Area and Data Description

The observed precipitation data are provided by Maurer et al. (2002). It has a temporal coverage of $1950-2005$, and a spatial resolution of $0.125°$(longtitude) $\times 0.125°$(latitude), making it 12km $\times 12$km gridded data. MIROC5 provides simulated data on several climate variables. It has a temporal coverage of $1859-2010$, and a spatial resolution of $1.4°$(longtitude) $\times 1.4°$(latitude), which is 150km $\times 150$km gridded data. MIROC5 data is ensemble averaged, and spatially interpolated to match the resolution of the observed data prior to our analysis.



(a) Median monthly observed precipitation over MRB from $1950-2005$

(b) Median monthly precipitation from MIROC5 over MRB from $1950-2005$

Figure 1: Median Precipitation

The area that we consider ranges from longitude $-115.5°$ to $-89.25°$, and latitude $36.5°$ to $49°$, which encompasses the entire Missouri River Basin. This rectangular region at the resolution of the observed data consists consists of $21,000$ locations. Monthly data from $1950-1994$ is used for training, and the data from $1995-2005$ is used for testing. This amounts to 540 time points for training, and 132 for testing at each location. Note that a spatio-temporal data set of this magnitude is typically not amenable to traditional modeling. Instead of fitting a spatio-temporal covariance structure in a parametric model, we fit a non-parametric regression model at each location using the lagged data, and extraneous data from a neighbordhood as covariates. This significantly reduces the computational burden especially when a parallel computing cluster is employed.

Figure 1 shows the median monthly observed, and MIROC5 simulated precipitation data from $1950-2005$ over the selected region. As can be seen MIROC5 precipitation is much more smoother with a narrower range. Since it is quite conceivable that there may be several months without rain at certain locations, we treat the monthly observed precipitation data as semi-continuous (point mass at 0, and continuous on the positive real line). On the other hand, since the daily precipitation provided by MIROC5 is strictly positive, the monthly counterpart of MIROC5 does not have zero values. In our data set, around 99% of the locations have a maximum of 5% proportion of dry months (0 values). Nevertheless, our proposed prediction method works for semi-continuous response. We use the following monthly variables as covariates in our model: precipitation, sea-level pressure, relative humidity, and maximum/minimum temperatures. For a given location s, we include 30 lags (current and previous 29 months) for each variable, as well as the current values of those variables, except precipitation, at the 8 neighboring locations of s in a regular grid, amounting to a total of 182 covariates.

3. Methodology

Let $Y(s,t)$ be the monthly observed precipitation at the location s, and time t, where $s \in D \subset R^2$, and $t = 1,..,T$. Here the set D is the MRB region bounded by the rectangle formed by $-115.5°$ to $-89.25°$, and latitude $36.5°$ to $49°$, and T is December 1994. Let $X_i(s,t)$, where $i = 1,..q$, $q = 5$ be the simulated monthly data on precipitation, sea-level pressure, relative humidity, maximum, and minimum temperatures provided by MIROC5. We assume that $Y(s,t)$ depends on the current, and lagged values of $X_i(s',t')$, where $s' \in D$, and $t' = t-1, t-2, .., 1$ as:

$$Y(s,t) = g(\mathbf{W}(s,t), e(s,t)),$$

where g is an unknown function, $e(s,t)$ is random noise, and $\mathbf{W}(s,t) = (\mathbf{X}_1(s,t), .., \mathbf{X}_q(s,t-p), \mathbf{X}_{s1}(s,t), .., \mathbf{X}_{sq}(s,t))'$, where $\mathbf{X}_i(s,t) = (X_i(s,t), .., X_i(s,t-p))'$, $\mathbf{X}_{si}(s,t) = (X_i(s_1,t), .., X_i(s_8,t))'$, $i = 1,..,q$ (except precipitation), p is the number of lags from t, and nodes $s_1 - s_8$ are the 8 neighbors of s as shown in Figure 2. For simplicity we arbitrarily set p to 30. With this choice of covariates, the vector $\mathbf{W}(s,t)$ is of dimension $r = 182$, which prohibits fitting a non-parametric regression model. Therefore, reducing the dimension of $\mathbf{W}(s,t)$ to something as low as 5 while preserving the regression information is desirable. In other words, we seek a matrix $B(s) \in R^{r \times d(s)}$ such that the distribution of $Y(s,t) \mid \mathbf{W}(s,t)$ is same as the distribution of $Y(s,t) \mid B(s)^T\mathbf{W}(s,t)$. Estimating such a matrix (rather the subspace spanned by the $d(s)$ columns of it) is sometimes known as sufficient dimension reduction (SDR) or effective dimension reduction (EDR). One such method to estimate $B(s)$ is Sliced Inverse Regression (SIR;Li (1991)).

The first step in our prediction method is to reduce the dimension at each location s in the MRB region D using the SIR method. Although in practice $d(s)$ must also be estimated, we fix $d(s) = 5$ for simplicity.

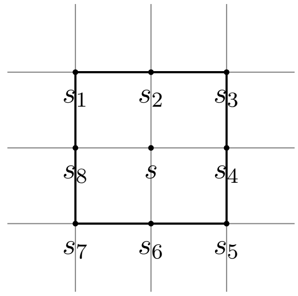


Figure 2: Eight neighbors in the spatial grid

Since $Y(s,t)$ is semi-continuous, we use a variant of SIR proposed in Li et al. (2006) shown in Algorithm 1. Let $\mathbf{W} = (\mathbf{W}(\mathbf{s},\mathbf{1})^{'}, .., \mathbf{W}(\mathbf{s},\mathbf{t})^{'})$ be the $n \times r$ matrix of design matrix, where $n = 540$ time points in the training set.

Algorithm 1 Semi-continuous (Tobit) SIR

Step 1: Normalize $\mathbf{w}$:

$$\mathbf{z}(s,t) = \hat{\Sigma}^{-1/2}[\mathbf{w}(s,t) - \bar{\mathbf{w}}],$$

where $\hat{\Sigma}$, and $\bar{\mathbf{w}}$ are the sample covariance matrix, and sample mean of $\mathbf{w}$ respectively.

Step 2: Divide the range of $y(s,t)$ into L slices I_l, $l = 1, .., L$, with $\{0\}$ being the first slice. Let n_l denote the number of observations in slice I_l.

Step 3: Compute the sample means of the normalized values of $\mathbf{w}(s,t)$ within each slice:

$$\bar{\mathbf{z}}_l(s,t) = \frac{1}{n_l} \sum_{i:y(s,t_i)\in I_l} \mathbf{z}(s,t_i)$$

Step 4: Compute the weighted matrix:

$$V = n^{-1} \sum_{l=1}^{L} n_l \bar{\mathbf{z}}(s,t_l) \bar{\mathbf{z}}^T(s,t_l)$$

Step 5: An estimate of the column space of $B(s)$ is given by the basis vectors:

$$\hat{\beta}(s)_i = \hat{\Sigma}^{-1/2} \hat{\eta}(s)_i,$$

where $i = 1, .., d(s)$, and $\hat{\eta}(s)_i$ is the eigenvector corresponding to the i^{th} largest eigenvalue of V. Therefore, the reduced subspace is represented by $\hat{B}(s) = (\hat{\beta}(s)_1, .., \hat{\beta}(s)_{d(s)})$.

In the second step of our prediction method, we use the reduced data set $(y(s,t), \mathbf{v}(s,t))$, where $\mathbf{v}(s,t)) = \hat{B}(s)^T \mathbf{w}(s,t)$, and $t = 1, .., T$, where the estimated basis matrix $\hat{B}(s)$ reduces the dimension of $\mathbf{w}(s,t)$ from r to $d(s)$. We modify the Nadaraya-Watson Estimator (NWE) (Siminoff (1996)) for the semi-continuous response $y(s,t)$, by fitting two separate NWEs as shown in Algorithm 2. Define $z(s,t)$ (not the same as $\mathbf{z}(s,t)$ in Algorithm 1) as

$$z(s,t) = \begin{cases} 0 & \text{if } y(s,t) = 0 \\ 1 & \text{if } y(s,t) > 0 \end{cases}$$

Let I be the index set of time points t such that $y(s,t) > 0$.

Algorithm 2 Semi-continuous Nadaraya-Watson Estimator

Step 1: Binary prediction using a new covariate $\mathbf{v}(s,t')$ at a future time point t':

1.
$$z^*(s,t') = \sum_{t=1}^{n} w_{i0} z(s,t), \tag{3.1}$$
where $w_{i0} = \frac{K_H(\mathbf{v}(s,t) - \mathbf{v}(s,t'))}{\sum_{t=1}^{n}(K_H(\mathbf{v}(s,t) - \mathbf{v}(s,t')))}$, K_H is a d-dimensional kernal (eg.: normal), and H is a smoothing parameter. Note that $z^*(s,t') \in [0,1]$.

2.
$$\hat{z}(s,t') = \begin{cases} 0, & \text{if } z^*(s,t') < 0.5 \\ 1, & \text{if } z^*(s,t') \geq 0.5 \end{cases}$$

Step 2: Prediction of the rain intensity (positive value) at a future time point t' using the covariate $\mathbf{v}(s,t')$:

$$\hat{y}^+(s,t') = \sum_{t \in I} w_{i0} y(s,t),$$

where $w_{i0} = \frac{K_H(\mathbf{v}(s,t) - \mathbf{v}(s,t'))}{\sum t \in I(K_H(\mathbf{v}(s,t) - \mathbf{v}(s,t')))}$. Note that the kernel K, and the parameter H need not be same as in equation 3.1. Also note that $\hat{y}^+(s,t')$ is strictly positive.

Step 3: Prediction of $y(s,t')$ at $\mathbf{v}(s,t')$:

$$\hat{y}(s,t') = \hat{y}^+(s,t') I(\hat{z}(s,t') = 1)$$

The novelty of the prediction method illustrated in Algorithms 1, and 2 is two-fold. Since the method is applied at each location, it is 'embarassingly parallel', and therefore can easily be implemented on a computing cluster. Secondly, the resulting predictions are semi-continuous with a point mass at 0.

4. Results

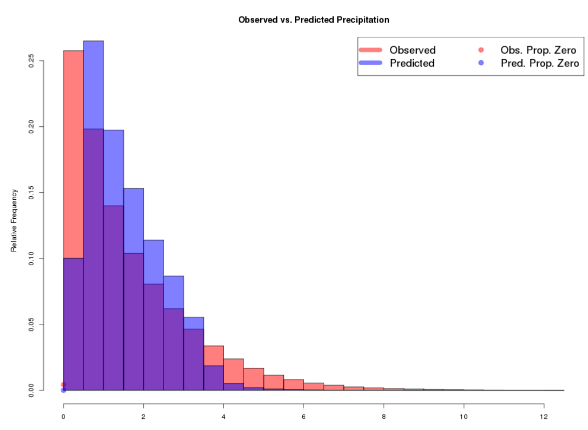
In order to assess the accuracy of the model, the data is divided into a training, and a testing set. The training set consists of entire MRB region from the time periods $1950 - 1994$, and the testing set covers the remaining data from the period $1995 - 2005$. As a measure of prediction accuracy, we use the mean squared error (MSE) as each location s defined as $mse(s) = \frac{1}{n_f} \sum_{t=T+1}^{T_f} (\hat{y}(s,t) - y(s,t))^2$, where T_f is December 2005, and $n_f (= 132)$ is the number of time points (months) in the testing period.

Figure 3a shows a histogram of $\hat{y}(s,t)$ from all the locations, and time periods in the testing set, overlaid with the histogram of the observed monthly precipitation $y(s,t)$. Since both the predictions, and the observed values are semi-continuous with a point mass at 0, we show the proportion of 0 values as points on the vertical axis on the histograms. The predicted proportion of 0 values (blue point) across all the locations, and time points is very small, and is close to the observed proportion of 0 values (red point). Since we used only the simulated data from MIROC5, which is much smoother with a narrower range compared to the observed data (Figure 1), our predictions too seem to be smoother with a narrower range as the histograms in Figure 3a indicate. Figure 3b shows the MSE values at each location across the MRB. The red/orange MSE values represent high accuracy predictions while yellow/green MSE values represent poorer accuracy. The figure illustrates higher accuracy in some of the more central, and arid regions of the MRB, and lower prediction accuracy in parts of south-east, and north-west MRB. These regions with low accuracy seem to correlate with regions of high altitude and high forestation (e.g., Yellowstone National Park in the upper lefthand corner of Wyoming, Flathead National Forest in western Montana and the Mark Twain national forest in Missouri). We suspect that, due to the varying altitudes and vegetation densities in these regions, overall weather variability is greater and therefore causing our model to produce inaccurate results.

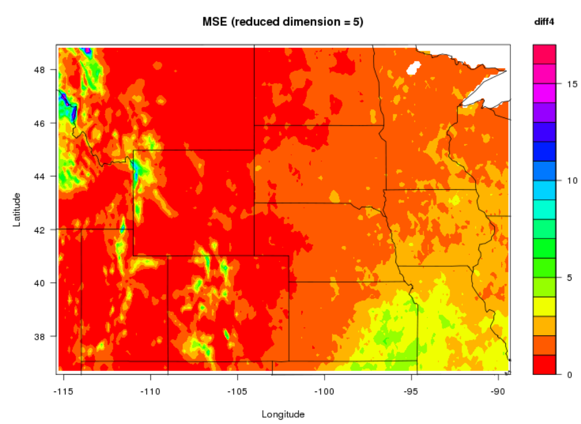
We compare the accuracy of our predictions with those from the linear regression model in equation 4.1 fitted at each location.

$$Y(s,t) = \beta \mathbf{X}(s,t) + e(s,t), \tag{4.1}$$

where $\mathbf{X}(s,t) = (1\, X_1(s,t)\, X_2(s,t)\, X_3(s,t)\, X_4(s,t)\, X_5(s,t))'$, $e(s,t)$ is random error, and $X_i(s,t)$, $i = 1,..,5$ are the simulated monthly data on precipitation, sea-level pressure, relative humidity, maximum, and minimum temperatures provided by MIROC5. Prediction based on the model in equation 4.1 at a future time point t_f is $E(Y(s,t_f) \mid \hat{\beta}\mathbf{X}(s,t_f))I(E(Y(s,t_f) \mid \hat{\beta}\mathbf{X}(s,t_f)) > 0)$, where I is the indicator function. Figures 4a, and 4b show the difference between the medians of predicted monthly precipitation, and the observed across the MRB over the testing period, and the analogous difference between predictions from the linear model in equation 4.1, and the observed data. Clearly, our prediction method based on SIR, and NWE is more accurate than the linear model, especially in south-west MRB. We further compare the prediction accuracy by dividing the MRB region in four quadrants: north-west, south-west, north-east, and south-east (figure 5). Figures 6, and 7 show the time plots of spatially averaged predictions from our method, the linear regression model, and the observed precipitation within each of the four quadrants. Both the models seem to miss the spikes in the observed data, which can possibly be improved by including the lagged values of the response $Y(s,t)$ in dimension reduction.

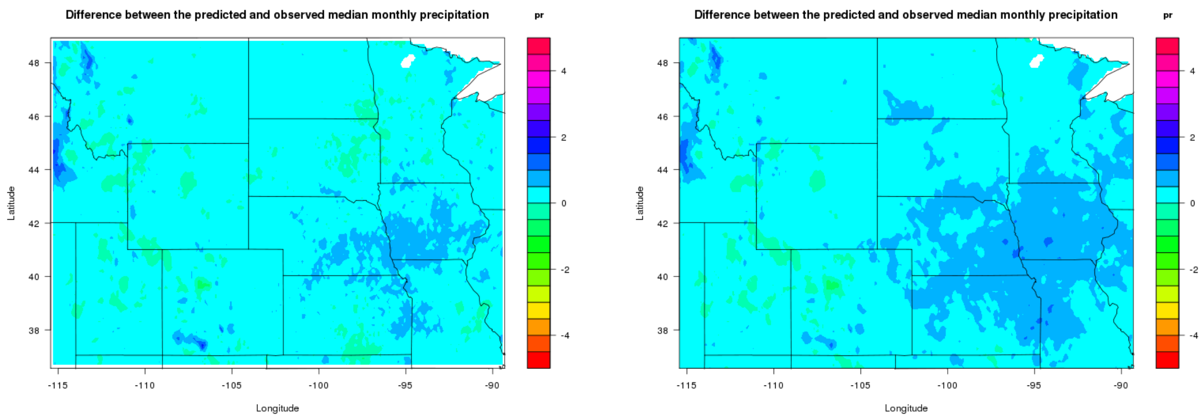


(a) Observed and predicted monthly precipitation, including the proportion of 0 values for both



(b) MSE from positive predictions for months with positive rainfall (true positives)

Figure 3: Prediction Results



(a) Prediction from SIR-NWE - Observed

(b) Prediction from liner regression - Observed

Figure 4: Difference in median Precipitation over MRB from $1950 - 2005$

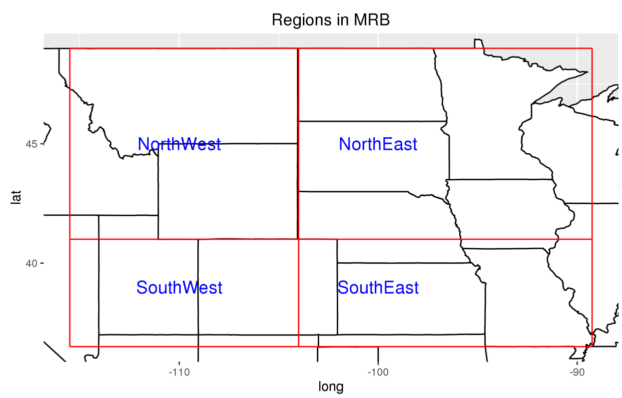
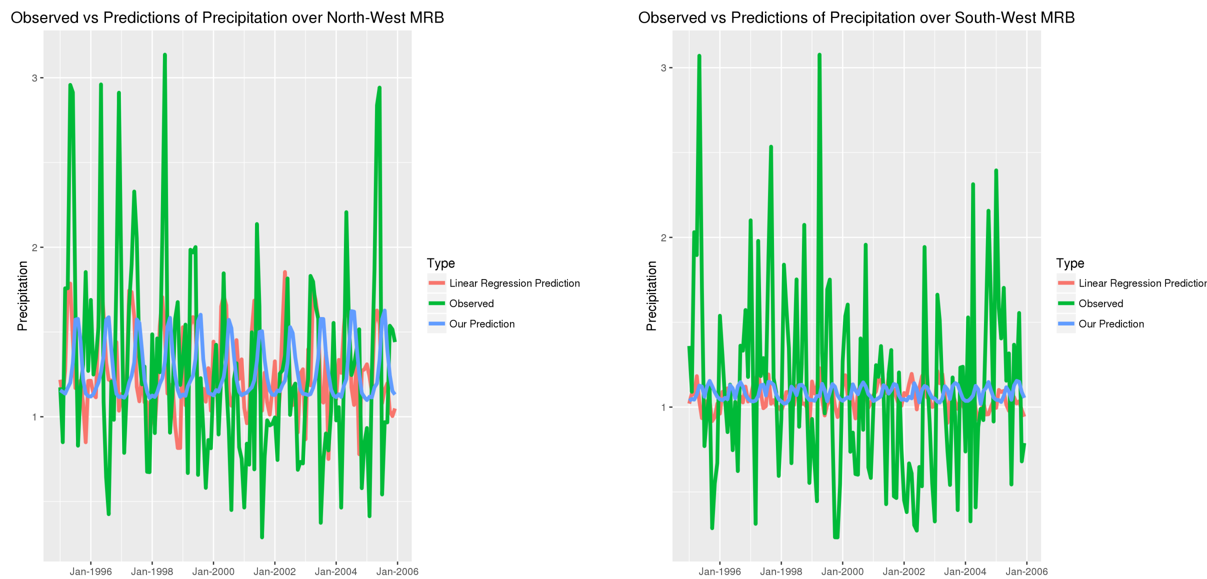
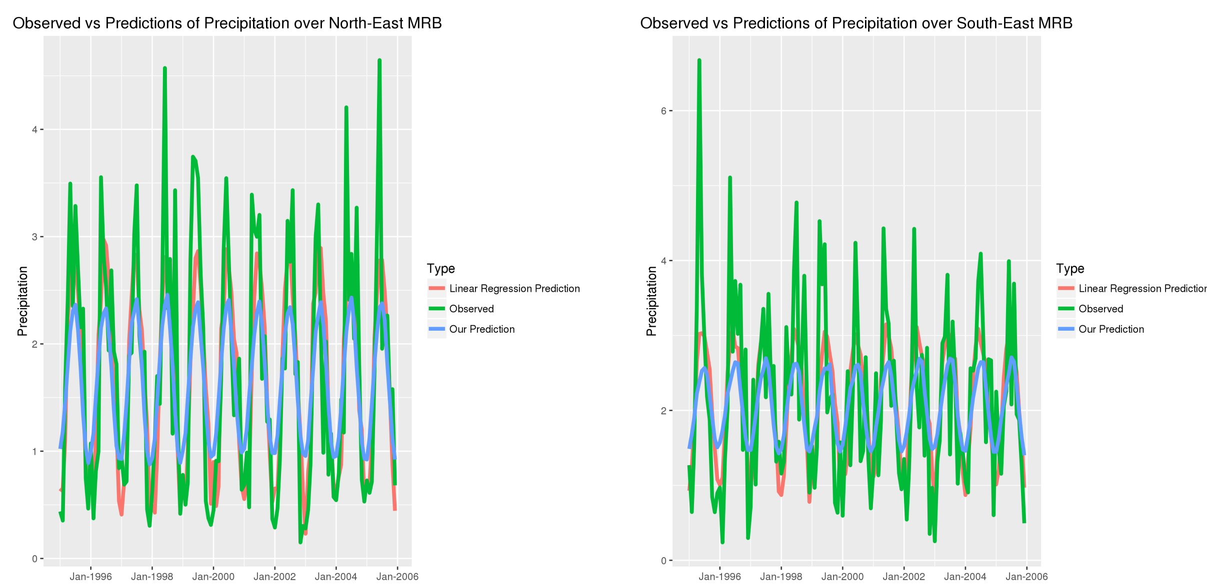


Figure 5: Regions in MRB



(a) North-West MRB (b) South-West MRB

Figure 6: Spatially averaged Observed vs Predictions in West MRB



(a) North-East MRB (b) South-East MRB

Figure 7: Spatially averaged Observed vs Predictions in East MRB

Monthly precipitation data in MRB for the period $1950-2005$ was also analyzed by Emelike et al. (2015). They fit multiple linear regression models at each location of MRB using a combination of simulated data on precipitation, sea-level pressure, relative humidity, and maximum/minimum temperatures (same covariates as we do in our analysis) from MIROC5. In order to assess overall accuracy of the model, and to compare with the models in Emelike et al. (2015), we calculate a standardized mean squared error value (smse) defined in 4.2. This quantity is equivalent to $1 - R^2$, so smaller values indicate a model with a better fit. Values above 1 indicate that the model does a poor job of predicting precipitation. Standardized mean squared error can be defined as:

$$smse = \frac{\sum(y(s,t) - \hat{y}(s,t))^2}{\sum(y(s,t) - \bar{y})^2}, \tag{4.2}$$

where the summations are over all the locations in MRB, and time points in the testing period, and $\bar{y}$ is the mean of all observed values. The smse value for our model is 0.6679, which is a 19% improvement over the corresponding value of 0.8262 computed based on the most successful model in Emelike et al. (2015).

Since the two steps of dimension reduction, and prediction is carried out at each location s in MRB, we can turn the computation into an 'embarassingly parallel' problem, and use a computing cluster to make predictions in parallel. This is a significant computational advantage over some of the traditional spatio-temporal models, which are often not readily parallelizable. We use a parallel computing protocal called Message-Passing Interface (MPI), implemented by the R package Rmpi(Yu (2016)). The procedure is shown in the following algorithm:

Algorithm 3 Parallel Implementation

Step 1: Load all the requisite data in on Process 0.
Step 2: Partition the data geo-spatially into "chunks".

- Think of the region we are modeling as a $lon \times lat$ grid of longitude and latitude values.
- If the region has more longitudes than latitudes, divide the region into chunks by row (else, by column).
- For models considering spatial dependency, append neighboring values on the borders of each chunk.

Step 3: Distribute each chunk of data to its own process (leaving all overflow values for Process 0 to manage).
Step 4: On each process, perform the predictions at each longitude-latitude combination across the sub-region provided by Process 0.
Step 5: From each process, send finished predictions back to Process 0 to be merged.
Step 6: From Process 0, merge the data back into one coherant array of prediction values and write the data to memory.

Table 1 shows the wall clock runtime (in HH:MM:SS) for a subregion in MRB, and for the entire MRB. The runtime speedup between the subregion, and the entire MRB is similar, indicating that the prediction method is scalable. Figure 8 shows the speedup plot against the optimal speedup with increasing number of processes. As the plot shows, we were able to achieve a near-optimal speedup in prediction across the subregion, and the entire MRB. Notice that the speedup is more prominent when predictions are made for the entire MRB indicating that our method is scalable, since the benefit from parallelizing is proportional to the size of the problem (number of locations).

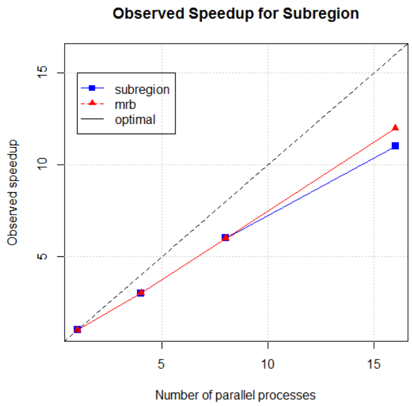


Figure 8: Speedup plot for the parallelized code

Processes	**Subregion***	**MRB**
1	01:31:02	35:00:28
4	00:27:29	09:48:53
8	00:14:20	05:08:00
16	00:07:45	02:50:10

Table 1: Performance Study Results

* denotes a region from latitude -101 to -97 and longitude 39.25 to 42.95.

5. Discussion

Predictions of monthly precipitation are a crucial input to several hydrological models to assess the impact of changes in climate on the availability of water in the Missouri River Basin. In this paper, we discuss a method that applies a semi-continuous variant of the Sliced Inverse Regression (SIR) to reduce the dimension of a large of number covariates on several climate variables from MIROC5 at each location in the MRB region, and uses a semi-continuous adaptation of Nadaraya-Watson estimator (NWE) for prediction. We successfully demonstrate a scalable implementation on a large spatio-temporal dataset using a parallel computing cluster. We further compare the prediction accuracy with results from a previous study that uses multiple regression models using the same set of covariates. There are several implementation details that can be improved upon. The dimension of the reduced subspace $d(s)$ was fixed to 5. $d(s)$ can be estimated, and it's spatial behavior can be studied. The kernel density K, and the smoothing parameter H were fixed to multivariate normal, and $I_{d(s)\times d(s)}$ respectively. These choices can be improved upon. We also note that the accuracy of the predictions can possibly be improved by including the information on the local geographical terrain as covariates. Note that the temporal dependence is modeled via the lags of covariates. We plan to extend this aspect to a more direct time series modeling by including the lagged response as a covariate. Finally, since our method works for semi-continuous data, we expect more prediction accuracy when applied to daily precipitation.

Acknowledgments

First author would like to thank Joint Center for Earth Systems Technology (JCET) for funding. We gratefully acknowledge The Center For Research on the Changing Earth System (CRCES) for providing us the data. The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF). The facility is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS-0821258 and CNS-1228778) and the SCREMS program (grant no. DMS-0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See hpcf.umbc.edu for more information on HPCF and the projects using its resources.

References

J.U. Emelike, D. Harper, C.Z. Mann, K. Owusu-Boaitey, S.K. Popuri, N.K. Neerchal, and A. Mehta. Asssessing Climate Impacts on Regional Water Resources in the Midwestern US. Technical Report HPCF–2015–22, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2015. (HPCF machines used: maya.).

P. W. Gassman, M. R. Reyes, C. H. Green, and J. G. Arnold. The Soil and Water Assessment Tool: Historical development, applications, and future research directions. `http://www.card.iastate.edu/environment/items/asabe_swat.pdf`, 2007.

K.C. Li. Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86(414), June 1991.

L. Li, J.S. Simonoff, and C. Tsai. Tobit Model Estimation and Sliced Inverse Regression. *Statistical Modelling*, 7(2), December 2006.

E. P. Maurer, A. W. Wood, J. C. Adam, and D. P. Lettenmaier. A long-term hydrologically based dataset of land surface fluxes and states for the conterminous united states. *Journal of Climate*, 15(22):3237–3251, 2002.

V. Mehta, C. Knutson, N. Rosenberg, J. Olsen, N. Wall, T. Bernasdt, and M. Hays. Decadal Climate Information Needs of Stakeholders for Decision Support in Water and Agriculture Production Sectors: A Case Study in the Missouri River Basin. *Weather Climate Society*, 5, 2013.

T. Nozawa, T. Nagashima, T. Ogura, T. Yokohata, N. Okada, and H. Shiogama. Climate change simulations with a coupled ocean-atmosphere gcm called the model for interdisciplinary research on climate: Miroc. *Center for Global Enrionment Research*, 2007.

United States Department of Agriculture Natural Resources Conservation Service. Assessment of the Effects of Conservation Practices on Cultivation Cropland in the Missouri River Basin, 2012. `http://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb104871\2.pdf`.

Sai K. Popuri, Nagaraj K. Neerchal, and Amita Mehta. Comparison of Linear and Tobit Modeling of Downscaled Daily Precipitation over Missouri River Basin using MIROC5. In *Machine Learning and Data Mining Approaches to Climate Science. Proceedings of the 4th International Workshop on Climate Informatics*, pages 35–49. Springer, 2015a. ISBN 978-3-319-17220-0.

Sai K. Popuri, Nagaraj K. Neerchal, and Amita Mehta. Bayesian and Transfer Function Estimation of a Tobit State-Space model for Daily Precipitation Data. In *JSM Proceedings, Statistics and the Environment Section*. ASA, 2015b.

J.S. Siminoff. *Smoothing Methods in Statistics*. Springer Series inn Statistics, 1996.

A. W. Wood, L. R. Leung, V. Sridhar, and D. P. Lettenmaier. Hydrologic implication of dynamical and statistical approaches to downscaling climate model outputs. *Climate Change*, 62:189–216, 2004.

H. Yu. Rmpi: Interface (wrapper) to MPI (Message-Passing Interface). `https://cran.r-project.org/web/packages/Rmpi/index.html`, 2016.