Dimensionality Reduction Using Sliced Inverse Regression in Modeling Large Climate Data

REU Site: Interdisciplinary Program in High Performance Computing

Ross Flieger-Allison[1], Lois Miller[2], Danielle Sykes[3], Pablo Valle[4],
Graduate assistants: Sai K. Popuri[3] and Nadeesri Wijekoon[3]
Faculty mentor: Nagaraj K. Neerchal[3], and Client: Amita Mehta[5]

[1]Department of Computer Science and Department of Statistics, Williams College
[2]Department of Mathematics, DePauw University
[3]Department of Mathematics and Statistics, UMBC
[4]Department of Mathematical Sciences, Kean University
[5]Joint Center for Earth System Technology (JCET)

Technical Report HPCF–2016–13, hpcf.umbc.edu > Publications

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. 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 significant speedup in the runtime when implemented in parallel.

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.: [3]) 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 [13]. 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.: [3]).

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 the Model of Interdisciplinary Research on Climate (MIROC5; [9]) as predictors. The MRB is the largest river basin in the United States. It is present in ten U.S. states and covers more than $500,000$ square miles, which amounts to about a sixth of the United States geographic area. The MRB is home to 12% of all U.S. farms and 28% of all land used for farming [10], making it a significant agricultural region. Although 90% of the basin is not irrigated, making the region dependent on rainfall, it provides 44% of the nation's wheat, 22% of grain corn and 34% of cattle. As a result, it is important to assess the impact of future changes in the climate on the availability of water in the region [8].

Prediction of daily maximum/minimum temperatures and precipitation in the MRB was studied by the 2014 UMBC REU team [2] using simulated data from two climate models: Hadley Center Coupled Model (HadCM3; [4]) and MIROC5. While predictions of daily maximum/minimum temperatures were found to be satisfactory, predictions of daily precipitation were not sufficiently accurate. The 2015 UMBC REU team [2] studied the prediction of precipitation at the monthly level instead using the monthly level simulated data from HadCM3 and MIROC5. Although predicting at the monthly level was computationally less expensive, the quality of predictions for the precipitation could not be improved, possibly because of the difficulty in predicting a semi-continuous response (point mass at 0) and heavy model dependency. Here we try to address that concern by recognizing the complicated, possibly non-linear, relationship between the observed and simulated data from MIROC5 and adopting a non-parametric approach. 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 adaptations of the Sliced Inverse Regression (SIR; [6]) and Nadaraya-Watson estimator (NWE) for prediction. Since the data can be fit at each location, the method is 'embarassingly parallel', which is a parallelization of the problem where processes do not communicate with each other during the computation. This offers significant computational advantage when implemented on a computing cluster in parallel.

The rest of the paper is organized as follows: Section 2 describes the MRB region, the data used, and the statistical and computational methods used. Section 3 outlines the results obtained. Finally, Section 4 summarizes the findings and offers suggestions for future work.

2 Methodology

2.1 Study Area and Data Description

The observed precipitation data are provided by [7]. It has a temporal coverage of $1950-2005$ and a spatial resolution of $0.125°$(longitude) $\times 0.125°$(latitude), making it 12km $\times$12km 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$150km gridded data. MIROC5 data is ensemble averaged and spatially interpolated to match the resolution of the observed data prior to our analysis.

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 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. 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 eight neighboring locations of s in a regular grid, amounting to a total of 182 covariates.

2.2 Sliced Inverse Regression for semi-continuous data

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 longitude $-115.5°$ to $-89.25°$ and latitude $36.5°$ to $49°$ and T is December 1994. Let $X_i(s,t)$, where $i = 1, \ldots, q$, $q = 5$ 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 $t' = t-1, t-2, \ldots, 1$ 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 $\mathbf{W}(s,t) = (\mathbf{X}_1(s,t) \ \ldots \ \mathbf{X}_q(s,t-p)\ \mathbf{X}_{s1}(s,t) \ \ldots \ \mathbf{X}_{sq}(s,t))'$ and $\mathbf{X}_i(s,t) = (X_i(s,t) \ \ldots \ X_i(s,t-p))'$, $\mathbf{X}_{si}(s,t) = (X_i(s_1,t) \ \ldots \ X_i(s_8,t))'$, $i = 1, \ldots, q$ (except precipitation), p is the number of lags from t, and nodes $s_1 - s_8$ are the eight neighbors of s as shown in Figure 2.1. 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; [5]).

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. Since $Y(s,t)$ is semi-continuous, we use a variant of SIR [6] shown in Algorithm 1.

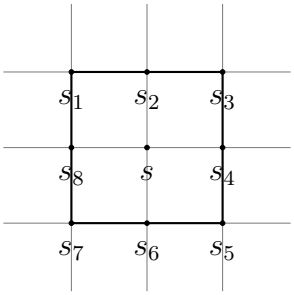


Figure 2.1: Eight neighbors in the spatial grid.

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:

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

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 $\hat{B}(s) = (\hat{\beta}(s)_1, .., \hat{\beta}(s)_{d(s)})$.

2.3 Nadaraya-Watson Estimator for semi-continuous data

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; [12]) 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.
$$z^*(s,t') = \sum_{t=1}^{n} w_{i0} z(s,t), \tag{2.1}$$

where $w_{i0} = \dfrac{K_H(\mathbf{v}(s,t) - \mathbf{v}(s,t'))}{\sum_{t=1}^{n}(K_H(\mathbf{v}(s,t) - \mathbf{v}(s,t')))}$, 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 $w_{i0} = \dfrac{K_H(\mathbf{v}(s,t) - \mathbf{v}(s,t'))}{\sum t \in I(K_H(\mathbf{v}(s,t) - \mathbf{v}(s,t')))}$. Note that the kernel K and the parameter H need not be same as in equation 2.1. Also note that $\hat{y}^+(s,t')$ is strictly positive.

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

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

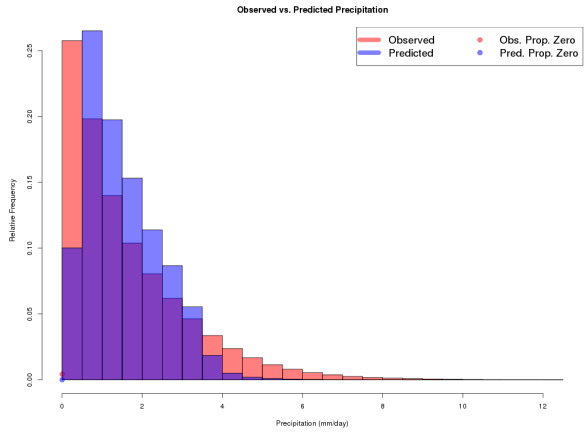
The novelty of the prediction method illustrated in Algorithms 1 and 2 is two-fold. Since the method is applied at each location, it is 'embarassingly parallel' and therefore can easily be implemented on a computing cluster. Secondly, the resulting predictions are semi-continuous with a point mass at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

3 Results

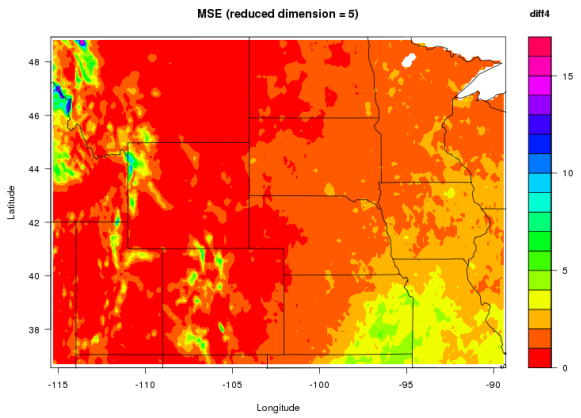
In order to assess the accuracy of the model, the data is divided into a training and a testing set. The training set consists of entire MRB region from the time periods $1950 - 1994$ and the testing set covers the remaining data from the period $1995 - 2005$. As a measure of prediction accuracy, we 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 3.1a 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). Figure 3.1b 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.



(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 3.1: Prediction Results.

Monthly precipitation data in MRB for the period $1950 - 2005$ was also analyzed by [1]. They fitted 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 [1], we calculate a standardized mean squared error value (smse) defined in equation 3.1. 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:

$$smse = \frac{\sum(y(s,t) - \hat{y}(s,t))^2}{\sum(y(s,t) - \bar{y})^2}, \tag{3.1}$$

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 [1].

3.1 Parallelization

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; [11]), implemented by the R package Rmpi([14]). 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 subregion 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 3.1 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. The plot in Figure 3.2 visualizes the numbers in Table 3.1. It shows the speedup plot against the optimal speedup with

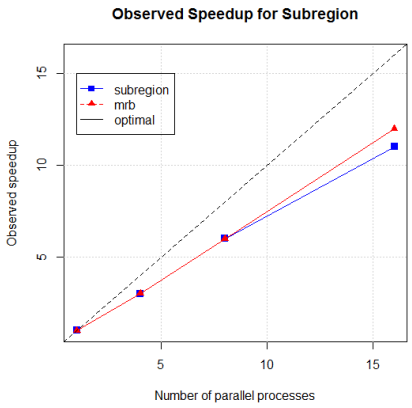


Figure 3.2: Speedup plot for the parallelized code.

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 3.1: Performance Study Results.

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

increasing number of processes. The speedup S_p for p number of processes is defined as the ratio $\frac{T_1}{T_p}$, where T_1 and T_p are the runtimes on a single process and on p number of processes, respectively. 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).

4 Conclusions

In this paper we discussed a prediction method for monthly level precipitation data in the MRB region using the semi-continuous Sliced Inverse Regression method for dimension reduction followed by a modified Nadaraya-Watson estimator. We successfully demonstrated that the method can be implemented to work on a large dataset. The accuracy of the predictions were shown to improve upon the method from the 2015 REU project [1]. We further showed that parallelization of the method greatly improved the computational efficiency on both the subregion and the entire Missouri River Basin, although the speedup is closer to optimal on the subregion. We note that although we improved upon the model from the 2015 UMBC REU [1], our standardized mean squared error is still relatively high, so further work is needed to continue improving the accuracy of rainfall predictions. Additionally, since the monthly precipitation data does not have a high proportion of zero values, the methodology described in this paper is more suitable for precipitation at the daily level.

Acknowledgments

These results were obtained as part of the REU Site: Interdisciplinary Program in High Performance Computing (hpcreu.umbc.edu) in the Department of Mathematics and Statistics

at the University of Maryland, Baltimore County (UMBC) in Summer 2016. This program is funded by the National Science Foundation (NSF), the National Security Agency (NSA), and the Department of Defense (DOD), with additional support from UMBC, the Department of Mathematics and Statistics, the Center for Interdisciplinary Research and Consulting (CIRC), and the UMBC High Performance Computing Facility (HPCF). HPCF 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 UMBC. Co-author Danielle Sykes was supported, in part, by the UMBC National Security Agency (NSA) Scholars Program through a contract with the NSA. Graduate assistants Sai K. Popuri and Nadeesri Wijekoon was supported by UMBC.

References

[1] 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.).

[2] C. Evans, A. Gartrell, L. Gomez, M. Mouyebe, D. Oxley, S.K. Popuri, N.K. Neerchal, and A. Mehta. Improving the Computational Efficiency of Downscaling GCM Data for Use in SWAT. Technical Report HPCF–2014–12, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2014. (HPCF machines used: maya.).

[3] 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.

[4] C. Gordon, C. Cooper, C.A. Senior, H. Banks, J.M. Gregory, T.C. Johns, J.F.B. Mitchell, and R.A. Wood. The simulation of sst, sea ice extents and ocean heat transports in a version of the hadley centre coupled model without flux adjustments. *Climate Dynamics*, 16:147–168, 2000.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] Peter S. Pacheco. *Parallel Programming with MPI.* Morgan Kaufmann, 1997.

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

[13] 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.

[14] Hao Yu. Rmpi: Parallel statistical computing in R. *R News*, 2(2):10–14, 2002.