Dimensionality Reduction Using Sliced Inverse Regression In Modeling Large Climate Data

UMBC REU Site: Interdisciplinary Program in High Performance Computing
Ross Flieger-Allison[1], Lois Miller[2], Danielle Sykes[3], Pablo Valle[4]
RAs: Sai K. Popuri[3], Nadeesri Wijekoon[3] Faculty: Nagaraj K. Neerchal[3] Client: Amita Mehta[5]
[1]Williams College [2]DePauw University [3]UMBC [4]Kean University [5]JCET

Background

- The Missouri River Basin (MRB) is a significant agricultural region that is not irrigated and thus dependent on rainfall
- Precipitation predictions are historically inaccurate due to the semi-continuous nature of observed precipitation data

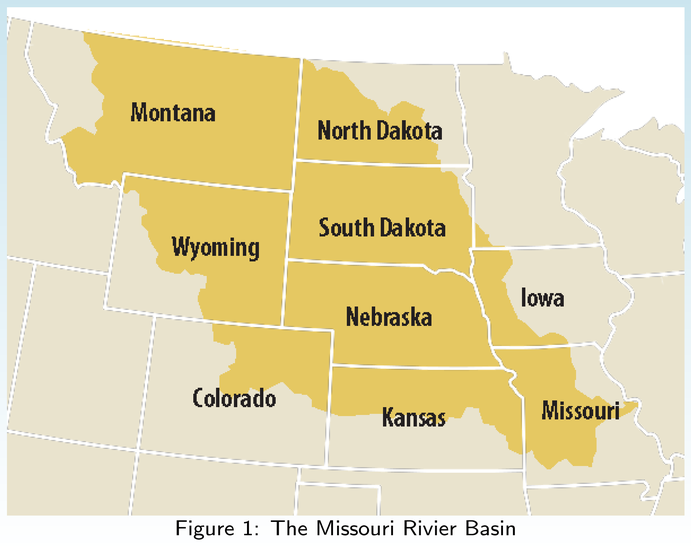

Figure 1: The Missouri Rivier Basin

Data

- Precipitation data is semi-continuous, meaning many observations are equal to 0 and all other observations are positive and follow a continuous distribution
- Predictions are calculated using observed data and data from MIROC5, a Global Climate Model (GCM), across 57 years (1949-2005) from over 270,000 locations
- Precipitation at any given location s depends on a large number of covariates: current and past values of monthly precipitation, sea-level pressure, relative humidity, and maximum/minimum temperatures at s and its neighboring locations

Methodology

- Sliced Inverse Regression (SIR) is a data-analytic tool that can be used to reduce the dimensionality of a large set of covariates
- Nadaraya-Watson Estimator (NWE) is a simple non-paramentric smoothing regression method:

$$\hat{m}_x = \frac{\sum_{i=1}^{n} K_h(x-x_i)y_i}{\sum_{i=1}^{n} K_h(x-x_i)}$$

- The NWE is adapted to account for the semi-continous nature of predictions

Results

- The following figures show a comparison between observed precipitation data, and their predictions

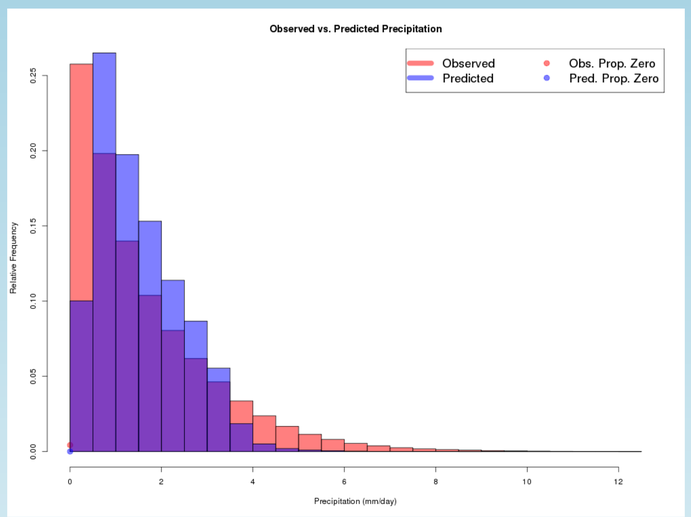

Figure 2: Observed and predicted monthly precipitation, including the proportion of 0 values for both

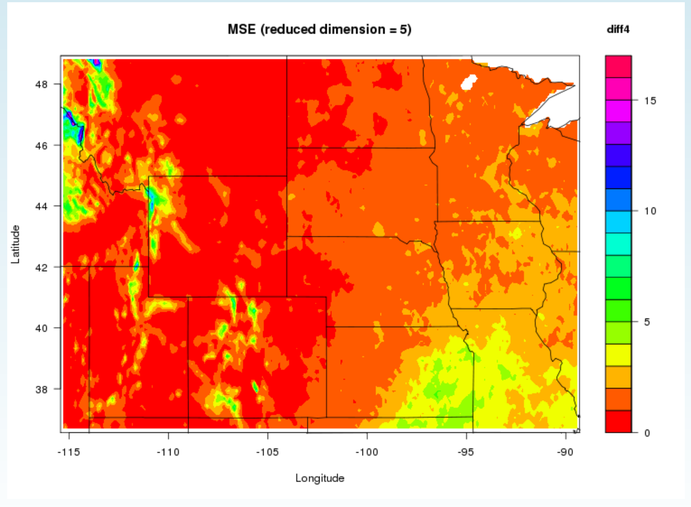

Figure 3: Mean Squared Error from positive predictions for months with positive rainfall (true positives)

Parallelization

Table 1: Performance Study Results

Processes	Subregion*	MRB
1	01:14:24	N/A
4	00:19:07	09:54:22
8	00:11:04	05:20:31
16	00:05:58	02:49:49
32	00:03:13	04:25:51
64	N/A	03:41:26

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

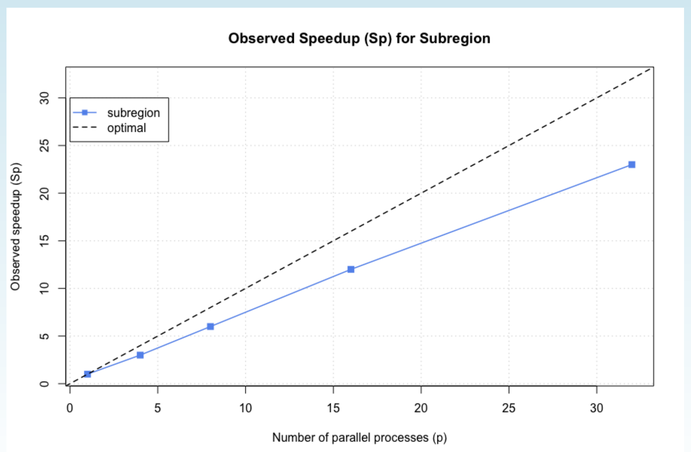

Figure 4: Observed speedup ($S_p = (T_1/T_p)$) of parallelization

Conclusions

- We have successfully demonstrated that SIR and NWE methods can be implemented to work on a large dataset
- Parallelization of SIR and NWE code greatly increases computational efficiency on the subregion, and also improves efficiency for the entire MRB region, up to 16 processes

References & Acknowledgments

[1] National Integrated Drought Information Systems. `https://www.drought.gov/drought/dews/missouri-river-basin/about`.

[2] S. K. Popuri, K. P. Adragni, *et al.*, "Spatio-Temporal Analysis of Daily Precipitation via a Sufficient Dimension Reduction." *(In Progress)*.

[3] Full Technical Report: HPCF-2016-13 `hpcf.umbc.edu` > Publications

REU Site: `hpcreu.umbc.edu`
NSF, NSA, DOD, UMBC, HPCF, CIRC