Bayesian and Transfer Function Estimation of a Tobit State-Space model for Daily Precipitation Data

Sai K. Popuri*[†], Nagaraj K. Neerchal[†], and Amita Mehta[‡]

[†]Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250
[‡]Joint Center for Earth Systems Technology, 5523 Research Park Dr, Baltimore, MD 21228

Abstract
We analyze the daily precipitation time series data at a location in the upper Missouri River Basin (MRB) with prediction as the objective using two approaches: a. Bayesian estimation of a standard Tobit state space model and b. a transfer function approach with an Expectation-Maximization (EM)-like method to "fill in" the zero values (dry days) in the observed series. We use the daily precipitation data simulated by MIROC5, a Global Climate Model (GCM), as an exogeneous predictor. The prediction methods based on the two models can predict zero values as valid predictions, which is desirable for daily precipitation. While the prediction of intensities of precipitation (positive precipitation on wet days) from both the methods are similar on average, the transfer function method was more successful at correctly predicting zero precipitation on days when there was no rain (dry days). A few other relative strengths and weaknesses of the two methods are also discussed.

Key Words: State-space, Bayesian, Tobit, EM, Transfer function, Statistical downscaling, Precipitation, MIROC5

1. Introduction

Daily precipitation data is an important variable in hydrological studies to assess the impact of decadal climate changes on crop and water yields at the regional scale. Since such impact studies are for future periods, the daily precipitation data must be forecasted at regional resolutions (ex.: $10\ \text{km}^2$). A popular method to generate predictions of daily rainfall is to use the simulated data provided by Global Climate Models (GCMs) (Wood et al. (2004)), often available at much coarser resolutions (ex.: $150\ \text{km}^2$), to build a statistical relationship with the observed data, and use the model to forecast using the retrospective simulations (also known as "hindcast data") from GCMs as covariates. This method of fitting a statistical model to the GCM simulated data is an example of what is known as "Statistical downscaling". Such methods are also known as "bias-correction" methods because hydrometeorological data, precipitation in particular, provided by GCMs is often not accurate enough to be used for applications that work at finer resolutions (Wood et al. (2004)).

In this paper we discuss and share preliminary findings from two statistical downscaling methods using the daily precipitation data simulated by the GCM MIROC5 (Model of Interdisciplinary Research on Climate) (Nozawa et al. (2007)) as the predictor. This research is part of a bigger project to assess the decadal climate changes on agricultural and water yields in the Missouri River Basin (MRB) (Mehta et al. (2013)). We use a hydrological software called Soil Water Assessment Tool (SWAT) (Gassman et al. (2007)) to run impact

*saiku1@umbc.edu

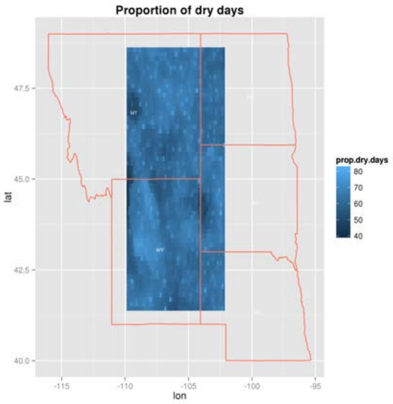

Figure 1: Proportion of dry days between 1949 and 2000

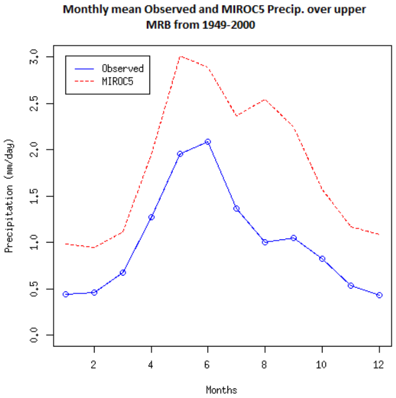

Figure 2: Avg. monthly Precipitation over upper MRB

studies on various climate scenarios, derived from forecasted data of various climate variables, including daily precipitation, at the resolution of 12 km^2.

There are several statistical downscaling methods used by researchers to bias-correct the GCM simulated data. Teutschbein and Seibert (2012) provide a review of some of the popular methods. We extend the Tobit model used in Popuri et al. (2015) to account for time dependence. We use a state-space model (see Shumway and Stoffer (2011) for details) with the standard Tobit model as the observation equation. The Tobit model is similar to a state-space model with a linear state process with MIROC5 data as the covariate and random noise, and an observation process of the Tobit form. We describe two methods to fit a Tobit state space model to the daily precipitation time series. The first method follows a Bayesian approach to estimate the state process and generates predictions from the predictive posterior distributions using a Gibbs sampling scheme. The second method uses a transfer function model to fit the observed series to the covariate series. The zero values (dry days) in the observed series are imputed using an Expectation-Maximization (EM)-like method.

Rest of the paper is organized as follows. In section 2 we describe the data. Section 3 discusses the Tobit state-space model used, the details of the Bayesian implementation, and the results. In Section 4 we describe the transfer function model with the EM-like algorithm used to “fill in” the zero observations. We conclude with some discussion in secion 5.

2. Data Description

The observed daily time series data are provided by Maurer et al. (2002). It has a temporal coverage of $1949 - 2005$, and a spatial resolution of $0.125°$(longtitude) $\times 0.125°$(latitude), making it 12km $\times$12km gridded data. MIROC5 provides daily simulated precipitation data, which has a temporal coverage of $1859 - 2010$, and are at $1.4°$(longtitude) $\times 1.4°$(latitude) spatial resolution, which is 150km $\times$150km gridded data. MIROC5 data is spatially interpolated to match the resolution of the observed data prior to our analysis. The data between 1949 and 2000 is used for model fitting and between 2001 and 2005 for evaluation. Figure 2 shows the systematic bias between the monthly observed and MIROC5 precipitation between 1949 and 2000 averaged over the upper MRB region. MIROC5 simulated daily data does not contain zero values, and the positive values for wet days are in general much lower than the observed. On the other hand, for more than 50% of the number of days on average at each location, observed precipitation is zero. As the proportion of dry days over the region in Figure 1 indicate, the observed precipitation

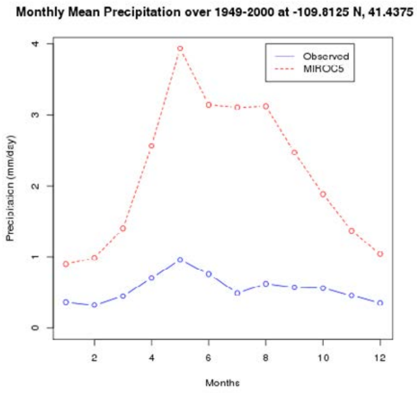

Figure 4: Monthly mean Precipitation at -109.8125N, 41.4375W

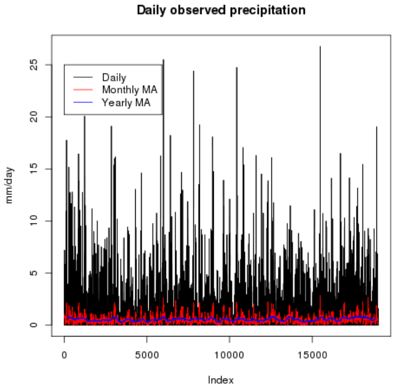

Figure 5: Observed daily and smoothed series

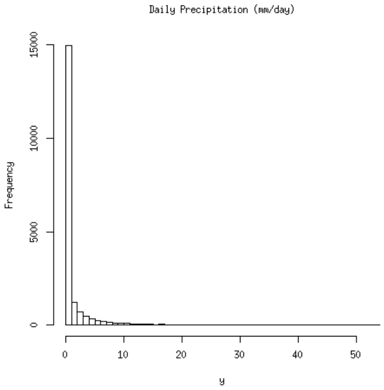

Figure 6: Histogram of daily precipitation at a location

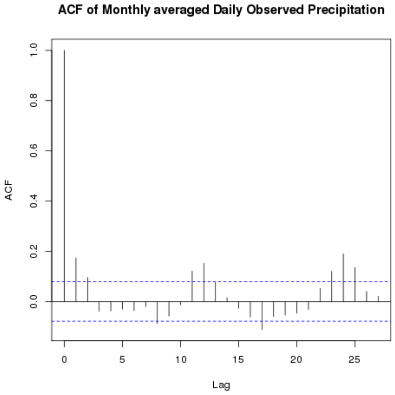

Figure 7: Sample autocorrelations of monthly mean series

in the region is heavily censored at zero. MIROC5 follows a 365-day calendar whereas the observed data follows the regular calendar with leap years. Prior to the analysis, such differences in calendars are resolved by aligning MIROC5 with the observed data. Figure 4 shows the monthly mean of observed and MIROC5 precipitation at -109.8125N, 41.4375W, which is to the west of Rock Springs, WY and is in the upper MRB region shown in figure 1. Clearly, on average MIROC5 seems to consistently overestimate precipitation each month. However, as figures 5, and 3 show, the large number of zero values in the observed data make the observed monthly averages lower than those of MIROC5. The large number of zero observed precipitation values can also be seen from the histogram of the observed data in figure 6. The sample autocorrelation function of the observed data averaged over each month in figure 7 indicates a seasonal component with a period of 1 year, as one would expect. This seasonal behavior can also be seen in the MIROC5 data as well. Notice the presence of a large number of zero values, which makes the monthly averages lower than those from MIROC5. In summary, the observed daily precipitation is heavily censored at zero and shows high volatility. On the other hand, MIROC5 data has low intensity strictly positive values.

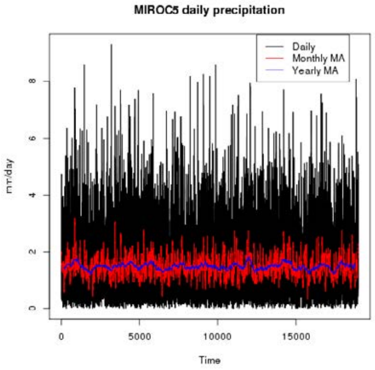

Figure 3: Daily and smoothed Precipitation by MIROC5

3. The Tobit State-space model

A State-space model consists of two processes: a state process, and an observation process. The state process is not observable but drives the observation process, which is directly observed. In general, the observation process is linearly related to the state process, and both the processes can include covariates (same or different), and noise components. In the model we use for our analysis, the observation process has the standard Tobit form, and does not include a noise term. The state process in our model is assumed to follow an AR(1) process with MIROC5 data as a covariate and a Gaussian white noise term. Brockwell et al. (2003) have used a similar model to analyze Internet traffic data. We closely follow their formulation with a few simplifying assumptions.

Let $\{Y_t\}$ be the observed daily precipitation series, and $\{X_t\}$ denote the unobserved state process. One could conceptualize $\{X_t\}$ as the build up of the underlying weather variable, which results in precipitation upon reaching an ideal condition. Thus, $\{Y_t\}$ can be thought of as the following hockey stick function of the unobserved $\{X_t\}$.

$$Y_t = g(X_t), \tag{1}$$

where

$$g(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases} \tag{2}$$

The latent process X_t can be modeled as a linear, causal stationary Gaussian time series satisfying the state equation

$$X_{t+1} = \phi X_t + \beta M_{t+1} + Z_t, \tag{3}$$

where ϕ is the autoregressive parameter, $\{M_t\}$ is the MIROC5 simulated precipitation on day t, and $\{Z_t\}$ is Gaussian white noise (Normal with mean 0) with variance σ^2.

For our analysis, we detrend both the observed and MIROC5 series by subtracting monthly means prior to fitting the model. Averages of monthly means from the last five years of the test period $1949 - 2000$ are used for the corresponding monthly means in the forecasting period, and to adjust the forecasted state variables to predict daily precipitation for the period $2001 - 2005$. Also, we assume an AR(1) model for the state process, which is defined as

$$X_t = \phi X_{t-1} + \beta M_t^* + Z_t, \tag{4}$$

where $\{M_t^*\}$ is detrended precipitation provided by MIROC5 on day $t = 1, 2, .., n$, and Z_t's are independent zero mean Normal with variance σ^2. M_t^* is calculated as $M_t^* = M_t - \mu^m_{m(t)}$, where $\mu^m_{m(t)}$ is the mean of precipitation from the (month, year) day t falls into. In a similar fashion, we detrend the $\{Y_t\}$ series as

$$Y_t^* = Y_t - \mu_{m(t)}, \tag{5}$$

where $\mu_{m(t)}$ is the mean of the observed precipitation from the month, year combination day t falls in. The observation equation after detrending is

$$Y_t^* = \begin{cases} X_t, & \text{if } X_t > -\mu_{m(t)} \\ -\mu_{m(t)}, & \text{if } X_t \leq -\mu_{m(t)}, \end{cases} \tag{6}$$

which is equivalent to

$$Y_t = g(X_t) + \mu_{m(t)} \tag{7}$$

In other words, the state process is the latent process driving the detrended observed series $\{Y_t^*\}$. The model is equations (4-7) is akin to the standard Tobit model (Takeshi (1985)). It's nonlinear nature complicates the model fitting a little.

3.1 Bayesian Approach

Prior to the analysis, the observed data is adjusted to 365-day calendar followed by MIROC5 for implementation convenience. The model in equations (4-7) has parameters $(\phi, \beta, \sigma^2, X_t, X_{t+f})$, where $t = 1, 2, .., n$, $f = 1, 2, .., n_f$, where $n = 18980$, and $n_f = 1825$, since there are 18980 days in the model testing period $1949 - 2000$, and 1825 days in the model evaluation period $2001 - 2005$. We use Markov chain Monte Carlo (MCMC) to estimate the parameters and perform forecasting. Let $\theta = (\phi, \beta, \sigma^2)$. Notice that the joint distribution of X_t and X_{t+1} conditional on X_{t-1} is

$$\begin{pmatrix} X_t \\ X_{t+1} \end{pmatrix} \mid X_{t-1}, \theta \;\sim\; N\left[\begin{pmatrix} \phi X_{t-1} + \beta M_{t-1} \\ \phi^2 X_{t-1} + \phi\beta M_t + \beta M_{t+1} \end{pmatrix}, \begin{pmatrix} \sigma^2 & \phi\sigma^2 \\ \phi\sigma^2 & (\phi^2+1)\sigma^2 \end{pmatrix}\right]$$

Therefore, for t= $2..(n-1)$, where n is the number of observations, the marginal distribution of X_t given X_{t+1} and X_{t-1} is given by

$$X_t \mid X_{t+1}, X_{t-1}, \theta \sim N(\star, \Delta), \tag{8}$$

where $\star = \phi X_{t-1} + \beta M_{t-1} + \frac{\phi}{(\phi^2+1)}(X_{t+1} - \phi^2 X_{t-1} - \phi\beta M_t)$, $\Delta = \frac{\sigma^2}{(\phi^2+1)}$.
For end points t= 1 and n, we have the following conditional distributions

$$X_1 \mid X_2, \theta \sim N(\frac{X_2}{\phi} - \frac{\beta M_2}{\phi}, \frac{\sigma^2}{\phi^2}) \tag{9}$$

$$X_n \mid X_{n-1}, \theta \sim N(\phi X_{n-1} + \beta M_n, \sigma^2) \tag{10}$$

Using a non-informative prior for β as $N(\mu_\beta = 0, \sigma_\beta^2 = 20)$, we get the full conditional posterior for β as

$$\beta \mid \mathbf{X}, \phi, \sigma^2 \sim N(\frac{B}{2A}, \frac{1}{2A}), \tag{11}$$

where $\mathbf{X} = (X_1, .., X_n)$, $A = \frac{1}{2\sigma^2}\sum_{t=2}^{n} M_t^2 + \frac{1}{2\sigma_\beta^2}$, $B = \frac{1}{\sigma^2}\sum_{t=2}^{n} M_t X_t + \frac{\mu_\beta}{\sigma_\beta^2} - \frac{\phi}{\sigma^2}\sum_{t=2}^{n} M_t X_{t-1}$.
We use a uniform prior on $(-1, 1)$ for ϕ in order to ensure causality of the state process $\{X_t\}$. Using this prior, the posterior for ϕ is given by

$$\phi \mid \mathbf{X}, \beta, \sigma^2 \sim \frac{N(B, \frac{\sigma^2}{\sum_{t=2}^{2} X_{t-1}^2}) I(\phi \in (-1, 1))}{\Phi(\frac{(1-B)\sqrt{\sum_{t=2}^{n} x_{t-1}^2}}{\sigma}) - \Phi(\frac{(-1-B)\sqrt{\sum_{t=2}^{n} x_{t-1}^2}}{\sigma})} \tag{12}$$

where $I(\phi \in (-1, 1)) = 1$, if $\phi \in (-1, 1)$ and 0 otherwise. The above closed form of the posterior is obtained from the conditional form of the autoregressive model where the first observation is treated as fixed. Considering the large number of data points and the simplicity this restricted likelihood achieves in terms of the ease of implementation, we think it is a fair trade-off. Using the full likelihood would result in a complicated posterior, which would warrant a Metropolis like step to sample.

For σ^2, we use a non-informative prior (conditional on ϕ) of scaled Inv-χ with hyperparameters 5 for degrees of freedom and Var(Y_t) for the scale parameter. This prior, along with the likelihood (restricted to $\mathbf{X}_{2:n} = (X_2, .., X_n)$) described earlier, results in a posterior of σ^2 as

$$\sigma^2 \mid \mathbf{X}_{2:n}, \phi, \beta \sim InvGa(\frac{\nu_0 + n - 1}{2}, \beta_1), \tag{13}$$

where $\beta_1 = \frac{\sum\limits_{t=2}^{n} X_t^2 + \beta^2 \sum\limits_{t=2}^{n} M_t^2 - 2\beta \sum\limits_{t=2}^{n} M_t X_t - B^2 \sum\limits_{t=2}^{n} X_{t-1}^2 + (\phi - B)^2 \sum\limits_{t=2}^{n} X_{t-1}^2 + \nu_0 \sigma_0^2}{2\sigma^2}$

and $B = \frac{\sum\limits_{t=2}^{n} X_t X_{t-1} - \beta \sum\limits_{t=2}^{n} M_t X_{t-1}}{\sum\limits_{t=2}^{n} X_{t-1}^2}$

3.2 Gibbs Sampling

Since the conditional distributions are fully specified for all the parameters, we estimate the model parameters using the Gibbs sampling algorithm. The Gibbs sampling algorithm produces samples that converge in distribution to a draw from the stationary joint distribution of $\mathbf{X}$, and θ by constructing a Markov chain $\{\mathbf{X}^{(\mathbf{k})}, \theta^{(\mathbf{k})}\}$ (see Gelman et al. (2003) for details). Our Gibbs sampling algorithm is given by

Algorithm 1 Gibbs sampler

```
Set $k = 1$. Choose some initial values for $\mathbf{X}^{(1)}$, $\phi^{(1)}$, $\beta^{(1)}$, and $\sigma^{2(1)}$ such that $Y_t = g(X_t^{(1)}) + \mu_{m(t)}$. Here $\mathbf{X}^{(1)}$ contains states $X_t^{(1)}$, $t = 1, 2, .., n$
Step 1: For $t = 1, 2, .., n$, replace $X_t^{(k)}$ using equations 8-10 using values from the $(k-1)^{st}$ step for rest of the parameters, and ensuring that $Y_t = g(X_t^{(1)}) + \mu_{m(t)}$. This is done by setting $X_t^{(k)}$ to $Y_t - \mu_{m(t)}$ if $Y_t > 0$ or to a sample from the corresponding truncated distribution, truncated at $-\mu_{m(t)}$ otherwise
Step 2: Update $\phi^{(k)}$ by drawing from equation 12
Step 3: Update $\beta^{(k)}$ by drawing from equation 11
Step 4: Update $\sigma^{2(k)}$ by drawing from equation 13
Step 5: Go to Step 2
```

Typically, the algorithm is run long enough ignoring samples for certain number of iterations (burn-in period) to collect samples after the burn-in period. By construction, these samples are dependent. One way to draw approximately independent samples is to collect a sample from every l^{th} iteration after the burn-in period. In our implementation, we used a burn-in period of 1000 iterations and collected every 2^{nd} sample to build 10000 draws for parameters $\mathbf{X}$, ϕ, β, and σ^2.

3.3 Forecasting

The MCMC scheme is extended to give predictive distributions of $Y_{n+1}, Y_{n+2}, ...,$ given $Y_1, Y_2, ..., Y_n$. After the step of drawing samples of $\mathbf{X}$ (Step 1 in the algorithm), we simply calculate the mean of the posteriod predictive distributions for the forecasting period using the current values of $\mathbf{X}^{(k)}$ and $\theta^{(k)}$, and MIROC5.

The distrbution of the state process for the forecasting period is given by

$$X_{n+f} \mid \theta \sim N(\phi E(X_{n+f-1} \mid X_n) + \beta M_{n+f}, \sigma^2) \tag{14}$$

$$Y_{n+f} = g(X_{n+f}) + \mu^*_{m(n+f)} \tag{15}$$

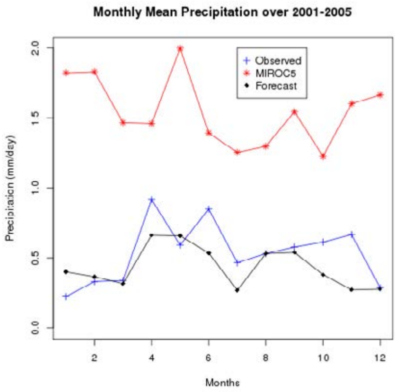

Figure 8: Monthly mean precipitation for 2001-2005

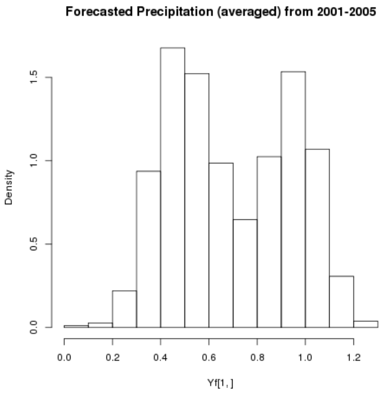

Figure 9: Predicted daily precipitation for 2001-2005

where $f = 1, 2, .., n_f$, and $\mu^*_{m(n+f)}$ is the mean of the observed precipitation from the month, year combination day $n + f$ falls into.

3.4 Results

Figure 8 shows the monthly mean of predicted daily precipitation for the period $2001-2005$ along with the observed and MIROC5 values. Figure 9 shows the histogram of predicted values for the same period, where the predicted value for the day f is the median of 10000 draws of Y_{n+f}. The rationale behind taking median instead of say, mean is that it can be interpreted as a consensus of dry/wet day from all the samples. In other words, if more than 50% of the samples are zeros, we can consider it as a dry day. Mean, on the hand, will always produce a strictly positive prediction (unless all the samples are zeros). Since we want our model to be able to predict zeros, median seems to be a better alternative. However, a downside of choosing median (or mean for that matter) is that the resulting predictions are close to independent whereas the observed values for each day are not. Figure 10 shows the histogram of the observed data for the period $2001 - 2005$. As our predictions in figure 9 indicate, our model has not predicted sufficient number of matched dry days. In other words, for most true dry days, more than 50% of the MCMC samples were strictly positive. Also, as the range of values in figure 9 show, our model predicted mostly low intensity positive values. This observation, in conjunction with reasonably good predictions at the monthly mean levels, implies that for several days (many of which are dry), predictions from our model are small positive quantities as the range of values in figure 9 shows.

Figures 11, 12, and 13 show histograms of the samples generated from the posteriors of ϕ, β, and σ^2 respectively. Low posterior mean for ϕ, and a close to zero estimate for β also suggest the low number of matched zero values in our predictions. Adding more auto-regressive terms and other covariate series to the state process could improve our predictions. We plan to investigate into these in future. We also note that instead of median, choosing $100p^{th}$ percentile for a suitable p could improve the predictions.

4. A Transfer Function model

In the Bayesian approach described in Section 3.1, the MIROC5 data is considered fixed, whereas in reality it is also a time series process. In this section we will discuss a transfer function model where the MIROC5 series will be used as a covariate in a lagged regression setting to predict the daily precipitation in two steps. In the first step an Expectation-Maximization (EM)-like method is used to impute the censored part of the state process,

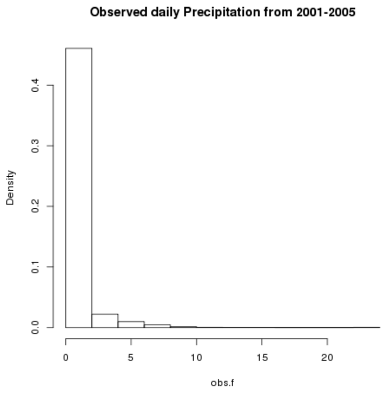

Figure 10: Histogram of daily prediction for 2001-2005

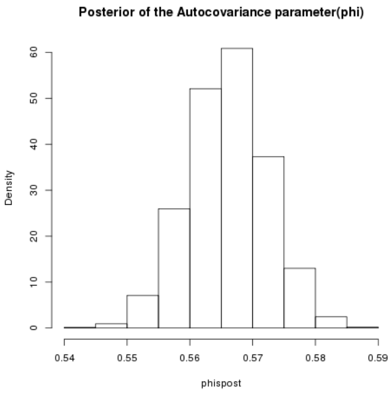

Figure 11: Posterior of the autoregressive parameter

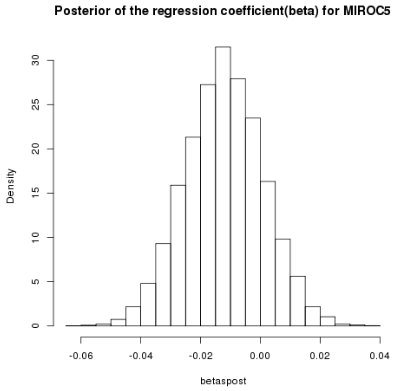

Figure 12: Posterior of the coefficient of MIROC5

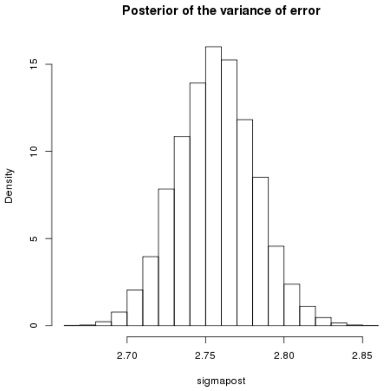

Figure 13: Posterior of the variance of error

which is partially observed as positive precipitation. In the second step the complete state process (with the zero values “filled in”) is used to fit a transfer function model using the MIROC5 simulated data as the covariate process.

4.1 Step One: An EM-like imputation method

The EM (Expectation-Maximimization) method is an iterative estimation procedure used when maximizing a likelihood function is difficult. In such a situation, often a latent variable, which is assumed to drive the observed data, is introduced to simplify the likelihood function. The simplified likelihood function, although often easier to maximize, is not completely known since the latent variables are not observed. Instead, the conditional expectation of the new likelihood function given the observed data, and the current value of the parameter is maximized to obtain an updated estimate of the parameter. These two steps of calculating conditional expectation of the simplified likelihood and maximizing it over the parameter space is iterated until the parameter estimate converges.

Following the notation in Hastie et al. (2009), let $\mathbf{Z}$ be the observed data and θ be the parameter to be estimated, with log-likelihood $\ell(\theta;\mathbf{Z})$. Let $\mathbf{Z}^m$ be the latent (or missing/censored) data with log-likelihood $\ell_1(\theta;\mathbf{Z}^m \mid \mathbf{Z})$ based on the conditional density of $\mathbf{Z}^m \mid \mathbf{Z}$, and $\mathbf{T} = (\mathbf{Z},\mathbf{Z}^m)$ be the complete data with log-likelihood $\ell_0(\theta;\mathbf{T})$. Applying the Bayes formula on $\mathbf{T} \mid \theta$, we have

$$\ell(\theta;\mathbf{Z}) = \ell_0(\theta;\mathbf{T}) - \ell_1(\theta;\mathbf{Z}^m \mid \mathbf{Z}) \tag{16}$$

Taking conditional expectations with respect to the distribution of $\mathbf{T} \mid \mathbf{Z}$ with a known parameter $\theta^{'}$ on both sides of equation 16, we get

$$\ell(\theta; \mathbf{Z}) = E[\ell_0(\theta; \mathbf{T}) \mid \mathbf{Z}, \theta^{'}] - E[\ell_1(\theta; \mathbf{Z}^m \mid \mathbf{Z}) \mid \mathbf{Z}, \theta^{'}] \equiv Q(\theta, \theta^{'}) - R(\theta, \theta^{'}) \qquad (17)$$

Note that we seek to maximize $\ell(\theta; \mathbf{Z})$ over θ. Since it is assumed to be a difficult function to maximize, it can be shown that maximizing $Q(\theta, \theta^{'})$, which is often an easier function to maximize, will never decrease the likelihood value of $\ell(\theta; \mathbf{Z})$. We show this is indeed the case in Appendix. Algorithm 2 shows the steps involved in the EM algorithm.

Algorithm 2 The EM Algorithm

```
Require: Initial estimates $\hat{\theta}^{(0)}$
1: E Step: At the $j^{th}$ step compute $E[\ell_0(\theta; \mathbf{T}) \mid \mathbf{Z}, \hat{\theta}^{(j)}]$ as a function of $\theta$
2: M Step: Maximize $E[\ell_0(\theta; \mathbf{T}) \mid \mathbf{Z}, \hat{\theta}^{(j)}]$ over $\theta$ to get the new estimate $\hat{\theta}^{(j+1)}$
3: Iterate steps E and M until convergence
```

The method we use to estimate the censored observations is different from the EM algorithm described above in two respects. Since we use the Yule-Walker method (Ch. 3- Shumway and Stoffer (2011)) to estimate the time series model, we do not work with the likelihood of the observed data and therefore the need to calculate the conditional expectation of $\ell_0(\theta; \mathbf{T})$ does not arise. Instead, the E Step in our method involves estimating or "filling in" the censored observations using the current estimate of θ and the M Step involves the Yule-Walker estimation of θ using the complete data (with the imputed values). These two steps are iterated with the E Step re-estimating only the censored part. Algorithm 3 shows the steps involved in our method. Similar approaches were used by Miller and Ferreriro (1984), and Bose and Neerchal (1997).

Algorithm 3 The EM-like Algorithm

```
Require: Initial Yule-Walker estimates $\hat{\theta}^{(0)}$ using $\mathbf{Z}$
1: E Step: At the $j^{th}$ step set $\mathbf{Z}^m = E[\mathbf{Z}^m \mid \mathbf{Z}, \hat{\theta}^{(j)}]$ to get $\hat{\mathbf{Z}}^m$
2: M Step: Update the Yule-Walker estimates using the $(\mathbf{Z}, \hat{\mathbf{Z}}^m)$ to get $\hat{\theta}^{(j+1)}$
3: Iterate steps E and M a few times
```

Let $\mathcal{O}$ be the set of time points when the precipitation is observed and $\mathcal{C}$ the set of time points when it is 0 (censored). Note that the latent state process X_t is same as Y_t $\forall t \in \mathcal{O}$. The first step of the method is to fully estimate the state process by imputing the censored ("missing") values for $t \in \mathcal{C}$. Figure 5 shows no significant time trend and indicates a seasonal pattern, as expected of daily rainfall data. We therefore fit a sinusoidal component as the mean process to de-seasonalise the observed series. We have used a scaled periodogram to identify prominent periods of 365 days (yearly), 182 days (half-yearly), 15 days, and 7 days. Therefore, our mean component can be represented as

$$g(t) = \sum_{f \in \mathbb{F}} (\beta_{1f} \sin(2\pi f t) + \beta_{2f} \cos(2\pi f t)), \qquad (18)$$

where $\mathbb{F} = \{\frac{1}{7}, \frac{1}{15}, \frac{1}{182}, \frac{1}{365.25}\}$. We fit this mean component to only the strictly positive observed values i.e., X_t, where $t \in \mathcal{O}$ to get the de-seasonalised process X_t^e, $t \in \mathcal{O}$. We assume that this process is stationary (weakly). Partial auto-correlations of the $\{X^e\}$ process in Figure 14 suggests an AR(1) model. Although the auto-correlations (not shown here) also indicated an MA(1) process, we chose not to use an ARMA(1,1) model for

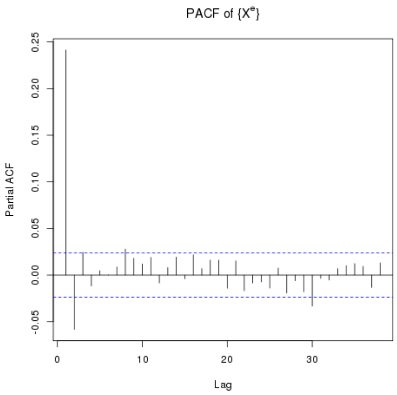

Figure 14: Partial Autocorrelations of $\{X^e\}$

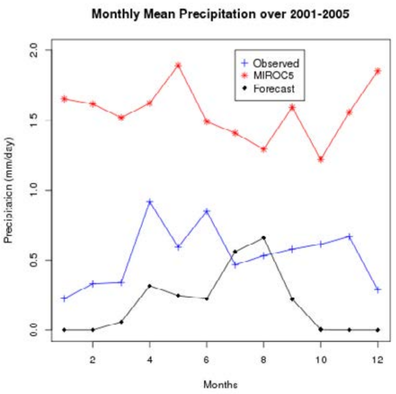

Figure 15: Monthly Observed vs Forecasted

simplicity. The Yule-Walker estimation method is used (M Step in Algorithm 3) as an initial estimate of the AR(1) model, which is used to “fill in” the values for the censored X_t, where $t \in \mathcal{C}$ (E Step). The M Step is repeated, this time using the complete $\{X_t^e\}$ series, with the imputed censored part. Note that the observed part is not changed. In the subsequent E Step, the censored part is re-imputed using the updated AR(1) parameter. These steps shown in Algorithm 3 are repeated a few times. We have not investigated into the convergence properties of the Yule-Walker estimates obtained in this fashion but we believe one or two iterations are sufficient. At the end of this estimation step, we get the complete latent process $\{X_t\}$, part of which is observed and the rest is imputed.

4.2 Step Two: Lagged Regression

Using the procedure described in Section 4.1, we have the complete detrended observed series $\{X_t^e\}$, $t = 1, 2, .., n$. In the second step of our method, we treat MIROC5 provided data as a time series process and use it as a covariate in a lagged regression setting. In a transfer function model we assume that the observed process can be explained by an exogenous time series process and it’s own lagged values. Consider the lagged regression model of the form

$$X_t^e = T(B)M_t^e + \eta_t, \tag{19}$$

where $\{X_t^e\}$ is the complete detrended observed series, $\{M_t^e\}$ the detrended MIROC5 series, $\{\eta_t\}$ is the noise series, and $T(B) = \sum_{j=0}^{\infty} t_j B^j$. Here, $\{M_t^e\}$, and $\{\eta_t\}$ are assumed stationary and mutually independent.

Estimating a transfer function model is usually done in a sequential manner (see Shumway and Stoffer (2011) for details) with the first step being fitting an ARIMA model to the detrended input $\{M_t\}$ series. Figure 3 indicates a seasonal pattern but no clear trend in the daily and smoothed series by MIROC5. Based on a scaled periodogram, we fit a sinusoidal component at one year time period to get the process $\{M_t^e\}$, which is assumed stationary. Based on ACF and PACF plots (not shown) from this series, an AR(9) model is fitted to $\{M_t^e\}$. The fitted AR(9) operator is used to transform the output series $\{X_t^e\}$ to get $\{\tilde{X_t^e}\}$, whose cross-correlation with the residual process from the AR(9) model for $\{M_t^e\}$ is used to suggest the regression form $T(B)$.

The operator T in equation 19 can be represented as the ratio of polynomial operators of the form $\frac{\gamma(B)}{\omega(B)}B^d$, where $\gamma(B) = \sum_{i=0}^{s} \gamma_i B^i$, and $\omega(B) = \sum_{i=0}^{r} \omega B^i$. The number of parameters s, and r, and the delay parameter d can be infered from the cross-correlation plot mentioned above between the whitened $\{M_t^e\}$ and the transformed output processes. Using the ratio representation of T, equation 19 can be written as

$$X_t^e = \sum_{i=1}^{r} \omega_i X_{t-i}^e + \sum_{j=1}^{s} \gamma_j M_{t-d-j}^e + e_t, \tag{20}$$

where $e_t = \omega(B)\eta_t$ A regression is performed using equation 20 to get the estimates of $\omega_1, .., \omega_r, \gamma_1, .., \gamma_s$, and the residuals $\hat{e}_t$, to which the estimated operator $\hat{\omega}(B)$ is applied to get the estimated noise $\hat{\eta}_t$. The last step involves fitting an ARMA(p_η, q_η) model to $\hat{\eta}_t$. These steps are shown in Algorithm 4.

Forecasting is done by first forecasting $\hat{\eta}_{n+f}$ using the estimated ARMA(p_η, q_η), followed by plugging in the MIROC5 values from the prediction period and the lagged values of X_t^e into the estimated equation 20 as

$$X_{n+f}^e = \sum_{i=1}^{r} \hat{\omega}_i X_{n+f-i}^e + \sum_{j=1}^{s} \hat{\gamma}_j M_{n+f-d-j}^e + \hat{\omega}(B)\hat{\eta}_{n+f} \tag{21}$$

Algorithm 4 Fitting a transfer function

```
Step 1: Fit an ARMA model to the stationary $\{M_t^e\}$.
Step 2: Transform the output series $\{X_t^e\}$ using the fitted ARMA operator from Step 1
    to get $\{\tilde{X}_t^e\}$.
Step 3: Using cross-correlation values between $\{\tilde{X}_t^e\}$, and the whitened process (resid-
    uals) from Step 1, infer a form for the operator $T$ in equation 19.
Step 4: Fit the multiple linear regression model in equation 20 to get the residuals $\hat{u}_t$.
Step 5: Transform the residuals $\hat{u}_t$ using the estimated MA operator $\hat{\omega}(B)$ in Step 4 to
    get the estimated noise process $\{\hat{\eta}_t\}$.
Step 6: Fit an ARMA model to $\{\hat{\eta}_t\}$.
```

4.3 Results

Figure 16 shows the time plot of the observed precipitation with the forecasts superimposed. Clearly, the predictions seem to underestimate the intensity of precipitation. Figure 17 shows the histogram of predictions from the transfer function method. The method is able to successfully predict zero values with around 52% matches in dry days, an overall 50% censoring at zero when the observed data has 62% zero values. The predicted intensity has an overall mean value of 0.38 mm/day, whereas the observed intensity has an average value of 0.82 mm/day.

5. Discussion

We have described two methods to predict daily precipitation at a location using the daily MIROC5 simulated data as the exogenous series. The first method uses the Bayesian approach to estimate a standard Tobit state space model, where the observation equation takes the Tobit form and the state process has MIROC5 data as the covariate. Since predictions are read off from the posterior predictive distributions as median values from the MCMC iterations, this method in principle can predict zero values. This is a natural attribute to

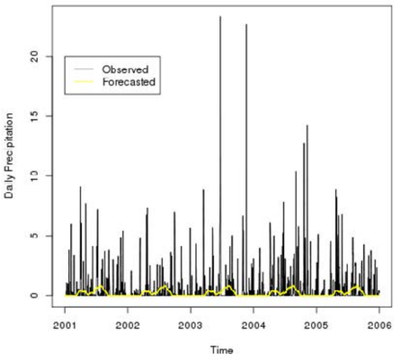

Figure 16: Observed vs Forecasted

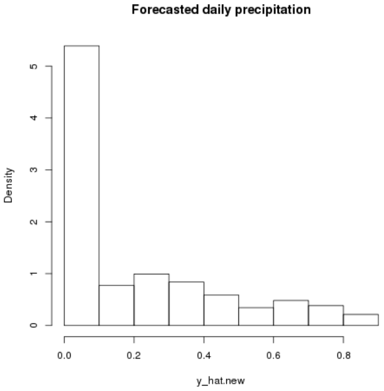

Figure 17: Forecasts from the Transfer function method

require of a prediction method for daily rainfall. The results however suggest several improvements to the state process. For example, including more auto-regressive terms and additional climate variables as covariates could be investigated into. As for implementation, since all the conditional distributions of parameters are fully specified, implementing a Gibbs scheme is straight-forward. However, including additional covariates could complicate implementing MCMC. Another downside of this method is the run time to fit the model. For a modest number of iterations in the Gibbs algorithm, the model has taken around 30 minutes. If the research problem demands fitting this model at several locations, the large run time might render the model infeasible to implement. However, in those situations, spatial dependence could be incorporated, possibly in a Hierarchical fashion, which would significantly alter the model described here.

The second method discussed in this paper is a transfer function model with the zero values in the output series imputed using an EM-like method prior to model fitting. This method too by design can predict zero values. Results indicate that the matched proportion of dry days is signficantly improved compared to the Bayesian method. However, intensities still seem to be underestimated. A possible reason for this could be a missing time dependence component in the residual process, which needs to be investigated closely. In future, we would like to improve these models by incorporating more covariates, and compare their predictive performance with the models from Teutschbein and Seibert (2012), Popuri et al. (2015), and SWAT's own Weather Generator.

Acknowledgments

We gratefully acknowledge the partial funding for this work by USDA-NIFA grant 2011-67003-30213 to The Center For Research on the Changing Earth System (CRCES), who we would like to thank for making the climate datasets available. Results were obtained using the computational systems at the UMBC High Performance Computing Facility (HPCF). First author would like to thank Joint Center for Earth Systems Technology (JCET) for funding and Dr. Anindya Roy for his valuable early feedback on the implementation of the Bayesian method.

Appendix

Below we show that maximizing $Q(\theta, \theta^{'})$ in Section 4.1 over θ never decreases the log likelihood $\ell(\theta; \mathbf{Z})$.

We reproduce the equation 17 below

$$\ell(\theta;\mathbf{Z}) = E[\ell_0(\theta;\mathbf{T}) \mid \mathbf{Z},\theta^{'}] - E[\ell_1(\theta;\mathbf{Z}^m \mid \mathbf{Z}) \mid \mathbf{Z},\theta^{'}] \equiv Q(\theta,\theta^{'}) - R(\theta,\theta^{'}) \tag{22}$$

Note that the Expectations in the above equation is with respect to $\mathbf{T} \mid \mathbf{Z},\theta^{'}$, whose distribution is same as $\mathbf{Z}^m \mid \mathbf{Z},\theta^{'}$. Therefore,

$$E[\ell_1(\theta;\mathbf{Z}^m \mid \mathbf{Z}) \mid \mathbf{Z},\theta^{'}] = E_{\mathbf{Z}^m|\mathbf{Z},\theta'}[\ell_1(\theta;\mathbf{Z}^m \mid \mathbf{Z})] = E_{\mathbf{Z}^m|\mathbf{Z},\theta'}[logPr_\theta(\mathbf{Z}^m \mid \mathbf{Z})]$$

Since $Pr(\mathbf{Z}^m \mid \mathbf{Z}) > 0$, $logPr_\theta(\mathbf{Z}^m \mid \mathbf{Z})$ is a strictly concave function. Therefore by Jensen's Inequality,

$$E_{\mathbf{Z}^m|\mathbf{Z},\theta'}[logPr_\theta(\mathbf{Z}^m \mid \mathbf{Z})] \leq log(E_{\mathbf{Z}^m|\mathbf{Z},\theta'}[Pr_\theta(\mathbf{Z}^m \mid \mathbf{Z})])$$

Therefore, $R(\theta,\theta^{'})$ in equation 27 is maximized when $\theta = \theta^{'}$.

Let $\theta^{(j)}$ is the estimate of θ at the M Step of jth iteration in 2, and $\theta^{(j+1)}$ the argmax of Q over θ at the $(j+1)^{st}$ step. Substituting $\theta^{(j)}$ for θ in equation 27, we get

$$\ell(\theta^{(j)};\mathbf{Z}) = Q(\theta^{(j)},\theta^{'}) - R(\theta^{(j)},\theta^{'})\forall\theta^{'} \tag{23}$$

Substituting $\theta^{(j)}$ for $\theta^{'}$ in equation 23, we get

$$\ell(\theta^{(j)};\mathbf{Z}) = Q(\theta^{(j)},\theta^{(j)}) - R(\theta^{(j)},\theta^{(j)}) \tag{24}$$

Substituting $\theta^{(j+1)}$ for θ in equation 27, we get

$$\ell(\theta^{(j+1)};\mathbf{Z}) = Q(\theta^{(j+1)},\theta^{'}) - R(\theta^{(j+1)},\theta^{'})\forall\theta^{'} \tag{25}$$

In particular, at $\theta^{(j)}$ for $\theta^{'}$, we get

$$\ell(\theta^{(j+1)};\mathbf{Z}) = Q(\theta^{(j+1)},\theta^{(j)}) - R(\theta^{(j+1)},\theta^{(j)})\forall\theta^{'} \tag{26}$$

Subtracting equation 24 from equation 26, we get

$$\ell(\theta^{(j+1)};\mathbf{Z}) - \ell(\theta^{(j)};\mathbf{Z}) = Q(\theta^{(j+1)},\theta^{j}) - Q(\theta^{(j)},\theta^{(j)}) + R(\theta^{(j)},\theta^{(j)}) - R(\theta^{(j)},\theta^{(j+1)}) \geq 0 \tag{27}$$

References

A. Bose and N. K. Neerchal. Estimation of Autoregressive Parameters with Systematic but Incomplete Sampling. *Proceedings of 2nd Triennial Symposium on Probability and Statistics, Calcutta, India*, pages 41–51, 1997.

A. E. Brockwell, N. H. Chan, and P. K. Lee. A class of models for aggregated traffic volume time series. *Appl. Statistics*, 52:417–430, 2003.

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.

A. Gelman, J. B. Carlin, H. S. Stern, and Rubin D. B. *Bayesian Data Analysis*. Chapman and Hall/CRC, 2003.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

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.

R. B. Miller and O. M. Ferreriro. A strategy to complete a time series with missing observations. In E. Parzen, editor, *Time Series Analysis of Irregularly Observed Data. Proceedings of a Symposium help at Texas A & M University, College Station, Texas.* Springer-Verlag, New York, 1984.

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.

S. K Popuri, N. K. Neerchal, and A. Mehta. Comparison of Linear and Tobit Modeling of Downscaled Daily Precipitation over Missouri River Basin using MIROC5. In V. Lakshmanan, E. Gilleland, A. McGovern, and M. Tingley, editors, *Machine Learning and Data Mining Approaches to Climate Science. Proceedings of the 4th International Workshop on Climate Informatics*. Springer, 2015.

R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications*. Springer, 2011.

A. Takeshi. *Advanced Econometrics*. Harvard University Press, Cambridge, Massachusetts, 1985.

C. Teutschbein and J. Seibert. Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods. *Journal of Hydrology*, 456-457:12–29, 2012.

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.