Hierarchical Bayes models for daily rainfall time series at multiple locations from different data sources (VERY PRELIMINARY: DO NOT CITE)

March 6, 2015

Abstract

We estimate a Hierarchical Bayesian models for daily rainfall that incorporates two novelties for estimating spatial and temporal correlations. We estimate the within site time series correlations for a particular rainfall site using multiple data sources at a given location, and we estimate the across site covariance in rainfall based on location distance. Previosu models have captured cross site correlations as a functions of site specific distances, but not within site correlations across multiple data sources, and not both aspects simultaneously. Further, we incorporte information on the technology used (satellite versus rain guage) in our estimations, which is also a novel addition. This methodology has far reaching applictions to other data contexts in which multiple datasources exist for a given event or variable for which within and between site covariances can be estimated over time.

JEL Classification: O160, G22, Q140
Keywords: micro finance, index insurance, credit constraints, financial education

1 Motivation

Weather simulations are used to study future weather patterns in the context of climate-sensitive systems, and to simulate climate change scenarios and study its repercussions. Extreme weather events, such as drought, are frequently the hardest to predict, the most damaging to agricultural livelihoods and the costliest events in terms of disaster response. However, multiple data sources may not always coincide with one another and may not coincide with human experience of extreme weather events making responsiveness difficult. In the case of such conflicting data, it is important that weather predictions are not drawn based on one information source alone, but with attention to multiple sources and the technology with which data are captured. Futhermore, more robust within site estimations can render more robust across site correlations. Both aspects can enable better simulation.

Past weather simulators have not simultaneously captured within site and across site correlation of weather time series. Initially, many weather simulators were estimated using one time series per location. Wilks and Wilby (1999) provide an overview of the developments in weather simulation. Weather series were first modeled with stochastic point process models in space and time. Both (Sanso) and (weather game) provide good overviews of this trajectory. First order Markov chains were first used to generate such point processes, modeling the occurrence of rainy versus non-rainy days dictated by transition probabilities. Transition probabilities were based on the relative frequency of precipitation in observed climatology (Gabriel and Neumann (1962)). Improvements to this first model then included modeling the non-zero rainfall days (Todorovic and Woolhiser, 1975), which followed with developments in the type of distribution for non-zero days: including an exponential distribution (Richardson (1981)) and then a gamma distribution [Thom, 1958; Katz, 1977; Buishand, 1977; 1978; Stern and Coe, 1984; Wilks, 1989; 1992 (1981)]. Other methodologies for simulating precipitation occurrences is the use of spell-length models. Rather than simulating rainfall occurrences day by day, spell-length models operate by fitting wet and dry periods to different probability distributions (Buishand, 1977; 1978; Roldan and Woolhiser,1982). Non-parametric simulation methods have also been developed which involved resampling data such that time correlations are captured in the resampling to estimate distribution parameters (Young (1994), Lall and Sharma (1996), Lall et al. (1996) and Rajagopalan et al. (1997)).

The objective in all of these approaches is to be able to simulate long synthetic time-series of weather that reflect key observed statistics in the region of interest e.g. means, dry spell lengths, inter annual rainfall variability, or probability of extreme events occurring. Capturing a spatial dependence between sites or series, as well as charactersitics of the series themselves, such as non-stationarity, can improve our predictive analytics.

Thus incorporating observed spatial correlations into a well calibrated stochastic approach is likely to lead to more realistic impact models. Several different techniques and methods have been suggested. The MarkSim weather generator uses spatially correlated input grids of statistics to allow one to allow estimation at a site where there are no observations (Jones and Thornton, 1993). Wilks (1999,1998) extended a Richardson weather generator by driving it with a grid of spatially correlated random numbers. More recently this approach was extended in the GIST weather generator, which incorporates spatial correlation through the use of correlation matrices to sample from a cumulative probability function at each location (Baigorria and Jones, 2010), or in Kleiber et al (2012) the use of latent Gaussian processes. In 2012. Greatrex (2012) proposed a geostatistical sequential simulation approach coupled to a markov generator, which would allow spatially correlated ensembles of maps of rainfall to be generated. It is important to note that many of these approaches rely on a large amount of observed and calibrated data, for example a dense station network to create variograms [HELEN EXPLAIN].

Hierchacal bayesian estimation is yet another approach equipped to model spatial as well as temporal relationships. It goes beyond the simplistic and often linear relationships assumed in markovian weather simulators. It models and estimates the functional relationship between weather series over time. The most recent development in the use of the hierarchal bayesian approach comes from Sanso and Guenni (2012), which improves upon Sanso and Duenni (1996), by 1) incorporating spatial correlation for the joint distribution of weather series from several stations and 2) the non-stationarity of rainfall data. The authors show that their model is able to simulate weather data and summary statistics for a large number of stations, some of which have poor quality data.

While we do not focus on the aspect of non-stationarity aspect in raifnall series, the novelty of our approach is our ability to account for two levels of the rainfall time series information at each site: 1) both the location of the measurement and 2) the instrument used to measure rainfall are incorporated. In addition, we 3) estimate the the spatial covariance between all these series and 4) the noise or error in recording rainfall due to the instrument itself. By incorporating several layers of information sources for each site in which we would like to predict weather, we are adding more information to the parameter estimates and, hopefully, improving out-of-sample predictions. Because we used daily data to fit our model, our simulated data and out-of-sample predictions are daily, which allows us to compute statistics sensitive to daily measurements: probability of rainfall, dry spells, and extreme rainfall.

The ability to incorporate spatial correlations has several policy implications. First, food insecurity is generally correlation in time and space in parallel with rainfall patterns. Thus, predictions that incorporate these patterns can allow for better response to anticipated droughts and subsequent food shortages. This is crucial for aid organizations and first responders to such crises. In addition, longer term tools such as weather insurance, are quite sensitive to spatial phenomena. Higher spatial correlations affect the price and responsiveness of such financial tools. Large covariate risks are the focus of index insurance, for example, and institutions offering such tools need adequate capital holdings to properly respond to the degree of and size of covariate risk.

The remainder the paper is as follows. In In Section 2 describes the data used and its preparation. In Section 3 shows how rainfall the and the processes are modeled in a hierarchical fashion. In Section 4 discusses the parameters to be estimated. In Section 5 exhibits the performance in recovering parameters with simulated data. In Section 6 exhibits the ability of our model, using our real historical data, in recovering deliberately removed portions of the data. In Section 7 we graph the posteriors distributions for select parameters using the full historical data set. In Section 8 concludes and describes next steps.

2 Procedure and Data

We model a set of 15 time series of daily rainfall in Ethiopia during the time period from 1992 to 2010, where these 15 time series come from six different locations, and each location has at least two time series of daily rainfall associated with it. The reason we observed multiple time series for each location is that for each location we have multiple sources of data, including rain station data and satellite-based rainfall proxy data. The statistical challenge to modeling such data is to separately estimate the spatial variability between locations and the within-location measurement-based variability. A standard hierarchical model with sets of 15 exchangeable parameters – one for each time series – would conflate these two sources of variation. The model we introduce here – a hierarchical Bayesian model with one level for locations, and another level for multiple data sources within a location – explicitly models each of these two sources of variation.

Figure 1 shows the six locations from which the daily rainfall time series are measured. The names of the locations are Hagere Selam, Maykental, Mekele, Abi Adi, Agibe, and Adi Ha.

The last of these, Adi Ha, is the location we are most interested in, because we wish to provide rainfall insurance for farmers who live there. Specifically, we want to model rainfall at one of the automated rain stations at Adi Ha, which is one of the 15 time series in our data set, but is also the series that has the least amount of observed data – only about 200 days worth of data from 2009.

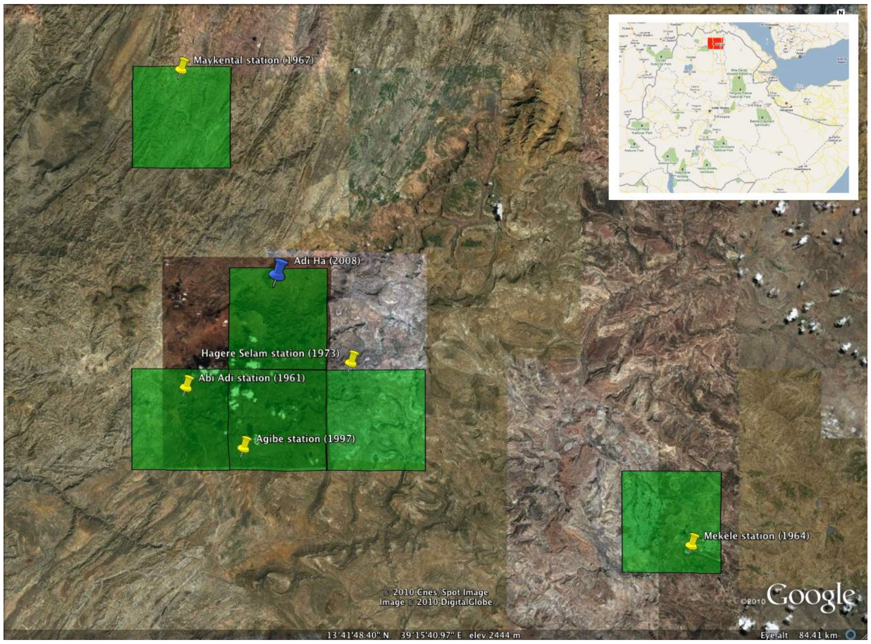

Figure 1: A map of the 6 locations, where the green squares denote ARC pixels and the pins denote rain station locations. The inset in the upper right corner shows the area of the map with a red rectangular box; this region is in north central Ethiopia.

The specifics of the data are as follows:

- For the first five locations, we have rain station data as well daily measurements from a satellite product called ARC, which is a rainfall proxy based on the temperature of the clouds over an area of about one hundred square kilometers. This comprises $5 \times 2 = 10$ time series.
- For the sixth location, Adi Ha, we have five separate data sources:
  1. One reliable rain station from which we only have 200 days of data from 2009-

2010; this is the time series in which we have the most interest, because it is a new, accurate rain station on which we want to base insurance contracts.

2. One unreliable rain station from which we have about 7 years of data from 2000-2009, with about 2 years of missing data interspersed.

3. The ARC satellite proxy.

4. Two additional satellite proxies that are different from ARC.

Figure 2 shows the range of observed and missing data for each of these 15 time series; note that the time scale goes back to 1961 for one of the rain stations, but for simplicity, we only consider the time span of 1992-2010 in our model fit, because this span contains most of the data.

Figure 2: A visualization of the observed data for each of the 15 time series we model. The black hash marks denote rain station data, and the red hash marks denote satellite-based data.

Table 1 contains some background information and summary statistics related to each time series of daily rainfall. For each time series we record the latitude, longitude, and elevation of the location where measurements were made, and the number of days of observed data. The maximum distance between locations is about 70 kilometers (between Mekele in the southeast and Maykental in the northwest).

Table 1: Background information about the 15 time series

	Site	Latitude	Longitude	Elev. (m)	Num. Obs
1	Hagere Salaam	13° 38' 49"	39° 10' 19"	2625	4887
2	Hagere Salaam (ARC)	"	"	"	5632
3	Maykental	13° 56' 13"	38° 59' 49"	1819	5620
4	Maykental (ARC)	"	"	"	5632
5	Mekele	13° 28' 1"	39° 31' 1"	2247	6205
6	Mekele (ARC)	"	"	"	5632
7	Abi Adi	13° 37' 19"	39° 0' 10"	1849	4205
8	Abi Adi (ARC)	"	"	"	5632
9	Agibe	13° 33' 43"	39° 3' 43"	1952	4722
10	Agibe (ARC)	"	"	"	5632
11	Adi Ha (ARC)	13° 43' 48"	39° 05' 38"	1713	5632
12	Adi Ha (Rain Station - Manual)	"	"	"	2769
13	Adi Ha (RFE2)	"	"	"	2920
14	Adi Ha (CMorph)	"	"	"	2190
15	Adi Ha (Rain Station - Automatic)	"	"	"	186

2.1 Exploratory Data Analysis

In this part of Ethiopia, the rainy season lasts roughly from June to October. Figure 3 shows the percentage of rainy days and the average amount of rain as a function of the time of year for each time series. The basic modeling strategy will be to use a set of periodic functions to model rainfall as a function of the season of the year.

We are also interested in the difference, on average, between the measurements of rainfall based on the ARC satellite proxy and the rain stations. Comparing rainfall frequencies pooled over 5-day periods, averaging across all parts of the year and all five locations with

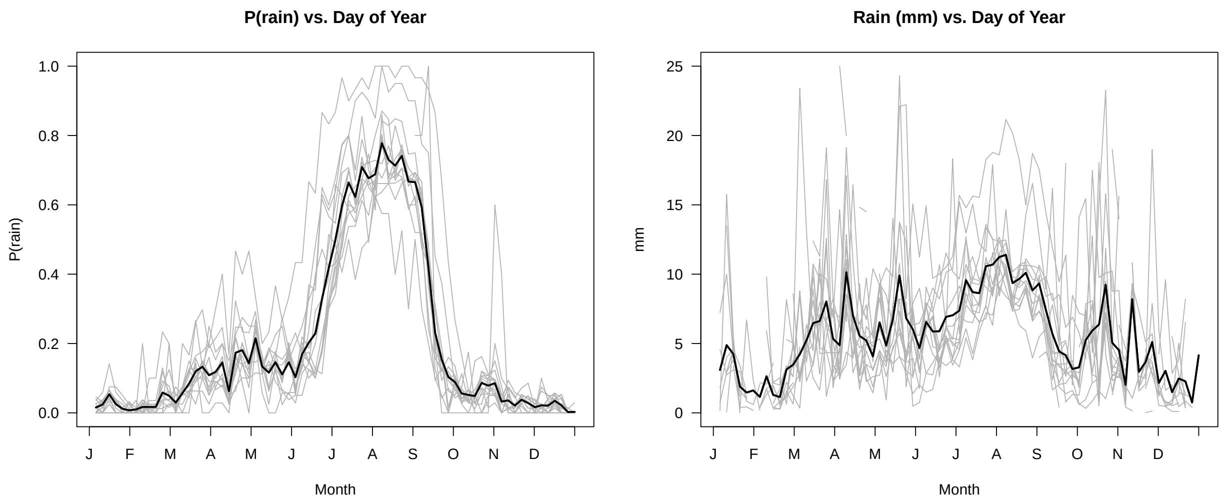

Figure 3: Plots of the percentage of rainy days (pooled into 5-day bins) and the average amount of rain as a function of the season. The gray lines are for each of the 15 individual time series, and the black lines are averaged across all 15 time series.

exactly one rain station and one ARC measurement, we find that the ARC records about 3% fewer days of rainfall than the rain stations. Across locations, this difference ranges from about -6% (Hager Selam) to +1% (Agibe).

3 Modeling Rainfall

We fit a tobit model for daily rainfall at multiple locations, with multiple time series observed at each site. Let us first set up some notation. Let $S = 6$ denote the number of locations where we measure rainfall, and $\boldsymbol{J} = \{2, 2, 2, 2, 2, 5\}$ is the vector denoting the number of daily rainfall time series observed for each of the S locations. The total number of days in our time series is $N = 6679$ days, from 1/1/1992 through 7/28/2010. Let Y_{stj} denote the amount of rainfall, measured in mm, for location $s \in (1, ..., S)$, day $t \in (1, ..., N)$, and time series $j = (1, ..., J_s)$. Last, let D_{ik} denote the Euclidian distance between site i and k, for $i, k \in (1, ..., S)$.

We model Y_{stj} using a hierarchical Bayesian tobit regression model, where the levels of the hierarchy correspond to different sources of variation:

$$Y_{stj} = \begin{cases} W_{stj} & \text{if } W_{stj} > 0 \\ 0 & \text{if } W_{stj} \leq 0, \end{cases} \qquad \text{Observed rainfall}$$

$$\boldsymbol{W}_{st} \sim \mathrm{N}_{J_s}(\boldsymbol{Z}_{st} + \boldsymbol{X}_s^{\mathrm{ARC}} \beta_s^{\mathrm{ARC}}, \tfrac{1}{\gamma_{st}} \Sigma_s), \qquad \text{Latent rainfall}$$

$$\boldsymbol{Z}_t \sim N_S(\boldsymbol{X}_t \boldsymbol{\beta}^Z, \tau^2 \boldsymbol{V}), \qquad \text{Spatial mean rainfall}$$

$$\beta_{ps}^Z \sim \mathrm{N}(\mu_p, \sigma_p^2),$$

$$\mu_p \sim \mathrm{N}(0, 5^2),$$

$$\sigma_p \sim \tfrac{1}{2}\mathrm{t}(0, 1, \mathrm{df} = 3),$$

$$\tau \sim \tfrac{1}{2}\mathrm{t}(0, 1, \mathrm{df} = 3),$$

$$\boldsymbol{V} = \{v_{ik}\}, v_{ii} = 1, v_{ik} = \exp(-\lambda d_{ik}) \qquad \text{for } i, k \in (1, .., S)$$

$$\lambda \sim \mathrm{gamma}(\mathrm{shape} = 50,\ \mathrm{scale} = 3),$$

$$\boldsymbol{X}_t = (1, t, t^2, \sin(2\pi t \omega_1), \cos(2\pi t \omega_1), ..., \sin(2\pi t \omega_4), \cos(2\pi t \omega_4), X_{\mathrm{Jan}_{[t]}}^{\mathrm{nino}}, X_{\mathrm{Feb}_{[t]}}^{\mathrm{nino}}, ..., X_{\mathrm{Dec}_{[t]}}^{\mathrm{nino}}),$$

$$\boldsymbol{X}_{sj}^{\mathrm{ARC}} = 1(\text{jth time series at site s is an ARC product}),$$

$$\beta_s^{\mathrm{ARC}} \sim \mathrm{N}(\mu^{\mathrm{ARC}}, \tau_{\mathrm{ARC}}^2),$$

$$\mu^{\mathrm{ARC}} \sim \mathrm{N}(0, 5^2),$$

$$\tau_{\mathrm{ARC}} \sim \tfrac{1}{2}\mathrm{t}(0, 1, \mathrm{df} = 3),$$

$$\Sigma_s \sim \text{Inv-Wish}(v_0 = J_s, \Lambda_0^{-1} = \text{diag}(J_s))$$

$$\gamma_{st} \sim \text{gamma}(\text{shape} = \tfrac{5}{2},\ \text{scale} = \tfrac{2}{5}),$$

The explanation of the model is as follows.

1. The first level of the model is a standard tobit regression, where we model the observed rainfall, Y_{stj}, as being equal to the j^{th} component of the latent rainfall vector, $\boldsymbol{W}_{st}$, if it is greater than zero, and equal to zero if the j^{th} component of the latent rainfall vector is less than or equal to zero.

2. Next, for each location s, the length-J_s vector of latent rainfall amounts, $\boldsymbol{W}_{st}$, is a multivariate t random variable centered at the spatial mean rainfall amount for that location, $\boldsymbol{Z}_{st}$, and offset by an ARC bias effect, β_s^{ARC} (where X_{sj}^{ARC} is an indicator variable of whether time series j at location s is an ARC satellite product). The latent rainfall, $\boldsymbol{W}_{st}$, is a multivariate-t random variable because it is a scale mixture of multivariate normals with a mixture weight, $\frac{1}{\gamma_{st}}$, for the covariance, Σ_s, that is drawn from a gamma distribution.

3. The location-specific covariance matrices Σ_s allow the multiple time series at each location to be correlated in unique ways. The mixing parameters γ_{st} determine the widths of the tails of the multivariate-t distributions, and are modeled with a gamma prior distribution with shape and scale parameters equal to 5/2 and 2/5, respectively, which implies that the multivariate-t distribution, $\boldsymbol{W}_{st}$, has 5 degrees of freedom. (In follow-up models, we could relax this assumption and estimate from the data how heavy the tails should be; the choice of 5 degrees of freedom is based on the fits of some simple, exploratory models).

4. The spatial mean rainfall amount for day t, $\boldsymbol{Z}_t$, is modeled as a multivariate normal random variable whose mean depends on the day, t (linearly, quadratically, and peri-

odically, with periods $\boldsymbol{\omega} = \frac{1}{365}(1,2,3,4))$, and also on effects from El Nino, where the El Nino effect is an additive constant that depends on the month (allowing El Nino to have different effects across the 12 months of the year).

5. The covariance matrix of $\boldsymbol{Z}_t$, $\tau^2\boldsymbol{V}$, depends on τ, a scaling factor, one known input, the Euclidean distance between locations, and one unknown parameter, λ. The spatial correlation in rainfall between locations is modeled separately from the noise inherent in the different measurement methods at each location, which is modeled by Σ_s. The model assumes that the covariances of pairs of Z_{st}'s decay exponentially with the Euclidian distance between the pairs of locations at the unknown rate λ, which we estimate from the data using a relatively flat prior.

6. The rest of the model is straightforward. We shrink the β^Z_{ps}'s for each location toward a common mean μ_p. We also model the ARC biases, β^{ARC}_s as normal random variables with an unknown mean, μ_{ARC}, and variance τ^2_{ARC}.

4 Fitting the Model

To fit our model, we need to estimate 224 parameters, which includes 184 weather coefficients, 30 covariance matrix coefficients, one scaling factor of the covariance matrix, the cross site correlation, and 8 ARC effects.

Regression coefficients describing the shape of the seasonal process, comprise the bulk of the estimation, and have a relatively simple linear estimation. For each of 6 locations there are 23 beta parameters, and each parameter's mean and standard deviation, totaling to 184 parameters (23*6+23 means and 23 standard deviations=184).

Regarding the X_t matrix comprising the base climatology, there are 23 vectors comprised of an intercept, time, time squared, 4 series of sine waves, 4 series of cosine waves, and 12 separate monthly betas for the El Nino effects. X_t and the distribution of it's parameters are established to correspond with the beta distribution. Time and time squared, sines, cosines, and ell nino effects are all centered around at 0 and scaled to have a standard deviation of 1. (Negative 1 for the time scale is 1992 and positive 1 is 2010.) The intercept we expect to be near 5. Therefore, a $N(0, 5^2)$ for the beta distribution will capture the estimated beta, where our beta estimates generally do not exceed -1 or 1. This weather pattern, with 4 main periodic components, is easy adaptable to other parts of the world, by choosing different frequencies for the sine and cosine waves, such as seasonalities occurring once every 2 years or once every four years. The 184 beta estimates can then capture the particular trends in seasonal rainfall.

Note that we are not sampling $\alpha, whichisthedegreesoffreedomforthemultivariatetdistribution, whichwe$

The next largest parameter estimation comes with estimating our covariance matrices. For Σ, the spatial correlation across sites, we estimate 6 matrices; 5 of these matrices are two by two symmetric matrices with 3 free parameters, the two variances and then the off-diagnol. The final matrix is five by five with 15 free parameters. This totals to 30 parameters to be estimated in the Sigma covariance matrix. τ, an unrestricted parameter, augments the spatial correlation matrix, making the product a covariance matrix.

λ is one parameter estimate capturing the across site correlation.

Estimation of Σ and λ differentiate our model from other weather simulation models, as it is generally difficult differentiate between within and across site correlations. By incorporating multiple time series that vary in their measurement of rainfall, we believe that we can differentiate the variability within series and the variability between series.

Our final innovation is in capturing the arc effect, which incorporates the effect of measuring rainfall with a satellite versus a rain gauge. Here we estimate 6 parameters for the effect at each site and a mean and variance, totaling to 8 parameters.

Finally, we set our hyper parameters in the model. First, τ_{ARC} and σ_p, and τ, the variances for the mean ARC effect and rainfall climatology, and the scaling factor for the spatial correlation matrix, are modeled as half t-distributions. Gelman (Bayesian Analysis 2006) describes the half t-distribution as a weakly informative prior that achieves faster convergence in MCMC than a uniform prior. A half t is a distribution truncated at 0, and exhibits fat tails, and thus represents variance. We use a metropolis hastings (random walk) sampler to sample these parameters as there are no conjugate methods to sample the half t distribution. The means μ_{ARC} and μ the means for the ARC effect and the rainfall climatology, are normally distributed.

4.1 Sampling

We sample our parameter estimates using a series of Gibbs and Metropolis Hastings steps. In the Appendix, we derive the conditional distributions for the variables for which we have no conjugate priors, namely Z_t and [ANY OTHERS?].

Note that in all of our draws, we maintain the structure of missing observations in our real data.

5 Simulation Study

We first ran a simulation study, where we simulated data from a set of known parameters. These parameter values were chosen to approximate values that could have produced our observed data, based on EDA. The size of the simulated data set was similar to the real data in both dimensions (the number of time series and the number of days).

We find that our model performs well with simulated data, in which are able to recover know parameters–in particular, our parameter for across and within site correlation. In addition, our model performs well in cross validation exercises with historic rainfall data, in which we reproduce historic time series for missing years or missing locations, perform well.

The results are encouraging: we were able to precisely estimate the spatial correlation between locations as well as the variability of different data sources within single locations. Figure 4 shows trace plots for λ, μ_α, and τ_α. All three parameters were well-estimated, and convergence was relatively quick with a burn-in period of 3000 draws. Furthermore, we were able to accurately estimate different ARC biases, α_s for each location, as well as different amounts of variability at each location, τ_s. Last, the estimates of β were accurate.

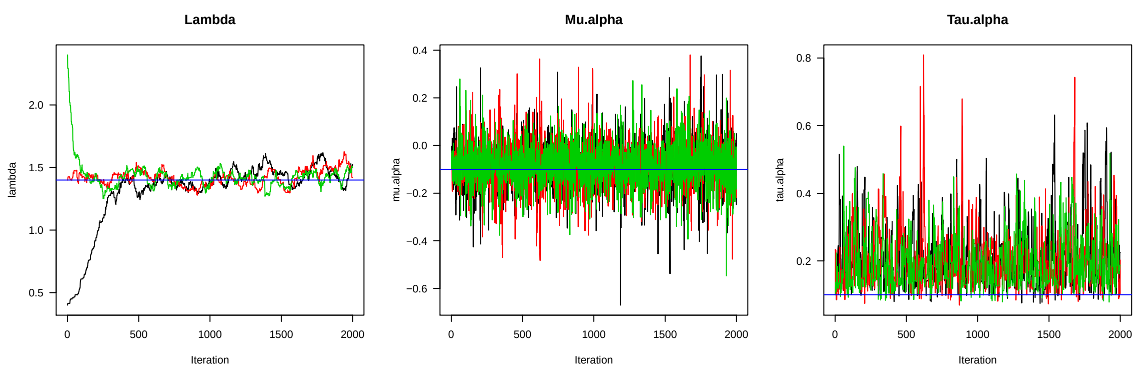

Figure 4: Simulation study results for λ, μ_α, and τ_α, where the horizontal blue lines represent the known true values of the parameters.

6 Validation

Wrap the mcmc into a function that you hand it 2 lists: experiments with dropping out 1 series and call out our mcmc function verctor of J can become 22225 to 22224 and we re-fit the model make the code pull in the dimension of the data

The way we show that we can estimate them both well is by doing clever experiments by for example removing one out of the two predicting the removed one. And then we will treat all 15 as if they are independent. So we will independently show that lambda matters and that sigma matters. That's the key.

7 Results

Our results are based on three chains after a burn-in of 5,000 iterations. We assessed convergence by running the chain for three different starting values, changing the initial seed of the random number generator, and monitoring the samples of 0. p, d. and A. The values of the Brooks, Gelman, and Rubin convergence diagnostic for each of the parameters, as well as the multivariate test, were all below 1.1, a threshold suggested by Gainerman (1997) as satisfactory, after 4,000 iterations. We noticed that our samples stabilized after few iterations (500), although a strong autocorrelation is present. We ran the Raftery and Lewis convergence test (Raftery and Lewis 1992) and obtained that for all four parameters, the suggested burn-in was less than 800, the suggested thinning was less than 20. Fewer than 50,000 iterations were needed to achieve an accuracy of Li.01, with probability .95, in the estimation of the .025 and .95 quantiles. All of these results where obtained running the set of routines BOA for R, developed by Brian Smith (http://www.pnblic-health.uiowa.edu/boa/)."

1. prob and mean rainfall historical vs. sim

2. dry spell historical vs sim

3. rainfall above a certain amount

3. metrics showing prob of rainfall in all locations?

8 Discussion

To do- 1-simulation 2-check some average rainfall measures 3-write out sampling strategy (Kenny) 4-check that the code works for conditional sampling (8-10 steps)