Hierarchical Bayes models for daily rainfall time series at multiple locations from different data sources

*

Kenneth Shirley [†] Kathryn Nadine Vasilaky [‡] Helen Greatrex [§]

Daniel Osgood [¶]

October 29, 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. Previous rainfall 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 incorporate information on the technology used (satellite versus rain gauge) in our estimations, which is also a novel addition. This methodology has far reaching applications in providing more accurate and complex weather insurance contracts based combining information from multiple data sources from a single site, a crucial improvement in the face of climate change. Secondly, the modeling extends to many other data contexts where multiple datasources exist for a given event or variable where both within and between series covariances can be estimated over time.

[*]THIS RESEARCH WAS SUPPORTED BY .....

[†]ATT Research Labs

[‡]Earth Institute and International Research Institute for Climate and Society, Columbia University, Lamont Campus 61 Route 9W, Lamont Hall, 2G (corresponding email: knv4columbia.edu)

[§]International Research Institute for Climate and Society Columbia University

[¶]International Research Institute for Climate and Society Columbia University

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

1 Motivation

The ability to simulate a realistic time-series of daily rainfall is crucial in modeling climate impacts, particularly in locations with limited weather observations. Simulated rainfall is widely used as an input into many end-user applications across agriculture, health, hydrology and ecology. Weather generators, or simulators, are often used "fill in" missing observations, examine uncertainty or to investigate situations where historical records are not sufficiently long to include extreme events. However, the current explosion of climate adaptation programs has substantially increased the practical demands placed on weather simulators, demands that often exceed the capabilities of state of the art.

One particularly salient example is the increase i weather index insurance in developing countries for which payouts are based on rainfall deficits. These projects have grown from a couple hundred farmers to tens of millions over the past decade (Greatrex et al., 2015). The price of the insurance must cover the payouts. Price is therefore is driven by the probability of a payout, which is driven by the probability of the rainfall deficits relevant to major crop losses. Naturally, insurance projects have relied heavily on rainfall models to drive insurance price calculations. Because payout formulas are closely tied to specific rainfall features relevant to the crops insured, it is important for rainfall models to accurately characterize the relevant details for infrequent, tail events. These rainfall features are often challenging for rainfall models, such as onset, inter-annual variability, consecutive dry days, climate teleconnections, and local spatial relationships. Because of these challenges, there are reports of initial applications of weather models leading to spurious results and unpredictable prices (Giannini et al., 2009). As a result, insurance companies often include substantial increases in the insurance price to protect against their lack of confidence in the rainfall models (Osgood and Shirley, 2012),.

Perhaps a greater challenge is data poverty; insurance projects must often be implemented

where there are few historical ground based rainfall observations. In some cases, networks of new gauges are installed, and in others, new satellite datasets are used. These alternative datasets often provide a great deal of additional knowledge about rainfall patterns at the site of interest. However, it is a non-trivial challenge to combine these multiple information sources, especially as satellites measure rainfall in a fundamentally different way to gauges, with different spatial statistics. In addition, in order to use these new sources of information, statistics and probabilities must be informed by any information available for the long histories necessary to characterize the extreme tail events that drive large insurance payouts. Often the historical record is captured over a region, but not by any individual data source. Typically, the data is in the form of a series of different rain gauges across the region in which only a subset of rain gauges are operational during different years or decades.

Projects such as insurance need tools that can utilize short histories of new gauges, as well as the overlapping experiences of historical gauges in the region while utilizing information from other sources, such as satellite estimates to provide the best distributions possible. They also need to be able to condition on climate processes to quantify impacts of decadal processes, climate change, and ENSO (Bell et al., 2013). This paper provides advances in methodology towards such a tool, based on a Hierarchical Bayesian approach. Our model incorporates two novel aspects to simultaneously take into account site-specific correlations across multiple data-sources, plus cross-site spatial correlations.

There have been a myriad of published methodologies on stochastic weather generation, as comprehensively described in Wilks and Wilby (1999), (Sansó and Guenni, 2000) and Verdin et al. (2015). We aim to give a brief overview to explain how our model fits in with weather generator literature. The concept of probabilistically modeling weather has existed since the early 1900s, leading to the first statistical model of rainfall in Gabriel and Neumann (1962). These statistical models were formalized as stochastic weather generators, conceived as a parametric, two step process as proposed by Richardson (1981). Daily rainfall occurrence

was modeled as a Markov chain, then other weather variables including rainfall intensity were randomly sampled from a pre-specified distributions (typically Gamma for rainfall). In recent years weather generator research has focused on four main areas: 1. Better capturing the rainfall occurrence and intensity distributions; 2. Incorporating spatial correlations to allow estimation at a point with little or no input data 3. Including multiple data sources and covariates; and 4. Increasing computational efficiency and tool development.

To better capture rainfall distributions, rainfall intensity models in the "standard" parametric approach have moved from exponential (Richardson, 1981), to gamma (Buishand, 1978; Coe and Stern, 1984; Katz, 1977; Thom, 1958; Wilks, 1992) and more recently to mixed gamma with generalized Pareto distributions (Lennartsson et al., 2008), or Weibull distributions (Furrer and Katz, 2008; Wilks, 1989). More complex Markov Chains have been suggested for rainfall occurrence (Dastidar et al., 2010; Jones and Thornton, 1993), followed by recent research utilizing probit GLMs, allowing simulated temperatures to be better conditioned on rainfall occurrence (Verdin et al., 2015). Semi-parametric and non parametric approaches have also been devised, including a mix of Markov chains and re-sampling (Apipattanavis et al., 2007), semi-empirical histograms to model wet/dry spell length and rainfall intensity (Racsko et al., 1991; Semenov, 2008) and fully non-parametric resampling (Lall and Sharma, 1996; Rajagopalan and Lall, 1999; Young, 1994), or ensemble reordering (Ghile and Schulze, 2009). In parallel to this research, new techniques have been utilized including Maximum Likelihood (Fassò and Finazzi, 2011) and Bayes Theorem (Sansó and Guenni, 2000).

Many of these techniques have allowed the incorporation of spatial information. Jones and Thornton (1993) used third order splines to interpolate their input parameters to locations with no observations. Wilks (1998) incorporated spatial correlations into a Richardson weather generator by driving it with a grid of spatially correlated random numbers, followed by Baigorria and Jones (2010) who developed this approach through the use of correlation

matrices and orthogonal Markov chains. Kleiber et al. (2012) utilized latent Gaussian processes to model spatial occurrence and rainfall amounts, an approach further extended in Kleiber et al. (2012) and Verdin et al. (2015) to include other variables. The approaches above would allow spatial correlations of statistics, but individual pixels-ensemble members would still remain independent. To overcome this issue, Greatrex (2012) proposed utilizing the geostatistical technique of sequential simulation to create stochastically generated, spatially correlated maps of rainfall.

There has been less research on the incorporation of multiple datasets. Furrer and Katz (2008) incorporated ENSO into a rainfall generator via the use of GLMs. Fassò and Finazzi (2011) used Maximum Likelihood Estimation to model heterogeneous environmental data, albeit to model air quality from satellite and ground based datasets, rather than to model rainfall. Hauser and Demirov (2013) utilized a Hidden Markov model linked with either polynomial regression or an artificial neural network to model rainfall conditioned on sea surface temperatures in the sub-polar North Atlantic. This approach was also used by Carey-Smith et al. (2014) to incorporate varying seasonality into a Richardson weather generator. In non-parametric research, Srivastav and Simonovic (2015) used a maximum entropy bootstrap to simulate daily weather data in Canada. To our knowledge, there has been no published research on the incorporation of multiple within-site realizations of rainfall, for example from gauges and satellites.

Hierarchical bayesian estimation is yet another approach equipped to model spatial as well as temporal relationships. These model and estimates the functional relationship between weather series over time. This was first proposed in Sans and Guenni (2015), then extended by Sansó and Guenni (2000) to incorporate 1) 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. Lima et al. (2015) extended

the methodology to model multi-site daily rainfall occurrence in order to investigate the distribution of the rainy season in Brazil.

The aim of this paper is to present an extension of the Hierarchical Bayes approach to incorporate two new novelties. While we do not focus on the aspect of non-stationarity aspect in rainfall 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. We focus on a case study in a data-sparse region, the Tigray region of Ethiopia.

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.

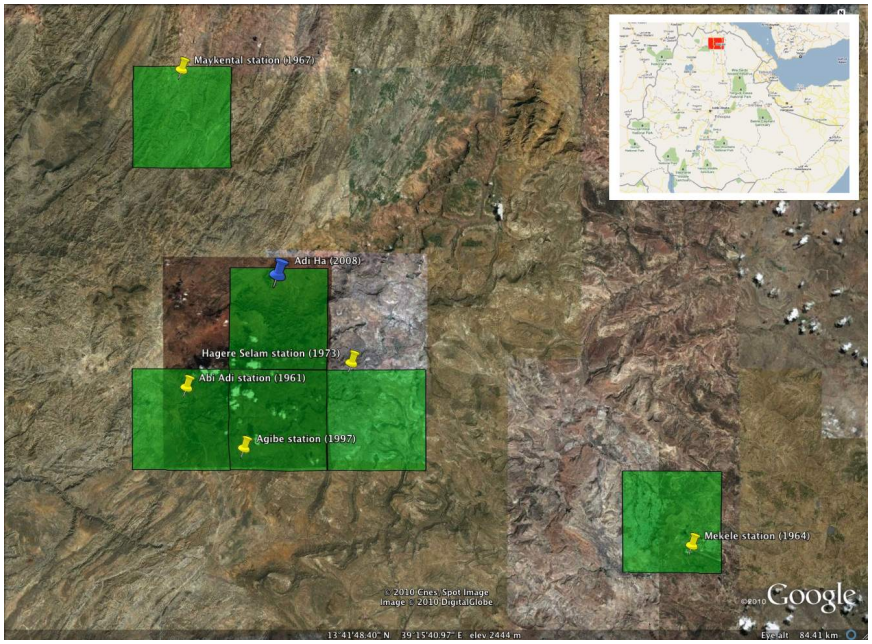

Figure 1: A map of the 6 locations, where the green squares denote ARC pixels and the pins denote rain station locations. The inset in the upper right corner shows the area of the map with a red rectangular box; this region is in north central Ethiopia.

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.

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).

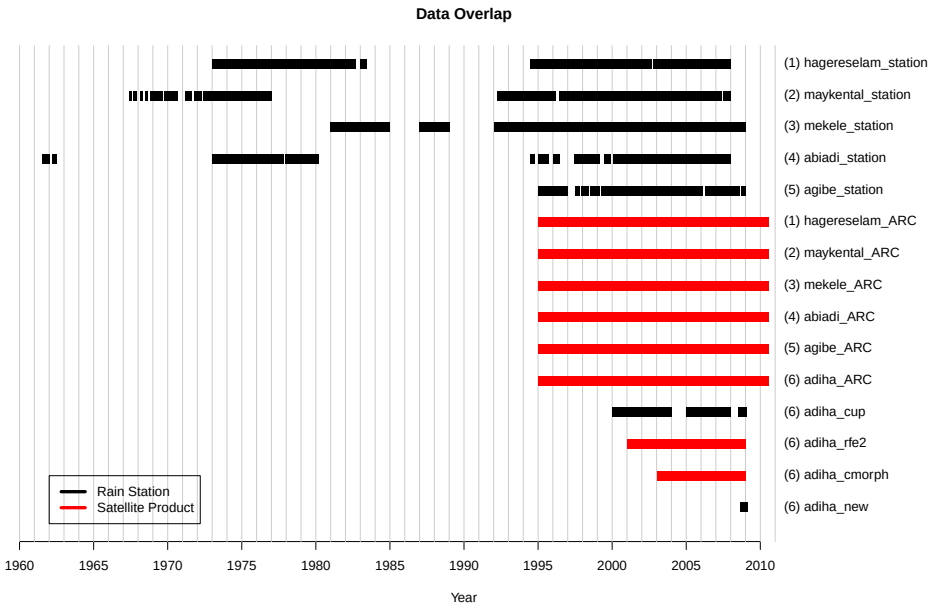

Figure 2: A visualization of the observed data for each of the 15 time series we model. The black hash marks denote rain station data, and the red hash marks denote satellite-based data.

Table 1: 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.

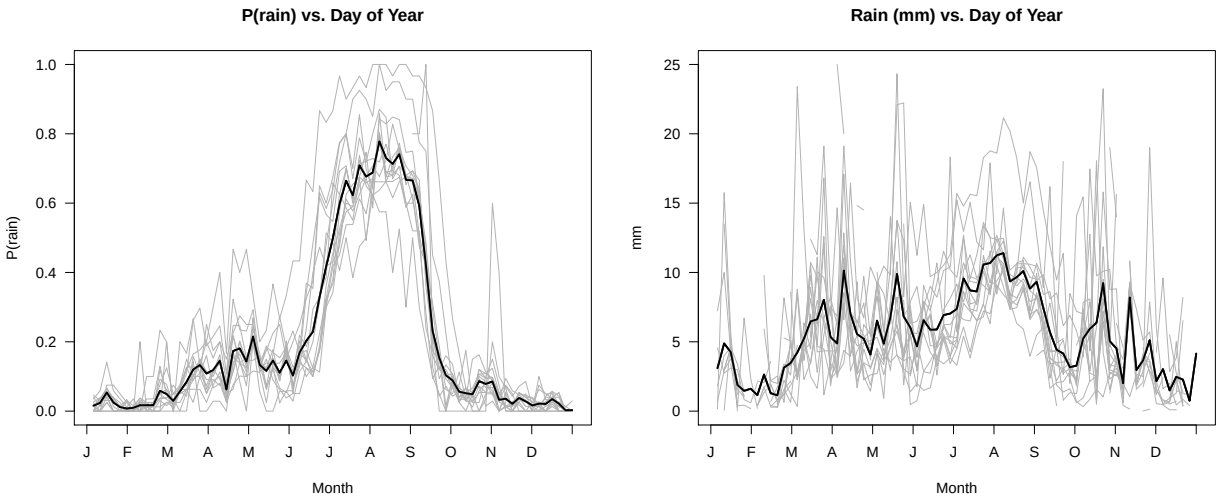

Figure 3: Plots of the percentage of rainy days (pooled into 5-day bins) and the average amount of rain as a function of the season. The gray lines are for each of the 15 individual time series, and the black lines are averaged across all 15 time series.

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 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 $\in= \{2, 2, 2, 2, 2, 5\}$ 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:

$$Y_{stj} = \begin{cases} W_{stj} & \text{if } W_{stj} > 0 \\ 0 & \text{if } W_{stj} \leq 0, \end{cases} \qquad \text{(1) Observed rainfall}$$

$$\boldsymbol{W}_{st} \sim \mathrm{N}_{J_s}(\mathbf{1} Z_{st} + \boldsymbol{X}_s^{\mathrm{ARC}} \beta_s^{\mathrm{ARC}}, \tfrac{1}{\gamma_{st}} \Sigma_s), \qquad \text{(2) Latent rainfall}$$

$$\boldsymbol{Z}_t \sim N_S(\boldsymbol{X}_t \boldsymbol{\beta}^Z, \tau^2 \boldsymbol{V}), \qquad \text{(3) Spatial mean rainfall}$$

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

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

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

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

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

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

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

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

$$\beta^{\text{ARC}}_s \sim \text{N}(\mu^{\text{ARC}}, \tau^2_{\text{ARC}}),$$

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

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

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

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

where $\mathbf{1}$ denotes the length-J_s vector of ones for location $s = 1, .., S$.

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 on day t, $\boldsymbol{W}_{st}$, is a multivariate t random variable centered at the spatial mean rainfall amount for that location and day, Z_{st}, and offset by an ARC bias effect, β^{ARC}_s (where X^{ARC}_{sj} 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 periodically, 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_{ps}^{Z}'s for each location toward a common mean μ_p. We also model the ARC biases, β_s^{ARC} 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 α, which is the degrees of freedom for the multivariate t distribution, which we set to 5 or 10, which is reasonable. That is, every day at every location we have a multivariate t with 5 degrees of freedom. 10 gives is fat tails to handle the large rainfalls.

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-diaganol. 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 describe the conditional distributions for each step.

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. We were able to precisely estimate the spatial correlation between locations as well as the variability of different data sources within single locations. Below we show trace plots for parameters that are lower in the hierarchical model and thus more difficult to estimate, but also the primary contribution of this model. This includes: λ, τ^{ARC}, and Σ. All three parameters were well-estimated, and convergence was relatively quick with a burn-in period of less than 5,000 draws. All other parameters were also well estimated and available upon request.

Results for Σ are equally as good, converging to the true parameter value before 5,000 iterations. Below we show the results for Site 1, Hagere Salaam. The remaining estimates are displayed in the Appendix.

6 Results

Our results are based on three chains of 20,000 after a burn-in of 15,000 iterations with an adaption parameter of 2,000. We assessed convergence by running the chain for three different starting values, 4 standard deviations above and below the historical means.The values of the Brooks, Gelman, and Rubin convergence diagnostic for each of the parameters, were mostly below 1.1 (lambda, tau, tau.arc, mu.arc), a threshold suggested by Gamerman and Lopes (1997) as satisfactory, except for the February Nino Effect ($\beta^{February}$). Results can be found in the Appendix.[1]

6.1 Trace Plots of Key Parameters

We ran trace plots for each parameter. Parameters for which are most interested in estimating and are typically difficult to identify λ, the distance correlation across sites, and Σ, the within site correlation matrix.

We begin by displaying λ, the parameter determining spatial correlation in our distance matrix. λ converges to an estimate of about 0.6, while τ, our parameter scaling λ in the covariance matrix of Z, converges to 8. λ exhibits more autocorrelation than in the simulated data.

Thus for two sites i and k that have an euclidean distance of .10 is approximately 10 meters, the v_{ik} entry of the V matrix will be equal to 0.5, and the v_{ik} entry of the correlation matrix will be the product of τ^2, or 8, and v_{ik}, 0.5, which is 4. That is, mean rainfall for a particular day where two sites' measurements for that day are a distance of 1,000 meters will result in a variance of 400 mm and standard deviation of 200 mm in rainfall. [THAT'S REALLY

[1] We noticed that some of our samples stabilized after under 5,000 iterations, but a strong autocorrelation is present, such as with tau.

BIG ACTUALLY. WORRISOME.]

7 Rainfall Metrics Measured

We have shown that our model performs well in terms of recovering parameters using simulated data, and it converges relatively quickly to its posterior estimates using real rainfall data. An additional motivation of this paper is to be able to predict higher level metrics such as: onset of rainfall within a season, inter-annual rainfall variability, and the number of consecutive dry days per year. In developing countries, where smallholder farmers have little savings and or capital to shield their risk, accuracy in predicting these measures can be the difference between a successful and devastating harvest. For example, yields can be heavily affected by planting a week too early or too late with respect to rainfall onset. Knowing consecutive dry days helps in preparing for a dry season, and inter-annual variability gives us a sense of the upper and lower bounds for rainfall over time.

We begin by plotting first some basic posterior estimates for mean rainfall and the proportion of wet days per month. We then follow with some of the harder to predict diagnostics: inter-annual rainfall variability, probability of consecutive dry days, probability of rainfall onset.

8 Discussion

We have considered a model where rainfall is a latent multivariate t-distribution, driven by seasonal patterns, satellite measurement error, and within site correlation between time series. The model is fitted using Monte Carlo methods with Gibbs and Metropolis Hastings sampling. Initial starting values in the three chains are set at the historical means, and then four standard deviations above and below the means. The model's accuracy is tested

on simulated data and performs well, recovering all of our predetermined parameters. The hierarchical structure allows for fitting within site correlations of various time series measured at a single location, as well across site spatial correlations. Both features enable us to incorporate information on rainfall patterns that can improve out of sample simulations of rainfall, as well as tougher to estimate rainfall characteristics: inter-annual rainfall variability, probability of consecutive dry days, probability of rainfall onset.

The ability to incorporate spatial, temporal, and measurement correlations has important policy implications. First, food insecurity is generally correlated over time and space in parallel with rainfall patterns. Thus, predictions that incorporate these spatial patterns can allow for better response to anticipated droughts and subsequent food shortages in a region. This is crucial for aid organizations and first responders to such crises. Second, different satellite measurements can detect and depict different weather patterns depending on their resolution and what they aim to measure: cloud cover, vegetation, lights, etc. Incorporating multiple information sets around a particular area can increase the accuracy with which we predict weather patterns by narrowing the variance around the mean trends. Although this research does not present a new weather generator for use by end-users, we hope that the methodological advances made will eventually lead to such a tool.

References

Apipattanavis, S., G. Podestá, B. Rajagopalan, and R. W. Katz (2007). A semiparametric multivariate and multisite weather generator. *Water Resources Research 43*(11), 19.

Baigorria, G. a. and J. W. Jones (2010). GiST: A Stochastic Model for Generating Spatially and Temporally Correlated Daily Rainfall Data. *Journal of Climate 23*(22), 5990–6008.

Bell, A. R., D. E. Osgood, B. I. Cook, K. J. Anchukaitis, G. R. McCarney, A. M. Greene, B. M. Buckley, and E. R. Cook (2013, August). Paleoclimate histories improve access and sustainability in index insurance programs. *Global Environmental Change 23*(4), 774–781.

Buishand, T. A. (1978). Some remarks on the use of daily rainfall models. *Journal of Hydrology 36*, 295–308.

Carey-Smith, T., J. Sansom, and P. Thomson (2014). A hidden seasonal switching model for multisite daily rainfall. *Water Resources Research 50*(1), 257–272.

Coe, R. and R. D. Stern (1984). Royal Statistical Society Analysis of Daily Rainfall Data. *Journal of the Royal Statistical Society. Series A (General) 147*(1).

Dastidar, A. G., D. Ghosh, S. Dasgupta, and U. K. De (2010). Higher order Markov chain models for monsoon rainfall over West Bengal, India. *Indian Journal of Radio AND Space Physics 39*(February), 39–44.

Fassò, A. and F. Finazzi (2011). Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. *Environmetrics 22*(6), 735–748.

Furrer, E. M. and R. W. Katz (2008). Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resources Research 44*(12), 1–13.

Gabriel, K. R. and J. Neumann (1962). A Markov chain model for daily rainfall occurrence at Tel Aviv. *Quarterly Journal of the Royal Meteorological Society 88*(375), 90–95.

Gamerman, D. and H. F. Lopes (1997). *Markov Chain Monte Carlo, Stochastic Simulation for Bayesian Inference.* London: Chapman and Hall.

Ghile, Y. and R. Schulze (2009). Use of an Ensemble Re-ordering Method for disaggregation of seasonal categorical rainfall forecasts into conditioned ensembles of daily rainfall for hydrological forecasting. *Journal of Hydrology 371*(1-4), 85–97.

Giannini, A., J. Hansen, E. Holthaus, A. Ines, Y. Kaheil, K. Karnauskas, M. Mclaurin, D. Osgood, A. Robertson, K. Shirley, and M. Vicarelli (2009). Desigingin Index Based Weather Insurance for Farmers in Central America. *IRI Technical Report 09-01* (March).

Greatrex, H. (2012). The application of seasonal rainfall forecasts and satellite rainfall estimates to seasonal crop yield forecasting for Africa. (May), 320.

Greatrex, H., J. Hansen, S. Garvin, R. Diro, S. Blakeley, M. L. Guen, K. Rao, and D. Osgood (2015). Scaling up index insurance for smallholder farmers : Recent evidence and insights. *CCAFS Report* (No.14), 1–32.

Hauser, T. and E. Demirov (2013). Development of a stochastic weather generator for the sub-polar North Atlantic. *Stochastic Environmental Research and Risk Assessment 27*(7), 1533–1551.

Jones, P. G. and P. K. Thornton (1993). A rainfall generator for agricultural applications in the tropics. *Agricultural and Forest Meteorology 63*(1-2), 1–19.

Katz, R. W. (1977). Precipitation as a Chain-Dependent Process.

Kleiber, W., R. W. Katz, and B. Rajagopalan (2012). Daily spatiotemporal precipitation simulation using latent and transformed Gaussian processes. *Water Resources Research 48*(January), 1–18.

Lall, U. and A. Sharma (1996). A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resources Research 32*(3), 679–693.

Lennartsson, J., A. Baxevani, and D. Chen (2008). Modelling precipitation in Sweden using multiple step markov chains and a composite model. *Journal of Hydrology 363*(1-4), 42–59.

Lima, C. H., U. Lall, T. J. Troy, and N. Devineni (2015). A climate informed model for nonstationary flood risk prediction: Application to Negro River at Manaus, Amazonia. *Journal of Hydrology 522*, 594–602.

Osgood, D. and K. E. Shirley (2012). The Value of Information in Index Insurance for Farmers in Africa. In R. Laxminarayan and M. K. Macauley (Eds.), *The Value of Information*, pp. 1–18. Dordrecht: Springer Netherlands.

Racsko, P., L. Szeidl, and M. Semenov (1991). A serial approach to local stochastic weather models. *Ecological Modelling 57*, 27–41.

Rajagopalan, B. and U. Lall (1999). A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water resources research 35*(10), 3089–3101.

Richardson, C. W. C. (1981). Stochastic simulation of daily precipitation, temperature, and solar radiation. *Water Resources Research 17*(1), 182.

Sans, B. and L. Guenni (2015). Royal Statistical Society Venezuelan rainfall data analysed by using a Bayesian space-time model. *48*(3).

Sansó, B. and L. Guenni (2000, December). A Nonstationary Multisite Model for Rainfall. *Journal of the American Statistical Association 95*(452), 1089–1100.

Semenov, M. a. (2008). Simulation of extreme weather events by a stochastic weather generator. *Climate Research 35*(3), 203–212.

Srivastav, R. K. and S. P. Simonovic (2015). Multi-site, multivariate weather generator using maximum entropy bootstrap. *Climate Dynamics 44*(11-12), 3431–3448.

Thom, H. C. S. (1958). a Note on the Gamma Distribution. *Monthly Weather Review 86*(3), 117–122.

Verdin, A., B. Rajagopalan, W. Kleiber, and R. W. Katz (2015). Coupled stochastic weather generation using spatial and generalized linear models. *Stochastic Environmental Research and Risk Assessment 29*(2), 347–356.

Wilks, D. (1992). Adapting Stochastic Weather Generation Algorithms for Cimate Chnage Studies. *Climatic Change 22*, 67–84.

Wilks, D. (1998). Multisite generalization of a daily stochastic precipitation generation model. *Journal of Hydrology 210*(1-4), 178–191.

Wilks, D. S. (1989). Rainfall Intensity, the Weibull Distribution, and Estimation of Daily Surface Runoff.

Wilks, D. S. and R. L. Wilby (1999, July). The weather generation game: a review of stochastic weather models. *Progress in Physical Geography 23*(3), 329–357.

Young, K. C. (1994). A Multivariate Chain Model for Simulating Climatic Parameters from Daily Data.

9 Appendix

The full posterior can be written as

$$p(\boldsymbol{\theta} \mid \boldsymbol{Y}) \propto \prod_{s=1}^{S}\prod_{t=1}^{T}\prod_{j=1}^{J_s} p(Y_{stj} \mid W_{stj}) p(W_{stj} \mid Z_{st}, \beta_s^{\text{ARC}}, \Sigma_s, \gamma_{st}) p(Z_{st} \mid \boldsymbol{\beta}_s^Z, \tau, \lambda)$$
$$p(\boldsymbol{\beta}_s^Z | \boldsymbol{\mu}, \boldsymbol{\sigma}) p(\beta_s^{\text{ARC}} \mid \mu^{\text{ARC}}, \tau^{\text{ARC}})$$
$$p(\boldsymbol{\mu}) p(\boldsymbol{\sigma}) p(\tau) p(\lambda) p(\mu^{\text{ARC}}) p(\tau^{\text{ARC}}) p(\boldsymbol{\Sigma}_s) p(\gamma_{st}),$$

where we set $\gamma_{st} = 5$ in our model fits, and $\boldsymbol{\theta}$ denotes the entire vector of parameters,

$$\boldsymbol{\theta} = \{\boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\beta}^{\text{ARC}}, \boldsymbol{\Sigma}, \boldsymbol{\gamma}, \boldsymbol{\beta}^Z, \tau, \lambda, \boldsymbol{\mu}, \boldsymbol{\sigma}, \mu^{\text{ARC}}, \tau^{\text{ARC}}\}.$$

To sample from this posterior, we use MCMC. First we set initial values of λ, τ, μ^{ARC}, τ^{ARC}, μ_p and σ_p (for $p = 1, ..., 23$), $\boldsymbol{\Sigma}_s$ and β_s^{ARC} (for $s = 1, ..., 6$), and β_{sp}^Z for $s = 1, ..., 6$ and $p = 1, ..., 23$. All initial values are set to be overdispersed around a rough estimate of the parameter based on fits of smaller models to the data.

Then, for iteration $g = 1, 2, ..., G$, where we set $G = 5000$, we perform the following steps:

1. Sample $\boldsymbol{W} \mid \boldsymbol{Y}, \boldsymbol{\gamma}, \boldsymbol{Z}, \boldsymbol{\beta}^{\text{ARC}}, \boldsymbol{\Sigma}$.

   The conditional posterior distribution of the length-J_s latent rainfall vector for location s and day t, $\boldsymbol{W}_{st}$, follows a multivariate normal distribution, and each element of this vector can be sampled from a univariate normal distribution according to the following three cases:

   (a) Case 1: If $Y_{stj} > 0$, then $W_{stj} = Y_{stj}$.

(b) Case 2: If $Y_{stj} = 0$, then (truncated normal distribution)

$$W_{stj} \sim \mathrm{N}(Z_{st} + X_s^{\mathrm{ARC}} \beta_s^{\mathrm{ARC}} ...).$$

(c) Case 3: If Y_{stj} is missing, then W_{stj} is drawn from the same normal distribution as in Case 2, except that the draw is not truncated to be less than zero (since the missing observed rainfall could have been greater than zero).

2. Sample $\boldsymbol{\gamma} \mid \boldsymbol{W}, \boldsymbol{Z}, \boldsymbol{\beta}^{\mathrm{ARC}}, \boldsymbol{\Sigma}, \alpha$ from

   $\gamma_{st} \sim \Gamma(\text{shape} = (J_s + \alpha)/2,\ \text{rate} = 2/((\boldsymbol{W}_{st} - \boldsymbol{\mu}_{st})^T \Sigma_s^{-1} (\boldsymbol{W}_{st} - \boldsymbol{\mu}_{st}) + \alpha))$,

   where $\boldsymbol{\mu}_{st} = Z_{st} + X_s^{\mathrm{ARC}} \beta_s^{\mathrm{ARC}}$, for $s = (1, ..., S)$ and $t = (1, ..., N)$.

3. Sample $\boldsymbol{Z} \mid \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{\beta}^{\mathrm{ARC}}, \boldsymbol{\Sigma}, \boldsymbol{\beta}^{Z}, \tau, \lambda$. The spatial mean rainfall vector can be sampled in closed for from a multivariate normal distribution. First, where

   $$Z_t \sim \mathrm{N}_S\left(\left((\Sigma_t^Z)^{-1} + (\sigma^2 V)^{-1} \right)^{-1} \left((\Sigma_t^Z)^{-1} \boldsymbol{\mu}_t^Z + (\sigma^2 V)^{-1} \boldsymbol{X}_t \boldsymbol{\beta}^Z \right), \left((\Sigma_t^Z)^{-1} + (\sigma^2 V)^{-1} \right)^{-1} \right).$$

4. Sample $\boldsymbol{\beta}^{\mathrm{ARC}} \mid \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{Z}, \boldsymbol{\Sigma}, \mu^{\mathrm{ARC}}, \tau^{\mathrm{ARC}}$, where

   $\beta_s^{\mathrm{ARC}} \sim ...$

5. Sample $\boldsymbol{\Sigma} \mid \boldsymbol{W}, \boldsymbol{\gamma}, \boldsymbol{Z}, \boldsymbol{\beta}^{\mathrm{ARC}}$, where

   $\Sigma_s \sim ...$

6. Sample $\boldsymbol{\beta}^{Z} \mid \boldsymbol{Z}, \lambda, \tau, \boldsymbol{\mu}, \boldsymbol{\sigma}$, where

   $\beta_{sp}^Z \sim ...$

7. Sample $\lambda \mid \boldsymbol{Z}, \boldsymbol{\beta}^Z, \tau$, where

   $\lambda \sim ...$

8. Sample $\tau \mid \boldsymbol{Z}, \boldsymbol{\beta}^Z, \lambda$, where

   $\tau \sim$ Inverse-Gamma(shape =)

9. Sample $\mu^{\mathrm{ARC}} \mid \beta^{\mathrm{ARC}}, \tau^{\mathrm{ARC}}$, where

   $\mu^{\mathrm{ARC}} \sim ...$

10. Sample $\tau^{\text{ARC}} \mid \beta^{\text{ARC}}, \mu^{\text{ARC}}$, where

    $\tau^{\text{ARC}} \sim ...$

11. Sample $\boldsymbol{\mu} \mid \boldsymbol{\beta}^Z, \boldsymbol{\sigma}$, where

    $\mu_p \sim ...$

12. Sample $\boldsymbol{\sigma} \mid \boldsymbol{\beta}^Z, \boldsymbol{\mu}$, where

    $\sigma_p \sim ...$