Spatio-temporal Bayesian modeling of precipitation using rain gauge data from the Hubbard Brook Experimental Forest, New Hampshire, USA

Sujit K. Sahu[1], K. Shuvo Bakar[2], Jinran Zhan[1], John L. Campbell[3] and Ruth D. Yanai[4] *

Abstract

Estimating precipitation volume over space and time is essential for many reasons such as evaluating air quality, determining the risk of flood and drought, making forest management decisions, and developing strategies for municipal water supplies. It is imperative to employ sound statistical methods for modeling data from a network of sparsely located rain gauges with known confidence. This paper proposes a spatio-temporal Bayesian model for estimating precipitation volumes using observations from a network of gauges. Based on Gaussian processes, the Bayesian model is able to interpolate at a high spatial resolution at each time point. Such interpolations are used to obtain various spatio-temporally aggregated statistics, such as annual precipitation volume in a local area. Markov chain Monte Carlo based model fitting, employed here, allows estimation of uncertainty that can be used in decision making. These methods are applied to a large data set of weekly precipitation volumes collected over the years 1997-2015 at the Hubbard Brook Experimental Forest (HBEF) in New Hampshire, USA. Using the proposed methodology we estimate trends in annual precipitation volumes spatially aggregated over nine gauged watersheds in the HBEF. The proposed modeling is also used to demonstrate a method for determining how to downsize a rain gauge network.

Key Words: Hierarchical spatio-temporal model, truncated Gaussian model, Bayesian inference.

1. Introduction

Measuring total precipitation volume aggregated over space and time is important for many environmental and ecological reasons, such as evaluating effects of air pollution, assessing factors influencing stream water quality, issuing flood and drought warnings, and making land management decisions. Practical decision making requires careful modeling and validation of precipitation data observed over space and time. Modeling of spatio-temporal data, especially in the Bayesian context, is well established in the literature, see e.g. Chapter 11 of the book by Banerjee, Carlin, and Gelfand 2015. Several statistical models exist specifically for modeling precipitation. Selected relevant articles include Sansó and Guenni 1999; Sansó and Guenni 2000; Jona Lasinio, Sahu, and Mardia 2007; Sigrist, Künsch, and Stahel 2012 and Bakar 2020. The proposed models in this paper are fitted using the software package `spTimer` (Bakar and Sahu 2015) which was extended by Bakar 2020 to include modeling of precipitation. The main focus of the current paper is to develop methods for spatio-temporal aggregation of various quantities of inferential interests, such as trend, and to illustrate the methods with a data set from the Hubbard Brook Experimental Forest (HBEF) in New Hampshire, USA.

*1: Mathematical Sciences, University of Southampton, Southampton, SO17 1BJ, United Kingdom; 2: Data61, CSIRO, Canberra 2601, Australia; 3: Northern Research Station, USDA Forest Service, Durham, NH, USA, 4: Department of Sustainable Resources Management, SUNY College of Environmental Science and Forestry, Syracuse, NY, USA

The HBEF was established by the United States Forest Service in 1955 as a major center for hydrology research in the northeastern US. The 3,160-ha site is located in the White Mountains of New Hampshire and has hilly terrain, ranging from 222 to 1,015 m altitude. Long-term data are central to the research program at the HBEF and include a diversity of environmental records, such as micro-meteorology, vegetation, soils, and hydrology.

The precipitation monitoring program at the HBEF is used to track water budgets to detect change over time Campbell et al. 2011 and the impact of forest manipulations (Hornbeck, Martin, and Eagar 1997; Campbell et al. 2011). Nine weirs were built to measure stream flow at catchment outlets and a network of 25 rain gauges was established to quantify precipitation (Figure 1). Precipitation data are also used to study annual wet deposition of elements such as calcium, see e.g. Campbell et al. 2016.

The Bayesian modeling methodology that we describe in this paper is intended to be used for a variety of scientific purposes. For example, the models allow us to study long-term spatio-temporal trends in precipitation volume in spatially aggregated ecological areas of interests, such as watersheds. We also demonstrate the value of the modeling methodology in deciding which gauges could be removed for the purposes of downsizing the monitoring program as discussed by Green et al. 2018. Green et al. 2018 used deterministic methods based on inverse distance weighting for performing spatial interpolation of precipitation volumes and identified the most influential gauges for specific watersheds by omitting each gauge individually, re-calculating the spatial interpolation, and then calculating the long-term average precipitation estimate for each watershed using data from 1998 to 2012. However, the deterministic spatial interpolation methods employed did not allow estimation of the uncertainties in the predicted values. In contrast, the proposed Bayesian model-based methodology described here can estimate the uncertainties, which is useful for evaluating and comparing estimates of precipitation.

The plan of the remainder of this article is as follows. Section 2 provides an exploratory analysis of the data set used. Specifically, it discusses the choice of the modeling scale and the value of the possible covariates used in regression modeling. This section also explores temporal variation in the data. Section 3 discusses the modeling and validation methods using the `spTimer` package in the `R` language (https://cran.r-project.org/). Section 4 develops and illustrates the model-based predictive inference methods for the data set. Specifically, Sub-section 4.1 illustrates the methods for selecting gauges for possible downsizing. Sub-section 4.2 illustrates the spatial pattern in 3-year rolling average annual precipitation levels. Catchment-specific trends in annual precipitation levels are illustrated in Sub-section 4.3. Sub-section 4.4 provides a few remarks regarding the enormous computing challenge overcome in this paper. The paper ends with a brief discussion of the results in Section 5 with some suggestions for future research.

2. Exploratory data analysis

In this analysis, we use weekly precipitation data from January 1, 1997 to 31 December 2015 collected at 24 rain gauges at the HBEF, see Figure 1. The 24 rain gauges (RG) include eleven, RG1-RG11, on a south-facing slope and twelve, RG12-17, 19-21, 23-25, on a north-facing slope. Data from gauge RG18 is not included in this analysis because it was decommissioned in 1975. RG22 is located near the administrative office, which is far from the other gauges. This makes it an ideal gauge to test the spatio-temporal modeling methods employed here. Values of elevation, slope, and aspect are available as covariates for modeling precipitation volumes.

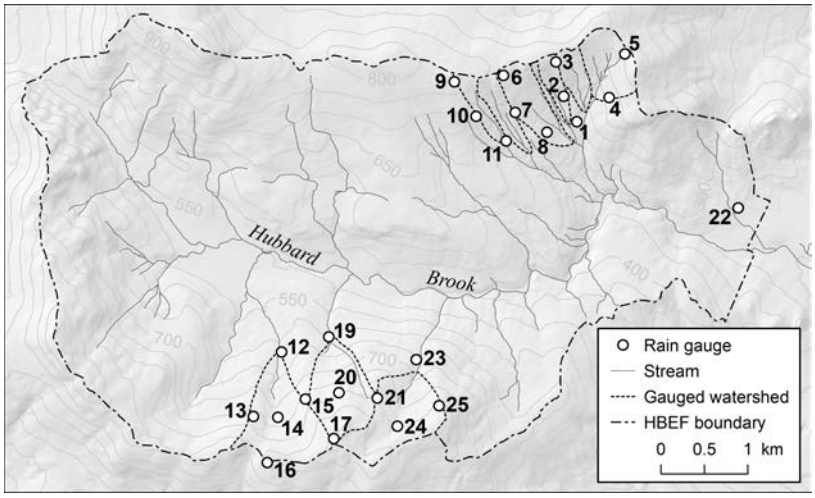


Figure 1: Map of the precipitation gauges and catchments

Over the 19 years of data in this analysis, there are $T = 991$ weeks, giving a total of $24 \times 991 = 23,784$ weekly observations from the 24 rain gauges. Table 1 provides the average weekly and annual precipitation volumes along with the standard deviations. There are no missing observations in the data set. The gauge RG22 records the lowest amount of precipitation. The highest volume is recorded in gauge RG24, see Table 1.

The weekly precipitation values range from 0 to 197 mm and the majority of the measurements are less than 120 mm, as shown in the left panel of Figure 2. The right hand panel of this figure shows that the observation distribution is more symmetric in the square-root scale. However, there is a spike in the zero value, reflecting weeks with zero precipitation. There are 1247 observations with zero precipitation, corresponding to 5.24% of the full data set. Thus, there is a positive probability of observing zero precipitation although the values are measured on a continuous scale. This causes a problem in regression modeling since a mixture distribution, accommodating the discrete values at zero and continuous values greater than zero, must be adopted to model these data. The spatio-temporal Bayesian modeling described in this article addresses this issue.

To identify the best modeling scale, we explore the mean-variance relationships of the weekly data (grouped by rain gauge and year) on the original scale (a) and on the square-root scale (b) in Figure 3. We do not use the log scale due to the presence of zero precipitation volumes during dry weeks. Panel (a) of this figure clearly indicates the presence of a mean-variance relationship on the original scale, namely, variance increases as the mean increases. However, there are no obvious patterns in Figure 3 (b) on the square-root scale. Therefore, we assume that mean and variance are independent on the square-root scale and use this transformation for modeling. These two plots also show clustering of annual means, which reflects similarity among catchments within years.

Figure 4 shows that the average square-root rainfall increases slightly as elevation increases However, RG3 and RG24 have lower and higher averages, respectively, compared to gauges at similar elevation. Figure 5 (a) plots the averages of weekly data by gauge against latitude, which shows two distinct clusters of gauges in the northern and southern parts of the map (Figure 1). Gauges at higher latitude (on south-facing slopes, at the northern boundary of the larger Hubbard Brook watershed) receive less precipitation. A similar pattern is visible in the plot of site-wise

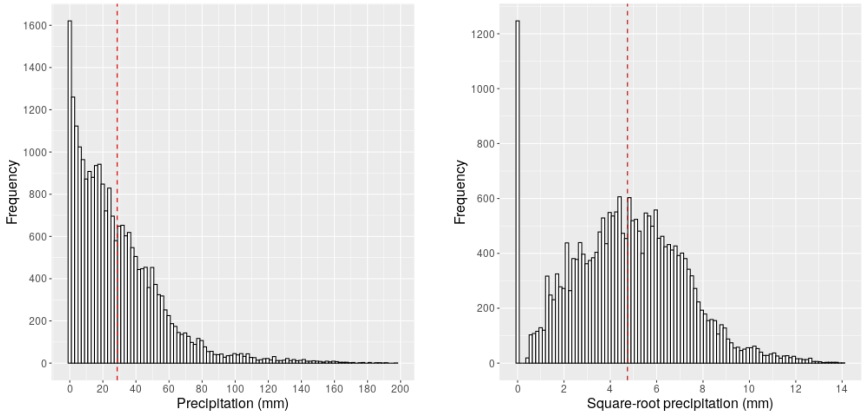


Figure 2: Histogram of the precipitation data on original scale (left) and square-root scale (right) with the mean indicated by the red dotted line.

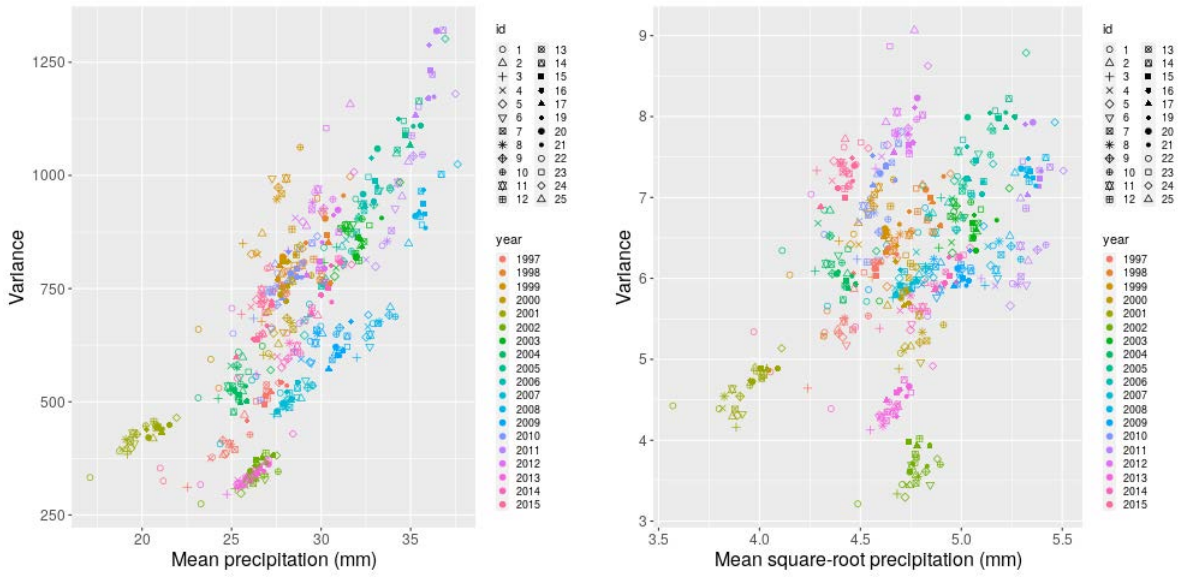


Figure 3: Weekly mean precipitation volume versus variance on original (left) and square-root scale (right). Different gauges and years are identified in the plot with different symbols and colors.

Table 1: Average weekly and annual precipitation volumes (mm) and the standard deviations at the 24 rain gauges over the years 1997 to 2015.

Id	Gauge	mean(weekly)	sd(weekly)	mean(annual)	sd(annual)
1	RG1	27.80	25.72	1450	176
2	RG2	28.43	26.30	1483	186
3	RG3	26.91	24.67	1403	165
4	RG4	27.48	25.12	1433	166
5	RG5	27.68	24.98	1444	164
6	RG6	28.39	25.76	1481	177
7	RG7	28.19	25.71	1470	171
8	RG8	28.18	25.85	1470	173
9	RG9	28.58	26.11	1491	188
10	RG10	29.23	26.57	1525	185
11	RG11	28.25	26.14	1473	175
12	RG12	29.14	26.48	1520	192
13	RG13	29.33	26.83	1530	193
14	RG14	29.72	27.38	1550	209
15	RG15	29.37	26.96	1532	197
16	RG16	29.59	26.69	1544	192
17	RG17	29.10	26.52	1518	193
19	RG19	29.29	27.45	1528	203
20	RG20	29.28	27.17	1527	195
21	RG21	29.81	27.13	1555	195
22	RG22	24.63	23.07	1285	161
23	RG23	29.61	27.62	1544	190
24	RG24	30.83	28.27	1608	207
25	RG25	28.87	26.33	1506	187

averages by gauge against longitude in panel (b) of this figure. The gauges on the south-facing slopes are located to the east of those on the north-facing slopes. However, the longitude values between the two sets of gauges are not as distinct. Hence, the north- and south-facing gauges are not much clustered in panel (b) compared to panel (a) of this figure.

To explore temporal variation, we plot the weekly data grouped by months within each of the years 1997-2015 in Figure 6. Precipitation generally peaks during the summer months with a higher level of variation than in the winter months. In particular, February is the driest and June is the wettest in most years. However, there appears to be no common monthly patterns over the years, although seasonality needs to be investigated in the modeling stage. Precipitation values are most consistent during the spring months of March and April. More extreme precipitation levels occur during the summer months of June-August and also during October and February.

There is much variation in annual precipitation volumes, see Figure 7. In this figure the plot of the 3-year rolling averages (where the rolling average for a particular year is defined as the average of that year and the previous two years) show trends in annual precipitation, where cycles of higher and lower precipitation volumes are observed. Model-based inference regarding these patterns is described below in the modeling section.

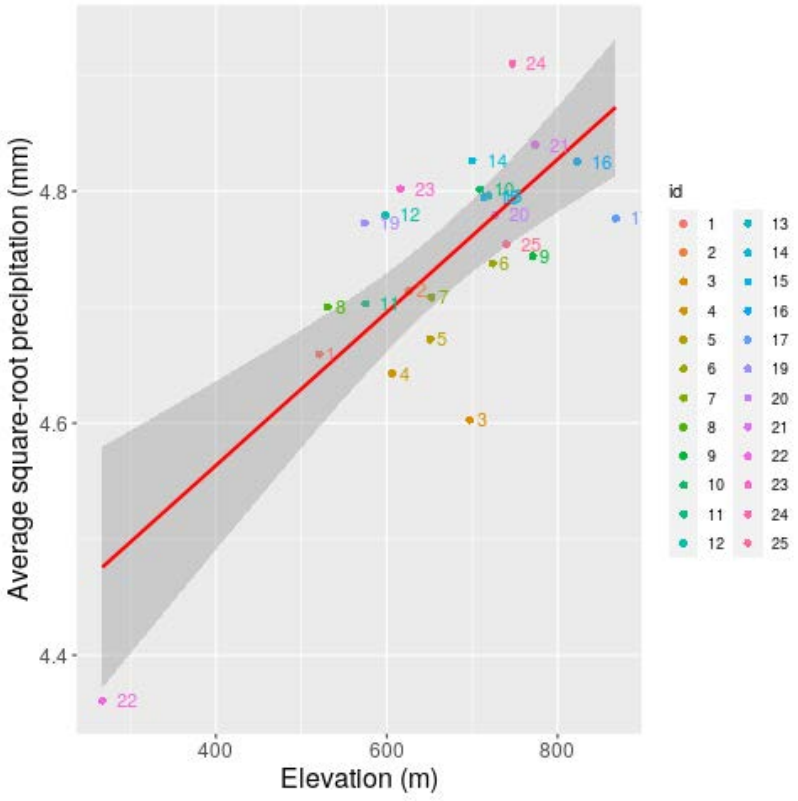
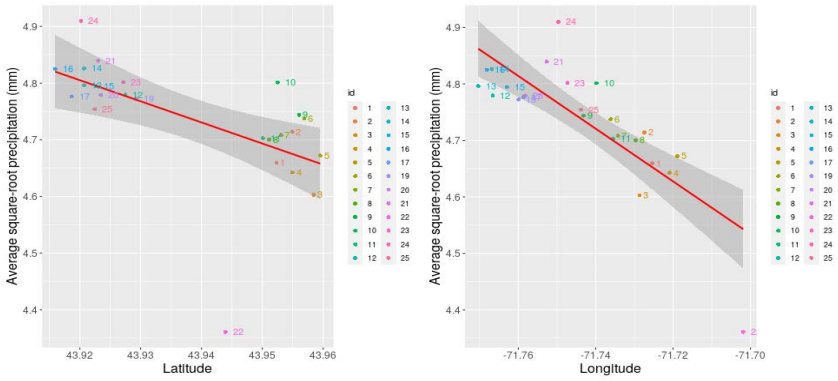


Figure 4: A plot of the square-root weekly rainfall averages at the 23 gauges against elevation. A linear regression line with a 95% confidence interval is superimposed.



(a) (b)

Figure 5: Plot of the square-root weekly rainfall averages at the 23 gauges against latitude (a) and longitude (b). A linear regression line with a 95% confidence interval is superimposed in both the plots.

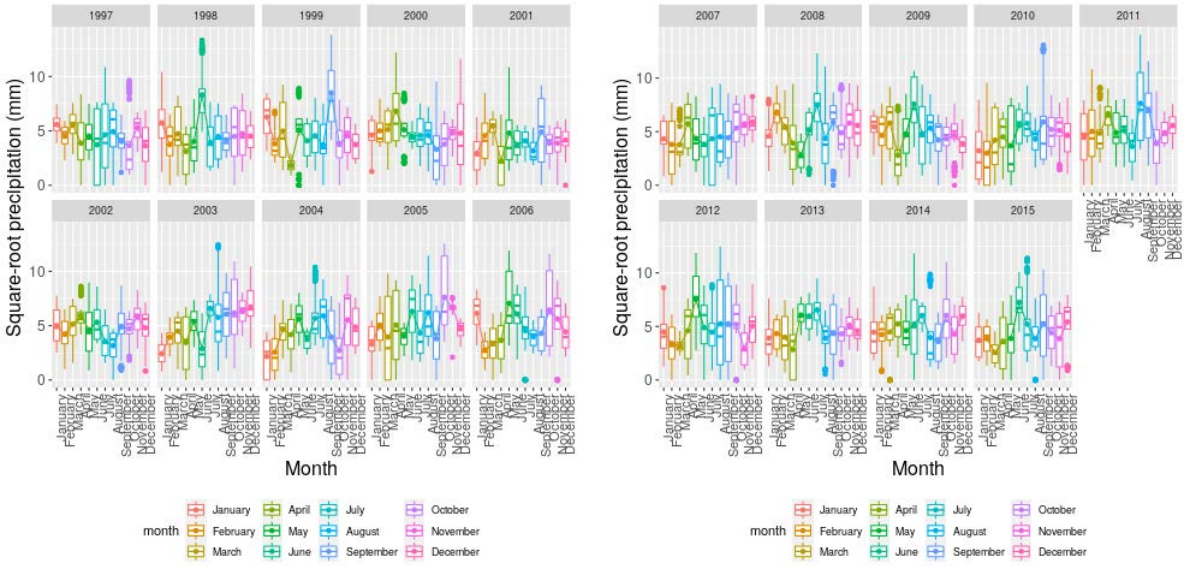


Figure 6: Boxplot of square-root precipitation against months over the years. The line running through the middle of the boxes connects the mean values.

3. Modeling methods

3.1 Spatio-temporal models

The problem of dry periods with no precipitation is a very common in Bayesian modeling of rainfall data (Sansó and Guenni 1999; Sansó and Guenni 2000). A standard approach to overcome this issue is to model the zeros by the values of a latent continuous variable below a threshold value (censoring) as has been done by Jona Lasinio, Sahu, and Mardia 2007. We now describe the general model by following the censoring method.

Let $Z(\mathbf{s}, t)$ denote the observed precipitation volume at location $\mathbf{s}$ at time t and let $g(z(\mathbf{s}, t))$ be a scaling transformation that stabilizes the variance. In the current example, we will use the square-root transformation as has been justified in the exploratory data analysis Section 2. Thus we take,

$$g(z(\mathbf{s}, t)) = \sqrt{z(\mathbf{s}, t)}.$$

To define the censoring mechanism, we introduce a latent random variable $Y(\mathbf{s}, t)$ whose observed value is $g(z(\mathbf{s}, t))$ when the observed $z(\mathbf{s}, t) > 0$ and the observation has not been rounded.

When $z(\mathbf{s}, t) = 0$, or the observation has been rounded, we assume that the corresponding $Y(\mathbf{s}, t)$ value is un-observed (i.e. latent) but we force it to lie in an appropriate interval so that the positive probability of $Y(\mathbf{s}, t)$ lying in that interval is the probability of $Z(\mathbf{s}, t)$ taking that discrete value (0 or the rounded value). This extension of the censoring mechanism allows the observed random variable to take multiple discrete values.

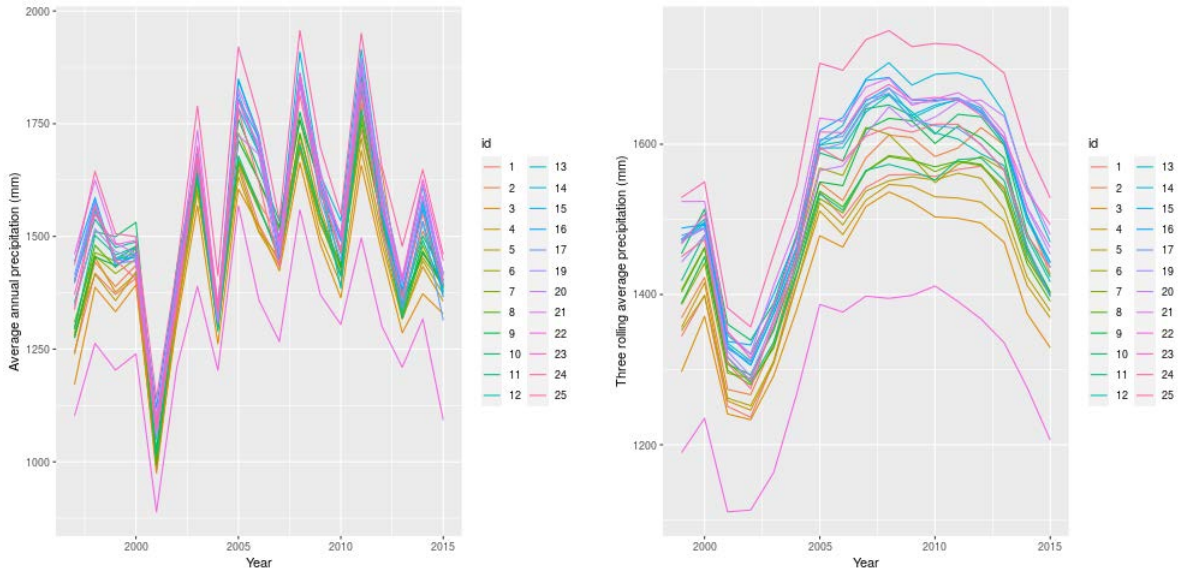


Figure 7: Mean annual total precipitation observed in the 24 gauges (left panel) and their 3-year rolling averages (right panel).

The above discussion is mathematically presented by the modeling formulation:

$$g(Z(\mathbf{s},t)) = \begin{cases} \lambda_1 & \text{if } Y(\mathbf{s},t) < c_1, \\ \lambda_2 & \text{if } c_1 \leq Y(\mathbf{s},t) < c_2, \\ \vdots & \vdots \\ \lambda_k & \text{if } c_{k-1} \leq Y(\mathbf{s},t) < c_k, \\ Y(\mathbf{s},t) & \text{otherwise.} \end{cases}$$

Here $k > 0$ is the number of discrete values that $Z(\mathbf{s},t)$ can take – for the current precipitation volume example we will assume $k = 1$. The constants $\lambda_1 < \lambda_2 < \cdots < \lambda_k$ are the transformed values of the k possible discrete values (in ascending order) that $Z(\mathbf{s},t)$ can take. The constants $c_1, \ldots, c_k$, such that $c_i > \lambda_i$ for all $i = 1, \ldots, k$, are the threshold values which can be set as the mid-point of $(\lambda_i, \lambda_{i+1})$ for $i = 1, \ldots, k-1$ and the last value c_k is chosen as the mid point of the interval with endpoints λ_k and the transformed value of the observed minimum positive precipitation value (Jona Lasinio, Sahu, and Mardia 2007).

In the current example, we have $k = 1$ and we take $c_1 = 0$ so that the $P(Z(\mathbf{s},t) = 0) = P(Y(\mathbf{s},t) < 0)$. Thus, the positive probability of having zero precipitation is the same as the transformed latent variable $Y(\mathbf{s},t)$ taking a negative value. Henceforth, we model the transformed and censored random variable $Y(\mathbf{s},t)$.

Different spatio-temporal linear models can be assumed for the random variable $Y(\mathbf{s},t)$. Here we assume that

$$Y(\mathbf{s}_i,t) = \mathbf{x}'(\mathbf{s}_i,t)\boldsymbol{\beta} + w(\mathbf{s}_i,t) + \epsilon(\mathbf{s}_i,t) \tag{1}$$

for all $i = 1, \ldots, n$ and $t = 1, \ldots, T$, where $\mathbf{x}'(\mathbf{s}_i,t)$ denotes the covariate values and components of $\boldsymbol{\beta}$ are the regression co-efficients. The residual error in model (1) is split into two parts: $w(\mathbf{s}_i,t)$

and $\epsilon(\mathbf{s}_i, t)$. We assume that the $\epsilon(\mathbf{s}_i, t)$ is the error term independent in space and time which follows the normal distribution with mean zero and variance σ^2_ϵ. The stochastic process $w(\mathbf{s}, t)$ is assumed to follow a zero mean temporally independent Gaussian process (Sahu 2012) with an exponential correlation function with unknown decay parameter ϕ. Thus, we assume that

$$\text{Cov}(w(\mathbf{s}_i, t), w(\mathbf{s}_j, t)) = \sigma^2_w \exp(-\phi d_{ij}) \tag{2}$$

where d_{ij} is the distance between the sites $\mathbf{s}_i$ and $\mathbf{s}_j$. The full Bayesian model is specified with suitable prior distributions, see e.g. Bakar and Sahu 2015 for a discussion of the prior distributions for this model.

We now discuss the specification of the regression component $\mathbf{x}'(\mathbf{s}_i, t)\boldsymbol{\beta}$ in model (1). Guided by the previously conducted exploratory data analysis and preliminary model exploration, we have chosen to keep elevation and longitude (represented by scaled UTMX values) in the model. Other site characteristics, such as latitude, slope, and aspect, did not contribute significantly to model fitting and validation after including elevation and latitude.

To capture seasonality we use the Fourier representations using pairs of sine and cosine functions as discussed in Chapter 8 of West and Harrison 1997. See also Sansó and Guenni 1999 and Jona Lasinio, Sahu, and Mardia 2007 who adopted those functions for practical spatio-temporal Bayesian modeling. Let m be the known periodicity of the data and define $K = m/2$ if m is even and $(m-1)/2$ otherwise. The seasonal term at time t and at any location $\mathbf{s}$ is assumed to be:

$$S_t(\mathbf{s}) = \sum_{r=1}^{K} c_r(\mathbf{s}) \cos\left(\frac{2\pi t r}{m}\right) + d_r(\mathbf{s}) \sin\left(\frac{2\pi t r}{m}\right), \tag{3}$$

where the unknown coefficients $c_r(\mathbf{s})$ and $d_r(\mathbf{s})$ may depend on the location $\mathbf{s}$ as well. When m is even we set $d_K(\mathbf{s}) = 0$ so that $m-1$ free seasonal parameters are kept in the model; the remaining parameter is obtained through the requirement that the seasonal effects cancel each other. An alternative to this approach is to adopt a sum-to-zero constraint. If there is no justification for having spatially varying seasonal effects, we work with the simpler model $c_r(\mathbf{s}) = c_r$ and $d_r(\mathbf{s}) = d_r$, as we do here in our example. In our example we choose $m = 365$ to adopt annual seasonal variation.

Not all the K pairs of terms in (3) necessarily contribute significantly to modeling and validation. Bayesian model selection criteria can be adopted to choose a minimum number of pairs to reduce model complexity. In our example we use $K = 1$ which corresponds to using just the first pair of sine and cosine functions. We scale these variables for ease in model fitting and to help achieve faster MCMC convergence. Thus the mean structure of the model has five parameters: an overall intercept and coefficients for scaled values of four covariates: latitude, elevation, $\sin\left(\frac{2\pi t}{m}\right)$ and $\cos\left(\frac{2\pi t}{m}\right)$ where $m = 365$ days.

3.2 Model fitting and validation

The joint posterior distribution of the parameters implied by the general model (1) and the adopted prior distribution are analytically intractable for the purposes of making posterior and predictive inference. Hence we must use suitable Bayesian computational methods for making inference. We also require a suitable software package that can fit and predict precipitation volumes using the spatio-temporal Bayesian model.

The only software packages, based on `R`, that can handle censoring and truncation are `spTimer` and `spTDyn` through the use of various truncated model options, see Bakar 2020. Below we illustrate only with the `truncatedGP` model in the `spTimer` package. The exploratory data analysis performed above dictates the use of the square root transformation as the choice for the $g(\cdot)$ function and the truncation is set at $c_1 = 0$ as mentioned above.

We fit the `truncatedGP` model in `spTimer` using the exponential covariance function. The default gamma $G(2, 1)$ prior distribution for the decay parameter ϕ leads to very slow convergence and hence we assume ϕ to be fixed and choose the optimal value of ϕ by validating 991 precipitation measurements in RG22 which is far away from the other 23 gauges. Figure 8 plots the values of root mean square error (RMSE), mean absolute error (MAE) and continuously ranked probability score (CRPS) (see e.g. Sahu, Bakar, and Awang 2015 for an explanation of these measures) for different values of ϕ in a discrete grid starting from 0.01 to 1.5. The implied range for the effective range parameter $3/\phi$ is between 2 to 300 kilometers. The three criteria do not agree on a single value of ϕ as they try to optimize on different predictive characteristics. However, it is clear that all three point to an optimal value of ϕ somewhere between 0.5 and 1. In fact, spatio-temporal interpolation is not very sensitive to a particular ϕ value in this range and in the rest of the illustration of this example we choose $\phi = 1$ which minimizes the RMSE. We note that the value of coverage of the 95% prediction intervals remains near the true value of 0.95 and does not change at all for different values of ϕ and hence coverage has been omitted from the discussion in this paragraph.

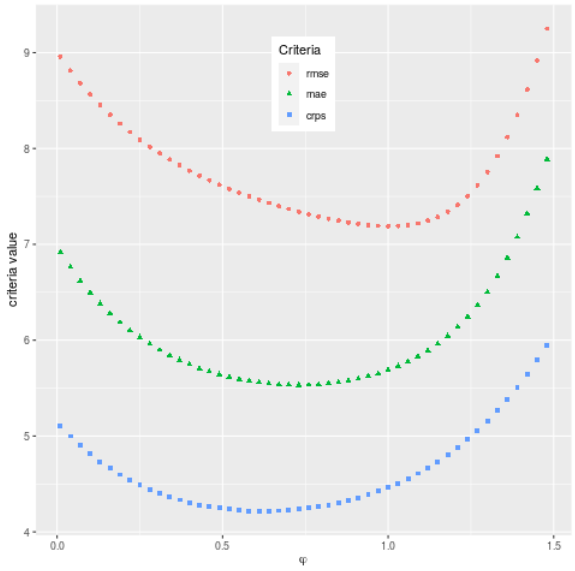


Figure 8: Plot of validation criteria for different values of ϕ for validating 991 weekly precipitation volumes.

Parameter estimates for the chosen model are presented in Table 2. The scaled variables for latitude and the seasonal terms (the sine and cosine pair) remain significant in the spatio-temporal model. Elevation is not significant in the model but we still keep it in the predictive model since it adds to the spatial variation in the predictions. The estimates of the variance components σ^2_ϵ and σ^2_w show a high level of precision because of the large numbers of the data points used to fit the model.

Table 2: Parameter estimates for the truncated GP model assuming $\phi = 1$.

	mean	sd	2.5%	97.5%
Intercept	10.68	0.01	10.66	10.70
Latitude	–0.05	0.01	–0.06	–0.03
Elevation	0.00	0.01	–0.01	0.02
$\sin\left(\frac{2\pi t}{365}\right)$	0.31	0.04	0.24	0.38
$\cos\left(\frac{2\pi}{365}\right)$	0.21	0.04	0.14	0.28
σ^2_ϵ	0.01	0.00	0.01	0.01
σ^2_w	0.51	0.00	0.50	0.52

4. Predictive inference from model fitting

Bayesian computation methods enable a vast range of predictive inference capabilities once a good model has been fitted and MCMC samples for parameters and missing data are available. In this section we illustrate Bayesian inference methods for solving the environmental decision-making problems discussed in the Introduction.

4.1 Selecting gauges for possible downsizing

One of the goals of our analysis was to select rain gauges for possible decommissioning as a cost-saving measure. Monitoring agencies are interested in reducing the number of gauges without losing much predictive accuracy. This reduction would require selection of gauges with precipitation volumes that can be predicted by the remaining gauges with the lowest possible error rates. We estimated the leave-one-out model validation criteria by removing each of the 24 gauges in turn. Table 3 shows the adopted four model validation criteria (Sahu, Bakar, and Awang 2015) by leaving out one gauge at a time. RG15 has the lowest RMSE and MAE values and would therefore be the first gauge to decommission based on these criteria. The next gauge to remove would be RG7 which has the second lowest RMSE value. This removal process may be continued to eliminate additional gauges from the network.

Note that RG22 has the highest RMSE, MAE, and CRPS values. This rain gauge is located away from the other rain gauges clustered in and around the monitored watersheds and therefore is the most difficult gauge for spatio-temporal interpolation. Because this rain gauge is at a lower elevation and receives less precipitation, the model suggests that it should be retained. This rain gauge was not included in the earlier analysis by Green et al. (2018), which focused on the gauges used to characterize precipitation in the monitored watersheds.

The approach used by Green et al. (2018) was to omit each gauge individually and calculate the effect on the long-term average (1998 to 2012) precipitation estimate for each watershed using inverse-distance weighting for the spatial interpolation. There is some agreement between the two approaches; for example, RG3 is shown to be important both in Table 3 and in Table 2 in Green et al. (2018), because this gauge consistently collects low amounts of precipitation relative to the others. However, the approaches are not directly comparable.

Another reason that our results do not correspond with the decisions to decommission gauges reported by (Green et al. 2018) is that other practical operational factors affect these decisions, such as ease or safety of access and the history of other measurements at the same location. If indeed

such practical considerations require definite retention (or removal) of a subset of gauges then the statistical method shown here should be modified accordingly. For example, the method will simply calculate the leave-one-out cross-validation statistics only for the subset of candidate gauges which are available to be removed from the network. In addition, it is possible to define and maximize a utility function that combines the leave-one-out cross-validation statistics and some measure of desirability for all the gauges. This approach to network optimization can be easily implemented and should be of value to monitoring design.

Table 3: Leave one out cross-validation statistics for the 24 rain gauges in Hubbard Brook experimental forest during the years 1997 to 2015.

Id	Gauge	RMSE	MAE	CRPS	CVG
1	RG1	2.83	2.15	2.91	94.95
2	RG2	3.27	2.07	2.89	94.75
3	RG3	3.72	2.67	3.11	94.85
4	RG4	2.85	2.14	3.02	95.06
5	RG5	3.16	2.41	3.23	94.95
6	RG6	3.05	2.19	3.01	94.95
7	RG7	2.31	1.73	2.79	95.16
8	RG8	2.60	1.97	2.87	95.06
9	RG9	3.04	2.27	3.14	95.06
10	RG10	3.05	2.10	3.08	95.06
11	RG11	2.78	2.06	2.96	94.75
12	RG12	5.33	2.49	3.27	94.55
13	RG13	2.77	2.02	3.08	94.55
14	RG14	2.81	2.07	2.99	94.65
15	RG15	2.24	1.70	2.82	94.65
16	RG16	3.06	2.34	3.19	94.55
17	RG17	2.78	2.20	2.96	94.75
19	RG19	3.87	2.56	3.19	94.25
20	RG20	2.82	1.96	2.86	94.35
21	RG21	2.39	1.80	2.90	94.45
22	RG22	7.19	5.69	4.46	94.75
23	RG23	3.88	2.58	3.29	94.25
24	RG24	3.75	2.68	3.24	94.25
25	RG25	3.43	2.58	3.29	94.35

4.2 Spatial patterns in 3-year rolling average annual precipitation

Recall the 3-year rolling averages plotted in Figure 7. This figure only shows the rolling averages for the 24 gauges. The fitted spatio-temporal model can be used to obtain predictions of these rolling averages at any location away from the 24 precipitation gauges. This in turn enables us to investigate spatial patterns in the de-trended annual averages as described subsequently.

At a new location $\mathbf{s}_0$ we first obtain the posterior predictive draws (Bakar and Sahu 2015) $Y^{(j)}(\mathbf{s}_0, t)$ given all the observed data for $j = 1, \ldots, J$ and for all the weeks in $t = 1, \ldots, 991$.

These basic weekly predictions are averaged to obtain annual predictions for each given year. For example,

$$\bar{Y}^{(j)}_{\text{year}}(\mathbf{s}_0) = \frac{1}{52}\sum_{t=1}^{52} Y^{(j)}(\mathbf{s}_0, t),\ j = 1, \ldots, J,$$

where the summation is over all the weeks in the given year. The 3-year rolling average for a particular year, say 2005 is then calculated as:

$$\bar{\bar{Y}}^{(j)}_{2005}(\mathbf{s}_0) = \frac{1}{3}\sum_{\text{year}=2003}^{2005} Y^{(j)}_{\text{year}}(\mathbf{s}_0),\ j = 1, \ldots, J.$$

Thus, the rolling average for a particular year is the average for that year and the two previous years. These calculated annual values, $\bar{\bar{Y}}^{(j)}_{\text{year}}(\mathbf{s}_0)$, for $j = 1, \ldots, J$ are averaged to estimate the prediction and the associated uncertainty for the 3-year rolling average at any location s0.

To illustrate, Figure 9 shows a spatially interpolated map of the 3-year rolling average annual precipitation volumes for 2010. There is a clear east-west gradient in this map which was expected based on model fitting results in Table 2. The Bayesian computation methods also allow estimation of uncertainties for this map. The uncertainty map is provided in the same figure and shows higher uncertainty levels for the locations that are farther away from the rain gauges.

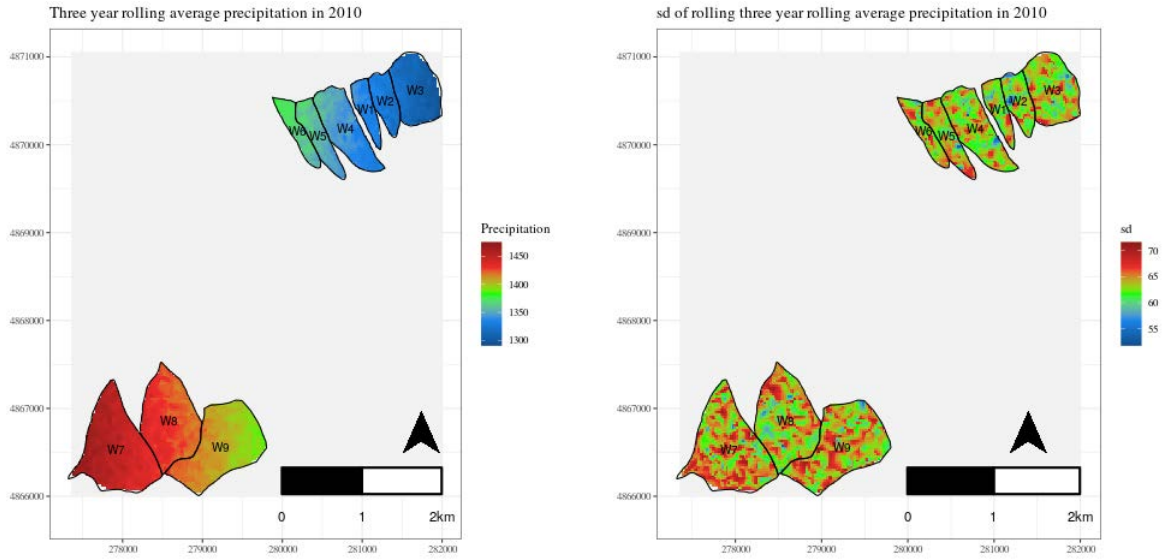


Figure 9: A predicted map of 3-year annual rolling average precipitation and their standard deviations in 2010.

Change over time in annual precipitation can be evaluated at a location $\mathbf{s}_0$ by using the quantities calculated $\bar{\bar{Y}}^{(j)}_{\text{year}}(\mathbf{s}_0)$ for any two years. For example, to estimate the annual percentage change in precipitation between 2005 and 2015 at a given location $\mathbf{s}_0$ we first evaluate

$$e^{(j)}(\mathbf{s}_0)_{2005,2015} = 100 \times \frac{\bar{\bar{Y}}^{(j)}_{2015}(\mathbf{s}_0) - \bar{\bar{Y}}^{(j)}_{2005}(\mathbf{s}_0)}{\bar{\bar{Y}}^{(j)}_{2005}(\mathbf{s}_0)},\ j = 1, \ldots, J.$$

These last MCMC iterates can be averaged to produce maps of the annual percentage change in precipitation volumes between any two years.

4.3 Catchment specific trends in annual precipitation

Catchment-specific trends in annual values can be estimated by spatially aggregating $e^{(j)}(\mathbf{s}_0)_{2005,2015}$ for any two specific years. For the jth catchment ($j = 1, \ldots, 9$) this can be calculated as:

$$h(W_j)^{(j)}_{2005,2015} = \frac{1}{n(W_j)} \sum_{i=1}^{n(W_j)} e^{(j)}(\mathbf{s}_{0i})_{2005,2015},\ j = 1, \ldots, J.$$

where the summation is over all $n(W_j)$ locations $\mathbf{s}_{0i}$, for example, within each watershed W_j, $j = 1, \ldots, 9$. Then $h(W_j)^{(j)}_{2005,2015}, j = 1, \ldots, J$ are averaged to estimate the trend and its uncertainty for the jth watershed, $j = 1, \ldots, 9$.

We complete our illustration by showing watershed specific trends and their uncertainties in Figure 10. This figure indicates a strong negative trend which seems plausible according to the 3-year rolling averages in Figure 9. Uncertainties in these watershed specific trends are plotted in the bottom panel of this figure.

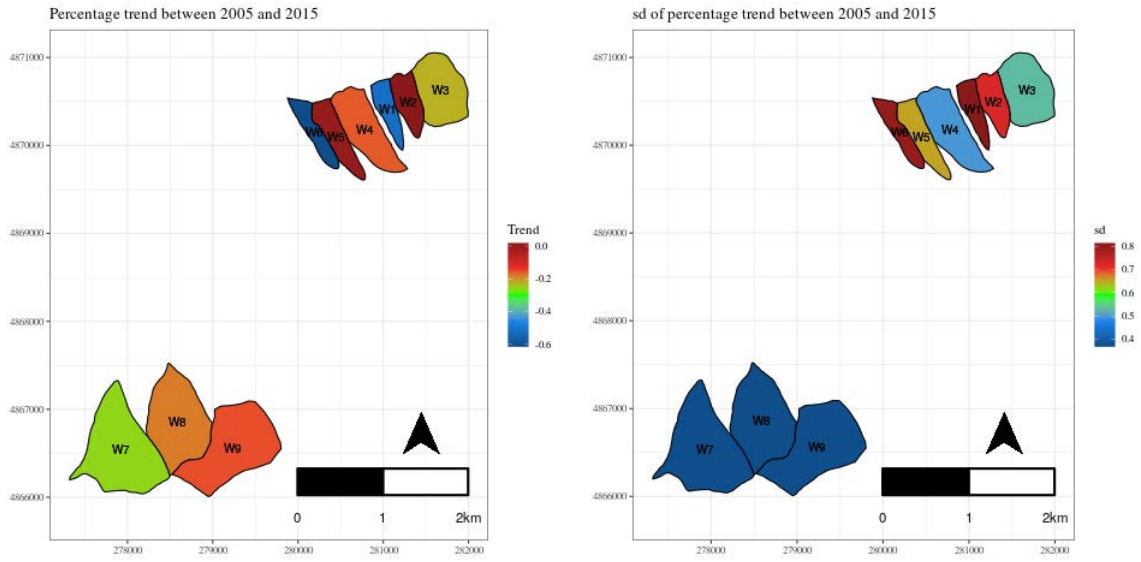


Figure 10: Trend for each watershed between 2005 and 2015 and standard deviation.

4.4 A note on computing

The large scale of the data set used in this analysis posed a challenge because of the massive computing requirements necessary to produce the results. The main issue was in generating the posterior predictive realizations for each of the 991 weekly measurements at each of the 1500 grid locations. The 200 MCMC replications implemented in this analysis resulted in a total of 297 million values.

In this example, we performed a relatively small number of MCMC iterations to demonstrate the methodology with a tractable amount of data in R without causing memory problems.

Our preliminary investigation shows that the truncated model in the spTimer package is the most viable tool currently available for solving this problem. The truncated models in the spTDyn package failed to converge, and modeling with INLA (Blangiardo and Cameletti 2015) is not possible because of the large size of the data set. The R modeling software package rstan (Stan Development Team 2020) could potentially be used; however, further coding would be required to produce acceptable results.

5. Discussion

This paper demonstrates how a large spatio-temporal precipitation data set from the Hubbard Brook Experimental Forest can be used to evaluate spatial patterns and trends in annual precipitation volume. We used a state-of-the-art Bayesian modeling and computational method to assess the uncertainty in the reported estimates.

The Bayesian methods used here pose some challenges but are not prohibitively difficult. Hence, the approach provides an example that can be used to analyze data sets of similar size and complexity. Methodological research that extends these Bayesian computational approaches based on suitable approximation techniques such as INLA (Blangiardo and Cameletti 2015) is worthy of consideration in future studies.

Additional topics worthy of future research include obtaining catchment-scale precipitation volumes at time scales of interest by summing the interpolated precipitation volumes, defined in Section 4.2, over a very fine grid. The MCMC methods can be used to obtain uncertainty estimates in any such statistics. This will require further running of the developed algorithms possibly in a parallel computing framework - a computational task not pursued in this methodology illustration paper. Exploring how these precipitation modeling methods could be used to evaluate trends in atmospheric deposition of chemical elements would also be worth investigating in future research.

References

Bakar, K. Shuvo (2020). "Interpolation of daily rainfall data using censored Bayesian spatially varying model". In: *Computational Statistics* 35, pp. 135–152. DOI: https://doi.org/10.1007/s00180-019-00911-0.

Bakar, Khandoker and Sujit Sahu (2015). "spTimer: Spatio-Temporal Bayesian Modeling Using R". In: *Journal of Statistical Software, Articles* 63.15, pp. 1–32. ISSN: 1548-7660. DOI: 10.18637/jss.v063.i15. URL: https://www.jstatsoft.org/v063/i15.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2015). *Hierarchical Modeling and Analysis for Spatial Data*. 2nd. Boca Raton: CRC Press.

Blangiardo, Marta and Michela Cameletti (2015). *Spatial and Spatio-temporal Bayesian Models with R - INLA*. Chichester: John Wiley and Sons.

Campbell, J. L. et al. (2011). "Streamflow responses to past and projected future changes in climate at the Hubbard Brook Experimental Forest, New Hampshire, United States". In: *Water Resources Research* 47. DOI: 10.1029/2010WR009438.

Campbell, John L. et al. (2016). "Uncertainty in the net hydrologic flux of calcium in a pairedwatershed harvesting study". In: *Ecosphere*. DOI: 10.1002/ecs2.1299.

Green, Mark B. et al. (2018). "Downsizing a long-term precipitation network: Using a quantitative approach to inform difficult decisions". In: *Plos One*. DOI: 10.1371/journal.pone.0195966.

Hornbeck, J. W., C. W. Martin, and C. Eagar (1997). "Summary of water yield experiments at Hubbard Brook Experimental Forest, New Hampshire". In: *Canadian Journal of Forest Research* 27, pp. 2043–2052. DOI: 10.1139/x97-173.

Jona Lasinio, G., S. K. Sahu, and K. V. Mardia (2007). "Modeling rainfall data using a Bayesian Kriged-Kalman model". In: *Bayesian Statistics and its Applocations*. Ed. by U. Singh S. K. Upadhyay and D. K. Dey. Anshan Ltd. London.

Sahu, S.K. (2012). "Hierarchical Bayesian models for space-time air pollution data". In: *Handbook of Statistics-Vol 30. Time Series Analysis: Methods and Applications*. Ed. by S Subba Rao T Subba Rao and C R Rao. Elsevier, pp. 477–495.

Sahu, Sujit K., Khandoker S. Bakar, and N. Awang (2015). "Bayesian Forecasting Using Hierarchical Spatio-temporal Models with Applications to Ozone Levels in the Eastern United States". In: *Geometry Driven Statistics*. Ed. by I. L. Dryden and J. Kent. John Wiley and Sons: Chichester, pp. 260–281.

Sansó, B. and L. Guenni (1999). "Venezuelan rainfall data analysed by using a Bayesian space-time model." In: *Journal of the Royal Statistical Society, Series C,* 48, pp. 345–362.

— (2000). "A nonstationary multisite model for rainfall." In: *Journal of the American Statistical Association* 95, pp. 1089–1100.

Sigrist, Fabio, Hans R. Künsch, and Werner A. Stahel (2012). "A dynamic nonstationary spatio-temporal model for short term prediction of precipitation". In: *Annals of Applied Statistics* 6.4, pp. 1452–1477.

Stan Development Team (2020). *RStan: the R interface to Stan*. R package version 2.21.2. URL: http://mc-stan.org/.

West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models*. 2nd. New York: Springer.