Using observed precipitation datasets to predict precipitation from climate model outputs using a Bayesian framework

Chase Dwelle, Michelle Lee, Zac Zarins

April 20, 2014

1 Introduction

The Amazon basin drains an area of approximately seven million square kilometers in South America. The majority of the basin (79.5%) is covered by the Amazon rainforest. Large swaths of these lands are uninhabited, but are still very important on an ecological and environmental level. For example, we may want to know amounts of precipitation so we can make inferences about the health of different ecosystems.

It would be cost prohibitive to install sensors in these uninhabited areas, so we would like to find a way of making predictions of rainfalls at certain locations. Much work has been done in predicting precipitation levels using various statistical models, including the recent spatial models developed by Berrocal, Raftery, and Gneiting (2008).[1] In our paper, we use a Bayesian spatial model, which would give us mean prediction values, but also probability densities for model parameters and output.

The available data were total accumulated rainfall at specific gage sites in the Amazon basin for January 2003 - reported in millimeters (mm). A plot showing these data are provided in Figure 1. This data served as the observed spatial process for a Bayesian spatial model. This paper investigates the accuracy of predictions from this Bayesian model against outputs of accumulated precipitation from published climate and reanalysis models of varying spatial resolution. We hope that this comparison can illuminate the strengths and shortcomings of using a spatial model to predict precipitation.

2 Methods

The data for the spatial model consists of January 2003 observed accumulated rainfall (mm) measured at 195 sites in the Amazon basin. January 2003 was chosen as a desirable time period because January falls within the wet season, and 2003 was an average year in weather for the Amazon basin. In addition to the observed data, we have forecasts from five different models that we test our model on. An overview of the data is given in Table 1.

In the interest of time, the spBayes package in R was used to construct the spatial model. One drawback of this package is that it only works on normal data, so we perform a square-root transform, as our original rainfall data is right-skewed. In our models, we predict the square-root of the accumulated rainfall in January 2003.

[1] Berrocal, V.J., Raftery, A.E., and Gneiting, T. *Probabilistic Quantitative Precipitation Field Forecasting Using a Two-Stage Spatial Model.* The Annals of Applied Statistics, Vol. 2, No. 4 (Dec., 2008), pp 1170-1193.

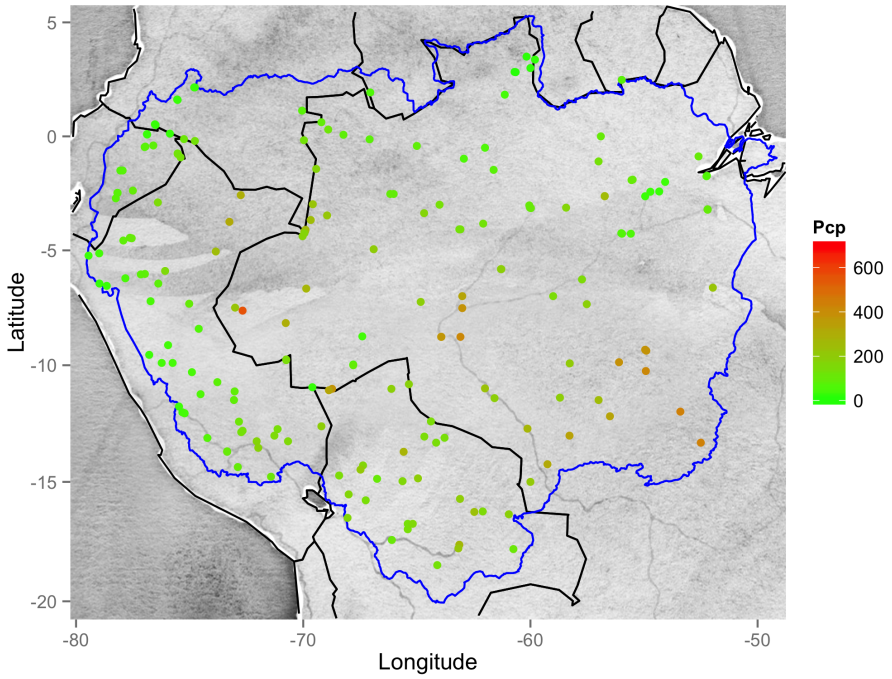


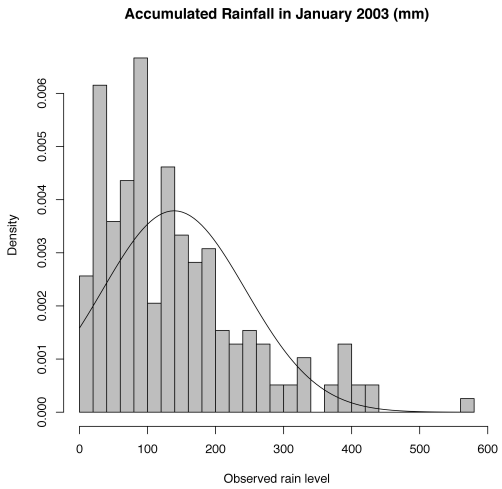
Figure 1: Total accumulated rainfall in mm (Pcp) for January 2003 in the Amazon basin (outlined in blue).

Table 1: Overview of gridded models used to compare predictions from Bayesian models.

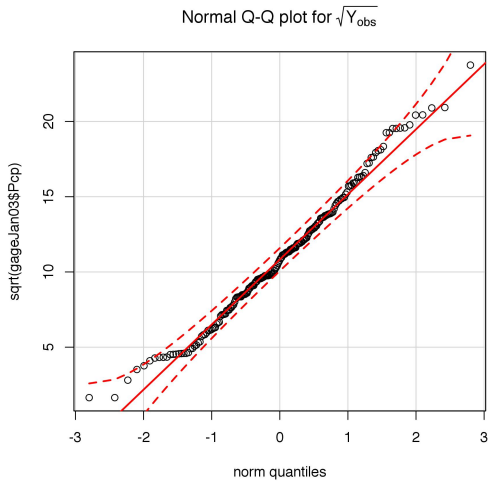
Model name	Type	Resolution
NCEP Climate Forecast System Reanalysis (CFSR)[2]	Global Reanalysis	38km
CMAP[3]	Satellite-based	2.5°
GPCP[4]	Satellite-based	1°
CMORPH[5] CRT	Satellite-based	25km
GPCC[6]	Station-dataset based	0.5°

Figure 2 shows a histogram of our observed data with a fitted normal curve, as well as a QQ plot of the square root transformed data. In addition, we transform our longitude and latitude data into 3-D coordinates.

We fit a non-spatial linear model using the X and Y coordinates as covariates: $\sqrt{R} \sim \beta_0 + \beta_1 X_i + \beta_2 Y_i$. We report our estimates from the linear model in Table 2. The Y covariate is much more significant than X, but because neither the X nor



(a) Plot of observed rainfall accumulation



(b) QQ Plot of square-root of rainfall accumulation

Figure 2: Transforming the data: (a) plots the observed rainfall data with a normal curve and (b) is a QQ plot of the square-root transformed data.

Y covariate was significant in explaining the spatial process, we posit that having extra covariates better explain the spatial process.

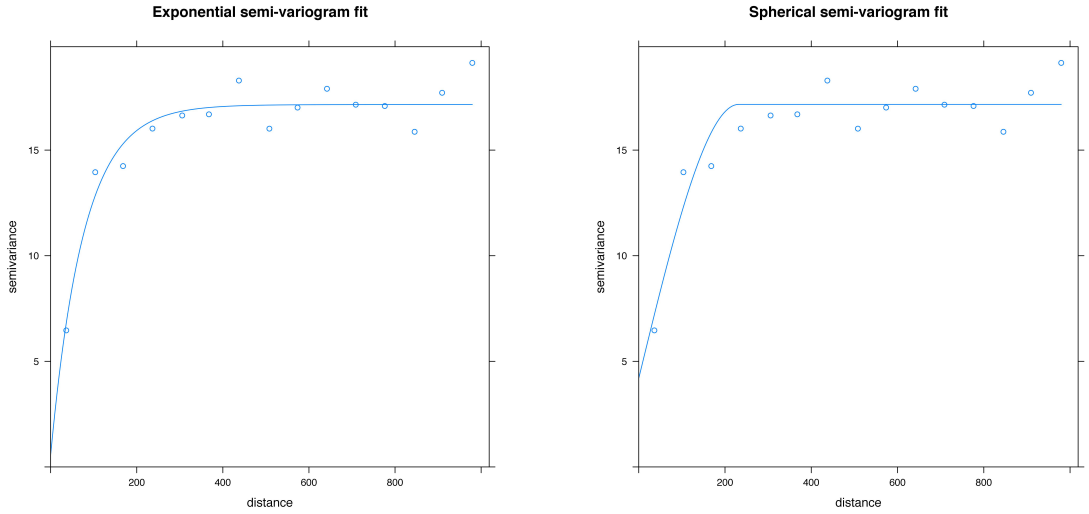
In order to use other covariates for the model, theses covariates needed to be present at every point in the domain that we wanted to make predictions. There were covariates available at the gage locations, but these weren't available at the rest of the prediction sites. Because of this, we decided to use covariates from a separate global climate model, MIROC4H, which has a 1/4-degree resolution. The physical covariates selected from this model were: air temperature [C], east and north wind velocity [m/s], evaporation [kg/m^2/month], and cloud cover [%]. Since the grids of the prediction and model sites didn't exactly overlap, the covariates were chosen from nearest grid point of the MIROC4H model.

Table 2: Parameter estimates from linear model using X and Y as covariates.

Parameter	Estimate	Std. Error	P-value
Intercept	51.029	22.996	0.028
X	-0.001	0.002	0.448
Y	0.006	0.003	0.056

We then plot the residuals from the model in a spatial semi-variogram. From our

empirical semivariogram, we can see that there is spatial dependence in the data. We fit both the exponential and spherical parametric semi-variogram models and plot them in Figure 3. Using those estimates, we fit restricted maximum likelihood (REML) models for the data, and chose the one with the lowest AIC and BIC, as shown in Table 3. We chose the exponential covariance function, despite the slightly high sill estimates, and use it in our Bayesian model with our REML estimates.



(a) Exponential Semi-variogram

(b) Spherical Semi-variogram

Figure 3: Comparing covariance functions: (a) plots the empirical semi-variogram with a exponential fit curve and (b) plots the empirical semi-variogram with a spherical fit curve and .

Table 3: REML estimates for exponential and spherical covariance functions.

Covariance Function	Nugget	Sill (Partial Sill)	Range	AIC	BIC
Exponential	2.37	30.6 (28.23)	630	981.8	1001
Spherical	2.46	22.0 (19.54)	231.7	985.4	1005

We then use the exponential covariance function to create a Bayesian hierarchical model:

$$\sqrt{R(s)} = X'(s) + \eta'(s) + \epsilon(s)$$

where $\eta'(s)$ is a Gaussian process with mean zero and covariance function, $\epsilon(s)$ are independent and identically distributed random variables with mean 0 and variance τ^2, and $\eta'(s)$ and $\epsilon(s)$ are independent. We use the priors of $IG(1, 0.015)$ for σ^2,

$IG(1, 0.01)$ for τ^2 and $Unif(1/3000, 1/150)$ for ϕ. We fit the model using an MCMC algorithm for 50,000 iterations, and use a burn-in of 25,000. Our trace plots and posterior densities eventually stabilize. After 50,000 iterations, we get a Metropolis sampling acceptance rate of approximately 70%.

3 Results

The spatial model was first validated with observed data. There were a total of 30 validation sites, selected randomly, which corresponds to 15.8% of the available observed data. A plot showing the location the the validation and model points is given in Figure 4.

A table giving summaries of the performance of the model for each prediction set is given in Table 4. It's clear that the spatial model works well for the validation set of data, with $> 95\%$ acceptance rate for the 95% predictive interval (PI) for both the spatial and physical covariate models, though there is a small improvement once the covariates are included. In this table, the coverage of the 95% PI indicates the percentage of the comparative model output fell within the 95% PI from the hierarchical model. A plot showing the coverage of the predictive interval for the GPCP data is given in Figure 5.

The spatial model does not do as well in predicting the climate models as it does the observed data. There is a 40-50% drop in the coverage rate when moving from observed to climate model data. The addition of the physical covariates improves predictions for the validation cases, lowering the error while increasing the coverage. The opposite is true for the prediction of climate model data - the introduction of the covariates decreases the size of the PI, which makes less of the supplied data fall in the PI.

Looking at Figure 5, we can see that there appears to be a clustering in the coverage of the 95% PI. The figure is for just one set of prediction locations, but the fact that there is a spatial trend holds true for the rest of the global models used, and these occur where we have validation cases. These results reinforce that predicting precipitation in large areas is difficult because of the scales involved; Figure 1 shows that there is a wide range of values for precipitation, with a minimum of 2 mm and a maximum of approximately 750 mm.

4 Conclusions

This project used observed precipitation data to attempt to predict precipitation output from climate models in the Amazon basin. The observed data were used to

Table 4: Mean Square Error (MSE), Mean Absolute Error (MAE), empirical coverage, average length of 95% predictave intervals (PI), and prediction variance for Bayesian predictions of square-root of rainfall values for different model types.

Locations predicted	Physical covariates used?	MSE	MAE	Coverage of 95% PI	Average length of 95% PI	Average predictive variance
Validation	N	56.8	1.656	96.67%	11.30	6.693
Validation	Y	48.3	1.323	100%	9.102	6.091
CFSR	N	110.5	4.917	65.52%	11.45	8.814
CMAP	N	134.4	5.419	55.12%	10.71	7.744
CMORPH	N	103.1	5.079	59.86%	11.42	8.780
GPCP	N	105.9	5.272	56.85%	10.90	8.157
GPCC	N	116.7	5.262	61.76%	11.34	8.724
CFSR	Y	110.6	4.91	58.10%	9.717	6.689
CMAP	Y	140.9	5.618	42.31%	8.729	5.429
CMORPH	Y	102.9	5.034	52.65%	9.661	6.659
GPCP	Y	105.2	5.246	46.63%	8.997	5.960
GPCC	Y	114.5	5.211	53.75%	9.579	6.610

construct two Bayesian models for precipitation, one with and one without physical covariates that could contribute to the amount of precipitation.

These models did a very good job at predicting the validation dataset for observed precipitation data, but performed poorly when attempting to predict precipitation output from climate models. We're confident in our models' ability to predict the observed rainfall, which leads to the conclusion that the climate models didn't make accurate predictions for rainfall in January 2003 in the Amazon basin.

This is understandable because this is not what the climate models were designed to do; they are designed to provide mean states over long periods of time. Also, looking at Figure 5, we can see that the clustering of errors occurs, showing that there are certain problem areas where these climate models may not be performing as well as we'd like.

Future work that could come from this project includes constructing Bayesian models that don't rely on R packages so that we could work with non-transformed data, also adding in a temporal component for the models to follow the seasonality of the precipitation process in the Amazon basin.

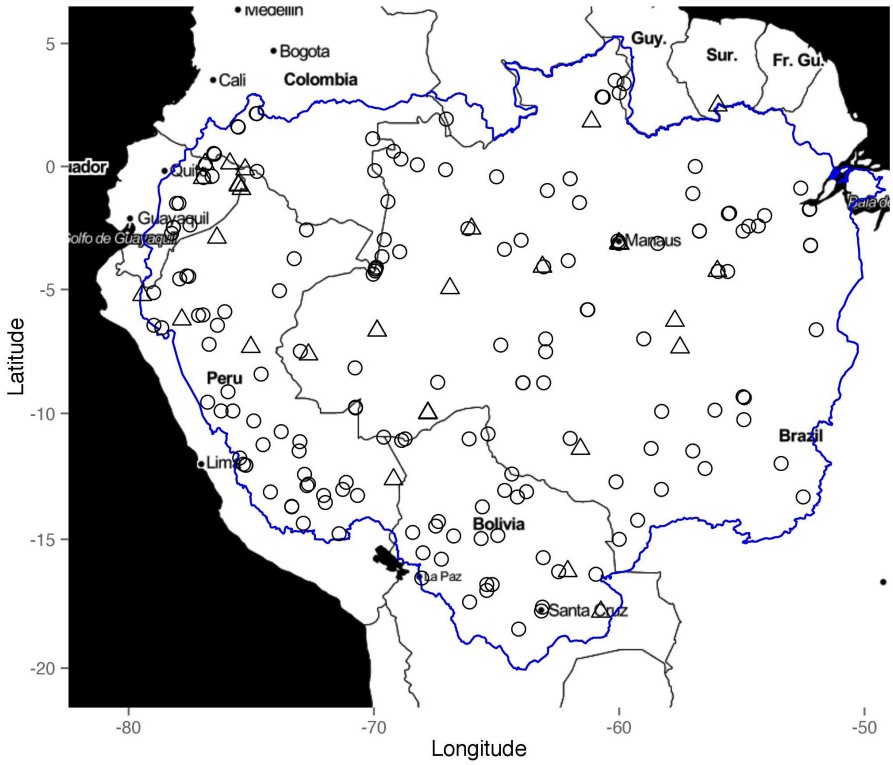
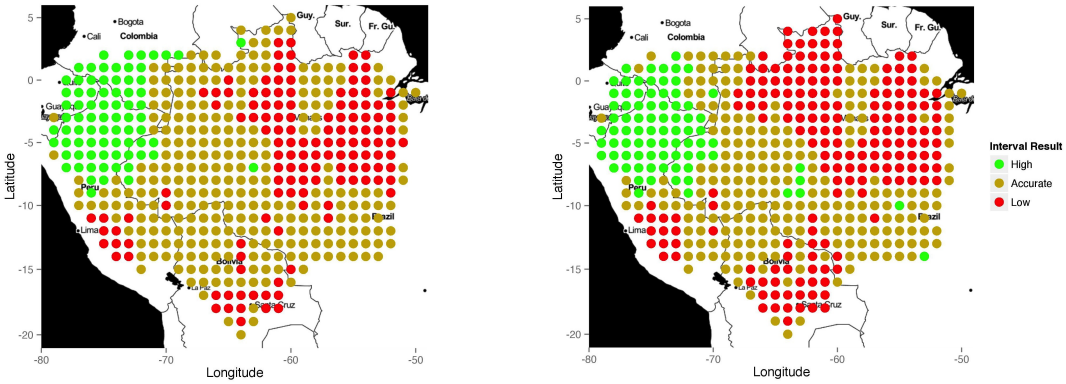


Figure 4: Plot of locations used for validation (hollow triangles) and for the spatial model (hollow circles).



(a) Purely spatial model

(b) Spatial model with covariates

Figure 5: Accuracy of the 95% predictive interval for the predictions of the GPCP data.