A Comparison of Bayesian Multivariate Versus Univariate Normal Regression Models for Prediction

Xun Li* Joyee Ghosh† Gabriele Villarini‡

Abstract

In many moderate dimensional applications we have multiple response variables that are associated with a common set of predictors. When the main objective is prediction of the response variables, a natural question is: do multivariate regression models that accommodate dependency among the response variables improve prediction compared to their univariate counterparts? Note that in this paper, by univariate versus multivariate regression models we refer to regression models with a single versus multiple response variables, respectively. We assume that under both scenarios, there are multiple covariates. Our question is motivated by an application in climate science, which involves the prediction of multiple metrics that measure the activity, intensity, severity etc. of a hurricane season. Average sea surface temperatures (SSTs) during the hurricane season have been used as predictors for each of these metrics, in separate univariate regression models, in the literature. Since the true SSTs are yet to be observed during prediction, typically their forecasts from multiple climate models are used as predictors. Some climate models have a few missing values so we develop Bayesian univariate/multivariate normal regression models, that can handle missing covariates and variable selection uncertainty. Whether Bayesian multivariate normal regression models improve prediction compared to their univariate counterparts is not clear from the existing literature, and in this work we try to fill this gap.

Keywords: Bayesian model averaging, Horseshoe priors, Linear regression, Markov chain Monte Carlo, Prediction intervals, Spike and slab priors.

1 Introduction

In this paper, one of the main goals is to address the question, whether Bayesian multivariate normal regression models yield improved predictions compared to univariate normal regression models. We empirically study the properties of Bayesian multivariate regression models, and compare them with those of univariate regression models under several variable selection priors, via extensive

*Xun Li is Modeler, Discover Financial Services.

†Joyee Ghosh is Associate Professor, Department of Statistics and Actuarial Science, The University of Iowa.

‡Gabriele Villarini is Professor, IIHR-Hydroscience & Engineering, 107C C. Maxwell Stanley Hydraulics Laboratory, The University of Iowa.

simulation studies. For moderate dimensional problems, our empirical results suggest that multivariate Bayesian methods can have a significant improvement in estimation; however, the gain in prediction is typically much less striking. The reason is that for most practical applications, the gain in estimation is typically small relative to the magnitude of the error variance, and as a result, the gain gets overshadowed by noise when considering prediction error. Thus, if the main goal is prediction, univariate modeling could be a reasonable choice in this scenario. We hope that our contribution will be useful for practitioners, when one needs to select between univariate and multivariate regression models for prediction.

The development of our methods and simulation studies was motivated by an application from climatalogy. Villarini and Vecchi (2012) and Villarini *et al.* (2019) have considered multiple quantities that measure different aspects of a hurricane season, such as the count of tropical storms, the count of hurricanes, North Atlantic Power Dissipation Index (PDI), and Accumulated Cyclone Energy (ACE). The first two quantities measure how active a hurricane season is, while the last two metrics are used to assess the frequency, duration and intensity of storms. Li *et al.* (2022), illustrated that Bayesian models have improved predictive performance compared to the hierarchical model of Villarini *et al.* (2019) for predicting the count of tropical storms. So, a natural goal is to examine the performance of Bayesian regression models for the three additional response variables.

Sea surface temperatures (SSTs) during the peak hurricane season have been shown to have reasonably good predictive performance for forecasting different quantities that measure tropical storm activity (Vecchi and Soden, 2007; Villarini *et al.*, 2010, 2019). Out of sample prediction was evaluated using different metrics such as Pearson/Spearman correlation coefficient, root mean square error (RMSE) and the mean absolute error (MAE) between the held out observations and the corresponding predicted values. In the current literature, the response variables (count of tropical storms/hurricanes, PDI, and ACE) have been modeled via separate univariate regression models with SSTs (or their forecasts) as predictors. In this application, the response variables exhibit moderate to strong pairwise correlations and some predictors have missing values; thus this motivates us to develop multivariate regression models that can accommodate dependency in response variables, missing covariates, and variable selection uncertainty.

The organization of this paper is as follows. In Section 2, we provide a brief literature review of methods for multivariate regression models, where one of the aims of multivariate modeling is to improve prediction. In Section 3, we provide a detailed description of the top level multivariate regression model for response variables. In Section 4, we conduct extensive simulation studies to compare the performance of univariate and multivariate Bayesian regression models. We design simulations to tease out the effect of modeling the dependence in the response variables, versus modeling the dependence in the prior. In Section 5, we present results from applying the methods in Section 3 to data from North Atlantic TC activity. In Section 6, we discuss the main contributions of this paper and directions for future work.

2 Review of Multivariate Regression Models for Prediction

In the frequentist literature, several authors have carried out a systematic comparison of univariate versus multivariate regression models for prediction, when the models have the same covariates, as in our application. Breiman and Friedman (1997) were one of the earliest authors who did such a comparison with their curds and whey method. The curds and whey method uses a linear combination of ordinary least squares estimates from separate regression models for each of the response variables. It is a multivariate shrinkage method based on cross validation. Breiman and Friedman (1997) concluded that multivariate methods can lead to improvement in prediction, for correlated response variables. More recently Rothman *et al.* (2010) proposed a multivariate regression with covariance estimation (MRCE) method, where they used a Lasso penalty on both the regression coefficients and the off-diagonal entries in the precision matrix (inverse of the covariance matrix), for sparse estimation in high dimensional regression models. In these papers, the authors looked at a quantity called model error in simulation studies, which is the error in estimating the mean of the predictive distribution for future covariates. Substantial gain was shown in reducing the model error with multivariate methods over their univariate counterparts. While model error is related to prediction error, it does not take into account the error variance. So, we evaluate both model error and prediction error in simulation studies, and leave one out prediction error for the application.

In the Bayesian framework, Brown *et al.* (1998, 1999) have proposed two approaches to extend variable selection for univariate regression (George and McCulloch, 1993) to a multivariate setting. Brown *et al.* (1998) assumed that a covariate is important or unimportant for all response variables. For the selected covariates, the resulting matrix of regression coefficients was assigned a matrix-normal prior. For the application, Brown *et al.* (1998) chose a generalization of Zellner's g-prior for univariate regression to multivariate normal regression. Brown *et al.* (1999) considered a joint normal distribution for the response variables and covariates, and showed improvement in prediction over a univariate regression method. However, the analysis was done for a particular dataset of interest, and the univariate regression method was based on a different non Bayesian approach. Richardson *et al.* (2010) relaxed the restriction of a covariate being included/excluded for all response variables and proposed priors to borrow information across the response variables. However, they assumed that the residual covariance matrix is diagonal. Recently Bottolo *et al.* (2021) generalized the previous model of Richardson *et al.* (2010) to allow nondiagonal covariance matrices for high dimensional model spaces and developed efficient posterior computation. However, their main focus is on high dimensional variable selection, whereas our focus in this paper is on prediction for moderate dimensional model spaces. Up to approximately 2^{25} models can be enumerated. By moderate dimensional model spaces we refer to cases where the number of models is large enough that it cannot be enumerated with standard computers, but modest enough so that MCMC algorithms are expected to converge reasonably well.

In recent years, global-local shrinkage priors (Polson and Scott, 2010) have gained popularity in univariate regression, because spike and slab priors (George and McCulloch, 1993) can be computationally very demanding for large model spaces, when one is interested in the entire posterior distribution. Global-local shrinkage priors are continuous priors that do not set a coefficient to exactly zero; instead they shrink smaller coefficients to nearly zero and keep the large coefficients almost unshrunk. Recently Bai and Ghosh (2018) and Kundu *et al.* (2021) have extended such priors to the multivariate regression scenario. Kundu *et al.* (2021) performed simulation studies to compare the effect of priors that borrow information across response variables, versus "naive" priors that do not. Their results show a small gain in prediction when using priors that borrow

information, when there is a shared pattern of covariates among the response variables. However, they assume the errors are correlated throughout the paper, thus their comparisons do not address whether it is useful to model the correlations among the response variables or not, for improvement in prediction.

Since we have a moderate dimensional model space in the Tropical Cyclone (TC) activity data, we prefer spike and slab priors that can set coefficients to exactly zero. Thus we develop models with spike and slab priors for variable selection in the top level of the model, which models the dependent response variables. In the second level, we have a sequence of regression models for the covariates, to accommodate missing covariates, following the approach of Mitra and Dunson (2010), which was also used by Li *et al.* (2022). Unlike Brown *et al.* (1998, 1999), where the covariates are assumed to be associated with all or none of the response variables, we propose a model space prior, which encourages but does not force a covariate to be included or excluded for all response variables.

3 Bayesian Multivariate Linear Regression Models with Variable Selection

Let $\boldsymbol{Y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_q)$ denote the $n \times q$ matrix containing q response variables, and let $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p)$ denote the $n \times p$ design matrix. We assume that the columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ have been standardized using the observed values to have mean 0 and standard deviation 1 for the observed part of each covariate and response variable.

Let $\boldsymbol{B}$ denote a $p \times q$ dimensional matrix of regression coefficients, and let $\boldsymbol{\Sigma}$ denote the $q \times q$ residual covariance matrix. Then the multivariate linear regression model is given as follows:

$$\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{B}, \boldsymbol{\Sigma} \sim MN_{n \times q}\left(\boldsymbol{X}\boldsymbol{B}, \boldsymbol{I}_n, \boldsymbol{\Sigma}\right), \tag{1}$$

where MN denotes the matrix-normal density in (1) and is given by

$$p(\boldsymbol{Y}) = \frac{\mid \boldsymbol{I}_n \mid^{-q/2} \mid \boldsymbol{\Sigma} \mid^{-n/2}}{(2\pi)^{nq/2}} e^{-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{I}_n^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{B})^T\right\}}.$$

We assume an inverse Wishart (IW) prior for the residual covariance matrix:

$$\boldsymbol{\Sigma} \sim IW\left(\nu, d_{\Sigma}\boldsymbol{I}_q\right), \tag{2}$$

where we set the hyperparameters at $\nu = q + 2$ and $d_{\Sigma} = 0.5$. The choice of the hyperparameter $\nu = q + 2$ offers a reasonably diffuse prior that ensures the existence of the prior mean, which is a commonly chosen value in the Bayesian literature. Here the prior mean is equal to $d_{\Sigma}\boldsymbol{I}_q$. There is less consensus regarding the choice of d_{Σ}. If the data are standardized to have variance 1, the variance of the residuals is expected to be smaller than 1. We specify $d_{\Sigma} = 0.5$, so that the prior mean denotes that the residual variance is 50% of the total variance, which seems like a reasonable choice, in the absence of any other information. Our choice can be regarded as a weakly informative prior. We did some additional simulation studies (reported in the Supplement) with $d_{\Sigma} = 1$, as in Kundu *et al.* (2021). The results concerning point estimates are very similar. The credible intervals under $d_{\Sigma} = 1$ are slightly wider and have slightly higher coverage than the nominal level. Under our hyperparameter choice $d_{\Sigma} = 0.5$, the prior probability of getting values of residual variance greater than 1 is smaller compared to the Kundu *et al.* (2021) prior. That is possibly why the results under $d_{\Sigma} = 0.5$ show a marginal improvement.

Regarding the prior for the regression coefficient matrix $\boldsymbol{B}$, we propose a spike-and-slab prior (George and McCulloch, 1993). A spike-and-slab prior is essentially a mixture of two components, of which one component corresponds to larger values of regression coefficients representing a signal and the other component corresponds to smaller values to denote a noise variable. These priors facilitate variable selection. The two components considered by George and McCulloch (1993) are normal distributions with zero means and different variances; a small variance corresponds to a noise variable and a large variance corresponds to a signal. In this paper, we consider a popular variation of this prior: a mixture of a point mass at zero for a noise variable and a normal prior for a signal. We propose a prior that promotes the inclusion/exclusion of a covariate to be similar (but not identical) across all response variables.

First, we define a q-dimensional row vector $\boldsymbol{\gamma}_j' = (\gamma_{j1}, \gamma_{j2}, \ldots, \gamma_{jq})$, for the jth row of the regression coefficients matrix $\boldsymbol{B}$, where $j = 1, \ldots, p$. If $\gamma_{jk} = 1$, the jth predictor is included in the

model for the kth response variable, and if $\gamma_{jk} = 0$, it is excluded from that model, for $k = 1, 2, \ldots, q$. Each γ_{jk} can take two possible values, 1 or 0, which leads to 2^q possible configurations for the vector $\boldsymbol{\gamma}_j'$. Let the $2^q \times q$ matrix $\boldsymbol{\Gamma}$ list all those configurations, and let $\boldsymbol{\gamma}^{\star\prime} = (\gamma_1^\star, \gamma_2^\star, \ldots, \gamma_q^\star)$ denote a specific row of the matrix $\boldsymbol{\Gamma}$.

Let $s_{\boldsymbol{\gamma}^{\star\prime}} = max\left(\sum_{k=1}^q \gamma_k^\star, (q - \sum_{k=1}^q \gamma_k^\star)\right)$, where $\sum_{k=1}^q \gamma_k^\star$ and $(q - \sum_{k=1}^q \gamma_k^\star)$ denote the number of ones and zeros in $\boldsymbol{\gamma}^{\star\prime}$. We propose $s_{\boldsymbol{\gamma}^{\star\prime}}$ as a similarity measure, which aims to capture the similarity (or lack of it) in the entries of $\boldsymbol{\gamma}^{\star\prime}$. If all the entries are ones or zeros, its value will be maximum. For example, if $q = 4$, the maximum value of $s_{\boldsymbol{\gamma}^{\star\prime}}$ will be 4, which represents the case when the predictor is either included or excluded for all response variables. For $q = 4$, the least value of $s_{\boldsymbol{\gamma}^{\star\prime}}$ is 2, when the predictor is included in the model for half of the response variables. Each $\boldsymbol{\gamma}_j'$ can be thought of being drawn randomly from the rows of the $\boldsymbol{\Gamma}$ matrix according to the following probability distribution based on the renormalized similarity measure:

$$p\left(\boldsymbol{\gamma}_j' = \boldsymbol{\gamma}^{\star\prime}\right) = \frac{s_{\boldsymbol{\gamma}^{\star\prime}}}{\sum_{\boldsymbol{\gamma}' \in \boldsymbol{\Gamma}} s_{\boldsymbol{\gamma}'}}. \tag{3}$$

We refer to this prior as the dependent spike-and-slab prior, to distinguish it from the independent spike-and-slab prior, under which the components of $\boldsymbol{\gamma}_j'$ are assumed to have independent Bernoulli distributions. The distributions of $\boldsymbol{\gamma}_j'$ are assumed to be *i.i.d.*, according to equation (3), for $j = 1, 2, \ldots, p$. There can be many variations of the similarity measure, and one can induce a dependence in other ways, such as through a Beta-Binomial prior. Conditional on $\boldsymbol{\gamma}_{jk}$, the prior for the regression coefficients is given by:

$$B_{jk} \mid \gamma_{jk} \sim (1 - \gamma_{jk})\, \delta_0 + \gamma_{jk} N\left(0, \lambda_{jk}\right), \; j = 1, \ldots, p, \; k = 1, \ldots, q, \tag{4}$$

where δ_0 denotes a degenerate distribution at zero and $N\left(.\,, .\right)$ denotes the normal distribution parameterized in terms of mean and variance, respectively. Since the predictors are standardized, we set $\lambda_{jk} = 1$.

A univariate version of the above model and prior is as follows:

$$
\begin{aligned}
\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}, \phi &\sim N_n\left(\boldsymbol{X}\boldsymbol{\beta}, I_n/\phi\right), \\
\beta_j \mid \gamma_j &\sim (1-\gamma_j)\,\delta_0 + \gamma_j N\left(0, \lambda_j\right), \ j = 1, \ldots, p, \\
\gamma_j &\overset{i.i.d.}{\sim} Bernoulli\left(\rho\right), \quad j = 1, \ldots, p, \\
\phi &\sim Gamma\left(\frac{\nu - q + 1}{2}, \frac{d_\Sigma}{2}\right),
\end{aligned}
\tag{5}
$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the $n \times 1$ vector of response variable, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the p-dimensional vector of regression coefficients, ϕ is the residual precision, ρ is the prior inclusion probability of a covariate, $N_n\left(.\,,.\right)$ is the n-dimensional multivariate normal distribution, and $Gamma\left(.\,,.\right)$ denotes the Gamma distribution parameterized in terms of shape and rate parameters, respectively. Since we have a moderate dimensional model space, we set ρ at 0.5, which corresponds to the discrete uniform prior on the model space.

We use the software JAGS for posterior computation for this model. An MCMC sampling algorithm is run to sample from the posterior predictive distribution approximately. The sample median of the posterior predictive distribution is used for point estimation and sample quantiles are used for interval estimation.

4 Simulation Study

We refer to the model and prior developed in the previous section as the dependent-data-dependent-prior (dep-data-dep-prior) spike-and-slab, where dependent data refers to the dependency among the residuals in the multivariate normal model, and dependent prior refers to the dependency in the inclusion indicators for a covariate to be associated with the q response variables. In this section, we perform an extensive simulation study to compare the performance of the multivariate Bayesian models with their univariate versions. The horseshoe prior proposed by Carvalho *et al.* (2009, 2010) has become a very attractive alternative to spike-and-slab priors, particularly for high dimensional model spaces. Unlike spike-and-slab priors, it is a continuous prior that makes it more appealing for posterior computation. It has Cauchy like tails that help to preserve large signals

and an infinitely tall spike at zero that helps to mimic the spike at zero of the spike-and-slab prior. Thus, in addition to spike-and-slab priors we also include horseshoe priors in the simulation study. We compare dep-data-dep-prior spike-and-slab with several other methods described below.

1. *ind-data-ind-prior spike-and-slab:* As a baseline, we consider this model/prior choice that fails to make use of the information about the shared pattern of important/unimportant covariates across response variables, and ignores the potential correlation between residuals. In this case, $\gamma_{jk} \overset{i.i.d.}{\sim} Bernoulli\,(\rho)$, and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$ where $1/\sigma_k^2 \overset{i.i.d.}{\sim} Gamma\,((\nu - q + 1)/2, d_\Sigma/2)$.

2. *dep-data-ind-prior spike-and-slab:* To separate the effect of accounting for the potential correlation between residuals from the effect of the dependency in the model space prior, we place independent spike-and-slab priors on the regression coefficients and an inverse Wishart prior on the residual covariance matrix $\boldsymbol{\Sigma}$. That is $\gamma_{jk} \overset{i.i.d.}{\sim} Bernoulli\,(\rho)$ and $\boldsymbol{\Sigma} \sim IW\,(\nu, d_\Sigma \boldsymbol{I}_q)$.

3. *ind-data-dep-prior spike-and-slab:* As the name suggests, here we take the similarity based dependent model space prior in (3), and $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$, with $1/\sigma_k^2 \overset{i.i.d.}{\sim} Gamma\,((\nu - q + 1)/2, d_\Sigma/2)$.

4. *dep-data-dep-prior horseshoe:* This corresponds to the MOHS prior of Kundu *et al.* (2021) where the prior on the regression coefficients are designed to shrink towards a common value across the response variables, and the model considers correlated residuals. Here $B_{jk} \sim N(0, \xi_j \tau_k)$, ${\xi_j}^{1/2} \overset{i.i.d.}{\sim} C^+(0,1)$, $\tau_k^{1/2} \overset{i.i.d.}{\sim} C^+(0,1)$, and an inverse Wishart prior is placed on the residual covariance matrix $\boldsymbol{\Sigma} \sim IW\,(\nu, d_\Sigma \boldsymbol{I}_q)$.

5. *ind-data-ind-prior horseshoe:* Similar to *ind-data-ind-prior spike-and-slab*, here we place independent horseshoe priors on the regression coefficients, conditional on τ_k. More specifically, $B_{jk} \sim N\,(0, \xi_{jk}\tau_k)$, ${\xi_{jk}}^{1/2} \overset{i.i.d.}{\sim} C^+\,(0,1)$, $\tau_k^{1/2} \overset{i.i.d.}{\sim} C^+\,(0,1)$, and $1/\sigma_k^2 \overset{i.i.d.}{\sim} Gamma\,((\nu - q + 1)/2, d_\Sigma/2)$, where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$.

6. *dep-data-ind-prior horseshoe:* Here we put independent horseshoe priors on the regression coefficients (conditional on τ_k) as above, $B_{jk} \sim N\,(0, \xi_{jk}\tau_k)$, ${\xi_{jk}}^{1/2} \overset{i.i.d.}{\sim} C^+\,(0,1)$, $\tau_k^{1/2} \overset{i.i.d.}{\sim}$

$C^+(0,1)$, but an inverse Wishart prior on the residual covariance matrix $\boldsymbol{\Sigma} \sim IW(\nu, d_\Sigma \boldsymbol{I}_q)$. This was referred to as the "Naive Horseshoe" by Kundu *et al.* (2021).

7. *ind-data-dep-prior horseshoe:* This corresponds to $B_{jk} \sim N(0, \xi_j \tau_k)$, ${\xi_j}^{1/2} \overset{i.i.d.}{\sim} C^+(0,1)$, $\tau_k^{1/2} \overset{i.i.d.}{\sim} C^+(0,1)$, and $1/\sigma_k^2 \overset{i.i.d.}{\sim} Gamma((\nu - q + 1)/2, d_\Sigma/2)$, where $\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2, \ldots, \sigma_q^2)$.

4.1 Data Generation

We generate datasets of one hundred observations. For each of the simulated datasets, we generate the covariates from a p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}_X$. We set $p = 10$, $\boldsymbol{\mu} = \boldsymbol{0}$, and $\boldsymbol{\Sigma}_{Xij} = 0.7^{|i-j|}$. The regression coefficient matrix $\boldsymbol{B}$ is chosen to be sparse. The third row of $\boldsymbol{B}$ is $(1.03, 0.00, 3.26, 3.55)$, the seventh row is $(-1.57, -2.76, 1.00, -3.79)$, and the ninth row is $(0.00, 0.00, 0.05, 0.00)$. All other rows have only zero entries, when rounded to two decimal places.

Each row of the matrix of residuals $\boldsymbol{\epsilon}$ is generated independently from a multivariate normal distribution $N_4(\boldsymbol{0}, \boldsymbol{\Sigma})$. We consider the fractional Gaussian noise covariance structure of Rothman *et al.* (2010) where the value of the Hurst index H determines the degree of dependence. We consider four values of H as follows:

1. *High correlation*: For $H = 0.95$, $\boldsymbol{\Sigma} = \begin{pmatrix} 1.00 & 0.87 & 0.80 & 0.77 \\ 0.87 & 1.00 & 0.87 & 0.80 \\ 0.80 & 0.87 & 1.00 & 0.87 \\ 0.77 & 0.80 & 0.87 & 1.00 \end{pmatrix}$.

2. *Moderate correlation*: For $H = 0.9$, $\boldsymbol{\Sigma} = \begin{pmatrix} 1.00 & 0.74 & 0.63 & 0.58 \\ 0.74 & 1.00 & 0.74 & 0.63 \\ 0.63 & 0.74 & 1.00 & 0.74 \\ 0.58 & 0.63 & 0.74 & 1.00 \end{pmatrix}$.

3. *Low correlation*: For $H = 0.8$, $\mathbf{\Sigma} = \begin{pmatrix} 1.00 & 0.52 & 0.37 & 0.31 \\ 0.52 & 1.00 & 0.52 & 0.37 \\ 0.37 & 0.52 & 1.00 & 0.52 \\ 0.31 & 0.37 & 0.52 & 1.00 \end{pmatrix}$.

4. *No correlation*: For $H = 0.5$, $\mathbf{\Sigma}$ is the Identity matrix.

Under each of the above four correlation scenarios, one hundred datasets are generated, each containing one hundred observations. For each dataset, we split the observations into two equal halves. The first fifty observations are used for estimation, and the last fifty for prediction. For all models with spike-and-slab and horseshoe priors, posterior computation is done using the software JAGS, and the MCMC sampling algorithm is run for six million iterations. The first twenty thousand samples are discarded as burn-in. The MCMC sample size is chosen so that the estimated medians of the posterior predictive distribution are (roughly) accurate up to two decimal places. The average MCMC standard errors for the estimated median of the posterior predictive distribution were calculated using the R package `mcmcse`. These values are approximately 0.001, which roughly ensures that estimates are accurate up to two decimal places, under multiple MCMC runs. The approximate running times needed to fit the model and predict for one simulated dataset for the spike-and-slab priors are 2.5, 8.3, 34, and 60.5 hours for ind-data-ind-prior, ind-data-dep-prior, dep-data-ind-prior, and dep-data-dep-prior, respectively. The corresponding times required for the horseshoe priors are 1.5, 1.5, 2.6, and 2.7 hours respectively. We also conducted some additional simulations studies with different initial values and other choices of hyperparameters. All computations were done on a cluster. The results in the additional simulation studies are similar to the ones below and are reported in the Supplement.

4.2 Results and Analysis

Under each correlation scenario, the results are averaged over the hundred datasets. The estimated posterior mean of the regression coefficient matrix, say $\hat{\boldsymbol{B}}$, is used for calculating Model Error

(ME) (Rothman *et al.*, 2010). To assess the predictive accuracy of each method, we use the median of the posterior predictive distribution for point estimation and compute the root mean squared error (RMSE). These quantities are reported in Tables 1 and 2. To show the variability, they are also plotted in Figure 1 for the high correlation scenario. They are calculated using the following formulas:

$$ME = \sum_{k=1}^{q} \left(\boldsymbol{B}_{.k} - \hat{\boldsymbol{B}}_{.k} \right)^T \boldsymbol{\Sigma}_X \left(\boldsymbol{B}_{.k} - \hat{\boldsymbol{B}}_{.k} \right), \tag{6}$$

$$RMSE = \left(\frac{1}{q \cdot n_{\text{test}}} \sum_{k=1}^{q} \left(\boldsymbol{Y}_{\text{test}.k} - \hat{\boldsymbol{Y}}_{\text{test}.k} \right)^T \left(\boldsymbol{Y}_{\text{test}.k} - \hat{\boldsymbol{Y}}_{\text{test}.k} \right) \right)^{\frac{1}{2}}, \tag{7}$$

where $\boldsymbol{B}_{.k}$ is the kth column of the regression coefficient matrix $\boldsymbol{B}$, $\hat{\boldsymbol{B}}_{.k}$ is the kth column of the estimate of coefficient matrix $\hat{\boldsymbol{B}}$, $\boldsymbol{\Sigma}_X$ is the covariance matrix of $\boldsymbol{X}$, n_{test} is the number of observations in the test set for prediction ($n_{\text{test}} = 50$), and $\boldsymbol{Y}_{\text{test}.k}$ and $\hat{\boldsymbol{Y}}_{\text{test}.k}$ denote the kth columns of the response matrix, and the matrix with corresponding medians of posterior predictive distributions, respectively. For assessing uncertainty, we also construct 90% equal-tailed intervals, for each response variable, based on the draws from the marginal posterior predictive distribution, which is based on the estimation/training data. The length and frequentist coverage, averaged over all response variables, are reported in Tables 3 and 4. In the Supplement, we report results from additional simulation studies where we evaluate 90% HPD intervals as well. They are marginally shorter than the equal tailed intervals and both have very similar coverage.

The results suggest that spike-and-slab and horseshoe priors behave similarly, within each class of priors, across the different correlation scenarios. However, spike-and-slab priors tend to be somewhat better than the corresponding horseshoe priors, in every scenario. This is likely due to the fact that horseshoe priors are more advantageous in high dimensional problems, but a regression coefficient matrix with forty entries and a sample size of fifty in our simulation study is a moderate dimensional problem, where spike-and-slab priors are still the gold standard.

So, in the following analysis of the results, we focus mainly on the spike-and-slab priors. Based on the results reported in Tables 1 and 2, ME and RMSE for dep-data-dep-prior and ind-data-dep-prior with both spike-and-slab prior and horseshoe prior are reduced in all cases, compared to

the corresponding ind-data-ind-prior cases. To assess the variability across hundred datasets, we use notched box plots (McGill *et al.*, 1978). These are like traditional box plots but have notches around the medians, which allow us to informally test whether the population medians are equal. When the notches of two box plots do not overlap, one concludes that the population medians are not equal. The notched box plots for ME and RMSE under the high correlation scenario are shown in Figure 1. For ME, the notches of the box plots for dependent data models do not overlap with those of the corresponding independent data models. This implies there can be a significant improvement in the estimation of the mean of the response variables, via modeling of the dependent data structure. However, the difference in RMSE between different model/prior combinations is not as prominent, and the notches overlap even under the high correlation scenario where the difference between dependent and independent models is the most pronounced. Based on Tables 3 and 4, dependent data methods have slightly shorter credible intervals than independent data methods, and all Bayesian methods have frequentist coverage close to 90%. To summarize, our dep-data-dep-prior structure with spike-and-slab prior, performs well in this simulation study, for reducing ME. However, the gain in predictive performance is not substantial.

Methods	$H = 0.95$	$H = 0.90$	$H = 0.80$	$H = 0.50$
ss-ind-ind	0.260	0.260	0.259	0.252
ss-dep-ind	0.124	0.151	0.204	0.275
ss-ind-dep	0.239	0.237	0.234	0.228
ss-dep-dep	0.124	0.147	0.191	0.249
hs-ind-ind	0.454	0.452	0.448	0.432
hs-dep-ind	0.202	0.269	0.364	0.450
hs-ind-dep	0.423	0.393	0.354	0.313
hs-dep-dep	0.214	0.245	0.291	0.331

Table 1: ME of all methods. Here "ss-ind-ind", "ss-dep-ind", "ss-ind-dep", "ss-dep-dep" refer to ind-data-ind-prior, dep-data-ind-prior, ind-data-dep-prior, and dep-data-dep-prior under spike-and-slab priors; similarly "hs-ind-ind", "hs-dep-ind", "hs-ind-dep", and "hs-dep-dep" refer to the methods under horseshoe priors.

Methods	$H = 0.95$	$H = 0.90$	$H = 0.80$	$H = 0.50$
ss-ind-ind	1.017	1.020	1.026	1.033
ss-dep-ind	1.002	1.009	1.020	1.035
ss-ind-dep	1.015	1.018	1.023	1.030
ss-dep-dep	1.001	1.008	1.019	1.032
hs-ind-ind	1.038	1.042	1.047	1.053
hs-dep-ind	1.010	1.021	1.037	1.055
hs-ind-dep	1.034	1.034	1.035	1.039
hs-dep-dep	1.011	1.019	1.028	1.041

Table 2: RMSE of all methods. Here "ss-ind-ind", "ss-dep-ind", "ss-ind-dep", "ss-dep-dep" refer to ind-data-ind-prior, dep-data-ind-prior, ind-data-dep-prior, and dep-data-dep-prior under spike-and-slab priors; similarly "hs-ind-ind", "hs-dep-ind", "hs-ind-dep", and "hs-dep-dep" refer to the methods under horseshoe priors.

Methods	$H = 0.95$	$H = 0.90$	$H = 0.80$	$H = 0.50$
ss-ind-ind	3.485	3.491	3.499	3.513
ss-dep-ind	3.432	3.447	3.473	3.510
ss-ind-dep	3.475	3.481	3.489	3.503
ss-dep-dep	3.430	3.445	3.469	3.502
hs-ind-ind	3.541	3.550	3.562	3.584
hs-dep-ind	3.476	3.500	3.533	3.577
hs-ind-dep	3.472	3.490	3.514	3.543
hs-dep-dep	3.477	3.489	3.510	3.539

Table 3: Length of 90% equal-tailed intervals for all methods. Here "ss-ind-ind", "ss-dep-ind", "ss-ind-dep", "ss-dep-dep" refer to ind-data-ind-prior, dep-data-ind-prior, ind-data-dep-prior, and dep-data-dep-prior under spike-and-slab priors; similarly "hs-ind-ind", "hs-dep-ind", "hs-ind-dep", and "hs-dep-dep" refer to the methods under horseshoe priors.

Methods	$H = 0.95$	$H = 0.90$	$H = 0.80$	$H = 0.50$
ss-ind-ind	0.907	0.907	0.906	0.908
ss-dep-ind	0.908	0.908	0.906	0.907
ss-ind-dep	0.908	0.908	0.907	0.908
ss-dep-dep	0.907	0.908	0.907	0.908
hs-ind-ind	0.904	0.905	0.906	0.908
hs-dep-ind	0.910	0.907	0.906	0.907
hs-ind-dep	0.900	0.902	0.904	0.910
hs-dep-dep	0.907	0.905	0.905	0.908

Table 4: Coverage of 90% equal-tailed intervals for all methods. Here "ss-ind-ind", "ss-dep-ind", "ss-ind-dep", "ss-dep-dep" refer to ind-data-ind-prior, dep-data-ind-prior, ind-data-dep-prior, and dep-data-dep-prior under spike-and-slab priors; similarly "hs-ind-ind", "hs-dep-ind", "hs-ind-dep", and "hs-dep-dep" refer to the methods under horseshoe priors.

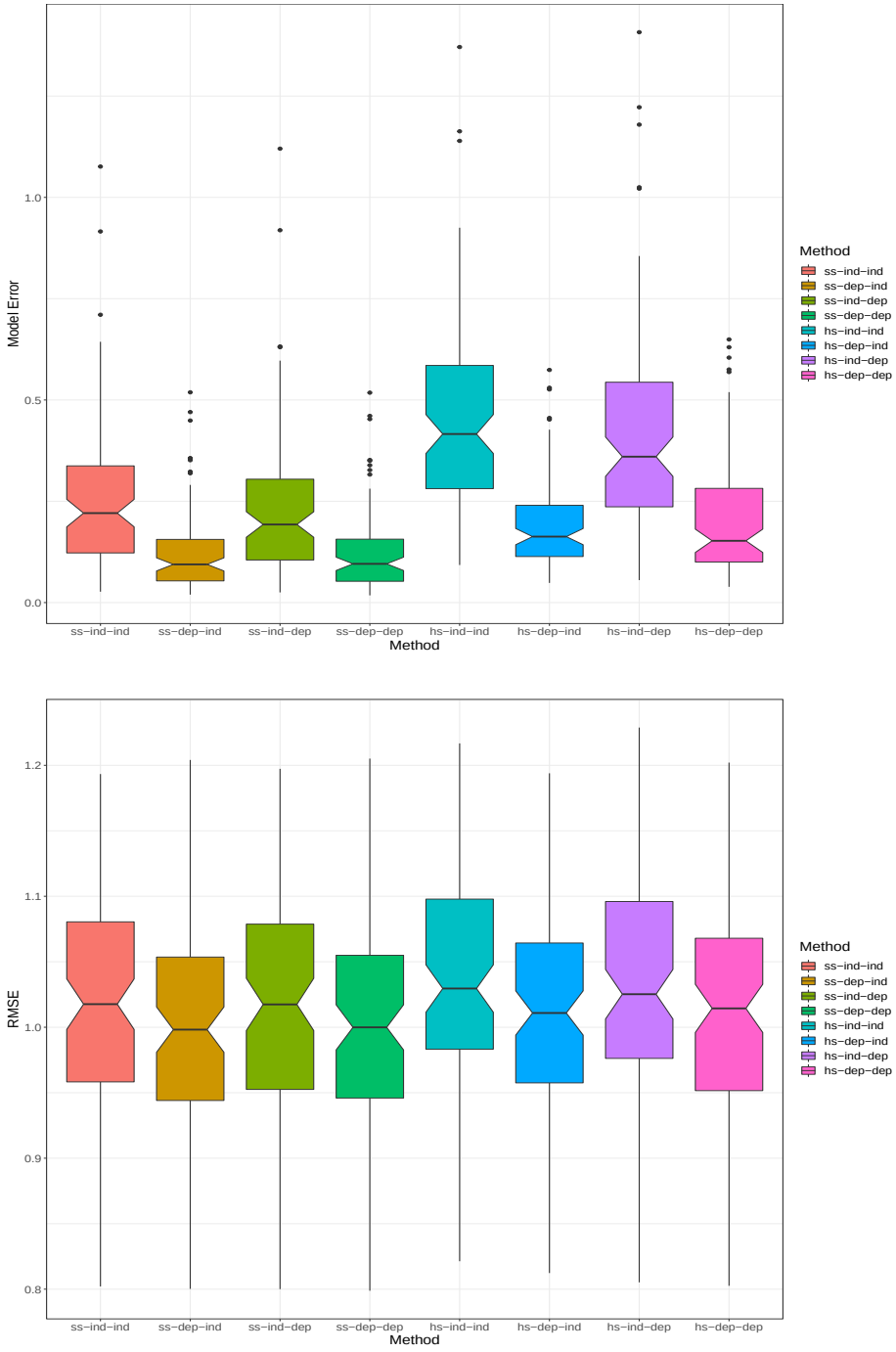

Figure 1: Box plots showing ME and RMSE of each method under the high correlation scenario when $H = 0.95$; results are shown for one hundred datasets.

5 Multivariate Normal Regression Models for the North Atlantic TC Activity Dataset

We have yearly data on the frequency of tropical storms and hurricanes, ACE, PDI, tropical Atlantic and tropical mean SSTs for the period 1958-2018. ACE and PDI are summary measures that take into account the duration, frequency and intensity of storms. They are computed using the duration and maximum sustained wind speed during the storm, from the HURDAT2 database (Villarini and Vecchi, 2012; Villarini *et al.*, 2019). The main difference between the two metrics is that ACE uses the square of the wind speed, while PDI uses the cube, in their respective formulas. We use data from 1982-2018, since the predictors in the Bayesian models are forecasts of SSTs from five climate prediction systems, which are unavailable before 1982.

We perform the analysis using the multivariate models with spike-and-slab priors described in Section 3. We use a square-root transformation for all response variables, prior to standardization. Due to the larger ME and RMSE of horseshoe priors in the simulation studies, we do not consider them here. However, for higher dimensional model spaces, we expect horseshoe priors to be an attractive alternative due to their much smaller running times. Because some of the covariates are missing, we follow the method in Mitra and Dunson (2010); Li *et al.* (2022), where a joint distribution is specified in the second level of the model as follows:

$$p\left(\boldsymbol{X}_i\right) = p\left(x_{i1}\right) \prod_{j=2}^{p} p\left(x_{ij} \mid x_{i1}, \ldots, x_{i(j-1)}\right),$$

and for $i = 1, 2, \ldots, n$,

$$\begin{aligned}
x_{i1} &\sim N\left(\theta_{10}, \frac{1}{\psi_1}\right), \\
x_{i2} \mid x_{i1} &\sim N\left(\theta_{20} + x_{i1}\theta_{21}, \frac{1}{\psi_2}\right), \\
&\vdots \\
x_{ip} \mid x_{i1}, \ldots, x_{i(p-1)} &\sim N\left(\theta_{p0} + x_{i1}\theta_{p1} + \ldots, + x_{i(p-1)}\theta_{p(p-1)}, \frac{1}{\psi_p}\right).
\end{aligned}$$

We specify the same prior distributions on the regression coefficients (θ_{jk}s) and the residual precisions (ψ_js), as Li *et al.* (2022):

$$\begin{aligned} \theta_{jk} &\sim N\left(0, \frac{\lambda_{jk}}{\psi_j}\right), \quad j = 1, 2, \cdots, p, \quad k = 0, 1, \cdots, j-1. \\ \psi_j &\sim Gamma\left(c, d\right), \; j = 1, 2, \cdots, p, \end{aligned} \tag{8}$$

where $\lambda_{j0} = 100$, $\lambda_{jk} = 1$ for coefficients other than the intercept, $c = 1$, and $d = \frac{1}{5}$ respectively.

All methods are run for six million iterations, with a burn-in of twenty thousand. For our models, we use Method 2 of Li *et al.* (2022), that discards covariates with missing/unobserved values in the year of prediction. This is because retaining or discarding such predictors was shown to have similar performance by Li *et al.* (2022), and Method 2 that discards those predictors is computationally less intensive. To be clear, we retain covariates that have missing values in past years, so the second level sequence of regression models for missing covariates is still relevant for estimation.

The median of the posterior predictive distribution is used as a point estimate for leave-one-out prediction during the period 2011-2018. Forecasts of SSTs are issued every month, starting from around nine months before the hurricane season. Following Li *et al.* (2022), we focus on SST forecasts issued in June, July, and August, as predictors of the model. We do an analysis for each of the three months (June-August) with its set of predictors, because the predictors are issued every month. Note that the response variables measure annual summary statistics related to the hurricane season, and those do not change across the analyses. Like Li *et al.* (2022), we assess the importance of predictors via ind-data-ind-prior spike-and-slab priors, and if a predictor has marginal posterior inclusion probability greater than 0.75 for any response variable, it is added to the model in the next month, as it can improve prediction. The predictors in the second level sequence of regression models are chosen based on exploratory data analysis so that the model assumptions are appropriate, and are given in the Supplement. The same hyperparameters are used as in the simulation study.

The accuracy of different methods in predicting the response variable is evaluated using RMSE, Mean Absolute Error (MAE), and correlation coefficients (Pearson/Spearman). The uncertainty

of point estimates is assessed using 90% equal-tailed intervals, for which we report the length and the frequentist coverage. We report the results for August in Table 5 and the results for June and July can be found in the Supplement. The dependent data models yield smaller RMSE and MAE than the independent models when predicting the frequency of hurricanes. The RMSE computed over all response variables for ind-data-ind-prior/dep-data-dep-prior spike-and-slab priors for June, July, and August are 1.95/1.80, 1.74/1.63, and 1.67/1.58 respectively, which shows a small decrease when using the multivariate model with our dependent spike-and-slab prior.

The estimated covariance matrix expressed as a correlation matrix, is provided in the Supplement, for June, July, and August. In general, we find that the dependent data model with dependent spike-and-slab prior tends to shrink more than the independent data model with independent spike-and-slab prior. The diagonals of the estimated $\mathbf{\Sigma}$ matrix are in the range 0.60-0.67 in June, and decrease to 0.50-0.53 in August. The results are on a standardized scale, where the response variables have marginal variance 1. Thus, around 50% of the variance in the response variables can be explained by the model in August. The correlations between response variables in the observed data are between 0.73 and 0.99, and the corresponding correlations between residuals range from 0.50 to 0.95, in August. We find that the correlations of PDI and ACE with the other variables show the most reduction. However, the correlation between frequency of tropical storms and hurricanes (on square root scale) decreases from 0.86 to 0.72 in the estimated residual correlation matrix, and for the PDI/ACE pair (on square root scale), it decreases from 0.99 to 0.95. The high correlations in the residual matrix suggest that there are other common factors between the response variables that the covariates in this model (SST forecasts) do not capture.

Response	Method	Cor.Pearson	Cor.Spearman	RMSE	MAE	Coverage	Length
TS	ss-ind-ind	0.74	0.69	1.93	1.45	1.00	11.28
	ss-dep-ind	0.67	0.72	2.10	1.74	1.00	11.41
	ss-ind-dep	0.75	0.69	1.94	1.39	1.00	11.19
	ss-dep-dep	0.65	0.68	2.16	1.81	1.00	11.50
Hurricane	ss-ind-ind	0.41	0.46	2.61	2.34	0.88	7.37
	ss-dep-ind	0.45	0.56	2.19	1.80	0.88	7.88
	ss-ind-dep	0.41	0.46	2.62	2.38	0.88	7.26
	ss-dep-dep	0.46	0.60	2.16	1.74	0.88	8.01
PDI	ss-ind-ind	0.62	0.67	0.65	0.52	0.75	1.85
	ss-dep-ind	0.53	0.50	0.68	0.54	0.75	2.05
	ss-ind-dep	0.61	0.55	0.66	0.53	0.75	1.83
	ss-dep-dep	0.52	0.38	0.69	0.55	0.75	2.07
ACE	ss-ind-ind	0.70	0.69	0.44	0.38	0.88	1.40
	ss-dep-ind	0.58	0.71	0.47	0.37	0.88	1.54
	ss-ind-dep	0.68	0.69	0.45	0.39	0.88	1.38
	ss-dep-dep	0.56	0.55	0.48	0.38	0.88	1.56

Table 5: Results for August with linear regression models.

6 Discussion

One of the main aims of this paper was to compare the predictive performance of Bayesian univariate and multivariate regression models. Our simulation results suggest that for moderate dimensional model spaces with a sparse coefficient matrix, multivariate Bayesian methods can have significantly improved performance in estimation of the mean of the predictive distribution compared to univariate methods. While there is also an improvement in prediction error, the gain can be relatively small. The reason is that the gain in estimation is typically quite small relative to the magnitude of error variance, and thus it does not lead to a significant reduction in the prediction error. A natural direction for future work is to consider higher dimensional model spaces.

Our results from the TC activity data are consistent with the simulation studies, in the sense that we see an overall small gain in point estimates for prediction using the multivariate methods. An interesting result for the application is that, while we find the SST forecasts can explain some of the correlations between the response variables, a large part still remains unaccounted for by this model. Perhaps, this suggests the need to look for additional covariates (which also have reliable forecasts) that could explain North Atlantic TC activity.

Acknowledgements

The authors are grateful to the Editor, the Associate Editor, and two referees for many suggestions that improved the paper. The authors thank Dr. Wei Zhang for his help with the data, and Dr. Luke Tierney for access to nodes on the ARGON cluster at The University of Iowa. This research was supported in part through computational resources provided by The University of Iowa, Iowa City, Iowa. Joyee Ghosh's research was supported by NSF Grant DMS-1612763. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation. The authors report there are no competing interests to declare.

References

Bai, R. and Ghosh, M. (2018). High-dimensional multivariate posterior consistency under global–local shrinkage priors. *Journal of Multivariate Analysis* **167**, 157 – 170.

Bottolo, L., Banterle, M., Richardson, S., Ala-Korpela, M., Jarvelin, M.-R., and Lewin, A. (2021). A computationally efficient Bayesian seemingly unrelated regressions model for high-dimensional quantitative trait loci discovery. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* .

Breiman, L. and Friedman, J. H. (1997). Predicting multivariate responses in multiple linear regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**, 1, 3–54.

Brown, B., Fearn, T., and Vannucci, M. (1999). The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. *Biometrika* **86**, 3, 635–648.

Brown, P. J., Vannucci, M., and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60**, 627–641.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). Handling sparsity via the horseshoe. In *Artificial Intelligence and Statistics*, 73–80. PMLR.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 2, 465–480.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association* **88**, 881–889.

Kundu, D., Mitra, R., and Gaskins, J. T. (2021). Bayesian variable selection for multioutcome models through shared shrinkage. *Scandinavian Journal of Statistics* **48**, 295–320.

Li, X., Ghosh, J., and Villarini, G. (2022). Bayesian negative binomial regression model with

unobserved covariates for predicting the frequency of North Atlantic tropical storms. *Journal of Applied Statistics* **to appear**.

McGill, R., Tukey, J. W., and Larsen, W. A. (1978). Variations of box plots. *The American Statistician* **32**, 1, 12–16.

Mitra, R. and Dunson, D. B. (2010). Two level stochastic search variable selection in GLMs with missing predictors. *International Journal of Biostatistics* **6**, 1, Article 33.

Polson, N. G. and Scott, J. G. (2010). Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics* **9**, 501-538, 105.

Richardson, S., Bottolo, L., and Rosenthal, J. S. (2010). Bayesian models for sparse regression analysis of high dimensional data. *Bayesian Statistics* **9**, 539–569.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics* **19**, 4, 947–962.

Vecchi, G. A. and Soden, B. J. (2007). Effect of remote sea surface temperature change on tropical cyclone potential intensity. *Nature* **450**, 1066–1070.

Villarini, G., Luitel, B., Vecchi, G. A., and Ghosh, J. (2019). Multi-model ensemble forecasting of North Atlantic tropical cyclone activity. *Climate Dynamics* **53**, 7461–7477.

Villarini, G. and Vecchi, G. A. (2012). North Atlantic power dissipation index (PDI) and accumulated cyclone energy (ACE): Statistical modeling and sensitivity to sea surface temperature changes. *Journal of Climate* **25**, 2, 625–637.

Villarini, G., Vecchi, G. A., and Smith, J. A. (2010). Modeling the dependence of tropical storm counts in the North Atlantic basin on climate indices. *Monthly Weather Review* **138**, 7, 2681–2705.