A Bayesian Hierarchical Model Framework to Quantify Uncertainty of Tropical Cyclone Precipitation Forecasts

Stephen A. Walsh∗ Marco A.R. Ferreira † David Higdon † Stephanie Zick ‡

July 19, 2022

Abstract

Tropical cyclones present a serious threat to many coastal communities around the world. Many numerical weather prediction models provide deterministic forecasts with limited measures of their forecast uncertainty. Standard postprocessing techniques may struggle with extreme events or use a 30-day training window that will not adequately characterize the uncertainty of a tropical cyclone forecast. We propose a novel approach that leverages information from past storm events, using a hierarchical model to quantify uncertainty in the spatial correlation parameters of the forecast errors (modeled as Gaussian processes) for a numerical weather prediction model. This approach addresses a massive data problem by implementing a drastic dimension reduction through the assumption that the MLE and Hessian matrix represent all useful information from each tropical cyclone. From this, simulated forecast errors provide uncertainty quantification for future tropical cyclone forecasts. We apply this method to the North American Mesoscale model forecasts and use observations based on the Stage IV data product for 47 tropical cyclones between 2004 and 2017. For an incoming storm, our hierarchical framework combines the forecast from the North American Mesoscale model with the information from previous storms to create 95% and 99% prediction maps of rain. For six test storms from 2018 and 2019, these maps provide appropriate probabilistic coverage of observations. We show evidence from the log scoring rule that the proposed hierarchical framework performs best among competing methods.

Keywords: Bayesian Statistics, Hurricane Forecasts, Massive Datasets, Meteorology, Spatial Statistics, Uncertainty Quantification

∗Corresponding author: Department of Statistics, Virginia Tech, walsh124@vt.edu
†Department of Statistics, Virginia Tech
‡Department of Geography, Virginia Tech
1 Introduction

Tropical cyclones (TCs) are some of the most costly and deadly natural disasters in the United States and across the world. The repercussions of these storms can take years to resolve, with TC rainfall being the primary culprit behind damage to inland communities (e.g., Cutter et al., 2014). Recent studies suggest that intensification rates of TCs may be increasing (e.g., Bhatia et al., 2019). Hurricanes are TCs that form in the Atlantic Ocean and achieve wind speeds of at least 33 meters per second; the Atlantic hurricane season of 2020 was the most active on record (Klotzbach et al., 2020). To effectively prepare communities and allocate emergency services, understanding and quantifying uncertainty in model-based TC forecasts is paramount. In this paper, we develop a Bayesian hierarchical model framework to quantify uncertainty of TC precipitation forecasts.

Our novel hierarchical model combines the numerical weather prediction (NWP) TC forecasts with historical data to predict rainfall for TCs and to more realistically characterize the uncertainty in future predictions. We focus on precipitation that takes place 24 hours after an individual TC landfall and consider the error fields for each storm to be defined as the observational data minus the NWP forecast (Gel et al., 2004). Our hierarchical model assumes that the error field for each TC follows a Gaussian process with a TC-specific set of parameters. We link the several TCs by assuming the parameters from each TC are realizations from a common linear model structure that may depend on known covariates. For example, in this paper we may assume that the prior distribution for the TC parameters depends on whether the storm makes landfall at the Atlantic, Florida, or Gulf of Mexico coastline.

We develop a Markov Chain Monte Carlo (MCMC) algorithm to explore the posterior distribution of the parameters of our hierarchical model. These include TC-specific parameters and hierarchical parameters related to the variability across different TCs. Here, we consider a training set of 47 TCs which made landfall between 2004 and 2017. Each TC’s set of data points ranges from 1,202 to 12,347, yielding a total of 308,013 data points for the 47 storms in the training set. In addition, analysis of spatial datasets with Gaussian processes is computationally expensive and usually scales cubically with sample size. Thus, a usual MCMC implementation of our hierarchical model would be infeasible. To deal with this massive data analysis problem, first we assume that the maximum likelihood estimate (MLE) of the parameters for each TC as well as the corresponding Hessian matrix contain all relevant information for that storm. We use these statistics to approximate the likelihood function for each TC. Finally, we develop an MCMC algorithm that uses these approximate likelihood functions to explore the posterior distribution of our proposed hierarchical model.

We use the output from the MCMC algorithm to obtain probabilistic forecasts for future storms. Specifically, we use the posterior sample of the hierarchical parameters to generate a sample from the prior distribution of the parameters of an incoming TC. With this sample, we use conditional sampling and the output of the NWP forecast
to generate a sample of the precipitation fields. With this latter sample, we can compute precipitation probability
maps to supplement the NWP forecast and better inform those affected by the incoming TC. We use storms from
2018 and 2019 as test data for this purpose. As we show in Section 5, our Bayesian hierarchical framework can also
describe the plausible ranges of volume of water over crucial geographical regions (watersheds, floodplains, urban
areas) that can adversely impact ecological systems (river flows and levels, flooding) and urban water systems.

NWP models use complex dynamical models based on atmospheric physics to describe future weather states.
Specifically, here we consider NWP models that provide a single forecast without any uncertainty quantification.
In particular, we do not consider probabilistic forecasting/data assimilation approaches (e.g. ensemble-based, vari-
ational) that might account for uncertainty in initial conditions, boundary conditions and parameterization choices
to produce estimates of prediction uncertainty (Hamill et al., 2012; Bannister, 2017). We note that, similarly to
what is discussed in Gneiting and Katzfuss (2014) and Li et al. (2017), our methods may be adapted to statistically
postprocess ensemble forecasts. For clarity of exposition we focus on NWP models that provide just one forecast.

Our Bayesian hierarchical framework may be considered a new statistical postprocessing approach. Statistical
postprocessors are methods that seek to remove biases from both deterministic and ensemble forecasts. In addition
to accounting for uncertainty in initial conditions and model uncertainty, statistical postprocessing offers a variety
of methods to calibrate biased model output and illustrate uncertainty with predictive distributions (see Li et al.,
2017, for an extensive review).

When only a single NWP forecast is available, one method of upgrading to a probabilistic forecast can be achieved
through the geostatistical output perturbation approach (Gel et al., 2004). Their work uses pairs of previous forecasts
and observations in a rolling window before the time of interest to learn about the discrepancies in the forecast and
characterize the uncertainty by generating error fields with spatial covariance estimated by the data pairs. Other
approaches that incorporate prior information in the form of climatological information include the Bayesian processor
of forecast approach (Krzysztofowicz and Evans, 2008) and work by Schaae et al. (2007) and Berrocal et al. (2008).
Using climatology or a rolling window assumes that the commonplace weather patterns for that location will suffice
to predict a future event. However, TCs are rare and extreme events and will likely not be characterized well by
these approaches. Rather than using a rolling window or climatological information, our proposed framework uses a
hierarchical model to account for forecast uncertainty for TCs.

Since the seminal work of Gel et al. (2004), a number of methodological extensions have been developed. Perhaps
the most prominent extensions account for multiple forecasting models; two of the most popular methods for Gaussian
variables are Bayesian model averaging (Raftery et al., 2005) and nonhomogeneous Gaussian regression (Gneiting
et al., 2005).

Additional extensions include accounting for non-Gaussian variables like precipitation (Sloughter et al., 2007),
and the inclusion of spatial components (Berrocal et al., 2007; Kleiber et al., 2011; Feldmann et al., 2015).

Recently, Villarini et al. (2022) proposed a method for generating probabilistic precipitation forecasts for TCs that made landfall in Louisiana. This method is applied to one particular region so the sample size of TCs (twelve) is relatively low. Our Bayesian framework provides spatially coherent uncertainty quantification (UQ) for rare and extreme weather events from a single NWP model output – a use case of interest to many forecasting centers (e.g., Ko et al., 2020). Additionally, our method can be applied generally for any TC in the contiguous United States (CONUS).

Many postprocessing approaches encounter some challenges when forecasting extreme events. Bishop and Shanley (2008) found that the Bayesian model averaging approach has a problematic treatment of extreme weather and propose including climatological information to alleviate this issue. Williams et al. (2014) compare postprocessing methods for extreme events and find that the methods become less reliable as more extreme events are considered. Williams et al. note the Lorenz 96 model (Lorenz, 1996) used in the study has short tails which will be easier to predict than real-world non-Gaussian variables and encourage development of postprocessing techniques specifically dedicated to extreme events. TC precipitation is some of the most extreme precipitation and will only exacerbate these challenges with extreme events. Another challenge related to many of these methods is the requirement of a rolling window (typically suggested to be around 30 days) of previous forecast/observation pairs to analyze characteristics of the model discrepancy. The rolling window is likely to contain errors for dry days and commonplace precipitation events that will not adequately characterize the errors corresponding to a TC.

In contrast, by explicitly modeling these extreme events, our Bayesian hierarchical model framework circumvents these difficulties by quantifying the uncertainty within a NWP TC precipitation forecast based on past storm events. To illustrate our approach, we select the North American Mesoscale (NAM) model for our NWP model and create error fields for 47 TC events from 2004 to 2017. We produce an estimate of systematic biases found from the NAM and supplement future forecasts with uncertainty quantification in a novel approach that does not require reevaluating the weather model.

Section 2 details the data preparation and processing procedure. We also introduce the hierarchical modeling framework and computations used to create our probability maps. Section 3 presents a simulation study which shows adequate estimation of spatial parameters used within the hierarchical model. Our modeling and computations are applied to the NAM and Stage IV data in Section 4. Section 5 details the UQ procedure for the test storms and compares our model to competing approaches, with the corresponding results from the logarithmic scoring rule. Section 6 provides discussion for future approaches and concluding remarks.
2 Uncertainty Quantification

2.1 Precipitation Forecasts and Observations

We study uncertainties of precipitation forecasts from the North American Mesoscale (NAM) model (Janjic, 2003; Rogers et al., 2009), which is a NWP model run by the National Centers for Environmental Prediction (NCEP). NAM forecasts are available for download at https://www.ncei.noaa.gov/products/weather-climate-models/north-american-mesoscale. Forecasts from the NAM have a resolution of approximately 12km over a domain which covers CONUS. The NAM produces forecasts for numerous meteorological variables; we will work specifically with quantitative precipitation forecasts (QPFs) at times which correspond to TCs making landfall in CONUS. The NAM was selected for this study given its relatively long historical record; hurricane forecasts from the NAM date back to 2004 (when including the earliest forecasts under the original name of Meso-ETA model) and it is still operational at the time of writing.

To assess the level of uncertainty in the NAM forecasts, we use the NCEP Stage IV product (Lin and Mitchell, 2005), hereafter identified as Stage IV, as ground truth observations. Stage IV is available to download from https://data.eol.ucar.edu/dataset/21.093. Stage IV is a quality controlled quantitative precipitation estimate (QPE) that synthesizes information from radars and rain gauges across the United States. Nelson et al. (2016) mention that, although there is a general underestimation for Stage IV at higher rain rates, the biases and fractional standard errors both decrease and correlation to rain gauges increases as the rain rate increases. Additionally, the eastern river forecast centers (which constitute our entire domain of interest) show the smallest fractional standard errors. For more comprehensive details of Stage IV and its many applications, see Nelson et al. (2016). Out of three competing QPEs, Villarini et al. (2011) found that Stage IV was the most effective for accurately estimating TC precipitation. Multiple studies have evaluated different QPEs during TCs or heavy rainfall by comparing their performances to that of Stage IV (Jiang et al., 2008; Habib et al., 2009a,b; Zagrodnik and Jiang, 2013). Studies by Clark et al. (2010) and Yan and Gallus Jr (2016) also evaluate NAM precipitation forecasts among other QPFs while using Stage IV as ground truth.

To study the most impactful portions of the storms, we focus on 24 hours of precipitation with NAM forecasts produced at either 0000 or 1200 Coordinated Universal Time (UTC). Specifically, if a storm makes CONUS landfall between 0600 UTC and 1800 UTC, we use the 1200 UTC forecast; otherwise, we use the 0000 UTC forecast. The same 24 hours of precipitation data are collected for Stage IV. The time and location of landfall for each of the storms are available in the second-generation hurricane database (HURDAT2; Landsea and Franklin, 2013) produced by the National Hurricane Center (NHC). Figure 1 shows the $N = 47$ TC landfall locations and intensities at time of landfall.
Figure 1: Landfall locations and intensities of 47 TCs with tropical storm (TS) strength or greater from 2004 to 2017. Shapes indicate the Saffir-Simpson scale at the time of landfall. Dashed lines partition the Gulf, Florida, and Atlantic landfall regions as suggested by Jagger and Elsner (2006).

To compare the NAM and Stage IV products, we interpolate both to a common coordinate reference system, the World Geodetic System 1984 (WGS84), with a spatial resolution of approximately 12km. To conserve the total volume of precipitation, we use a nearest neighbor interpolation scheme (Accadia et al., 2003). This interpolated grid is a rectangular array comprised of grid points both over CONUS and the neighboring oceans. Stage IV precipitation data collected over the ocean is inherently less reliable than data collected over land (Nelson et al., 2016). To account for this, we use a land-sea mask that excludes grid points over the ocean so that we only analyze precipitation over CONUS.

To focus solely on precipitation from the TC of interest, we employ two circular buffers with radii of 700km that remove all precipitation outside of the buffer. Commonly a 600km buffer is used (e.g., Marchok et al. 2007; Zick and Matyas 2016); we choose to increase this to 700km to include precipitation for some of the larger or faster-moving storms. The two buffer centers are chosen to approximate the eye of the storm at 6 and 18 hours after landfall based on HURDAT2 information to adequately encompass the 24 hours of precipitation. We define the ith storm’s buffer region $B_i$ to be all grid points contained in either of the two overlapping circular buffers. Thus, all grid points in $B_i$ will contain the ith storm’s accumulated precipitation for the 24 hour time period; all other grid points are removed. We denote by $n_i$ the number of grid points in $B_i$ which, as noted earlier, ranges from 1,202 to 12,347 for the 47 TCs in our training set. The union of all buffer regions establishes our common domain, $D = \bigcup_{i=1}^{47} B_i$, the set of all grid points over CONUS that are contained within at least one buffer region. For our training dataset, $D$ has $n_D = 26,399$. 
Precipitation information for both NAM and Stage IV are available in millimeters (mm). We originally used the log transformation to normalize and variance-stabilize the data, but this produced unrealistically high precipitation values when exponentiation was performed in the prediction phase. The square root transformation resolves this issue and is used instead. Let $m_i$ and $o_i$ be, respectively, the vector of square roots of NAM forecasts and the vector of square roots of Stage IV data for the $i$th storm within $B_i$. We subtract the NAM forecast from the Stage IV observation to define the error field for each storm in our training dataset. Then, the vectorized error field for storm $i$ within $B_i$ is defined as $y_i = o_i - m_i$. The data processing procedure is summarized in Figure 2. We now have an error field for each of the 47 storms that we use to analyze biases and uncertainty within the NAM forecasts. Plots for $m_i, o_i, y_i$ for $i = 47$ are shown in Figure 3, the data from Hurricane Nate in October 2017, the final storm in our training set. The plots for all storms are available in Figure S4 in the Supplementary Material (Walsh et al., 2022a).

### 2.2 Hierarchical Modeling for Error Fields

Here we propose a hierarchical model for uncertainty quantification of the NAM forecasts. Specifically, we model the error field of the NAM forecast for each storm as a realization of a Gaussian process with storm-specific parameters. We then connect the different storms by assuming that the storm-specific parameters of the different storms are realizations from a common linear model structure.

We model $y_i$ using a Gaussian process with a vector of covariance parameters $\lambda_i = [\sigma_i^2, \phi_i]^T$ such that $y_i \sim N(A_i\mu, A_i\Sigma(\lambda_i)A_i^T)$. Based upon coverage probabilities for prediction storms shown in Section 5, we find the
Gaussian assumption to provide reasonable results. The length of $y_i$ is $n_i$, $\sigma_i^2$ is the marginal variance of the $i$th error field, and $\phi_i$ is the scale parameter establishing the empirical range of the $i$th error field (or the distance required for correlation to become negligible). We include $\mu$, a vector of length $n_D$, to model the mean process of these error fields, representing systematic biases that may be present within the TC forecasts. $A_i$ is an incidence matrix (with dimension $n_i \times n_D$) for the $i$th error field, indicating whether or not a particular grid point is within $B_i$. That is, $A_i$ is the mathematical equivalent to applying a buffer, reducing the domain of $\mu$ from $D$ to $B_i$.

Originally, we modeled $\Sigma(\lambda_i)$ with the more flexible Matérn covariance structure. We also included a nugget effect to describe microscale variation of the spatial process or measurement error. However, 21 of the 47 storms had an estimated nugget of 0 and only 7 of the storms had an estimated nugget above 0.1 (with a maximum nugget estimate of 0.25). For this reason we choose to continue with a more parsimonious model and set the nugget variance to 0. We also estimated the additional smoothness parameters for the Matérn model and found these estimates to have an average of 0.51 and all fell within [0.35, 0.74]. Given that the exponential covariance function is a special case of the Matérn with smoothness 0.5, we opt to use this more computationally efficient covariance structure. Thus, the covariance matrix for the error field $\Sigma(\lambda_i)$ is modeled by the exponential covariance function (Banerjee et al., 2014):

$$
\Sigma(\lambda_i)_{j,k} = \text{Cov}(s_j, s_k) = \sigma_i^2 \exp\left(-\frac{||s_j - s_k||}{\phi_i}\right),
$$

(1)

where $||s_j - s_k||$ is the Euclidean distance between two grid points with locations $s_j$ and $s_k$.

We wish to model the storms’ spatial parameters with a Gaussian hierarchical model. Upon simulating 2,350 error fields with an exponential covariance function and a common true value for $\lambda$, we see a correlation of 0.996 for the MLEs of $\sigma^2$ and $\phi$ as well as right-skewed distributions for each (see Figure S2 in the Supplementary Material, Walsh et al. (2022a)). To reduce the correlation, we reparameterize our model and use $\theta_i = [\theta_{i1}, \theta_{i2}]^T = [\log(\sigma_i^2/\phi_i), \log(\sigma_i^2)]^T$ to model the covariance parameters. This reparameterization is motivated by results from Zhang (2004) and allows us to convert from a poorly behaved likelihood to one which is reasonably well approximated by a Gaussian distribution (see Figure S12 in the Supplementary Material, Walsh et al. (2022a)). In our simulation study, this transformation makes the distributions approximately Gaussian and the correlation is -0.036 for $\theta_{i1}$ and $\theta_{i2}$; for details see Section 3 and Appendix C in the Supplementary Material (Walsh et al., 2022a). Therefore, $\Sigma_{\theta}$ will be nearly diagonal and our posterior distributions for $\theta_i$ are well approximated by Gaussian distributions.

Our resulting framework includes $\theta_i \sim N(Bx_i, \Sigma_{\theta})$, where $B$ is a matrix of regression coefficients, $x_i$ is a vector of known regressors, and $\Sigma_{\theta}$ is the corresponding covariance matrix for $\theta_i$. One option for modeling the mean structure of $\theta_i$ is to assume a common mean for each of the TCs. In this setting, $B$ reduces to a vector of length
two, with elements corresponding to the mean components of $\theta_i$ and our known regressor reduces to a scalar with $x_i = 1$. In this setting we can define $\mu_\theta \equiv Bx_i$ to represent a common mean across all TCs. $\Sigma_\theta$ represents the covariance amongst components of each $\theta_i$ across the storms. To avoid overfitting, we choose to estimate $\Sigma_\theta$ as a common covariance matrix across the different TC error fields.

Following work from Jagger and Elsner (2006), we may categorize each of our storms by their landfall location, with three categories of Atlantic, Florida and Gulf storms delineated in Figure 1. This allows us to incorporate the influence of landfall location in the model. In the training dataset, we have 9 Atlantic, 21 Florida and 17 Gulf TCs. Here $x_i$ is a vector of regressors indicating the landfall region; the first entry is a 1 which represents the Atlantic region as a baseline, with indicators in the second and third positions corresponding to effects from storms being in the Florida and Gulf regions, respectively. For this model $B$ is a matrix of regression coefficients with the first column expressing the expected values of $\theta_i$ if the $i$th storm were an Atlantic storm. The second and third columns represent the differences in these expectations if the storm were in the Florida or Gulf regions respectively. The rows of $B$ correspond to the spatial parameters contained in $\theta_i$.

We implement the following general hierarchical model for the TC error fields:

$$y_i | \mu, \theta_i \sim N(A_i \mu, A_i \Sigma(\theta_i) A_i^T),$$

$$\mu | m, C \sim N(m, C),$$

$$\theta_i | B, \Sigma_\theta \sim N(Bx_i, \Sigma_\theta),$$

$$\pi(B) \propto 1,$$

$$\Sigma_\theta \sim IW(\nu_0, S_0).$$

Including Equation (3) allows us to model variability of the mean process $\mu$. After employing two different prior specifications for $\mu$, we found that setting $\mu = 0$ provided the best scores based on the logarithmic scoring rule (Gneiting and Raftery, 2007). Therefore, for this application we specify $\mu = 0$. For more details on this, see Section 5 and Appendix B in the Supplementary Material (Walsh et al., 2022a). Equation (4) is the distribution of the spatial parameters for each error field conditional on the prior distributions (5) and (6). Both priors for $B$ and $\Sigma_\theta$ constitute vague hyperprior specifications, with a flat prior for $B$ and a conjugate inverse Wishart prior for $\Sigma_\theta$, which permits the use of an efficient Gibbs sampler. For hyperparameters of $\Sigma_\theta$, we set $\nu_0 = p + 1$ where $p$ represents the dimension of $\theta_i$; this contains reasonably vague prior information while ensuring the prior distribution is proper. We explore options for prior settings of $S_0$ through sensitivity analysis and find that the results are rather sensitive to
the choice of $S_0$. Therefore we implement an empirical Bayes method of setting $S_0 = \nu_0 \text{Cov}(\hat{\Theta})$, where the $i$th row of $\hat{\Theta}$ is $\hat{\theta}_i^T$, the maximum likelihood estimate (MLE) of $\theta_i^T$. This prior specification has been shown to outperform other common prior specifications for $S_0$ when the variances of the parameters (e.g. $\theta$) are small (Schuurman et al., 2016). From this hierarchical model, we can learn the posterior distributions for $\mathbf{B}, \Sigma_{\theta},$ and $\theta_i, i \in \{1, \ldots, N\}$.

### 2.3 Computations

We perform computations in two steps. In the first step, we compute the MLEs of $\theta_1, \ldots, \theta_N$ as well as the corresponding Hessian matrices. After that, we use the Gaussian approximation to the distribution of $\hat{\theta}_1, \ldots, \hat{\theta}_N$ to construct approximate likelihood functions for each storm. In the second step, we use these approximate likelihoods combined with the hyperpriors (4), (5), and (6) to build a Gibbs sampler for $\theta_1, \ldots, \theta_N, \mu_\theta,$ and $\Sigma_\theta$. As we explain in Section 5, we use the posterior sample of $\mu_\theta$ and $\Sigma_\theta$ and the hierarchical model specification to obtain probabilistic forecasts for future storms.

For each error field $y_i$, we estimate $\hat{\theta}_i = [\hat{\theta}_{i1}, \hat{\theta}_{i2}]^T$ using the maximum of the profile likelihood (e.g., Diggle and Ribeiro, 2007, Chapter 5). Let $H_i$ denote the negative of the Hessian from the full likelihood function for the $i$th storm. We derive an approximate asymptotic covariance matrix with $H_i^{-1}$, which can be calculated analytically (see Appendix A in the Supplementary Material, Walsh et al. (2022a)) or with numerical approximations derived from the pracma package in R (Borchers, 2019). The Hessians from the two methods match to the third decimal place and the numerical approximation was found to be 2.5-10 times faster. Therefore we implement the numerical approximation within our framework. From the simulation study results in the Supplementary Material (Walsh et al., 2022a), we see empirical evidence that the distributions of the MLEs of $\theta_1$ and $\theta_2$ are very well approximated by a normal distribution. In this application, we obtain good coverages for the prediction storms’ precipitation, so these approximations are reasonable.

We obtain posterior draws for each of the parameters of interest by using a Gibbs sampler. The dimensionality of the $i$th storm, $n_i$, carries the computational burden of $A_i \Sigma(\theta_i) A_i^T$ within the Gibbs sampler. To address this, we assume that the MLEs $\hat{\theta}_i$ and their corresponding asymptotic covariance matrices $H_i^{-1}$ contain all useful information for the $i$th storm’s data. This achieves a massive data reduction, where the dimension of the problem is reduced from 308,013 to just $N \times p = 94$. Therefore, we have that $\mathcal{L}(y_i|\theta_i) = \mathcal{L}(\hat{\theta}_i|\theta_i, H_i) \sim N(\theta_i, H_i^{-1})$ which enables the Gibbs sampler to complete 10,000 iterations after burn-in in 150 seconds. The joint posterior is shown below, where $y_i$ is a zero-mean Gaussian process.
\[
\pi(\theta_i, B, \Sigma_\theta | y_i, x_i) = \pi(\theta_i, B, \Sigma_\theta | \hat{\theta}_i, x_i)
\]
\[
= \mathcal{L}(\hat{\theta}_i | \theta_i, H_i^{-1}) \pi(\theta_i | B x_i, \Sigma_\theta) \pi(B) \pi(\Sigma_\theta)
\]
\[
\propto \left\{ \prod_{i=1}^{N} \exp \left( -\frac{1}{2} (\hat{\theta}_i - \theta_i)^T H_i (\hat{\theta}_i - \theta_i) \right) \right\}
\]
\[
\times \left\{ \prod_{i=1}^{N} \Sigma_\theta^{-1/2} \exp \left( -\frac{1}{2} (\theta_i - B x_i)^T \Sigma_\theta^{-1} (\theta_i - B x_i) \right) \right\}
\]
\[
\times 1 \times |\Sigma_\theta|^{-(\nu_0 + p + 1)/2} \exp(-\text{tr}(S_0 \Sigma_\theta^{-1})/2)
\]

The full conditional distributions for the Gibbs sampler are shown below; derivations are in Appendix A in the Supplementary Material (Walsh et al., 2022a). The dash (−) indicates all inputs of the joint posterior with the exception of the particular variable for which the full conditional is defined. These distributions are multivariate normal, inverse Wishart and matrix-variate normal, respectively.

\[
\pi(\theta_i | -) \equiv N \left( (H_i + \Sigma_\theta^{-1})^{-1} (H_i \hat{\theta}_i + \Sigma_\theta^{-1} B x_i), (H_i + \Sigma_\theta^{-1})^{-1} \right)
\]
\[
\pi(\Sigma_\theta | -) \equiv IW \left( N + \nu_0, \sum_{i=1}^{N} (\theta_i - B x_i)(\theta_i - B x_i)^T + S_0 \right)
\]
\[
\pi(B | -) \equiv MN \left( \sum_{i=1}^{N} \theta_i x_i^T (\sum_{i=1}^{N} x_i x_i^T)^{-1}, \Sigma_\theta, (\sum_{i=1}^{N} x_i x_i^T)^{-1} \right)
\]

This method is feasible in an online forecasting context; the Gibbs sampler can run in about 5 minutes on a standard laptop (MacBook Pro 2.3 GHz Intel Core i5), and the prediction simulations take about 5 minutes as well. If more nodes/cores were available, prediction simulations can be computed in parallel and further decrease wait time. Calculation of the MLEs and Hessians can be done offline; their combined calculations average around 2 minutes per storm, with a maximum of about 11 minutes. Thus, the total computational time of our approach is the sum of the computational time for the calculation of MLEs and Hessian matrices with the computational time for the Gibbs sampler. Hence, if the computation of the MLEs is performed in parallel as we do here, then the computational time is less than 17 minutes. Given the large computational burden of ensemble prediction systems, which can take hours, we believe this method provides an efficient alternative. The code is available in the Supplementary Material (Walsh et al., 2022b).
3 Simulation Study

We present here the results of a simulation study to evaluate the statistical properties of the proposed methods to estimate $B, \Sigma_\theta, \theta_1, \ldots, \theta_N$. To illustrate adequate coverage of true spatial parameter values based on normal approximations of the MLEs, we simulate error fields with mean zero and random exponential covariance parameters $\theta_l$. The true $\theta_l$ values are draws from $\theta_l \sim N(\tilde{B} x_i, \tilde{\Sigma}_\theta)$ where $\tilde{B}$ and $\tilde{\Sigma}_\theta$ are provided in Section 4. Given that the spatial resolution and size of the spatial domain are critical factors governing the coverage rates, we choose to use the buffer regions $B_i$ for each of the storms in the training set as well as the corresponding $x_i$ values to generate $\theta_l$.

We simulate error field values for each $B_i$ and $x_i$ 50 times for a total of 2,350 simulated error fields. To create 95% confidence intervals, we obtain MLEs and approximate the covariance matrix with $H_l^{-1}$ as described above, with $l \in \{1, \ldots, 2350\}$. Our results show that we have adequate coverage for both $\theta_{1l}$ and $\theta_{2l}$ in these intervals, with coverage occurring for 95.7% and 93.1% of all simulations, respectively. The coverage for $\theta_{2l}$ is lower than 95% as a result of the likelihood for $\theta_{2l}$ having a slightly heavier right tail which results in a slight decrease in coverage when using the normal approximation. There was very weak correlation between $n_i$ and the corresponding average coverage rate of the 50 simulations, implying that the coverages were not heavily influenced by the number of grid points.

To illustrate the utility of the Gibbs sampler, we evaluate the credible intervals for $\theta_1, \ldots, \theta_N, B$ and $\Sigma_\theta$ by looking at the coverage of the true values in each interval. The intervals contain the true $\theta_{1l}$ and $\theta_{2l}$ values in 94.9% and 92% of the simulations, respectively. The true generating values for $B$ and $\Sigma_\theta$ are covered 95.7% and 94%, respectively.

4 Application to Error Fields

After the data processing (see Section 2.1), we interpolate, transform and apply buffers to each of the $N = 47$ training storms. We obtain $y_i$ for each TC and calculate $\hat{\theta}_i$ and $H_i$ for $i \in \{1, \ldots, N\}$. Using the MLEs and Hessian matrices as inputs, we run the Gibbs sampler and analyze the output.

Alongside each of these posterior estimates we have a measure of variance derived from the collection of posterior draws after burn-in. This allows us to compare the variability in our posterior estimates with the variability of the MLEs determined by the asymptotic covariance matrices derived from the Hessians. As illustrated in Figure 4, we see heteroskedasticity in the MLEs, as the estimates’ variability tends to increase as the estimates increase for $\theta_{2l} = \log(\hat{\sigma}_l^2)$. That is, larger values of $\hat{\theta}_{2l} = \log(\hat{\sigma}_l^2)$ tend to have larger variances so the influence of the prior is most noticeable here. MLE estimates for $\theta_{1l} = \log(\sigma_i^2/\phi_i)$ have very high precision and, as a result, the posterior
estimates do not show notable differences from the corresponding MLEs. The difference in the precisions of the MLEs of \( \theta_{i1} \) and \( \theta_{i2} \) is explained by theoretical and simulation results given in Zhang (2004), which shows that under in-fill asymptotics both MLEs are asymptotically unbiased but their variances exhibit different behaviors as the sample size increases; while the variance of the MLE of \( \theta_{i1} \) converges to zero, the variance of the MLE of \( \theta_{i2} \) decreases to a positive value and stops decreasing for larger sample sizes. Additionally, the within-TC uncertainty for each \( \theta_i \) (based on the Hessian matrices) differs from that of the between-TC variability. Our hierarchical framework models this variability of parameters between TCs through \( \Sigma_{\theta} \) estimated from the data, which is crucial for generating well-calibrated probabilistic predictions from deterministic forecasts for incoming storms.

The posterior medians for each entry in the regression coefficient matrix \( B \) and covariance matrix \( \Sigma_{\theta} \) are:

\[
\tilde{B} = \begin{bmatrix}
1.036_{(0.709,1.357)} & 0.258_{(-0.133,0.643)} & 0.017_{(-0.391,0.413)} \\
1.052_{(0.721,1.396)} & 0.229_{(-0.171,0.637)} & 0.206_{(-0.201,0.627)}
\end{bmatrix}
\]

\[
\tilde{\Sigma}_{\theta} = \begin{bmatrix}
0.235_{(0.160,0.367)} & 0.064_{(-0.008,0.155)} \\
0.064_{(-0.008,0.155)} & 0.209_{(0.129,0.347)}
\end{bmatrix},
\]

with subscripts indicating the 95% credible interval for each element. The columns of \( B \) represent the Atlantic baseline, the Florida effects and the Gulf effects. Rows correspond to the elements of \( \theta \). The second and third columns of \( B \) all contain 0 in the 95% credible intervals, suggesting the model with a common mean across landfall regions will likely suffice for this dataset.
Calculate MLEs and Hessian matrices for each TC
Use Gibbs sampler to obtain posterior distributions
Draws from posterior distributions for $\theta$
For each draw, conditionally simulate an error field
Combine simulations and forecast to obtain UQ

Figure 5: Framework to obtain inference and uncertainty quantification results.

4.1 Model Selection

We use the Laplace-Metropolis estimator to estimate the integrated likelihood for each of three different modeling regimes for $\theta$ (Lewis and Raftery, 1997). We consider three competing models with Model 1 assuming $\theta_i \sim N(Bx_i, \Sigma_\theta)$, with $B$ and $x_i$ specified to model the effects of the Atlantic, Florida and Gulf landfall regions. Model 2 assumes a common mean across all TCs, such that $\theta_i \sim N(\mu_\theta, \Sigma_\theta)$ and Model 3 assumes $\theta_i = \mu_\theta$, where $\mu_\theta$ is a common mean for $\theta_i$ across all landfall regions. Note that Model 3 is the only model that drops the hierarchical component of the model corresponding to $\Sigma_\theta$.

The integrated log-likelihood estimates for Models 1, 2 and 3 are $-75.42$, $-74.49$ and $-13044.51$, respectively. This shows strong support for the hierarchical model framework; under the non-hierarchical Model 3 the uncertainty in the Hessians do not sufficiently explain the variability of the parameters’ estimates across different TCs. Model 2 has the most support from the training data; this can be explained by the similarity of average spatial parameter estimates between landfall locations (see Figure S2 in the Supplementary Material, Walsh et al. (2022a)) and the penalization for 4 additional hyperparameters to estimate for Model 1. Therefore, we will use Model 2 in the following UQ applications.

5 Uncertainty Quantification for Future Storms

With Gibbs sampler output, we validate our method with TCs from 2018 and 2019, using the NAM forecasts to quantify uncertainty related to precipitation that will not yet be observed in a real-time setting. Prior to landfall, we obtain current track forecast data from the NHC’s tropical cyclone forecast advisories (https://www.nhc.noaa.gov/archive/). Details of these advisories and other NHC products can be found at https://www.nhc.noaa.gov/aboutnhcprom.shtml. For each forecast initialization, a corresponding advisory is reported three hours afterward; we use this information to construct the buffer region for an incoming TC.

Assume that we have observed $N$ storms and a new storm is coming. Let $y_{N+1}$ be the error field for the $(N+1)$th storm and $\theta_{N+1}$ be the corresponding vector of parameters. Then, according to our hierarchical model, the predictive density of $y_{N+1}$ given the data from the previous storms is
\[ p(y_{N+1} | y_1, \ldots, y_N) = \int \int \int p(y_{N+1} | \theta_{N+1}) p(\theta_{N+1} | \mu_\theta, \Sigma_\theta) d\theta_{N+1} \]

\[ p(\mu_\theta, \Sigma_\theta | y_1, \ldots, y_N) d\mu_\theta d\Sigma_\theta. \] (7)

Guided by Equation (7), we use the output from the MCMC algorithm proposed in Section 2.3 to simulate a sample from the predictive distribution of \( y_{N+1} \). Let \((\mu_\theta^{(g)}, \Sigma_\theta^{(g)}), g \in \{1, \ldots, G\}\) be the sample from the posterior distribution of \((\mu_\theta, \Sigma_\theta)\) obtained with the MCMC algorithm outlined in Section 2.3. We can then simulate a sample from the prior distribution of \( \theta_{N+1} | y_1, \ldots, y_N \) using conditional sampling as \( \theta_{N+1}^{(g)} = \mu_\theta^{(g)} + \omega^{(g)} \), where \( \omega^{(g)} \sim N(0, \Sigma_\theta^{(g)}) \). We then generate \( y_{N+1}^{(g)} \) from its conditional distribution \( y_{N+1}^{(g)} | \theta_{N+1}^{(g)} \). With this, we can generate 1000 corresponding error fields and create 95% prediction maps by adding the pointwise 95th percentiles of the generated error fields to the available NAM forecast. We can see if the prediction maps are greater than approximately 95% of grid points for the corresponding Stage IV data. The results for each of the three models are shown in Table 1. For Model 2, the average of coverages for the 95% prediction maps for the six TCs in the test set is about 96.86% and the 99% prediction maps show average coverages of 98.81%. These maps of precipitation totals illustrate potential worst-case scenarios based on the uncertainty in the forecast (see Figure 6b). We also look at coverages based on the extreme low and high forecasted precipitation. We define low values to be below 2.5mm (0.098 in) of precipitation and above the 95th percentile for each particular storm, respectively. Subsetting by these thresholds can show the coverages for these extremes. For the low extremes, 95th and 99th upper bound coverages of Model 2 are 98.87% and 99.85%, and for high extremes we have 95.55% and 97.35% coverage.

For the six prediction storms, Florence and Dorian were Atlantic storms, Alberto and Michael were Florida storms, and Gordon and Barry were Gulf storms. Thus, if we compare coverage rates by landfall location, we obtain the results shown in the bottom of Table 1. Models 1 and 2 assume \( \theta \), follow a normal distribution, where landfall region is and is not considered in the mean, respectively. From the table, we see approximately the equal average coverages, regardless of whether a location-specific mean is used; this further supports the idea that Model 2 is the best model (which agrees with the results from the Laplace-Metropolis estimators of the integrated likelihoods).

The prediction maps provide precipitation totals for a given percentile. Alternatively, if one is interested in the probability of precipitation surpassing a particular threshold (e.g., 2 inches) at different locations, then a probability map can be produced. Adding an error field to the NAM forecast will create one potential realization for the observed precipitation; an indicator map can be constructed to check whether or not a given grid point has surpassed the given threshold. Repeating this 1000 times and aggregating indicator plots will provide probabilistic information regarding which locations are most likely to experience severe rain.
Table 1: Coverage rates for the Stage IV precipitation of the six TCs in the test dataset for the 95% and 99% prediction maps based on 1000 error fields added to the corresponding NAM forecast. Models 1, 2, and 3 are defined as in Section 4.1. Each of the three landfall regions contained two prediction storms.

| Storm       | Model 1 95% | Model 1 99% | Model 2 95% | Model 2 99% | Model 3 95% | Model 3 99% |
|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Alberto, 2018 | 0.9753 0.9951 | 0.9768 0.9960 | 0.9651 0.9890 |
| Florence, 2018 | 0.9356 0.9651 | 0.9492 0.9715 | 0.9356 0.9610 |
| Gordon, 2018 | 0.9616 0.9816 | 0.9633 0.9821 | 0.9537 0.9737 |
| Michael, 2018 | 0.9421 0.9810 | 0.9478 0.9837 | 0.9224 0.9676 |
| Barry, 2018  | 0.9832 0.9958 | 0.9840 0.9954 | 0.9772 0.9897 |
| Dorian, 2019 | 0.9967 1.0000 | 0.9991 1.0000 | 0.9937 1.0000 |
| Atlantic Average | 0.9613 0.9798 | 0.9701 0.9835 | 0.9600 0.9774 |
| Florida Average | 0.9607 0.9889 | 0.9641 0.9906 | 0.9463 0.9796 |
| Gulf Average | 0.9716 0.9882 | 0.9730 0.9883 | 0.9647 0.9811 |
| Overall Average | 0.9647 0.9865 | 0.9687 0.9881 | 0.9561 0.9796 |

Figure 6: Square root precipitation fields for (a) the NAM forecast, (b) the 95% prediction map, and (c) the Stage IV data product for Hurricane Michael, 2018.

One approach to evaluating our prediction intervals is comparing their widths with the observed forecast errors in the test storms. Obtaining the 95% prediction interval at each grid point for each prediction storm, the average length of these intervals is 1.606 inches, which indicates a margin of error of 0.803in. If we look at the 95% upper bound of these 95% prediction intervals, the margin of error is 2.161in. The maximum margin of error from every 95% prediction interval over all prediction storms was 6.607in. We note that the 95% upper bound, 99% upper bound and maximum absolute difference between the forecasted values (from NAM) and observed values (from Stage IV) in the prediction storms were 1.54in, 3.93in and 16.109in respectively. For the grid points with the largest forecasted precipitation totals in the 24 hour window, we obtain margins of error that are approximately half of this total (see Figure S11 in the Supplementary Material Walsh et al. (2022a)). Therefore, we find our prediction intervals to be reasonable and useful.

To illustrate another application of our method, we study the variability of a TC’s accumulated precipitation for a hydrologic subregion. The United States Geological Survey defines hydrologic regions, subregions and other areas
based upon the drainage locations for CONUS. We select the Ochlockonee hydrologic subregion as an example since it is contained within the buffer region of Hurricane Michael and it also contains the city of Tallahassee (see Figure 7). Using the aforementioned prediction output, we transform the potential precipitation realizations to millimeters and aggregate over the hydrologic subregion. From this, a predictive distribution for the accumulated precipitation for the subregion can be obtained, for example by using kernel density estimation. This distribution can be further transformed to obtain potential volumes of water that will pass through the subregion’s rivers as a result of the TC. In the example for Hurricane Michael, we can see that the predictive distribution successfully captures the observed value based upon Stage IV data (see Figure 7c).

Figure 7: (a) The Ochlockonee subregion, (b) the interpolated Ochlockonee subregion showing 24 hour accumulated precipitation (in square root mm) from Hurricane Michael, and (c) the predictive density for the accumulated precipitation (in mm) for the subregion in the 24 hour time window. The solid line indicates the total from the NAM forecast and the dashed line indicates the total from the Stage IV data.

5.1 Scores for competing UQ approaches

The Ochlockonee subregion described above is one of 90 different watersheds we study to evaluate our proposed methodology. Across the six prediction storms, we consider watersheds located across seven states (Alabama, Florida, Georgia, Louisana, Mississippi, North Carolina and South Carolina). Over 100 of these watersheds are contained within one or more of the six prediction storms’ buffer regions. We set a requirement that at least 30 grid points of the buffer region must be contained in a watershed to qualify, and we proceed by analyzing TC precipitation for the six prediction storms at these 90 watersheds that qualify.

To assess the performance of our proposed method, we compare performance with multiple candidate models. We consider multiple approaches for modeling the spatial parameters $\theta$, including the three models previously mentioned in subsection 4.1. Additionally, we include a nonparametric approach for modeling $\theta$, Model 4, where a bootstrap sample is drawn from the 47 MLEs of the training storms. Model 5 does not consider spatial dependence between grid points, and thus reduces $\theta$ to a scalar $\theta = \log(\sigma^2)$. In Models 1-5, there is no bias adjustment performed (i.e.,
\( \mu = 0 \)). Models 6-9 will correspond to Models 1-4 except that we consider a bias adjustment such that the pointwise mean has been subtracted from each NAM forecast (see Appendix B in the Supplementary Material, Walsh et al. (2022a)).

We use the logarithmic scoring rule to evaluate the predictive densities from each of the nine candidate models for the 90 watersheds (Good, 1952; Gneiting and Raftery, 2007). The logarithmic score has many desirable properties: it is a proper scoring rule that takes spatial dependence into account and, because it is equivalent to the log of the predictive density, it is related to posterior model probabilities and Bayes factors.

![Table 2](image)

Table 2: For each of the nine candidate models, the sum of the log predictive densities over the 90 prediction basins from the 6 prediction storms is presented. Models 1 and 2 use a hierarchical model with a Gaussian distribution for \( \theta \), where the mean does (\( Bx_i \)) and does not (\( \mu_\theta \)) vary by landfall location (\( x_i \)), respectively. Model 3 is similar to Model 2, but removes modeling of \( \Sigma_\theta \), so the only variability of \( \theta \) is from the asymptotic precision matrices derived from the Hessians. Model 4 uses a bootstrap sample of the MLEs \( \hat{\theta} \). Model 5 is a nonspatial version of Model 1, which reduces \( \theta \) to a scalar, \( \log(\sigma^2) \). Models 6-9 are equivalent to Models 1-4, respectively, except for a bias adjustment using the posterior mean of \( \mu \) (Appendix C). The best score is shown in boldface.

![Table 3](image)

Table 3: For each of the nine candidate models, the number of times the model had the highest score for each of the prediction basins. These results are subset by landfall region (selected by a majority vote of grid points), and also if the basin was coastal or not. The best performer for each category is set in boldface.

When we sum the log predictive densities from each watershed, we obtain a total score for each candidate model. These sums are found in Table 2. Here, we see that Model 2, a hierarchical model with a common mean \( \mu_\theta \) for each \( \theta \) and no bias adjustment, has the best performance. When we consider each basin individually, we see that Models 1 and 2 (both hierarchical models with normal distributions for \( \theta \) and no bias adjustment) each are selected as the best about the same number of times (see Table 3). If we subset basins by the three landfall regions or by whether the basin is coastal or not, this persists, although it seems Florida basins prefer Model 1 and Gulf basins generally prefer Model 2. So, we don’t have a definite answer as to whether or not different watersheds should have different
distributions for the model parameters. It is worth noting that even when we subset basins by landfall region or their coastal status, the best model overall by the log score is still Model 2 (Table 3). This is likely due to the fact that Model 2 generally has slightly wider tails than Model 1, so it is not as heavily penalized when there is a larger error between the forecasted and observed values. Model 4 is outperformed by Models 1 and 2 providing evidence that, in this application, the Bayesian approach outperforms the use of MLEs alone.

One of the most appealing aspects of Bayesian methods is the natural updating process as new data becomes available. After an incoming storm is processed by the UQ algorithm and the Stage IV products become available, MLEs can be calculated for the storm and the posterior distributions from the Gibbs sampler will be updated accordingly. As the number of storms increase, we obtain more information with which to quantify the uncertainty for future storms. Although the updating process could be done by simple weighting (linear combination of the old and current forecasts), we prefer our approach because the updates are based on the data (via $\theta$) and hence will not depend on some chosen vector of weights as in a simple weighting scheme.

6 Discussion

Many NWP models produce precipitation forecasts with limited information regarding their uncertainty. When dealing with extreme events like TCs, it is important to quantify this uncertainty to better inform those in the storm’s path. With our Bayesian framework, we propose a novel approach to analyze the variability in NWP forecast errors and provide UQ for these extreme weather events with a limited number of storms. Our approach maintains spatial coherence and allows modeling of these rare events where previous postprocessing methods would deteriorate.

We study the operational NAM TC forecasts and obtain the corresponding QPEs available from Stage IV. We also learn about the amount of uncertainty we can expect from a given NAM forecast for TC precipitation through estimation of $\theta_i$ and the implementation of our UQ algorithm. This general framework can be implemented with other NWP models (eg: Global Forecast System or Hurricane Weather Research and Forecast models) to explore these results and also compare their performances with that from the NAM.

This work inspires avenues for future research. The effects of using alternative, or even nonparametric, distributions for $\theta$ or $y_i$ should be explored. Additionally, there is uncertainty not only in the amount of precipitation, but also in the track that the eye of the TC follows. Including uncertainty within the storm track can help improve our results by having less dependence on the particular NWP model forecast being studied. From here, additional uncertainty with respect to the landfall region of the storm could be pursued as well. We plan to create a more complex model by introducing nonstationarity for each error field, as well as a spatiotemporal component to allow
for changes in the structure of the error fields over time, e.g., the subsequent 24 hour window of time for each TC.

Our methodology can be expanded to ensembles. Also, by incorporating our UQ methodology within the Bayesian model averaging or nonhomogeneous Gaussian regression frameworks (Raftery et al., 2005; Gneiting et al., 2005), or adapting the Kennedy and O’Hagan (2001) framework, this approach may illuminate the types of uncertainties not generally characterized in most ensemble approaches.

In conclusion, we provide a novel framework for quantifying the uncertainty of tropical storm and hurricane forecasts. This technique can help to illuminate systematic biases in forecasts as well as better understand the variability within a particular forecast. The framework can also be applied to other spatial fields, such as wind forecasts. By implementing the UQ algorithm, one is able to have a clearer understanding of the potential variability in a TC precipitation forecast. With greater understanding of the abilities and limits of a QPF, we hope to inform research scientists of new approaches for assessing accuracy and reliability of their products.

Acknowledgments

The authors would like to thank the anonymous referees, the Associate Editor and the Editor for their conscientious efforts and constructive comments, which improved the quality of this paper.

SUPPLEMENTARY MATERIAL

Supplement to “A Bayesian Hierarchical Model Framework to Quantify Uncertainty of Tropical Cyclone Precipitation Forecasts” (DOI:10.1214/[provided by typesetter]; .pdf) The supplementary material contains Appendices A, B, C, and D. Appendix A contains derivations for the full conditional distributions used in the Gibbs sampler and analytical calculations for the Hessian matrices of the parameters of the exponential covariance function. Appendix B provides details on modeling the systematic bias of the NAM forecast. Appendix C contains simulation study results, and Appendix D provides plots for each of the tropical cyclone landfalls within the training set and 95% upper bounds for uncertainty of each TC in the test set. Additionally, plots of the margins of error for each test storm, and an example log-likelihood surface for the original and reparameterized parameter space are provided.

Supplement to “A Bayesian Hierarchical Model Framework to Quantify Uncertainty of Tropical Cyclone Precipitation Forecasts” Code for “A Bayesian Hierarchical Model Framework to Quantify Uncertainty of Tropical Cyclone Precipitation Forecasts” (DOI:10.1214/[provided by typesetter]; .zip) This file contains the code and some processed data to reproduce the results from the manuscript. It is also available at https://github.com/stevewalsh124/NAM-Model-Validation.
References

Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A., and Speranza, A. (2003). “Sensitivity of Precipitation Forecast Skill Scores to Bilinear Interpolation and a Simple Nearest-Neighbor Average Method on High-Resolution Verification Grids.” *Weather and Forecasting*, 18, 5, 918–932.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data (2nd ed.)*. CRC press.

Bannister, R. (2017). “A review of operational methods of variational and ensemble-variational data assimilation.” *Quarterly Journal of the Royal Meteorological Society*, 143, 703, 607–633.

Berger, J. O., De Oliveira, V., and Sansó, B. (2001). “Objective Bayesian analysis of spatially correlated data.” *Journal of the American Statistical Association*, 96, 456, 1361–1374.

Berrocal, V. J., Raftery, A. E., and Gneiting, T. (2007). “Combining spatial statistical and ensemble information in probabilistic weather forecasts.” *Monthly Weather Review*, 135, 4, 1386–1402.

— (2008). “Probabilistic quantitative precipitation field forecasting using a two-stage spatial model.” *Annals of Applied Statistics*, 2, 4, 1170–1193.

Bhatia, K. T., Vecchi, G. A., Knutson, T. R., Murakami, H., Kossin, J., Dixon, K. W., and Whitlock, C. E. (2019). “Recent increases in tropical cyclone intensification rates.” *Nature Communications*, 10, 1, 1–9.

Bishop, C. H. and Shanley, K. T. (2008). “Bayesian model averaging’s problematic treatment of extreme weather and a paradigm shift that fixes it.” *Monthly Weather Review*, 136, 12, 4641–4652.

Borchers, H. W. (2019). *pracma: Practical Numerical Math Functions*. R package version 2.2.5.

Clark, A. J., Gallus Jr, W. A., and Weisman, M. L. (2010). “Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM.” *Weather and Forecasting*, 25, 5, 1495–1509.

Cutter, S. L., Emrich, C. T., Mitchell, J. T., Piegorsch, W. W., Smith, M. M., and Weber, L. (2014). *Hurricane Katrina and the forgotten coast of Mississippi*. Cambridge University Press.

Diggle, P. and Ribeiro, P. (2007). *Model-based Geostatistics*. Springer Series in Statistics. Springer.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L. (2015). “Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression.” *Monthly Weather Review*, 143, 3, 955–971.
Gel, Y., Raftery, A. E., and Gneiting, T. (2004). “Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation method.” Journal of the American Statistical Association, 99, 467, 575–583.

Gneiting, T. and Katzfuss, M. (2014). “Probabilistic forecasting.” Annual Review of Statistics and Its Application, 1, 125–151.

Gneiting, T. and Raftery, A. E. (2007). “Strictly proper scoring rules, prediction, and estimation.” Journal of the American Statistical Association, 102, 477, 359–378.

Gneiting, T., Raftery, A. E., Westveld III, A. H., and Goldman, T. (2005). “Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation.” Monthly Weather Review, 133, 5, 1098–1118.

Good, I. J. (1952). “Rational decisions.” Journal of the Royal Statistical Society: Series B (Statistical Methodology), 14, 1, 107–114.

Gramacy, R. B. (2020). Surrogates: Gaussian process modeling, design, and optimization for the applied sciences. Chapman and Hall/CRC.

Habib, E., Henschke, A., and Adler, R. F. (2009a). “Evaluation of TMPA satellite-based research and real-time rainfall estimates during six tropical-related heavy rainfall events over Louisiana, USA.” Atmospheric Research, 94, 3, 373–388.

Habib, E., Larson, B. F., and Graschel, J. (2009b). “Validation of NEXRAD multisensor precipitation estimates using an experimental dense rain gauge network in south Louisiana.” Journal of Hydrology, 373, 3-4, 463–478.

Hamill, T. M., Brennan, M. J., Brown, B., DeMaria, M., Rappaport, E. N., and Toth, Z. (2012). “NOAA’s future ensemble-based hurricane forecast products.” Bulletin of the American Meteorological Society, 93, 2, 209–220.

Jagger, T. H. and Elsner, J. B. (2006). “Climatology Models for Extreme Hurricane Winds near the United States.” Journal of Climate, 19, 13, 3220–3236.

Janjic, Z. (2003). “A nonhydrostatic model based on a new approach.” Meteorology and Atmospheric Physics, 82, 1-4, 271–285.

Jiang, H., Halverson, J. B., and Simpson, J. (2008). “On the differences in storm rainfall from Hurricanes Isidore and Lili. Part I: Satellite observations and rain potential.” Weather and Forecasting, 23, 1, 29–43.

Kennedy, M. C. and O’Hagan, A. (2001). “Bayesian calibration of computer models.” Journal of the Royal Statistical Society: Series B (Statistical Methodology), 63, 3, 425–464.
Kleiber, W., Raftery, A. E., and Gneiting, T. (2011). “Geostatistical model averaging for locally calibrated probabilistic quantitative precipitation forecasting.” *Journal of the American Statistical Association*, 106, 496, 1291–1303.

Klotzbach, P. J., Bell, M. M., and Jones, J. (2020). “Summary of 2020 Atlantic Tropical Cyclone Activity and Verification of Authors’ Seasonal and Two-Week Forecasts.” Tech. rep., Colorado State University.

Ko, M.-C., Marks, F. D., Alaka, G. J., and Gopalakrishnan, S. G. (2020). “Evaluation of hurricane Harvey (2017) rainfall in deterministic and probabilistic HWRF forecasts.” *Atmosphere*, 11, 6, 666.

Krzysztofowicz, R. and Evans, W. B. (2008). “Probabilistic forecasts from the national digital forecast database.” *Weather and Forecasting*, 23, 2, 270–289.

Landsea, C. W. and Franklin, J. L. (2013). “Atlantic hurricane database uncertainty and presentation of a new database format.” *Monthly Weather Review*, 141, 10, 3576–3592.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Lewis, S. M. and Raftery, A. E. (1997). “Estimating Bayes factors via posterior simulation with the Laplace—Metropolis estimator.” *Journal of the American Statistical Association*, 92, 438, 648–655.

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., and Di, Z. (2017). “A review on statistical postprocessing methods for hydrometeorological ensemble forecasting.” *Wiley Interdisciplinary Reviews: Water*, 4, 6, e1246.

Lin, Y. and Mitchell, K. E. (2005). “1.2 the NCEP stage II/IV hourly precipitation analyses: Development and applications.” In 19th Conf. Hydrology, American Meteorological Society, San Diego, CA, USA, 1–4. Citeseer.

Lorenz, E. N. (1996). “Predictability: A problem partly solved.” In *Proc. Seminar on predictability*, vol. 1, 1–18. ECMWF.

Marchok, T., Rogers, R., and Tuleya, R. (2007). “Validation Schemes for Tropical Cyclone Quantitative Precipitation Forecasts: Evaluation of Operational Models for U.S. Landfalling Cases.” *Weather and Forecasting*, 22, 4, 726–746.

Nelson, B. R., Prat, O. P., Seo, D.-J., and Habib, E. (2016). “Assessment and implications of NCEP Stage IV quantitative precipitation estimates for product intercomparisons.” *Weather and Forecasting*, 31, 2, 371–394.

Petersen, K. B. and Pedersen, M. S. (2012). “The Matrix Cookbook.” Version 20121115.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). “Using Bayesian model averaging to calibrate forecast ensembles.” *Monthly Weather Review*, 133, 5, 1155–1174.
Roger, E., DiMego, G., Black, T., Ek, M., Ferrier, B., Gayno, G., Janjic, Z., Lin, Y., Pyle, M., Wong, V., Wu, W. S., and Carley, J. (2009). “The NCEP North American Mesoscale Modeling System: Recent changes and future plans.” In 23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction, Omaha, NE, Amer. Meteor. Soc., 2A.4.

Schaeke, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X., and Seo, D. (2007). “Precipitation and temperature ensemble forecasts from single-value forecasts.” Hydrology and Earth System Sciences Discussions, 4, 2, 655–717.

Schuurman, N., Grasman, R., and Hamaker, E. (2016). “A comparison of inverse-wishart prior specifications for covariance matrices in multilevel autoregressive models.” Multivariate Behavioral Research, 51, 2-3, 185–206.

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., and Fraley, C. (2007). “Probabilistic quantitative precipitation forecasting using Bayesian model averaging.” Monthly Weather Review, 135, 9, 3209–3220.

Villarini, G., Smith, J. A., Baek, M. L., Marchok, T., and Vecchi, G. A. (2011). “Characterization of rainfall distribution and flooding associated with US landfalling tropical cyclones: Analyses of Hurricanes Frances, Ivan, and Jeanne (2004).” Journal of Geophysical Research: Atmospheres, 116, 1–19.

Villarini, G., Zhang, W., Miller, P., Johnson, D. R., Grimley, L. E., and Roberts, H. J. (2022). “Probabilistic rainfall generator for tropical cyclones affecting Louisiana.” International Journal of Climatology, 42, 3, 1789–1802.

Walsh, S. A., Ferreira, M. A., Higdon, D., and Zick, S. (2022a). “Supplement to “A Bayesian Hierarchical Model Framework to Quantify Uncertainty of Tropical Cyclone Precipitation Forecasts”.”

— (2022b). “Supplement to “A Bayesian Hierarchical Model Framework to Quantify Uncertainty of Tropical Cyclone Precipitation Forecasts”.”

Williams, R., Ferro, C., and Kwasniok, F. (2014). “A comparison of ensemble post-processing methods for extreme events.” Quarterly Journal of the Royal Meteorological Society, 140, 680, 1112–1120.

Yan, H. and Gallus Jr, W. A. (2016). “An evaluation of QPF from the WRF, NAM, and GFS models using multiple verification methods over a small domain.” Weather and Forecasting, 31, 4, 1363–1379.

Zagrodnik, J. P. and Jiang, H. (2013). “Investigation of PR and TMI version 6 and version 7 rainfall algorithms in landfalling tropical cyclones relative to the NEXRAD stage-IV multisensor precipitation estimate dataset.” Journal of Applied Meteorology and Climatology, 52, 12, 2809–2827.
Zhang, H. (2004). “Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics.” *Journal of the American Statistical Association*, 99, 465, 250–261.

Zick, S. E. and Matyas, C. J. (2016). “A shape metric methodology for studying the evolving geometries of synoptic-scale precipitation patterns in tropical cyclones.” *Annals of the American Association of Geographers*, 106, 6, 1217–1235.
SUPPLEMENTARY MATERIAL

Supplementary Material for A Bayesian Hierarchical Model Framework to Quantify Uncertainty of Tropical Cyclone Precipitation Forecasts

A Derivations

A.1 Full Conditionals for Gibbs Sampler

Below is the joint posterior distribution from which the full conditionals are derived. Recall the prior choices of $\pi(B) \propto 1$ and $\pi(\Sigma_\theta) = IW(\nu_0, S_0)$.

\[
\pi(\theta_i, B, \Sigma_\theta| y_i^*, x_i) = \pi(\theta_i, B, \Sigma_\theta| \hat{\theta}_i, x_i) \\
= \mathcal{L}(\hat{\theta}_i| \theta_i, H_i^{-1}) \times \pi(\theta_i| B x_i, \Sigma_\theta) \times \pi(B) \times \pi(\Sigma_\theta)
\]

\[
\propto \left\{ \prod_{i=1}^{N} \exp \left( -\frac{1}{2} (\hat{\theta}_i - \theta_i)^T H_i (\hat{\theta}_i - \theta_i) \right) \right\}
\]

\[
\times \left\{ \prod_{i=1}^{N} |\Sigma_\theta|^{-1/2} \exp \left( -\frac{1}{2} (\theta_i - B x_i)^T \Sigma_\theta^{-1} (\theta_i - B x_i) \right) \right\}
\]

\[
\times 1 \times |\Sigma_\theta|^{-(\nu_0+p+1)/2} \exp(-\text{tr}(S_0 \Sigma_\theta^{-1})/2)
\]

We begin by obtaining the full conditional for $\theta_i$.

\[
\pi(\theta_i| -) \propto \left\{ \prod_{i=1}^{N} \exp \left( -\frac{1}{2} (\hat{\theta}_i - \theta_i)^T H_i (\hat{\theta}_i - \theta_i) \right) \right\}
\]

\[
\times \left\{ \prod_{i=1}^{N} |\Sigma_\theta|^{-1/2} \exp \left( -\frac{1}{2} (\theta_i - B x_i)^T \Sigma_\theta^{-1} (\theta_i - B x_i) \right) \right\}
\]

\[
\propto \exp \left( -\frac{1}{2} \sum_{i=1}^{N} [(\hat{\theta}_i - \theta_i)^T H_i (\hat{\theta}_i - \theta_i) + (\theta_i - B x_i)^T \Sigma_\theta^{-1} (\theta_i - B x_i)] \right)
\]

\[
\propto \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \left[ \theta_i^T (H_i + \Sigma_\theta^{-1}) \theta_i - 2 \theta_i^T (H_i \hat{\theta}_i + \Sigma_\theta^{-1} B x_i) \right] \right)
\]

\[
\equiv N \left( (H_i + \Sigma_\theta^{-1})^{-1} (H_i \hat{\theta}_i + \Sigma_\theta^{-1} B x_i), (H_i + \Sigma_\theta^{-1})^{-1} \right)
\]

Next we find the full conditional for $\Sigma_\theta$, which is a $p \times p$ matrix, with $p$ being the number of elements contained
in the vector of parameters $\theta$.

\[
\pi(\Sigma_\theta | -) \propto \left\{ \prod_{i=1}^{N} |\Sigma_\theta|^{-\frac{1}{2}} \exp \left( -\frac{1}{2}(\theta_i - Bx_i)^T \Sigma_\theta^{-1}(\theta_i - Bx_i) \right) \right\} \\
\times |\Sigma_\theta|^{-\nu_0 + p + 1 \over 2} \exp(-\text{tr}(S_0 \Sigma_\theta^{-1})/2) \\
\propto \left\{ |\Sigma_\theta|^{-\frac{N}{2}} \exp \left( -\frac{1}{2} \sum_{i=1}^{N} (\theta_i - Bx_i)^T \Sigma_\theta^{-1}(\theta_i - Bx_i) \right) \right\} \\
\times |\Sigma_\theta|^{-\nu_0 + p + 1 \over 2} \exp(-\text{tr}(S_0 \Sigma_\theta^{-1})/2) \\
\propto \left\{ |\Sigma_\theta|^{-\frac{N}{2}} \exp \left( -\frac{1}{2} \text{tr} \left( \sum_{i=1}^{N} (\theta_i - Bx_i)(\theta_i - Bx_i)^T \Sigma_\theta^{-1} \right) \right) \right\} \\
\times |\Sigma_\theta|^{-\nu_0 + p + 1 \over 2} \exp(-\text{tr}(S_0 \Sigma_\theta^{-1})/2) \\
\propto |\Sigma_\theta|^{-(N + \nu_0 + p + 1) / 2} \exp \left( -\frac{1}{2} \text{tr} \left( \sum_{i=1}^{N} (\theta_i - Bx_i)(\theta_i - Bx_i)^T + S_0 \right) \Sigma_\theta^{-1} \right) \\
\equiv IW \left( N + \nu_0, \sum_{i=1}^{N} (\theta_i - Bx_i)(\theta_i - Bx_i)^T + S_0 \right)
\]

So, the full conditional of $\Sigma_\theta$ is distributed inverse Wishart.

The following equation is the probability density function of the matrix-variate normal distribution:

\[
p(X|M, U, V) = \frac{\exp \left( -\frac{1}{2} \text{tr} \left[ V^{-1}(X - M)^T U^{-1}(X - M) \right] \right)}{(2\pi)^{np/2} |V|^{n/2} |U|^{p/2}}
\]

Note $X \in \mathbb{R}^{n \times p}$ is the random variable, $M$ is the mean matrix, and $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{p \times p}$ represent the among-row and among-column variance, respectively. Expanding the kernel of the matrix-variate distribution we find the following equivalent expressions that will help determine the full conditional for $B$.

\[
p(X|M, U, V) \propto \exp \left( -\frac{1}{2} \text{tr} \left[ V^{-1}(X - M)^T U^{-1}(X - M) \right] \right) \\
\propto \exp \left( -\frac{1}{2} \text{tr} \left[ V^{-1}(X^T U^{-1} X - X^T U^{-1} M - M^T U^{-1} X + M^T U^{-1} M) \right] \right) \\
\propto \exp \left( -\frac{1}{2} \text{tr} \left[ V^{-1} X^T U^{-1} X - V^{-1} X^T U^{-1} M - V^{-1} M^T U^{-1} X + V^{-1} M^T U^{-1} M \right] \right)
\]
Let us obtain the full conditional for $B$:

$$
\pi(B|\cdot) \propto \prod_{i=1}^{N} \exp \left( -\frac{1}{2} \left( \theta_i - Bx_i \right)^T \Sigma_{\theta}^{-1} \left( \theta_i - Bx_i \right) \right)
$$

$$
\propto \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \left[ x_i^T B^T \Sigma_{\theta}^{-1} Bx_i - \theta_i^T \Sigma_{\theta}^{-1} Bx_i - x_i^T B^T \Sigma_{\theta}^{-1} \theta_i \right] \right)
$$

$$
\propto \exp \left( -\frac{1}{2} \sum_{i=1}^{N} \text{tr} \left[ x_i^T B^T \Sigma_{\theta}^{-1} Bx_i - \theta_i^T \Sigma_{\theta}^{-1} Bx_i - x_i^T B^T \Sigma_{\theta}^{-1} \theta_i \right] \right)
$$

$$
\propto \exp \left( -\frac{1}{2} \left( \text{tr} \left[ \sum_{i=1}^{N} x_i^T B^T \Sigma_{\theta}^{-1} B \right] - \text{tr} \left[ \sum_{i=1}^{N} \theta_i^T B^T \Sigma_{\theta}^{-1} \theta_i \right] \right) \right)
$$

So from the first portion of the expression we obtain $V^{-1} = \sum_{i=1}^{N} x_i x_i^T$ and $U^{-1} = \Sigma_{\theta}^{-1}$. The second and third show that $MV^{-1} = \sum_{i=1}^{N} (\theta_i x_i^T)$, which implies that $M = (\sum_{i=1}^{N} \theta_i x_i^T)(\sum_{i=1}^{N} x_i x_i^T)^{-1}$. Therefore, we have the full conditional for the matrix $B$:

$$
\pi(B|\cdot) \equiv MN \left( (\sum_{i=1}^{N} \theta_i x_i^T)(\sum_{i=1}^{N} x_i x_i^T)^{-1}, \Sigma_{\theta}, (\sum_{i=1}^{N} x_i x_i^T)^{-1} \right)
$$

### A.2 Analytical Hessian calculations for Exponential Covariance Function

To obtain a Hessian matrix based on the MLEs of $\theta = (\log(\sigma^2/\phi), \log(\sigma^2))^T$, we can use the multivariate delta method (Lehmann and Casella, 2006) after calculating the MLEs and corresponding Hessian matrix with respect to $\lambda = [\sigma^2, \phi]^T$. Let’s assume we have a zero-mean spatial process with no nugget and exponential covariance function with marginal variance $\sigma^2$ and scale parameter $\phi$, that is $\Sigma(\phi)_{i,j} = \Sigma(\phi, Y)_{i,j} = \exp(-\frac{||Y_i - Y_j||}{\phi})$ with $||Y_i - Y_j||$ represents the Euclidean distance between the $i$th and $j$th observations. Then, if we have a spatially correlated $n$–vector $Y_n \equiv Y \sim N(0, \sigma^2 \Sigma(\phi))$, we can express the likelihood in the following manner:

$$
\mathcal{L} = \mathcal{L}(\phi, \sigma^2|Y) = (2\pi\sigma^2)^{-n/2} |\Sigma(\phi)|^{-1/2} \exp \left( -\frac{1}{2\sigma^2} Y^T \Sigma^{-1}(\phi) Y \right)
$$

$$
l = l(\phi, \sigma^2|Y) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log |\Sigma(\phi)| - \frac{1}{2\sigma^2} Y^T \Sigma^{-1}(\phi) Y
$$
Taking the derivative twice with respect to $\sigma^2$, we get the following:

\[
\frac{\partial l}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} Y^T \Sigma^{-1}(\phi) Y
\]

\[
\frac{\partial^2 l}{(\partial \sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} Y^T \Sigma^{-1}(\phi) Y \approx \frac{n}{2(\sigma^2)^2} - \frac{n}{2(\sigma^2)^2} = -\frac{n}{2(\sigma^2)^2}
\]

Before we look at the derivatives with respect to $\phi$, let’s consider some important rules that we can use:

\[
\frac{\partial \Sigma^{-1}(\phi)}{\partial \phi} = -\Sigma^{-1}(\phi) \frac{\partial \Sigma(\phi)}{\partial \phi} \Sigma^{-1}(\phi)
\]

\[
\frac{\partial \log |\Sigma(\phi)|}{\partial \phi} = \text{tr} \left[ \Sigma^{-1}(\phi) \frac{\partial \Sigma(\phi)}{\partial \phi} \right]
\]

We can see that each of these require $\frac{\partial \Sigma(\phi)}{\partial \phi} = \Sigma'(\phi) = \Sigma(\phi) \odot \frac{D}{\partial \phi}$, where $\odot$ represents the Hadamard (element-wise) product of two $m \times n$ matrices (in this case $m = n$), $D$ is the $n \times n$ matrix of (Euclidean) distances between all of the $N$ locations where $Y$ is observed (e.g., Gramacy, 2020, Chapter 5). Let us begin with the derivatives with respect to $\phi$:

\[
\frac{\partial^2 l}{\partial \phi \partial \sigma^2} = \frac{\partial}{\partial \phi} \left[ \frac{-n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} Y^T \Sigma^{-1}(\phi) Y \right]
\]

\[
= \frac{1}{2(\sigma^2)^2} Y^T \frac{\partial}{\partial \phi} \left[ \Sigma^{-1}(\phi) \right] Y
\]

\[
= -\frac{1}{2(\sigma^2)^2} Y^T \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) Y
\]

Now we can start working to obtain the final derivative which will be used to calculate the Hessian, $\frac{\partial^2 l}{\partial \sigma^2}$:

\[
\frac{\partial l}{\partial \phi} = \frac{\partial}{\partial \phi} \left[ -\frac{1}{2} \log |\Sigma(\phi)| - \frac{1}{2\sigma^2} Y^T \Sigma^{-1}(\phi) Y \right]
\]

\[
= -\frac{1}{2} \text{tr} \left[ \Sigma^{-1}(\phi) \Sigma'(\phi) \right] + \frac{1}{2\sigma^2} Y^T \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) Y
\]

Recall $\frac{\partial \Sigma(\phi)}{\partial \phi} = \Sigma'(\phi) = \Sigma(\phi) \odot \frac{D}{\partial \phi}$. Let’s start the second derivative:
\[
\frac{\partial^2 l}{\partial \phi^2} = \frac{\partial}{\partial \phi} \left[ -\frac{1}{2} \text{tr} \left[ \Sigma^{-1}(\phi) \Sigma'(\phi) \right] + \frac{1}{2\sigma^2} Y^T \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) Y \right]
= -\frac{1}{2} \text{tr}(A) + \frac{1}{2\sigma^2} Y^T BY
\]

Here,

\[
A = \frac{\partial}{\partial \phi} \left[ \Sigma^{-1}(\phi) \Sigma'(\phi) \right]
B = \frac{\partial}{\partial \phi} \left[ \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \right]
\]

Before finding more specific expressions for A and B, let’s take a moment to find the second derivative of \(\Sigma(\phi)\), that is \(\frac{\partial^2 \Sigma(\phi)}{\partial \phi^2}\). We need to use \(\partial (X \odot Y) = \partial (X) \odot Y + X \odot \partial (Y)\) (Petersen and Pedersen, 2012):

\[
\Sigma''(\phi) = \frac{\partial^2 \Sigma(\phi)}{\partial \phi^2} = \frac{\partial}{\partial \phi} \Sigma'(\phi) = \frac{\partial}{\partial \phi} \left[ \Sigma(\phi) \odot \frac{D}{\sigma^2} \right]
= \Sigma'(\phi) \odot \frac{D}{\sigma^2} - \Sigma(\phi) \odot \frac{2D}{\sigma^3}
\]

So, let’s evaluate A:

\[
A = \frac{\partial}{\partial \phi} \left[ \Sigma^{-1}(\phi) \Sigma'(\phi) \right]
= -\Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \Sigma'(\phi) + \Sigma^{-1}(\phi) \Sigma''(\phi)
= -\left( \Sigma^{-1}(\phi) \Sigma'(\phi) \right)^2 + \Sigma^{-1}(\phi) \Sigma''(\phi)
\]

Here we evaluate B:
\[ B = \frac{\partial}{\partial \phi} \left[ \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \right] \]
\[ \begin{align*} 
&= \frac{\partial}{\partial \phi} \left[ \Sigma^{-1}(\phi) \right] \Sigma'(\phi) \Sigma^{-1}(\phi) + \Sigma^{-1}(\phi) \frac{\partial}{\partial \phi} \left[ \Sigma'(\phi) \right] \Sigma^{-1}(\phi) + \Sigma^{-1}(\phi) \Sigma'(\phi) \frac{\partial}{\partial \phi} \left[ \Sigma^{-1}(\phi) \right] \\
&= \left[ -\Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \right] \Sigma'(\phi) \Sigma^{-1}(\phi) + \Sigma^{-1}(\phi) \left[ \Sigma''(\phi) \right] \Sigma^{-1}(\phi) \\
&\quad - \Sigma^{-1}(\phi) \Sigma'(\phi) \left[ \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \right] \\
&= \Sigma^{-1}(\phi) \left[ \Sigma''(\phi) \right] \Sigma^{-1}(\phi) - 2 \left( \Sigma^{-1}(\phi) \Sigma'(\phi) \right)^2 \Sigma^{-1}(\phi) \\
&= \Sigma^{-1}(\phi) \left[ \Sigma''(\phi) \right] \Sigma^{-1}(\phi) - 2 \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \\
&= \Sigma^{-1}(\phi) \left[ \Sigma''(\phi) \right] \Sigma^{-1}(\phi) - 2\Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) 
\end{align*} \]

With our second derivatives in hand, we can take the expectation of each of these expressions to simplify them. This is using the result from Appendix A of Berger et al. (2001), where it states that if \( X \sim N(\mu, \Sigma) \) and \( A \) is a symmetric matrix, then \( \mathbb{E}(X^TAX) = \text{tr}(A\Sigma) + \mu^T A \mu \).

So, that will give us the following:

\[ \mathbb{E} \left( \frac{\partial^2 l}{(\partial \sigma^2)^2} \right) = \mathbb{E} \left[ \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} Y^T \Sigma^{-1}(\phi) Y \right] \]
\[ = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \mathbb{E} \left[ Y^T \Sigma^{-1}(\phi) Y \right] \]
\[ = \frac{n}{2(\sigma^2)^2} - \frac{1}{(\sigma^2)^3} \text{tr} \left( \sigma^2 \Sigma^{-1}(\phi) \Sigma(\phi) \right) + 0^T \Sigma^{-1}(\phi) 0 \]
\[ = \frac{n}{2(\sigma^2)^2} - \frac{n}{(\sigma^2)^2} \]
\[ = \frac{-n}{2(\sigma^2)^2} \]

Since we have that \( \mu = 0 \), we can always ignore the \( \mu^T A \mu \) term.
\[ E\left( \frac{\partial^2 l}{\partial \sigma^2 \partial \phi} \right) = E \left[ - \frac{1}{2(\sigma^2)^2} Y^T \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) Y \right] \]
\[ = - \frac{1}{2(\sigma^2)^2} E \left[ Y^T \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) Y \right] \]
\[ = - \frac{1}{2(\sigma^2)^2} \text{tr} \left( \sigma^2 \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \Sigma(\phi) \right) \]
\[ = - \frac{1}{2\sigma^2} \text{tr} \left( \Sigma^{-1}(\phi) \Sigma'(\phi) \right) \]

In the following calculation, note that neither \( A \) nor \( B \) depend on \( Y \).

\[ E\left( \frac{\partial^2 l}{\partial \phi^2} \right) = E \left[ - \frac{1}{2} \text{tr}(A) + \frac{1}{2\sigma^2} Y^T BY \right] \]
\[ = - \frac{1}{2} \text{tr}(A) + \frac{1}{2\sigma^2} E \left[ Y^T BY \right] \]
\[ = - \frac{1}{2} \text{tr}(A) + \frac{1}{2\sigma^2} \text{tr} \left( \sigma^2 B \Sigma(\phi) \right) \]
\[ = - \frac{1}{2} \text{tr}(A) + \frac{1}{2} \text{tr} \left( \Sigma^{-1}(\phi) \Sigma'(\phi) \right)^2 - \Sigma^{-1}(\phi) \Sigma'(\phi) \]
\[ + \frac{1}{2} \text{tr} \left( \Sigma^{-1}(\phi) \Sigma'(\phi) \right)^2 - \Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \Sigma(\phi) \Sigma'(\phi) \]
\[ = \frac{1}{2} \text{tr} \left( (\Sigma^{-1}(\phi) \Sigma'(\phi))^2 \right) - \Sigma^{-1}(\phi) \Sigma'(\phi) + \Sigma^{-1}(\phi) \Sigma'(\phi) \]
\[ - 2\Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \Sigma'(\phi) \]
\[ = \frac{1}{2} \text{tr} \left( (\Sigma^{-1}(\phi) \Sigma'(\phi))^2 \right) - 2\Sigma^{-1}(\phi) \Sigma'(\phi) \Sigma^{-1}(\phi) \Sigma'(\phi) \]
\[ = - \frac{1}{2} \text{tr} \left( (\Sigma^{-1}(\phi) \Sigma'(\phi))^2 \right) \]

B Modeling the systematic bias of the NAM forecast

Within our hierarchical framework, we specified a flat prior for \( \mu \) by specifying a precision matrix of zeros: \( C^{-1} = 0_{n \times n} \). This setting allows the data to speak for itself. We also employed an informative prior, with a prior mean of 0 and a covariance matrix modeled with an exponential covariance function, such that its prior spatial parameters were estimated by maximum likelihood from the grid points of \( \hat{\mu} \) which had data for 20 or more storms.
Upon implementing both of these priors for $\mu$ and obtaining the corresponding posterior distributions, both were outperformed by specifying no systematic bias, i.e., $\mu = 0$.

Therefore, we begin by obtaining an estimate $\hat{\mu}$ of the mean process to approximate the bias-adjusted $y^*_i$ for $i \in \{1, \ldots, N\}$. This is achieved through an empirical average over all error fields: $\hat{\mu}(s) = (1/n_M(s))\sum_{i \in M(s)} y_i(s)$, where $s$ is a grid point location in CONUS, $M(s)$ is the set of storms for which the buffer region contains $s$, and $n_M(s)$ is the cardinality of $M(s)$. This allows us to approximate a zero-mean Gaussian process with $\hat{y}^*_i = y_i - A_i \hat{\mu} \sim N(0, A_i \Sigma_i A_i^T)$. For computational reasons, we choose to treat $\hat{\mu}$ as constant here to avoid simulating $\mu$ within the MCMC loop.

We also use each $\hat{\theta}_i$ to learn about the variability of the mean process $\mu$. For each error field, we estimate the covariance matrix $\Sigma_i \equiv A_i \Sigma(\hat{\theta}_i) A_i^T$ detailing the correlation between grid points for a particular bias-adjusted error field. The $N$ precision matrices $\Sigma_i^{-1}$ along with the prior precision $C^{-1}$ allow us to generate $\Sigma_\mu$, the posterior covariance of the mean process $\mu$:

$$
\Sigma_\mu = \left[ C^{-1} + \sum_{i=1}^{N} \Sigma_i^{-1} \right]^{-1}.
$$

(8)

This additionally informs $m_\mu$, the posterior mean for $\mu$:

$$
m_\mu = 0_n + \Sigma_\mu \sum_{i=1}^{N} \left[ \Sigma_i^{-1} y_i \right].
$$

(9)

Given the inherent interdependence between calculating $\hat{\theta}_i$, $i \in \{1, \ldots, N\}$ and $m_\mu$, we implement an algorithm with the style of Expectation Maximization. The goal of this algorithm is to jointly optimize both the spatial parameter estimates and the mean process in a balanced and equitable manner. We begin by estimating $\hat{\mu}$, the MLEs $\hat{\theta}_i$ and finally $m_\mu$ as described above. We replace our empirical estimate $\hat{\mu}$ with the posterior mean $m_\mu$ and use this to redefine our bias-adjusted error fields: $\hat{y}^*_i = y_i - A_i m_\mu$. Upon redefining $\hat{y}^*_i$ we update our MLEs $\hat{\theta}_i$ and then update $m_\mu$ similarly. Repeating this process until convergence allows for joint optimization between both $\hat{\theta}_i, i \in \{1, \ldots, N\}$ and $m_\mu$.

With estimate $\hat{\mu}$, we obtain $\hat{y}^*_i$ and calculate $\hat{\theta}_i$ for $i \in \{1, \ldots, N\}$. We implement calculations for $m_\mu$, the posterior mean of the mean process $\mu$ as described in Section 2.3 (see Figure S1a). Upon comparing $\hat{\mu}$ and $m_\mu$, we find that the two are identical down to 12 decimal places, indicating that we obtain convergence between $m_\mu$ and $\theta_i$ after one iteration when we use an objective prior on $\mu$ as described in Section 2.2 of the paper.

We can illustrate the uncertainty for each grid point in $m_\mu$ by creating a map with the corresponding posterior standard deviations, available as the square root of the diagonal elements of $\Sigma_\mu$ (Figure S1b). Using $m_\mu$ and $\Sigma_\mu$, we
calculate the standardized mean map of errors, as shown in Figure S1c. These plots suggest there may be locations where the NAM is systematically biased, with notable areas of overestimation along the Appalachian mountains and southern Texas. We can also see that while southwestern Florida is typically overestimated, there is a tendency for the NAM to underestimate TC precipitation in northeastern Florida.

We illuminate potentially systematic biases of NAM with respect to TC precipitation by producing a posterior mean map for the error fields with \( \mathbf{m}_{\mathbf{\mu}} \). Given the results of the model comparisons in subsection 5.1 of the paper, models that do not account for systematic bias performed better in this application. More complex models, such as systematic bias that changes through time, are conceptually possible but may create substantial computational burden. More complex models for systematic bias offer an avenue for future research.

\section{Simulation Study Results}

To illustrate adequate coverage of true spatial parameter values based on normal approximations of the MLEs, we simulate error fields with mean zero and exponential covariance parameters of \( \sigma^2 = 4 \) and \( \phi = 1.5 \). These true values for the simulations are chosen as they are similar to the average MLE values over all of the error fields in the test set (4.14 and 1.42, respectively). Upon reparameterizing, this is equivalent to a true parameter vector of \( \mathbf{\theta} = [\log(\sigma^2/\phi), \log(\sigma^2)]^T = (0.981, 1.386)^T \). In the second row of Figure S2, histograms for \( \sigma^2 \) and \( \phi \) both show a right skew. There is also evidence for a ridge in the likelihood surface of \( \mathbf{\lambda} \) (as described by Zhang, 2004) with a correlation of 0.996 between \( \sigma^2 \) and \( \phi \). Conversely, the histograms in the first row show the components of \( \mathbf{\theta} \) are both approximately normal and the scatter plot shows approximate independence between the two elements of \( \mathbf{\theta} \).

In the last two rows of Figure S2, the true \( \mathbf{\theta}_i \) values are random samples from \( \mathbf{\theta}_i \sim N(\bar{\mathbf{B}} \mathbf{x}_i, \bar{\Sigma}_\mathbf{\theta}) \) where \( \bar{\mathbf{B}} \) and
\( \tilde{\Sigma}_\theta \) are provided in Section 4 of the paper. Accounting for the variation of estimates across the different TCs, we see that the correlation of the MLEs for \( \sigma^2 \) and \( \phi \) decreases but is still relatively strong at 0.64. The correlation for \( \theta_{l1} \) and \( \theta_{l2} \) increases but is still relatively weak at 0.232.
Figure S2: Parameter estimates for 2,350 simulated error fields for the reparameterization $\theta$ (first row) as well as the original parameterization $\lambda$ (second row) when true values are fixed at $\sigma^2 = 4, \phi = 1.5$ for each simulation (shown by the blue lines and diamonds). Parameter estimates for 2,350 simulated error fields for the reparameterization $\theta$ (third row) as well as the original parameterization $\lambda$ (fourth row) when true values are generated randomly where $\theta_i \sim N(Bx_i, \Sigma_\theta)$ shown in Section 4 of the paper. Blue, red and yellow lines and diamonds represent the true parameter values generated by values of Atlantic, Florida and Gulf landfall regions, respectively.
D Plots

D.1 Plots of Tropical Cyclone Landfalls and Other Model Output

Figure S3: Plots of landfalls for a subset of the 47 training TCs, where 6 of each are selected from each landfall region. The first column represents 24 hour accumulated square root precipitation of the Stage IV data, the second column is the corresponding NAM forecast, and the third column is the difference between Stage IV and NAM. The remaining columns follow this pattern.
Figure S4: Same as S3, but for all 47 landfalls for the training TCs. The number and letter in each plot represents the storm number (1-47) and landfall region (A for Atlantic, F for Florida, G for Gulf).
D.2 95% Upper Bounds for Uncertainty of Prediction Storms

These images show the NAM forecast (left), the 95% upper bound from our UQ method (middle), and the observed precipitation based on Stage IV (right) for each of the six prediction storms from 2018-2019.

- Figure S5: First prediction storm: Alberto 2018.
- Figure S6: Second prediction storm: Florence 2018.
- Figure S7: Third prediction storm: Gordon 2018.
Figure S8: Fourth prediction storm: Michael 2018.

Figure S9: Fifth prediction storm: Barry 2019.

Figure S10: Sixth prediction storm: Dorian 2019.
Figure S11: Margins of error for each grid point across each of the six test storms. These are calculated as half of the length of the predictive interval for each grid point of each storm.
Figure S12: Log likelihood on the (a) original and (b) transformed parameter space for the first training storm. Plots are similar across all training storms.