A Bayesian Regression Approach for Predicting Seasonal Tropical Cyclone Activity over the Central North Pacific

PAO-SHIN CHU

Department of Meteorology, School of Ocean and Earth Science and Technology, University of Hawaii at Manoa, Honolulu, Hawaii

XIN ZHAO

Department of Information and Computer Science, University of Hawaii at Manoa, Honolulu, Hawaii

(Manuscript received 9 November 2005, in final form 17 November 2006)

ABSTRACT

In this study, a Poisson generalized linear regression model cast in the Bayesian framework is applied to forecast the tropical cyclone (TC) activity in the central North Pacific (CNP) in the peak hurricane season (July–September) using large-scale environmental variables available up to the antecedent May and June. Specifically, five predictor variables are considered: sea surface temperatures, sea level pressures, vertical wind shear, relative vorticity, and precipitable water. The Pearson correlation between the seasonal TC frequency and each of the five potential predictors over the eastern and central North Pacific is computed. The critical region for which the local correlation is statistically significant at the 99% confidence level is determined. To keep the predictor selection process robust, a simple average of the predictor variable over the critical region is then computed. With a noninformative prior assumption for the model parameters, a Bayesian inference for this model is derived in detail. A Gibbs sampler based on the Markov chain Monte Carlo (MCMC) method is designed to integrate the desired posterior predictive distribution. The proposed hierarchical model is physically based and yields a probabilistic prediction for seasonal TC frequency, which would better facilitate decision making. A cross-validation procedure was applied to predict the seasonal TC counts within the period of 1966–2003 and satisfactory results were obtained.

1. Introduction

The tropical cyclone (TC) is one of the most destructive natural catastrophes that cause loss of lives and enormous property damage on the eastern coast and in the gulf states of the United States. Even located far away in the central Pacific, Hawaii is not immune to hurricane perils. For instance, the Hawaiian Islands were directly struck by Hurricanes Iniki in 1992 and Iwa in 1982. Estimates for damage were about $2.5 billion for Iniki and $250 million for Iwa. Given the soaring property values in the past few years in Hawaii, the damage would have been much higher if it were adjusted to the current value. After Iniki, a minimum category-4 hurricane on the Saffir–Simpson scale, hurricanes continued to pose a threat to the islands. In 1994, three intense, category-5 hurricanes tracked westward just to the south of Hawaii; this is the first time that such intense hurricanes were reported in the central North Pacific (Fig. 1), let alone three in a single season. Because of their profound socioeconomic repercussions, understanding climate factors that are instrumental for the year-to-year TC variability, and developing a sound and modern method for predicting seasonal TC counts before the peak season are becoming increasingly important.

Based on a two-sample permutation procedure, Chu and Wang (1997) noticed that the mean annual number of tropical cyclones in the vicinity of Hawaii is higher during El Niño years than during non–El Niño years, and this difference is statistically significant at the 95% confidence level. Clark and Chu (2002) investigated large-scale circulation patterns related to tropical cyclone genesis and development over the central North Pacific in association with warm and cold El Niño–Southern Oscillation (ENSO) events. One of the salient findings is the marked enhancement of the 1000-hPa

Corresponding author address: Dr. Pao-Shin Chu, Department of Meteorology, University of Hawaii at Manoa, 2525 Correa Rd., Honolulu, HI 96822.
E-mail: chu@hawaii.edu

DOI:10.1175/JCLI4214.1

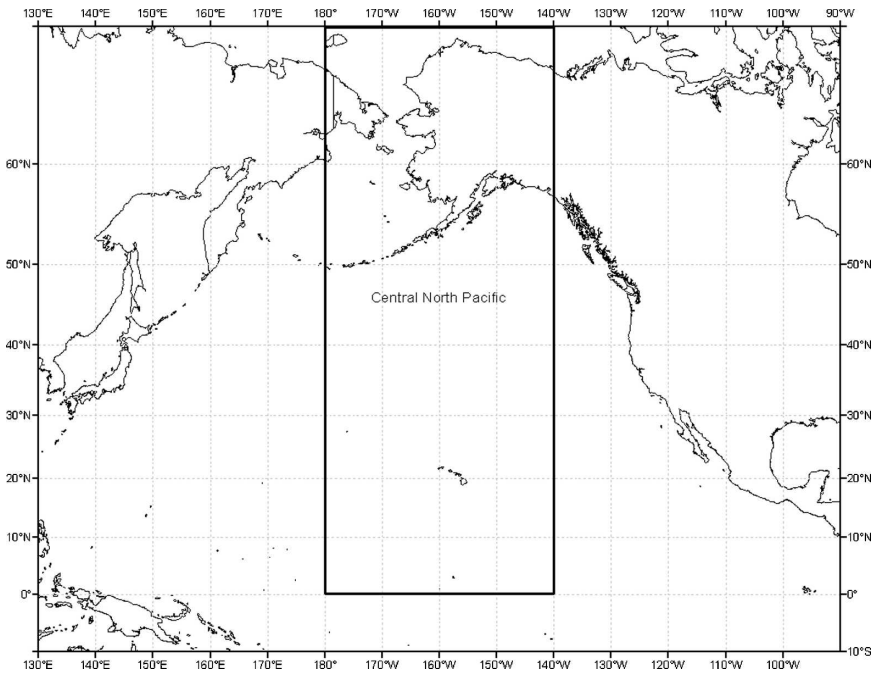

FIG. 1. Orientation map for the CNP.

relative vorticity values just to the south of Hawaii during an El Niño autumn relative to La Niña autumn. Tropospheric vertical wind shear (VWS) to the south of Hawaii also shows a two- to threefold reduction when an El Niño composite is compared to a La Niña one.

The seasonal hurricane prediction enterprise using regression-based linear statistical models was pioneered by Gray et al. (1992, 1993, 1994). They showed that nearly half of the interannual variability of hurricane activity in the North Atlantic could be predicted in advance. Klotzbach and Gray (2004) have continued to revise their forecasts as peak seasons approach, and they operationally issue seasonal forecasts for the Atlantic basin (more information available online at http://hurricane.atmos.colostate.edu/Forecasts). Seven parameters are routinely predicted at long lead times. These include the number of hurricanes, number of named storms, number of hurricane days, number of named storm days, intense hurricanes, intense hurricane days, and net tropical cyclone activity (NTC). NTC is a combined measure of the aforementioned other six parameters normalized by their climatological averages. When verified for a 52-yr record (1950–2001), hindcast skill is lowest for named storms but is better for intense hurricanes and NTC if the forecast is issued in early December, 6–11 month prior to the Atlantic hurricane season (Klotzbach and Gray 2004). As an example, assuming a climatological forecast of 100 NTC for each year, their statistical model exhibits a 27% reduction over climatology in errors for the test period.

Elsner and Schmertmann (1993) considered a different approach to predict intense annual Atlantic hurricane counts. Specifically, the annual hurricane occurrence is modeled as a Poisson process, which is governed by a single parameter, the Poisson intensity. The intensity of the process is then made to depend upon a set of covariates such as the stratospheric zonal winds and the west Sahel rainfall via a multiple regression equation. Parameters of the regression are estimated by maximum likelihood. Recently, Elsner and Jagger (2004) introduced a Bayesian approach to this Poisson linear regression model so that the predicted annual hurricane numbers could be cast in terms of probability distributions. This is an advantage over the deterministic forecasts because the uncertainty inherent in forecasts is quantitatively expressed in the probability statements. They especially addressed the issue regarding the unreliable records by introducing an informative prior for the coefficient parameters of the model via a bootstrap procedure. With a similar Bayesian regression model, more recently Elsner and Jagger (2006) attempted to predict annual U.S. hurricane counts. The model includes predictors representing the North Atlantic Oscillation (NAO), the Southern Oscillation (SO), the Atlantic multidecadal oscillation, as well as an indicator variable that is either 0 or 1 depending on the time period specified. As a baseline for comparison,

a climatology model is used, which contains only the regression constant (i.e., intercept) and the indicator variable. In an out-of-sample cross-validation test, the Bayesian model appears to have a lower mean-squared error relative to climatology for years in which there are exactly zero, three, five, or more hurricanes observed.

Chu and Zhao (2004) applied a Bayesian analysis to detect changepoints in the TC series over the central North Pacific (CNP). More recently, Zhao and Chu (2006) developed a more advanced method for detecting multiple changepoints in hurricane time series for the eastern North Pacific. In both of these two studies, the annual TC counts are described by a Poisson process where the Poisson intensity is conditional on a gamma distribution. A hierarchical Bayesian approach is applied to make inferences, in terms of the posterior probabilities, for shifts in the TC time series. In view of the probabilistic nature of the Bayesian paradigm, this study will apply the similar Bayesian framework advocated by Elsner and Jagger (2004, 2006) to forecasting the seasonal TC activity over the CNP prior to the peak hurricane season. To complete the Bayesian model, we apply noninformative priors to the model parameters.

The structure of this paper is as follows. The data used and the mathematical model of the TC counts are covered in sections 2 and 3, respectively. The key concepts of the Markov chain Monte Carlo (MCMC) and Gibbs sampler are introduced in section 4, based on which algorithm is developed to fulfill the Bayesian inference of our proposed probabilistic model. Section 5 describes the procedure to select the appropriate predictors for the TC count series in the CNP case. Results are presented in section 6. The conclusions are found in section 7.

2. Data

The tropical storm (maximum sustained surface wind speeds between 17.5 and 33 m s^{-1}) and hurricane (wind speed at least 33 m s^{-1}) records over the CNP from the National Hurricane Center's best-track data are used (Chu 2002). Here TC refers to only tropical storms and hurricanes. The period of analysis is 1966–2003. The records prior to 1966 are thought to be less reliable because satellite observations were not in sufficient quantities. The domain of the CNP coincides with the area of responsibility of the Central Pacific Hurricane Center, an entity of the National Weather Service in Honolulu, Hawaii. Two types of TCs appear in the CNP. The TC counts include storms that form within the CNP as well as those that form in the eastern North Pacific and subsequently propagate into the CNP.

Monthly mean sea level pressure (SLP), wind data at the 1000-, 850-, and 200-hPa levels, relative vorticity data at the 850-hPa level, and total precipitable water (PW) are derived from the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis dataset (Kistler et al. 2001). The horizontal resolution of the reanalysis dataset is 2.5° latitude–longitude. Tropospheric VWS is computed as a square root of the sum of the squared difference of the zonal wind component between 200 and 850 hPa and the squared difference of the meridional wind component between 200 and 850 hPa (Clark and Chu 2002). The monthly mean sea surface temperatures (SSTs) over the North Pacific are taken from Reynold's reconstruction of the Comprehensive Ocean–Atmosphere Data Set, as detailed in Smith et al. (1996). SST data are available on a 2° latitude–longitude resolution. Chu (2002) used the reanalysis and the reconstructed SST datasets to investigate circulation features associated with decadal variations of tropical cyclone activity over the central North Pacific.

3. The Bayesian regression model for TC counts

The Poisson process is a proper probability model for describing independent, rare event counts. Given the Poisson intensity parameter λ (i.e., the mean seasonal TC rates), the probability mass function (PMF) of h TCs occurring in a unit of observation time (e.g., one season) is (Epstein 1985)

$$P(h|\lambda) = \exp(-\lambda)\frac{\lambda^h}{h!}, \quad \text{where}$$

$$h = 0, 1, 2, \ldots \quad \text{and} \quad \lambda > 0. \tag{1}$$

The Poisson mean is simply λ, thus, so is its variance.

In the context of building a regression model, through which one can develop the relationship between the target response variable, the seasonal TC counts, and the selected predictors, the Poisson rate λ is usually treated as a random variable that is conditional on the predictors.

In this study, we adopt the Poisson linear regression model. Assume there are N observations and for each observation there are K relative predictors. We define a latent random N-vector $\mathbf{Z}$, such that for each observation h_i, $i = 1, 2, \ldots, N$, $Z_i = \log\lambda_i$, where λ_i is the relative Poisson intensity for the ith observation. Here N denotes the sample size, which is 38 in this study (1966–2003). The link function between this latent variable and its associated predictors is expressed as $Z_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$, where $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_K]'$ is a random

vector; noise ε_i is assumed to be identical and independent distributed (IID) and normally distributed with zero mean and σ^2 variance; $\mathbf{X}_i = [1, X_{i1}, X_{i2}, \ldots, X_{iK}]$ denotes the predictor vector. In the vector form, the general Poisson linear regression model can be formulated as below:

$$P(\mathbf{h}|\mathbf{Z}) = \prod_{i=1}^{N} P(h_i|Z_i), \text{ where } h_i|Z_i \sim \text{Poisson}(h_i|e^{Z_i})$$

$\mathbf{Z}|\boldsymbol{\beta}, \sigma^2, \mathbf{X} \sim \mathbf{Normal}(\mathbf{Z}|\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N)$, where, specifically $\mathbf{X}' = [\mathbf{X}_1', \mathbf{X}_2', \ldots, \mathbf{X}_N']$, $\mathbf{I}_N$ is the $N \times N$ identity matrix, and $\mathbf{X}_i = [1, X_{i1}, X_{i2}, \ldots, X_{iK}]$ is the predictor vector for h_i, $i = 1, 2, \ldots, N$

$$\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_K]'. \quad (2)$$

Here, Normal and Poisson stand for the normal distribution and the Poisson distribution, respectively. In Eq. (2), β_0 is referred to as the intercept.

It is worth noting that the Poisson rate λ is a real value while the TC counts h is only an integer. Accordingly λ contains more information relative to h. Furthermore, because h is conditional on λ, λ is subject to less variations than h. Taken together, λ should be preferred as the forecast quantity of the TC activity than h for decision making. We also notice the fact that this hierarchical structure essentially fits well for Bayesian inference.

4. MCMC approach to the Bayesian inference

a. General idea of MCMC and Gibbs sampler

In general, we assume the model is given and denote the set of parameters of this model by the vector $\boldsymbol{\theta}$. The data for training analysis is symbolized by $\mathbf{h}$. Thus, the basic Bayesian formula is described as

$$P(\boldsymbol{\theta}|\mathbf{h}) = \frac{P(\mathbf{h}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int P(\mathbf{h}|\boldsymbol{\theta})P(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \propto P(\mathbf{h}|\boldsymbol{\theta})P(\boldsymbol{\theta}), \quad (3)$$

where "$\propto$" means "proportional" since $\boldsymbol{\theta}$ in the denominator is only a dummy variable. In Eq. (3), $P(\mathbf{h}|\boldsymbol{\theta})$ is the conditional distribution of data $\mathbf{h}$ given the model parameters $\boldsymbol{\theta}$ (i.e., the likelihood) and $P(\boldsymbol{\theta})$ is a prior distribution. Equation (3) provides the inference for posterior distribution $P(\boldsymbol{\theta}|\mathbf{h})$, the probability of $\boldsymbol{\theta}$ after the data $\mathbf{h}$ are observed. It is clear that data affect the posterior distribution only through the likelihood function $P(\mathbf{h}|\boldsymbol{\theta})$. To make predictive inference, we rely on the posterior predictive distribution:

$$P(\tilde{\mathbf{h}}|\mathbf{h}) = \int P(\tilde{\mathbf{h}}|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{h})\, d\boldsymbol{\theta}, \quad (4)$$

where $\tilde{\mathbf{h}}$ denotes the prediction (Gelman et al. 2004). Here $P(\tilde{\mathbf{h}}|\mathbf{h})$ is the posterior predictive distribution since it is conditional on the observed data $\mathbf{h}$ and provides a prediction for the unknown observable $\tilde{\mathbf{h}}$. This formula is at the heart of Bayesian analysis.

The MCMC approach is one of the efficient algorithms for Bayesian inference. The general Bayesian analysis method described above essentially involves integrating the posterior expectation:

$$E[a|\mathbf{h}] = \int_{\boldsymbol{\theta}} a(\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathbf{h})\, d\boldsymbol{\theta},$$

where $a(\boldsymbol{\theta})$ can be of any function conditional on the model parameters $\boldsymbol{\theta}$. This expectation, however, is very difficult to integrate in most models. Alternatively, a numerical way to calculate such an expectation is to use Monte Carlo integration by

$$E[a|\mathbf{h}] \approx \frac{1}{L}\sum_{i=1}^{L} a(\boldsymbol{\theta}^{[i]}),$$

where $\boldsymbol{\theta}^{[1]}, \boldsymbol{\theta}^{[2]}, \ldots, \boldsymbol{\theta}^{[L]}$ are independently drawn from $P(\boldsymbol{\theta}|\mathbf{h})$. When the sample size L is large enough, this approximation will converge to its analytical integral.

This method is straightforward, but practically it is often infeasible to generate such an independent series $\boldsymbol{\theta}^{[1]}, \boldsymbol{\theta}^{[2]}, \ldots, \boldsymbol{\theta}^{[L]}$ when $P(\boldsymbol{\theta}|\mathbf{h})$ is complicated. Nonetheless, in many applications, it may be possible to generate a series of dependent values by using a Markov chain (MC) that has $P(\boldsymbol{\theta}|\mathbf{h})$ as its stationary distribution. The MC is defined by giving an initial distribution for the first state of the chain $\boldsymbol{\theta}^{[1]}$ and a set of transition probabilities for a new state $\boldsymbol{\theta}^{[i+1]}$ that is conditional on the current state $\boldsymbol{\theta}^{[i]}$. Under very general conditions (i.e., the MC is ergodic), the distribution for the state will converge to a unique stationary distribution. Obviously, if this stationary distribution is $P(\boldsymbol{\theta}|\mathbf{h})$, the Monte Carlo integration described above still gives an unbiased estimate for $E[a|\mathbf{h}]$ (Ripley 1987).

One of the most widely used MCMC algorithms is known as the Gibbs sampler. Suppose there are p components of the model parameter vector, defined as $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]'$. (Note that θ_i could be a vector.) Presumably, directly sampling from the posterior distribution $P(\boldsymbol{\theta}|\mathbf{h})$ is unlikely; however, we can generate a value from the conditional distribution for one part of the $\boldsymbol{\theta}$ given the values of the rest of other parts of $\boldsymbol{\theta}$. In detail, Gibbs sampling involves successive drawing from the complete conditional posterior densities $P(\theta_k|\mathbf{h}, \theta_1, \ldots, \theta_{k-1}, \theta_{k+1}, \ldots, \theta_p)$ for k from 1 to p.

The Gibbs sampling algorithm is described as follows:

1) Choose arbitrary starting values: $\boldsymbol{\theta}^{[0]} = [\theta_1^{[0]}, \theta_2^{[0]}, \ldots, \theta_p^{[0]}]$.
2) Start at $j = 1$ and complete the single cycle by drawing values from the p distributions given by
3) Set $j = j + 1$ and go back to step 2 until meeting the required number of iterations.

$$\begin{aligned}
\theta_1^{[j]} &\sim P(\theta_1|\mathbf{h}, \theta_2^{[j-1]}, \theta_3^{[j-1]}, \ldots, \theta_{p-1}^{[j-1]}, \theta_p^{[j-1]}),\\
\theta_2^{[j]} &\sim P(\theta_2|\mathbf{h}, \theta_1^{[j]}, \theta_3^{[j-1]}, \ldots, \theta_{p-1}^{[j-1]}, \theta_p^{[j-1]}),\\
\theta_3^{[j]} &\sim P(\theta_3|\mathbf{h}, \theta_1^{[j]}, \theta_2^{[j]}, \ldots, \theta_{p-1}^{[j-1]}, \theta_p^{[j-1]}),\\
&\ldots\\
\theta_{p-1}^{[j]} &\sim P(\theta_{p-1}|\mathbf{h}, \theta_1^{[j]}, \theta_2^{[j]}, \ldots, \theta_{p-2}^{[j]}, \theta_p^{[j-1]}),\\
\theta_p^{[j]} &\sim P(\theta_p|\mathbf{h}, \theta_1^{[j]}, \theta_2^{[j]}, \ldots, \theta_{p-2}^{[j]}, \theta_{p-1}^{[j]}). \qquad (5)
\end{aligned}$$

Once convergence is reached, we can approximate $E[a|\mathbf{h}]$ by $(1/L)\sum_{i=1}^{L} a(\boldsymbol{\theta}^{[i]})$ with a large enough sample size L, where $\boldsymbol{\theta}^{[i]}$ is the ith sample drawn from the Gibbs sampler descried by (5) within each iteration after convergence.

b. Gibbs sampler for the Bayesian inference of the TC count model

Let us first derive the posterior distribution for the model given by (2). Since we do not have any credible prior information for the coefficient vector $\boldsymbol{\beta}$ and the variance σ^2, it is reasonable to choose the noninformative prior. In formula, it is (Gelman et al. 2004, p. 355)

$$P(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2}. \qquad (6)$$

This is not a proper distribution function; however, it leads to a proper posterior distribution.

Equation (A2) in the appendix implies that, the posterior distribution of any hidden variable Z is conditionally independent from each other given the model parameters $\boldsymbol{\beta}$ and σ^2. Therefore, with the newly observed predictor set $\tilde{\mathbf{X}} = [1, \tilde{X}_{i1}, \tilde{X}_{i2}, \ldots, \tilde{X}_{iK}]$, the predictive distribution for the latent variable $\tilde{Z}$ and TC counts $\tilde{h}$ will be

$$P(\tilde{Z}|\tilde{X}, \mathbf{X}, \mathbf{h}) = \int\int_{\boldsymbol{\beta},\sigma^2} P(\tilde{Z}|\tilde{X}, \boldsymbol{\beta}, \sigma^2) P(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{h}) d\boldsymbol{\beta} d\sigma^2, \qquad (7a)$$

$$P(\tilde{h}|\tilde{X}, \mathbf{X}, \mathbf{h}) = \int_{\tilde{Z}} \frac{\exp(-e^{\tilde{Z}} + \tilde{Z}\tilde{h})}{\tilde{h}!} P(\tilde{Z}|\tilde{X}, \mathbf{X}, \mathbf{h}) d\tilde{Z}. \qquad (7b)$$

Even with the noninformative prior, the posterior distribution for the model parameter set $(\boldsymbol{\beta}, \sigma^2)$ in (7) is still not standard and directly sampling from it is difficult. In this section, we will design a Gibbs sampler, which has $P(\boldsymbol{\beta}, \sigma^2|\mathbf{X}, \mathbf{h})$ as its stationary distribution, and then we can use an alternative approach to integrate (7) by

$$P(\tilde{Z}|\tilde{X}, \mathbf{X}, \mathbf{h}) = \frac{1}{L}\sum_{i=1}^{L} P(\tilde{Z}|\tilde{X}, (\boldsymbol{\beta}, \sigma^2)^{[i]}), \qquad (8a)$$

$$P(\tilde{h}|\tilde{X}, \mathbf{X}, \mathbf{h}) = \frac{1}{L}\sum_{i=1}^{L} \frac{\exp(-e^{\tilde{Z}^{[i]}} + \tilde{Z}^{[i]}\tilde{h})}{\tilde{h}!}, \qquad (8b)$$

where $(\boldsymbol{\beta}, \sigma^2)^{[i]}$ is the ith sampling from the proposed Gibbs sampler after the burn-in period, $\tilde{Z}^{[i]}$ is sampled from $\tilde{Z}^{[i]}|\tilde{X}, (\boldsymbol{\beta}, \sigma^2)^{[i]} \sim \text{Normal}(\tilde{Z}^{[i]}|\tilde{X}\boldsymbol{\beta}^{[i]}, \sigma^{2[i]})$ subsequently, and L is a large enough number.

Based on the inference analysis derived in the appendix, our proposed Gibbs sampler yields the following sequence:

1) Select proper initial value for $\mathbf{Z}^{[0]}$, $\boldsymbol{\beta}^{[0]}$, $\sigma^{2[0]}$ and set $i = 1$.
2) Draw $Z_j^{[i]}$ from $Z_j^{[i]}|\mathbf{h}, \boldsymbol{\beta}^{[i-1]}, \sigma^{2[i-1]}$ for $j = 1, 2, \ldots, N$ via (A3).
3) Draw $\boldsymbol{\beta}^{[i]}$ from $\boldsymbol{\beta}^{[i]}|\mathbf{h}, \mathbf{Z}^{[i]}, \sigma^{2[i-1]}$ via (A6).
4) Draw $\sigma^{2[i]}$ from $\sigma^{2[i]}|\mathbf{h}, \mathbf{Z}^{[i]}, \boldsymbol{\beta}^{[i]}$ via (A7).
5) Set $i = i + 1$ then go back to step 2 until meeting the required number of iterations. (9)

With the observation data $\mathbf{h}$ and following the algorithm presented in (9), after the burn-in period, one can sample set $\mathbf{Z}$, $\boldsymbol{\beta}$, σ^2 within each iteration, which will have the desired posterior distribution that facilitates the numerical computation of (8a) and (8b).

A practical issue in the step 2 of algorithm (9) is that the distribution governed by Eq. (A3) is not standard. We resort to the Metropolis–Hasting algorithm in this study, which is relatively computationally expensive. Some other approaches can be considered here. For example, based on our simulation results, the estimated posterior PDFs for the hidden variables are all Gaussian like, which theoretically can also be easily proven log-concave. Therefore, using Laplace approximation in this context should also be a sound choice.

5. Procedures for selecting predictors

One of the challenges in this study is to find appropriate predictors to be applied in the Bayesian regression model. For predicting the North Atlantic seasonal hurricane counts, predictor variables such as the stratospheric zonal wind component, the El Niño index, and

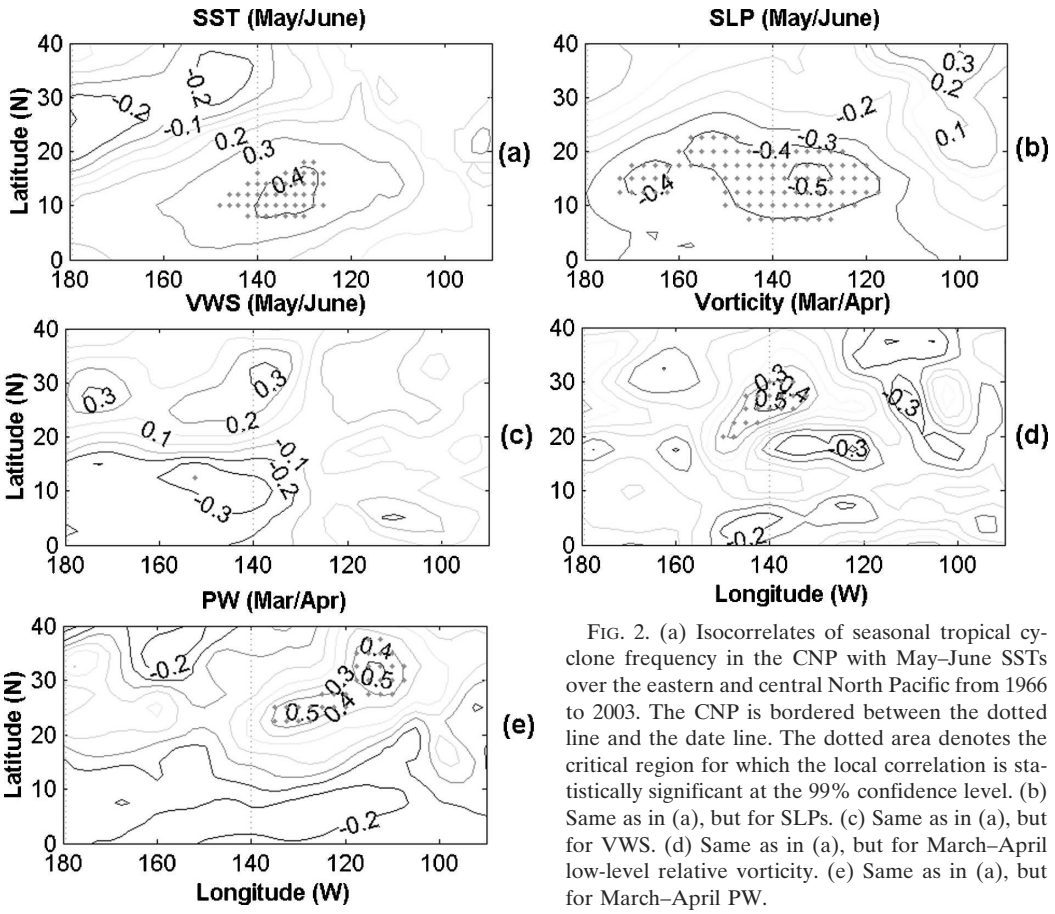

FIG. 2. (a) Isocorrelates of seasonal tropical cyclone frequency in the CNP with May–June SSTs over the eastern and central North Pacific from 1966 to 2003. The CNP is bordered between the dotted line and the date line. The dotted area denotes the critical region for which the local correlation is statistically significant at the 99% confidence level. (b) Same as in (a), but for SLPs. (c) Same as in (a), but for VWS. (d) Same as in (a), but for March–April low-level relative vorticity. (e) Same as in (a), but for March–April PW.

the west Sahel rainfall, among others, have been suggested by many researchers (e.g., Gray et al. 1992; Elsner and Schmertmann 1993; Klotzbach and Gray 2004). For the CNP, it is less clear as to which predictors are vital. Without a priori knowledge of the information needed, we resort to calculate the Pearson correlation between the TC counts in the peak season (July–September) and the preseason environmental variables, including SST, SLP, PW, low-level relative vorticity, and VWS over the central and eastern North Pacific (0°–40°N, 180°–90°W). Historically, there is not a single TC occurrence on record over the CNP in May or June.

For each of the candidate environmental variables, we identify a critical region using the following procedure. For any grid point over the eastern and central North Pacific, if the Pearson correlation between the predictor and the seasonal TC frequency is statistically significant, this point is adopted. Based on the linear regression theory, for a sample size of 38, the critical value for a correlation coefficient is 0.38 at the 99% confidence level (Bevington and Robinson 2003). Accordingly, a correlation coefficient with its absolute value greater than 0.38 at a grid point will be deemed locally significant and be selected as a critical region. To keep the results robust, a simple average of the predictor variable over the critical region is chosen.

a. SSTs

SSTs are known to be important for TC formation and intensification. Warmer SSTs are expected to fuel the overlying atmosphere with additional warmth and moisture, thereby reducing atmospheric stability and increasing the likelihood of deep tropical convection. The isocorrelate map of seasonal TC frequency over the CNP and SSTs during the antecedent May–June is displayed in Fig. 2a, in which the area with strong positive correlations (with a maximum of 0.44) is found over the tropical eastern North Pacific (near 14°N, 132°W) and the identified critical region is marked by dots. Thus, the average of the SST series over this critical region is chosen as a predictor.

b. SLPs

The contour plot for the correlation between the TC frequency and the SLPs during the precedent May–

June is shown in Fig. 2b, in which strong negative correlations (with a minimum of −0.52 near 15°N, 132.5°W) are found over the tropical eastern Pacific. That is, lower SLPs over the eastern Pacific in the preceding May–June correspond to high TC frequency over the CNP. This result is physically reasonable. Lower SLP implies decreased subsidence, which would result in weaker trade wind inversion (Knaff 1997). Because the trade wind inversion acts as a lid to atmospheric convection, weaker inversion would promote deeper convection. The occurrence of deep convection is important for TC formation because it provides a vertical coupling between the upper-level outflow and lower-tropospheric inflow circulations.

c. VWS

Strong VWS disrupts the organized deep convection (the so-called ventilation effect) that inhibits intensification of the TCs. Negative and significant correlations (with minimum −0.38) exist between TC frequency and May–June VWS in the low latitudes with a center near 12.5°N, 152.5°W (Fig. 2c).

d. Relative vorticity

Figure 2d shows the correlations between TC frequency and the relative vorticity in the preceding March–April. It displays positive and significant correlations near the area 25°N, 140°W (with a maximum of 0.52). For May–June, the correlation between TC frequency and relative vorticity at the aforementioned area are dropped noticeably. Thus, only the March–April predictor is used.

e. PW

In Fig. 2e, positive and significant correlations between TC frequency and PW in the antecedent March–April are found in the eastern North Pacific, where the correlation coefficient reaches as high as 0.51 at 25°N, 130°W. Collins and Mason (2000) also found strong relationships between TC indices and PW in the western portion of the eastern North Pacific. Adequate moisture in the atmosphere provides a fundamental ingredient for deep convection. Conversely, drier atmosphere tends to suppress deep convection and inhibits TC activity. Just like the vorticity, the correlation between TC counts and PW in May–June in the critical region becomes much lower; therefore, the average of the March–April PW over the identified critical area in Fig. 2e is employed.

f. An overall model

With the predictors selected through correlation analysis, the overall regression model for this study can be exactly specified. From (2), the predictor vector $\mathbf{X}_i$ and the associated coefficient parameter vector $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \ldots, \beta_K]'$ is given by

$$\mathbf{X}_i = [1, \mathrm{SST}_i, \mathrm{SLP}_i, \mathrm{VWS}_i, \mathrm{Vorticity}_i, \mathrm{PW}_i],$$
$$i = 1, 2, \ldots, N \quad \text{and}$$
$$\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5]'. \tag{10}$$

In practical applications, it is desirable to normalize each predictor series before further analysis to avoid the scaling problem among the different predictors.

6. Results

As mentioned in section 2, there are a total 38 yr (1966–2003) of TC counts in the CNP. We apply the Gibbs sampler outlined in (9) to this dataset. For simplicity, for all the simulations in this study, we take the first 2000 samples as burn-in and use the following 10 000 samples as the output of the Gibbs sampler.

In all, we have five predictors (besides the intercept term, which does not vary within different seasons). A general way to verify the effectiveness of a regression method is to apply a strict cross-validation (CV) test for the relevant dataset. Considering the fact that the TC variation is approximately independent from year to year, it is proper to apply the leave-one-out cross validation (LOOCV) in this context. That is, a target year is chosen and a model is developed using the remaining 37-yr data as the training set. The observations of the selected predictors for the target year are then used as inputs to forecast the missing year. This process is repeated successively until all 38 forecasts are made.

Because cross validation applied in the Bayesian regression model is relatively novel (Elsner and Jagger 2006), it warrants more explanations. For a target year, say the ith year, we use the other $N - 1$ yr's observation as the training data for the input of algorithm (9) and obtain the output as the posterior sampling of the model parameter $\boldsymbol{\beta}$ and σ^2 for each iteration after the burn-in period. Given the predictor set for the target year, (SST, SLP, VWS, Vorticity, PW), we can sample the corresponding latent variable Z, which is the natural logarithm of the relative TC rate λ, from a normal distribution with $\beta_0 + \beta_1(\mathrm{SST}) + \beta_2(\mathrm{SLP}) + \beta_3(\mathrm{VWS}) + \beta_4(\mathrm{Vorticity}) + \beta_5(\mathrm{PW})$.This leads to the posterior sample of $\lambda = \exp(Z)$. Subsequently, we use Eq. (8b) to calculate the posterior predictive distribution of the TC counts (i.e., h) for this target year.

With all the samples drawn, we can estimate any statistic deemed as important. To demonstrate this, the median, upper, and lower quartiles (the upper 75% and lower 25%) of the predicted TC rates, through a

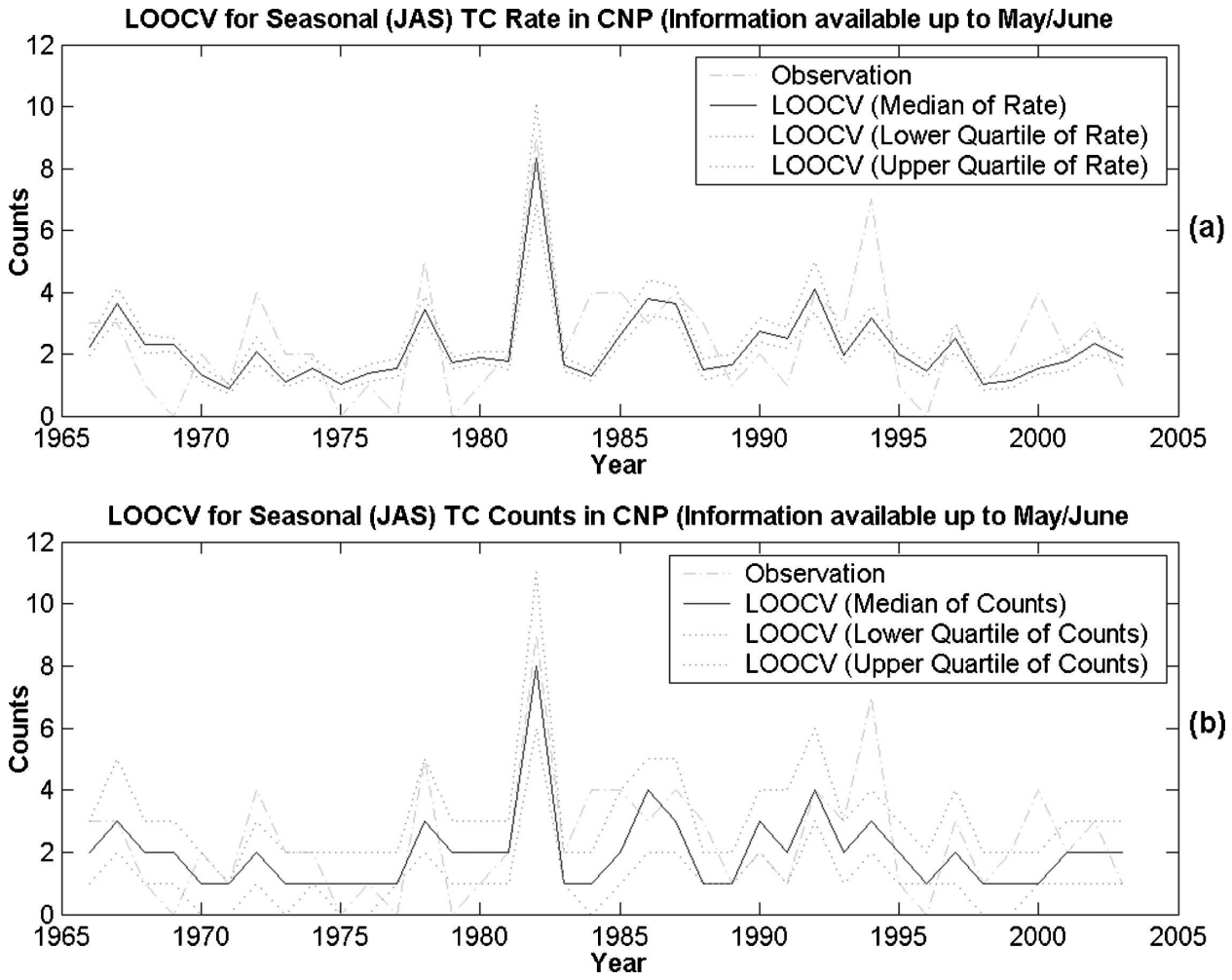

FIG. 3. (a) The median (solid line), upper, and lower quartiles (dotted line) of the LOOCV-predicted TC rate are plotted together with the actual observed TC counts (dash–dotted line) during 1966–2003. (b) Same as in (a), but for the actual observations (dash–dotted line) during 1966–2003.

LOOCV, are plotted together with the actual observation for each year in Fig. 3a. The distance between the upper and lower quartile locates the central 50% of the predicted TC variations. The Pearson correlation between the median of predictive rate and independent observations is 0.63. In Fig. 3b, the median, upper, and lower quartiles of the predicted TC counts are plotted together with the actual observation for each year. Out of a total of 38 yr, there are only 9 yr in which the actual TC counts lie outside the predictive central 50% boundaries.

Furthermore, using all 38-yr observations as training data, we run the proposed algorithm again. As an illustration, we plot the first 5000 samples of the coefficient parameters, β_1, β_2, β_3, β_4, β_5, in Figs. 4a,c,e,g,i, respectively; and their relative autocorrelations are displayed in Figs. 4b,d,f,h,j, respectively. The autocorrelation of the samples for each parameter reaches zero very quickly, implying the output of the proposed Gibbs sampler is the unbiased samples drawn from their joint posterior distribution.

With all the samples, we also calculated the kernel estimated marginal probability density function (PDF) for the parameter set, $\boldsymbol{\beta}$ and σ, by convolving the resulting frequency of the target samples with a smoothing filter. The marginal posterior PDF for each model parameter, as shown in Fig. 5, also yields some useful information. The relative contribution of each regression coefficient in the Bayesian strategy can be judged approximately by the so-called p value. This can be evaluated by the ratio of the number of samples that lie to the left of zero to the total number of iterations if the predictor is expected to have a positively orientated impact on the forecast quantity (e.g., SST). Conversely,

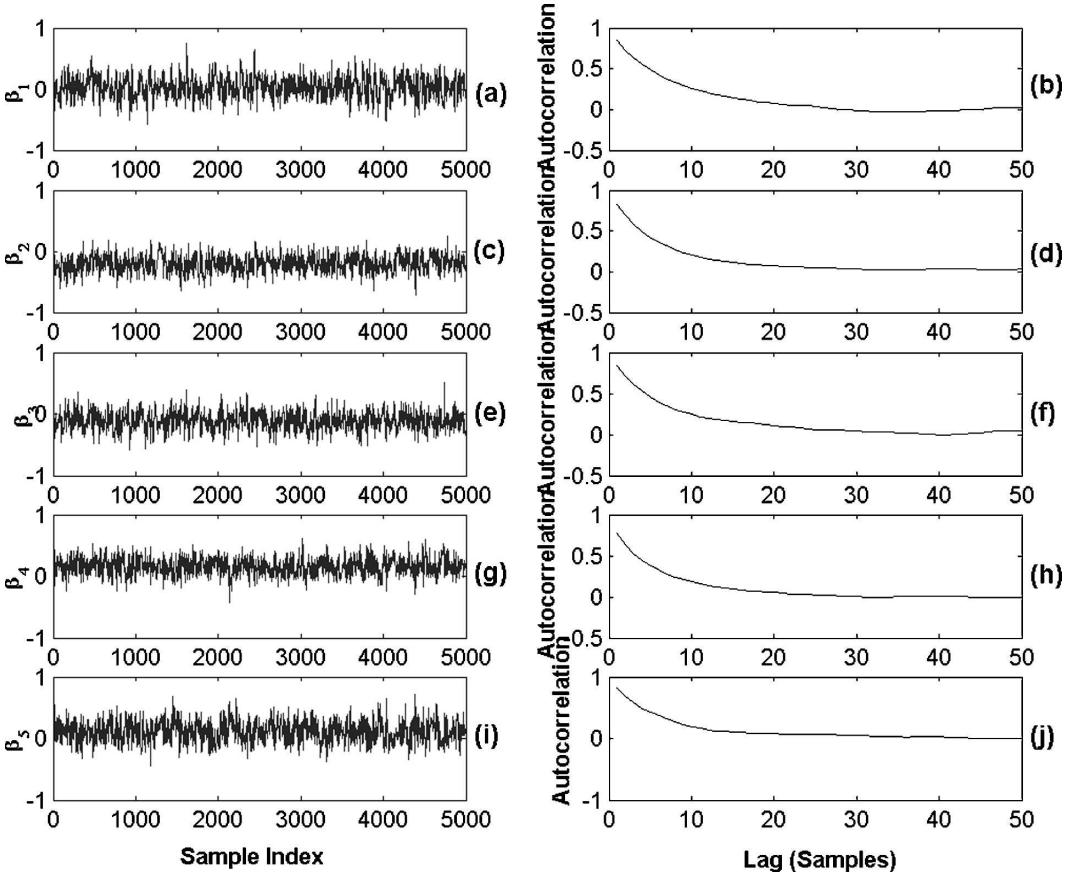

FIG. 4. Plots of the first 5000 samples of (a),(c),(e),(g),(i) each coefficient parameter and (b),(d),(f),(h),(j) their associated autocorrelation coefficients: (a), (b) SST; (c), (d) SLP; (e), (f) VWS; (g), (h) relative vorticity; and (i), (j) PW.

if the predictor is to have a negatively orientated impact (e.g., SLP), the number of samples that lie to the right of zero to the total number of iterations is of concern. Graphically, the smaller the area to the left (positively oriented predictors) or right (negatively oriented predictors) of zero in the PDF plots, the more important this predictor is in the regression model. Figure 5 indicates the SLP and, to a lesser extent, the vorticity as key predictors.

Further insight into the quantitative relationship between TC activity and environmental variables can be gained by computing the TC rate changes given a unit change in each of the predictors. From Eqs. (2) and (10), the influence of the ith climate variable on TC rate is indicated by the value of its coefficient parameter β_i. Because the latent variable Z is expressed as a natural log of the TC rate λ, a unit increase in the predictor variable will marginally lead to an $\exp(\beta_i)$ fold change in TC rate. To illustrate this, we will consider the variable SST. The sample mean of the SST predictor is 26.52°C and its sample standard deviation is 0.46°C (Table 1). If the predicted SST is increased over the mean by 1°C, the predictive TC rate will increase by 6.9%. Sometimes, one would like to consider the scale of predictors as well, in which case the variation of a predictor is preferably measured by its standard deviation. For the SST predictor, an increase of the TC rate by 3.1% is anticipated if the new observed SST predictor is 26.98°C, which is one standard deviation over the mean. Likewise, for SLP, an increase of 1 hPa over the mean results in a 30.1% decrease in TC rates. Although the 1-hPa change seems small, it is considerable when viewed in the context of the seasonal mean condition in the Tropics. As gleaned from columns 4 and 5 in Table 1, SST, relative vorticity, and PW are more likely to appear as positive factors modulating the seasonal TC frequency, while SLP and VWS act to affect the TC activity in a negative way.

For operational settings, it is desirable to have predictors selected prior to May–June so decision makers in relevant agencies could have longer lead times to respond to potential hazards. In this regard, we have

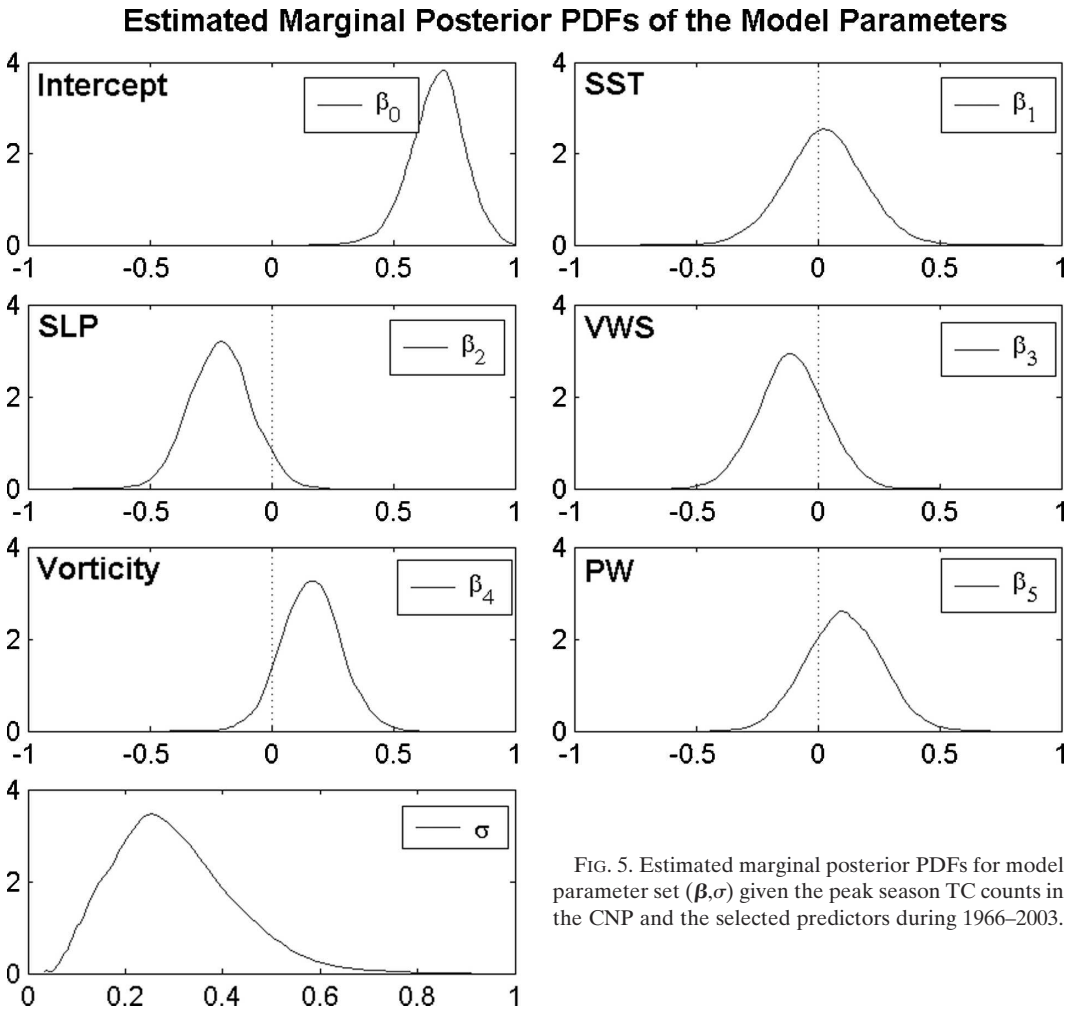

FIG. 5. Estimated marginal posterior PDFs for model parameter set ($\boldsymbol{\beta}$,σ) given the peak season TC counts in the CNP and the selected predictors during 1966–2003.

performed a separate analysis using only predictors of March–April while keeping the July–September TC frequency as the predictand. The Pearson correlation between the median of predictive rate and independent observations is 0.50. Out of 38 yr, there are 13 yr in which the true observations lay outside the lower/upper quartile bounds of the predictions. For both measures, predictions using data only for March–April are not as good as those that considering all predictors up to May–June. This result is reasonable as predictors immediately preceding the peak hurricane season normally yield more useful forecast skill.

7. Summary

Being able to forecast seasonal TC counts accurately before its peak season is important. In this study, we apply a Poisson generalized linear regression model to the historical, seasonal TC counts over the central North Pacific (CNP) and select the preseason SST, SLP, PW, relative vorticity, and VWS as predictors. Using a simple correlation analysis, critical regions over the eastern and central North Pacific are identified for each environmental variables. We then derive a Bayesian inference for this model by assuming a noninformative prior. An MCMC approach is adopted to numerically analyze the data. Through a Gibbs sampler, we are able to forecast the probabilistic distribution of TC activity over the CNP prior to the peak season. When tested for the period 1966–2003, the leave-one-out cross-validation correlation test delivers satisfactory results as seen in section 6.

The Bayesian regression model developed in this study is valuable. First, it is physically based so it is perhaps easier to interpret the success or failure of the forecast. Second, forecasts of seasonal TC counts are presented in probabilistic format, which is preferred since it gives the uncertainty of the prediction. In addition, the proposed hierarchical probabilistic model can

TABLE 1. Bayesian analysis results. The name of each predictor is labeled in the first column. The second and the third columns refer to the mean and the std dev of each predictor for the period 1966–2003, respectively. The fourth column (%/Unit) denotes the percentage change in TC rates corresponding to change by one unit relative to its mean for each predictor. Similarly, the fifth column (%/Std dev) denotes the percentage change in TC rates corresponding to change by one std dev for each predictor. The positive (or negative) sign in the fourth and the fifth columns refers to increasing (or decreasing) TC rates. The units for SST, SLP, PW, relative vorticity, and VWS are °C, hPa, kg m^{-2}, 10^{-6} s^{-1}, and m s^{-1}, respectively. For SLP, the raw values have been subtracted by 1000 hPa.

Predictor	Mean	Std dev	%/Unit	%/Std dev
SST	26.52	0.46	6.9%	3.1%
SLP	14.17	0.54	−30.1%	−17.5%
VWS	27.19	3.48	−3.0%	−10.1%
Vorticity	−8.66	1.94	9.1%	18.5%
PW	13.87	1.12	8.7%	9.8%

serve as the perfect platform for further researches because any probabilistic model can be treated as an independent modulo and seamlessly plugged into it under Bayesian framework. For example, in this study we assume the link function between the natural logarithm of the TC rate and predictors be linear, which, however, is not necessarily the best assumption. In principle, this model can be extended to a nonlinear link function via a proper nonlinear probabilistic model such as kernel-based Gaussian processes. Obviously, the predictor selection procedures are also needed to be revised accordingly in this regard. This promising approach, nevertheless, is beyond the scope of this study.

Currently, the National Oceanic and Atmospheric Administration (NOAA) Climate Prediction Center (CPC) is issuing their central Pacific hurricane outlook but the method is a mix based on guidance and experience (i.e., subjective). The Bayesian probabilistic model outlined in this study could serve as an objective and additional tool to forecasters at the CPC and the NWS Forecast Office in Honolulu, Hawaii. This tool, together with others, will enable operational forecasters and researchers working together to finalize the official NOAA outlook for the central Pacific hurricane season. In the meantime, a close dialogue with forecast users (e.g., State Civil Defense) is envisioned so researchers may explain the limitations and challenges of the predictive research to users. For example, would the central 50% range of the predictive TC variations shown in Fig. 3 be useful for users? If not, what would be the alternative? Through this user involvement, it is hoped that the utility of hurricane climate forecasts could be enhanced by systematic efforts to bring scientific research and users' needs together.

Acknowledgments. Constructive criticisms from two anonymous reviewers helped to improve the presentation of this paper. Thanks to Di Henderson for technical editing. This study was partially supported by a grant from NOAA to the Hawaii Coastal Zone Management program of the Department of Business, Economic Development, and Tourism.

APPENDIX

Conditional Posterior Distribution for a Poisson Regression Model

For the sake of simplicity, in the following derivation, we will drop the notation of the predictor matrix $\mathbf{X}$, which is always given by default.

Based on Eqs. (3) and (2), it is obvious that

$$P(\mathbf{Z}|\mathbf{h}, \boldsymbol{\beta}, \sigma^2) \propto P(\mathbf{h}|\mathbf{Z}, \boldsymbol{\beta}, \sigma^2)P(\mathbf{Z}|\boldsymbol{\beta}, \sigma^2)$$
$$= P(\mathbf{h}|\mathbf{Z})P(\mathbf{Z}|\boldsymbol{\beta}, \sigma^2). \quad \text{(A1)}$$

Substituting the probability model (2) into (A1) and ignoring the constant part yields

$$P(\mathbf{Z}|\mathbf{h}, \boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^N}\prod_{i=1}^{N} \exp\left[-e^{Z_i} + Z_i h_i - \frac{1}{2\sigma^2}(Z_i - \mathbf{X}_i\boldsymbol{\beta})^2 \right]. \quad \text{(A2)}$$

This is not a standard density distribution, but we can design a Gibbs sampler through which the output of each of its iteration will be of the distribution given by (A2).

From (A2), it is easy to see that Z_i is conditionally independent from each other for $i = 1, 2, \ldots N$ given $\boldsymbol{\beta}$ and σ^2; therefore, sampling from $Z_i|\mathbf{h}, \boldsymbol{\beta}, \mathbf{Z}_{-i}, \sigma^2$, where $\mathbf{Z}_{-i} = [Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_N]'$, is equivalently sampling from $Z_i|\mathbf{h}, \boldsymbol{\beta}, \sigma^2$. We ignore the constant part and obtain

$$P(Z_i|\mathbf{h}, \boldsymbol{\beta}, \sigma^2) \propto \exp\left[-e^{Z_i} + Z_i h_i - \frac{1}{2\sigma^2}(Z_i - \mathbf{X}_i\boldsymbol{\beta})^2 \right], \quad i = 1, 2, \ldots N. \quad \text{(A3)}$$

To sample Z_i from (A3), in this paper we apply the Metropolis–Hasting algorithm. One can refer to Ripley (1987), Gelman et al. (2004), or originally Hastings (1970) for the details of this algorithm.

After the latent vector $\mathbf{Z}$ is obtained, the model is exactly the same as the so-called ordinary linear regression and its Bayesian inference derivation is straightforward. The joint posterior distribution for $(\mathbf{Z}, \boldsymbol{\beta}, \sigma^2)$ can be expressed as

$$P(\mathbf{Z}, \boldsymbol{\beta}, \sigma^2|\mathbf{h}) \propto P(\mathbf{h}|\mathbf{Z}, \boldsymbol{\beta}, \sigma^2)P(\mathbf{Z}, \boldsymbol{\beta}, \sigma^2) = P(\mathbf{h}|\mathbf{Z})P(\mathbf{Z}|\boldsymbol{\beta}, \sigma^2)P(\boldsymbol{\beta}, \sigma^2). \quad \text{(A4)}$$

With (A4) and under the noninformative prior for the parameter given by Eq. (6), we have

$$P(\boldsymbol{\beta}, \sigma^2|\mathbf{Z}, \mathbf{h}) \propto P(\mathbf{Z}, \boldsymbol{\beta}, \sigma^2|\mathbf{h}) \propto P(\mathbf{Z}|\boldsymbol{\beta}, \sigma^2)P(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-(N/2+1)} \exp\left[-\frac{(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right]. \quad \text{(A5)}$$

From (A5), if σ^2 is given, the conditional posterior distribution for $\boldsymbol{\beta}$ obviously is multivariate Gaussian:

$$\boldsymbol{\beta}|\mathbf{Z}, \mathbf{h}, \sigma^2 \sim \text{Normal}(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2),$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$. (A6)

Alternatively, if $\boldsymbol{\beta}$ is given, the conditional posterior distribution for σ^2 is a scaled-inverse-χ^2 distribution. That is,

$$\sigma^2|\mathbf{Z}, \mathbf{h}, \boldsymbol{\beta} \sim \text{Inv} - \chi^2(\sigma^2|N, s^2),$$

where

$$s^2 = \frac{1}{N}(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Z} - \mathbf{X}\boldsymbol{\beta}). \quad \text{(A7)}$$

In (A7), Inv $-$ χ^2 refers to the scaled-inverse-χ^2 distribution. With (A3), (A6), and (A7), we have completed the proposed Gibbs sampler, and its stationary output within each iteration will be equivalently sampled from the joint posterior distribution of set $(\mathbf{Z}, \boldsymbol{\beta}, \sigma^2)$ from the model given by Eq. (2).

REFERENCES

Bevington, P. R., and D. K. Robinson, 2003: *Data Reduction and Error Analysis for the Physical Sciences.* 3d ed. McGraw-Hill, 320 pp.

Chu, P.-S., 2002: Large-scale circulation features associated with decadal variations of tropical cyclone activity over the central North Pacific. *J. Climate,* **15,** 2678–2689.

——, and J. Wang, 1997: Tropical cyclone occurrences in the vicinity of Hawaii: Are the differences between El Niño and non-El Niño years significant? *J. Climate,* **10,** 2683–2689.

——, and X. Zhao, 2004: Bayesian change-point analysis of tropical cyclone activity: The central North Pacific case. *J. Climate,* **17,** 4893–4901.

Clark, J. D., and P.-S. Chu, 2002: Interannual variation of tropical cyclone activity over the central North Pacific. *J. Meteor. Soc. Japan,* **80,** 403–418.

Collins, J. M., and I. M. Mason, 2000: Local environmental conditions related to seasonal tropical cyclone activity in the northeast Pacific basin. *Geophys. Res. Lett.,* **27,** 3881–3884.

Elsner, J. B., and C. P. Schmertmann, 1993: Improving extended-range seasonal predictions of intense Atlantic hurricane activity. *Wea. Forecasting,* **8,** 345–351.

——, and T. H. Jagger, 2004: A hierarchical Bayesian approach to seasonal hurricane modeling. *J. Climate,* **17,** 2813–2827.

——, and ——, 2006: Prediction models for annual U.S. hurricane counts. *J. Climate,* **19,** 2813–2827.

Epstein, E. S., 1985: *Statistical Inference and Prediction in Climatology: A Bayesian Approach*. *Meteor. Monogr.,* No. 42, Amer. Meteor. Soc., 199 pp.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2004: *Bayesian Data Analysis*. 2d ed. Chapman & Hall/CRC, 668 pp.

Gray, W. M., C. W. Landsea, P. W. Mielke, and K. J. Berry, 1992: Predicting Atlantic seasonal hurricane activity 6–11 months in advance. *Wea. Forecasting,* **7,** 440–455.

——, ——, ——, and ——, 1993: Predicting Atlantic basin seasonal tropical cyclone activity by 1 August. *Wea. Forecasting,* **8,** 73–86.

——, ——, ——, and ——, 1994: Predicting Atlantic basin seasonal tropical cyclone activity by 1 June. *Wea. Forecasting,* **9,** 103–115.

Hastings, W. K., 1970: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika,* **57,** 97–109.

Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50-Year Reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc.,* **82,** 247–267.

Klotzbach, P., and W. M. Gray, 2004: Updated 6–11 month prediction of Atlantic basin seasonal hurricane activity. *Wea. Forecasting,* **19,** 917–934.

Knaff, J. A., 1997: Implications of summertime sea level pressure anomalies in the tropical Atlantic region. *J. Climate,* **10,** 789–804.

Ripley, B. D., 1987: *Stochastic Simulation*. John Wiley, 237 pp.

Smith, T. M., R. W. Reynolds, R. E. Livezey, and D. C. Stokes, 1996: Reconstruction of historical sea surface temperatures using empirical orthogonal functions. *J. Climate,* **9,** 1403–1420.

Zhao, X., and P.-S. Chu, 2006: Bayesian multiple change-point analysis of hurricane activity in the eastern North Pacific: A Markov chain Monte Carlo approach. *J. Climate,* **19,** 564–578.