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

**Pao-Shin Chu and Xin Zhao**
**Department of Meteorology**
**School of Ocean and Earth Science and Technology**
**University of Hawaii**

## Abstract

In this study, a Poisson generalized linear regression model is applied to the TC counts series in the Central North Pacific. The SLP, PW, NINO34, SOI, and CLIPER are chosen as the predictors. With a non-informative 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 and predictive distribution. The proposed hierarchical model is physically based and a probability distribution is shown for predicting TC frequency in 1982 and 1992 . A cross-validation procedure is applied to the TC series and reasonable forecast results are achieved.

## 1. Introduction

W. Gray pioneered the seasonal hurricane prediction enterprise using regression based linear statistical models (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. This is amazing because hurricane is a small system and physical mechanisms governing its formation are complicated and still not well understood. Gray and his associates have constantly revised their forecasts as time approaches the peak season and operationally issued seasonal forecasts for the Atlantic basin.

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, Elnser and Jagger (2004) introduced a Bayesian strategy to the Poisson regression model so that the predicted annual hurricane numbers could be cast in terms of probability distributions. This is certainly an advantage over the deterministic forecasts because the uncertainty inherent in forecasts is quantitatively expressed in the probability statements.

Also recently, Chu and Zhao (2004) applied a Bayesian analysis to detect change points in the tropical cyclone (TC) series over the central North Pacific. In Chu and Zhao (2004), 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, about shifts in the TC time series. In view of the probabilistic nature of the Bayesian paradigm, this study will use the Poisson regression cast in the Bayesian framework to predict the seasonal TC activity over the central North Pacific prior to the peak hurricane season.

## 2. Data

Monthly mean sea level pressure (SLP), wind data at the 1000-, 850-, and 200-hPa levels, relative vorticity data at the 1000-hPa level and total precipitable water (PW) are derived from the National Centers for Environmental Prediction – National Center for Atmospheric Research reanalysis dataset (Kistler et al., 2001). The horizontal resolution of the reanalysis dataset is 2.5° latitude-longitude. Tropospheric vertical wind shear 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. SST data are available on a 2° latitude-longitude. 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 (Fig. 1). Nino 3.4 region sea surface temperature (SST) and the Southern Oscillation Index (SOI) data are obtained from the NCEP/Climate Prediction Center.

## 3. Predictor Selection Procedure

Initially, plausible predictor candidates include SST, SLP, low-level relative vorticity, vertical wind shear, two El Niño indices (i.e., the Southern Oscillation Index and Nino3.4 index), precipitable water in the atmosphere, and the inherent short-term oscillations in TC series. Pearson correlations between TC counts at the peak season and each one of those potential predictors for each bi-month from January through June are computed. Because the peak TC season in the CNP is July, August, and September (Chu, 2002), TC counts from these three months are summed to produce a seasonal value. If correlations between the seasonal TC frequency and one of the predictor candidates are statistically significant over a particular region of the North Pacific, then this predictor variable over this area for this bi-month is retained. For the sake of simplicity, if there is more than one region that is significantly correlated with the peak season TC frequency, we'll only choose the one with the highest value.

For seasonal TC frequency over the CNP and SLPs over the North Pacific during the antecedent March/April, a strong negative correlation is found over the tropical/subtropical eastern Pacific. That is, lower SLPs over the eastern Pacific in the preceding March/April are correlated with high TC frequency over the CNP and vice versa. This result is reasonable physically. 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 deep convection to grow. 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. Likewise, precipitable water content over the tropical/subtropical eastern North Pacific in the

preceding March/April is positively and significantly correlated with TC counts. Adequate moisture in the atmosphere provides a fundamental ingredient for deep convection. Not surprisingly, the area where the correlation is high between precipitable water and TC is also approximately the region of high, negative correlation between SLP and TC. We have also computed correlations between seasonal TC frequency and some dynamic variable (e.g., vertical wind shear, low-level relative vorticity) of the preceding months over the North Pacific but do not find them to be statistically significant.

As previously mentioned, the El Niño influence on TC activity over the CNP is very pronounced (Chu and Wang, 1997; Clark and Chu, 2002). This influence is mainly reflected by an increase in the low-level cyclonic vorticity as induced by the eastward displacement of the monsoon trough from the western Pacific and by a concomitant reduction of the vertical wind shear over the CNP. Because tropical cyclones over the eastern North Pacific tend to form farther westward during El Niño years (Chu, 2004), they may propagate further west and perhaps enter the CNP when vertical wind shear is less. Due to the simultaneous change in large-scale circulation over both the eastern and western North Pacific, there are more TCs observed over the CNP during El Niño years. Because El Niño is a coupled ocean and atmosphere phenomenon, we use one atmospheric index and one ocean index to represent El Niño. The atmospheric index is the standard SOI, which is the difference in normalized sea level pressures between Tahiti and Darwin in northern Australia. Large and negative SOI corresponds to the El Niño condition. On a seasonal time scale, Clark and Chu (2002) found a strong correlation between the summer SOI and TC counts over the CNP. This correlation reaches -0.54, significant at the 1% level after taking climatological persistence into account. For the ocean index, a common one is the SSTs in the Nino 3.4 region, which covers an area between 5°N - 5°S and 170°W - 120°W. Accordingly, a Nino 3.4 index is chosen.

The analysis of variance (ANOVA) is a traditional statistical technique to reveal the existence of hidden periods in a time series (e.g., Chan et al. 1998). The basic idea is to repeatedly divide the data batch into groups with a given period $T$ and calculate the ratio of the within-group variance and the among-group variance. If this ratio exceeds the critical value at a given confidence level, then the data series is regarded as having a significant period $T$. We do this analysis to all possible hidden periods and only 2-year period is significant.

Once the hidden periods for the TC series are determined from the variance analysis, these variations can be used as climatology and persistence (CLIPER) predictors. Calculating CLIPER predictor is basically a cross-validation procedure. We first re-group the time series according to the hidden period. In our study, the TC series has a significant period of two years. Second, we use the group mean, excluding this data itself, as its cross-validation prediction. We do this procedure to all the data and the resulted series will be the CLIPER predictor.

## 4. Mathematical model for TC counts

A Poisson process is a proper probability model for describing independent, rare event counts. Given the Poisson intensity parameter $\lambda$ (i.e., the mean seasonal TC rates), the probability mass function (PMF) of $h$ TCs occurring in $T$ years is (Epstein, 1985)

$$P(h \mid \lambda, T) = \exp(-\lambda T)\frac{(\lambda T)^h}{h!},$$

where $h = 0,1,2,...$ and $\lambda > 0$, $T > 0$ (1)

In many cases, a TC time series cannot simply be described by a constant rate Poisson process. Thus, the Poisson intensity, $\lambda$, should not be treated as a determinant single-value constant but as a random variable.

In this study, we apply the Poisson generalized linear 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 $h_i$, $i = 1,2,...,N$, $Z_i = \log\lambda_i$, where $\lambda_i$ is the relative Poisson rate. The link between this latent variable and the predictors is expressed as $Z_i = \mathbf{X}_i\boldsymbol{\beta} + \varepsilon_i$, where $\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, ..., \beta_K]'$ is a random vector, $\varepsilon_i$ is assumed to be identically and independently distributed (IID) and is normally distributed with zero mean and $\sigma^2$ variance and $\mathbf{X}_i = [1, X_{i1}, X_{i2}, ..., X_{iK}]$ is the predictor vector for $h_i$. In the vector form, this model can be formulated as below:

$$P(\mathbf{h} \mid \mathbf{Z}) = \prod_{i=1}^{N} P(h_i \mid Z_i)$$

where $h_i \mid Z_i \sim Poisson(e^{Z_i})$

$\mathbf{Z} \mid \boldsymbol{\beta}, \sigma^2, \mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_N)$,

$\mathbf{X}' = [\mathbf{X}_1', \mathbf{X}_2', ..., \mathbf{X}_N']$,

$\mathbf{I}_N$ is $N \times N$ identity matrix,

$\mathbf{X}_i = [1, SOI_i, PW_i, SLP_i, NINO_i, CLIPER_i]$

$i = 1,2,...,N$,

$\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5]'$

(2)

## 5. MCMC approach to the Bayesian Inference

### 5.1 General idea of MCMC

In general, let us assume $\boldsymbol{\theta}$ be the set of the model parameters and $\mathbf{h}$ be the data for the analysis. The basic Bayesian formula is described as:

$$P(\boldsymbol{\theta} \mid \mathbf{h}) = \frac{P(\mathbf{h} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})}{\int P(\mathbf{h} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})d\boldsymbol{\theta}} \propto P(\mathbf{h} \mid \boldsymbol{\theta})P(\boldsymbol{\theta})$$

(3)

where "$\propto$" means "proportional" since $\boldsymbol{\theta}$ in the denominator is only a dummy variable. In this formula, $P(\mathbf{h} \mid \boldsymbol{\theta})$ is the conditional distribution of data $\mathbf{h}$ given the model parameters $\boldsymbol{\theta}$ (i.e., the likelihood given the model) and $P(\boldsymbol{\theta})$ is a prior

distribution. Formula (3) provides the inference for posterior distribution $P(\boldsymbol{\theta} \mid \mathbf{h})$, the probability of $\boldsymbol{\theta}$ after the data $\mathbf{h}$ are observed. It is also clear that data affect the posterior distribution only through the likelihood function $P(\mathbf{h} \mid \boldsymbol{\theta})$. To make predictive inference, we rely on the posterior predictive distribution

$$P(\tilde{\mathbf{h}} \mid \mathbf{h}) = \int P(\tilde{\mathbf{h}} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta} \mid \mathbf{h}) d\boldsymbol{\theta} \qquad (4)$$

where $\tilde{\mathbf{h}}$ denotes the prediction (Gelman et al., 2004). $P(\tilde{\mathbf{h}} \mid \mathbf{h})$ is the posterior predictive distribution since it is conditional on the observed data **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

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

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

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

where $\boldsymbol{\theta}^{[1]}, \boldsymbol{\theta}^{[2]}, \ldots, \boldsymbol{\theta}^{[L]}$ are independently sampled from $P(\boldsymbol{\theta} \mid \mathbf{h})$. When $L$ goes to infinity, this approximation will converge to its analytical integral under very general condition.

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 $P(\boldsymbol{\theta} \mid \mathbf{h})$ is complicated. Nonetheless, in most of applications, it may be possible to generate a series of dependent values by using a Markov chain (MC) that has $P(\boldsymbol{\theta} \mid \mathbf{h})$ 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 current state $\boldsymbol{\theta}^{[i]}$. Under very general condition (i.e., the MC is ergodic), the distribution for the state will converge to a unique stationary distribution.

## 5. 2 Gibbs sampler for the Bayesian inference of TC model

A common MCMC integration is known as the Gibbs sampler. 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 $\sigma^2$, it is reasonable to choose the non-informative prior. In formula, it is (Gelman et al., 2004)

$$P(\boldsymbol{\beta}, \sigma^2) \propto \sigma^{-2} \qquad (5)$$

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

With the new observed predictor set $\tilde{\mathbf{X}} = [1, \tilde{X}_{i1}, \tilde{X}_{i2}, \ldots, \tilde{X}_{iK}]$, if we have the posterior distribution of the parameter, the predictive distribution for the latent variable $\tilde{Z}$ and TC count $\tilde{h}$ will be

$$P(\tilde{Z} \mid \tilde{X}, \mathbf{X}, \mathbf{h}) = \iint_{\boldsymbol{\beta}, \sigma^2} P(\tilde{Z} \mid \boldsymbol{\beta}, \sigma^2) P(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{h}) d\boldsymbol{\beta} d\sigma^2 \qquad (6a)$$

$$P(\tilde{h} \mid \tilde{X}, \mathbf{X}, \mathbf{h}) = \int_{\tilde{z}} \frac{\exp(-e^{\tilde{Z}} + \tilde{Z}\tilde{h})}{\tilde{h}!} P(\tilde{Z} \mid \tilde{X}, \mathbf{X}, \mathbf{h}) d\tilde{Z} \qquad (6b)$$

However, even with the non-informative prior assumption, the posterior distribution for the model parameter set $(\boldsymbol{\beta}, \sigma^2)$ is still not a standard density distribution and directly sampling from it is very difficult. In this section, we will design a Gibbs sampler, which has $P(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{X}, \mathbf{h})$ as its stationary distribution, and then we can use an alternative approach to integrate (6a) by

$$P(\tilde{Z} \mid \tilde{X}, \mathbf{X}, \mathbf{h}) = \frac{1}{L} \sum_{i=1}^{L} P(\tilde{Z} \mid (\boldsymbol{\beta}, \sigma^2)^{[i]}) \qquad (7)$$

where $(\boldsymbol{\beta}, \sigma^2)^{[i]}$ is the i-th sampling from this proposed Gibbs sampler after the burn-in period.

The overall Bayesian inference for this TC model is as below:

$$P(Z_i \mid \mathbf{h}, \boldsymbol{\beta}, \mathbf{Z}_{-i}, \sigma^2) \propto \exp\left\{\exp(-e^{Z_i}) + Z_i h_i - \frac{1}{2\sigma^2}(Z_i - \mathbf{X}_i \boldsymbol{\beta})^2\right\} \quad i = 1,2,\ldots N \qquad (8a)$$

$$\boldsymbol{\beta} \mid \mathbf{Z}, \mathbf{h}, \sigma^2 \sim N(\hat{\boldsymbol{\beta}}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2) \quad \text{where } \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} \qquad (8b)$$

$$\sigma^2 \mid \mathbf{h}, \mathbf{Z} \sim \text{Inv} - \chi^2(N-K, s^2) \quad \text{where } s^2 = \frac{1}{N-K}(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Z} - \mathbf{X}\hat{\boldsymbol{\beta}}) \text{ and } \text{Inv} - \chi^2 \text{ refers to the scaled inverse-} \chi^2 \text{ distribution.} \qquad (8c)$$

The Gibbs sampling algorithm is briefly outlined.

1. Select proper initial value for $\mathbf{Z}^{[0]}, \boldsymbol{\beta}^{[0]}, \sigma^{2[0]}$ and set $t = 1$.

2. Draw $Z_i^{[t]}$ from $Z_i^{[t]} \mid \mathbf{h}, \boldsymbol{\beta}^{[t-1]}, \sigma^{2[t-1]}$ for $i = 1,2,...,N$ via Eq. (8a)

3. Draw $\sigma^{2[t]}$ from $\sigma^{2[t]} \mid \mathbf{h}, \mathbf{Z}^{[t]}$ via Eq. (8c).

4. Draw $\boldsymbol{\beta}^{[t]}$ from $\boldsymbol{\beta}^{[t]} \mid \mathbf{h}, \mathbf{Z}^{[t]}, \sigma^{2[t]}$ via Eq. (8b).

5. Set $t = t + 1$ then go back to step 2 until meeting the required number of iterations.

(9)

With the observation data $\mathbf{h}$ and following Eq. (9), one thereby can sample a set $\mathbf{Z}, \boldsymbol{\beta}, \sigma^2$ for each iteration .

## 6. Bayesian prediction

There are a total of 38 years (1966 – 2003) of TC counts in the CNP. We apply them in the framework given above. The model is detailed as formula (2) and as described in formula (3), we have 5 predictors.

In order to verify the effectiveness of the proposed method, we design a cross-validation test for this dataset. Cross-validation is a generalization of the common technique of repeatedly omitting a few observations from the data, reconstructing the model, and then making estimates for the omitted cases. In this study, only one point is removed from the data set repeatedly and developmental data sets contain size n-1.

We apply these datasets to the designed Gibbs sampler, and use its output as the posterior sampling of the model parameter. With these sampling sets, after plugging in the target year's predictor observation, we use the formula (8) to determine the posterior predictive distribution of the latent variable Z which is equivalently to the natural logarithm of the TC rates, $\lambda$. In Fig. 2, the predicted TC rates (solid line) are plotted together with the actual observation (dotted line) for each year. We plot the prediction based on the predictor variables in March/April in Fig. 2a, while in Fig. 2b we plot the prediction based on predictor variables in May/June. The reason that we test the case with information up to April is because a 2-month lead time is considered useful in an operational mode. The predicted TC count rate in Fig. 2a is very close to the observation and the Pearson correlation between them is 0.65. In Fig. 2b, the Pearson correlation between observations and forecasts is as high as 0.79.

Also, one of the significant advantages of the Bayesian analysis, comparing to a conventional regression model, is that rather than predicting a single point value, the former can give out a predictive distribution of the TC counts for each individual year. For example, in Fig. 3, we present the posterior predictive distribution for the seasonal TC counts through a cross-validation procedure for 1982 and 1992. For both cases, the observed TC frequency is in line with the mode (most frequent) of the predictive distribution.

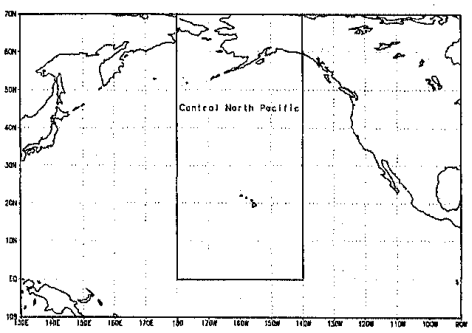


Fig. 1

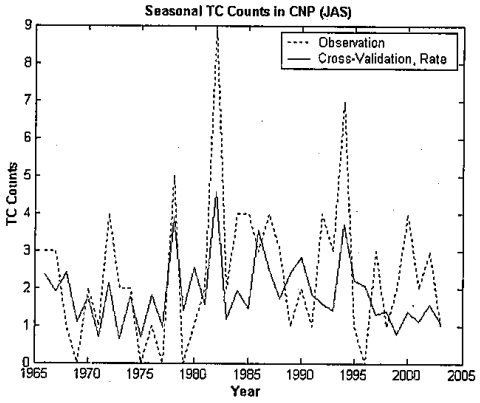


Fig. 2a.

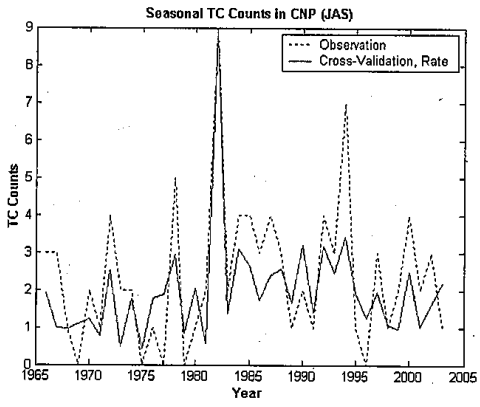


Fig. 2b

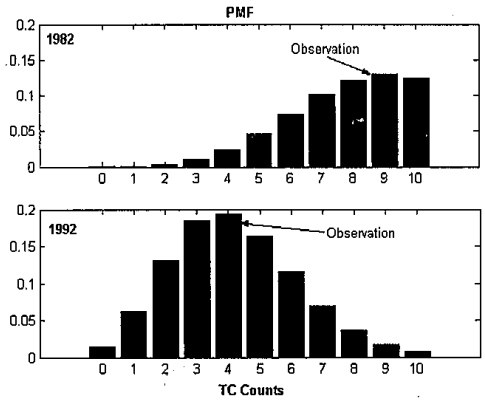


Fig. 3

## References


Chan, J.C.L., J.-E. Shi, and C.-M. Lam, 1998: Seasonal forecasting of tropical cyclone activity over the western North Pacific and the South China Sea. *Wea. Forecasting*, **13**, 997-1004.

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.

Chu, P.-S., 2004: ENSO and tropical cyclone activity. *Hurricanes and Typhoons: Past, Present, and Potential*, R.J. Murnane and K.-B. Liu, Eds., Columbia University Press, 297-332.

Chu, P.-S., and J. Wang, 1997: Tropical cyclone occurrences in the vicinity of Hawaii: Are the differences between El Nino and non-El Nino years significant? J. Climate, 10, 2683-2689.

Chu, P.-S., 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.

Gelman, A., J.B. Carlin, H.S. Stern, and D.B. Rubin, 2004: Bayesian data analysis. 2nd edition, Chapman & Hall/CRC, 668 pp.

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

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

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

Gray, W.M., C.W. Landsea, P.W. Mielke, and K.J. Berry, 1993: Predicting Atlantic basin seasonal tropical cyclone activity by 1 August. *Wea. Forecasting*, **8**, 73-86.

Gray, W.M., C.W. Landsea, P.W. Mielke, and K.J. Berry, 1994: Predicting Atlantic basin seasonal tropical cyclone activity by 1 June. *Wea. Forecasting*, **9**, 103-115.

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

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