Modeling Seasonal Tropical Cyclone Activity in the Fiji Region as a Binary Classification Problem

SAVIN S. CHAND
Centre for Australian Weather and Climate Research, Bureau of Meteorology, Melbourne, Victoria, Australia

KEVIN J. E. WALSH
School of Earth Sciences, University of Melbourne, Melbourne, Victoria, Australia

(Manuscript received 31 August 2011, in final form 15 January 2012)

ABSTRACT

This study presents a binary classification model for the prediction of tropical cyclone (TC) activity in the Fiji, Samoa, and Tonga regions (the FST region) using the accumulated cyclone energy (ACE) as a proxy of TC activity. A probit regression model, which is a suitable probability model for describing binary response data, is developed to determine at least a few months in advance (by July in this case) the probability that an upcoming TC season may have for high or low TC activity. Years of “high TC activity” are defined as those years when ACE values exceeded the sample climatology (i.e., the 1985–2008 mean value). Model parameters are determined using the Bayesian method. Various combinations of the El Niño–Southern Oscillation (ENSO) indices and large-scale environmental conditions that are known to affect TCs in the FST region are examined as potential predictors. It was found that a set of predictors comprising low-level relative vorticity, upper-level divergence, and midtropospheric relative humidity provided the best skill in terms of minimum hindcast error. Results based on hindcast verification clearly suggest that the model predicts TC activity in the FST region with substantial skill up to the May–July preseason for all years considered in the analysis, in particular for ENSO-neutral years when TC activity is known to show large variations.

1. Introduction

The southwest tropical Pacific islands, such as Fiji, Samoa, and Tonga (the “FST region” defined as the area between 5°S and 25°S, 170°E and 170°W), are very vulnerable to catastrophic impacts of tropical cyclone (TC) activity. Advance warning of TC strike on these island nations is therefore important to minimize property destruction and to possibly avoid loss of life. TCs in the FST region exhibit a marked year-to-year variability due to the El Niño–Southern Oscillation (ENSO) phenomenon (Chand and Walsh 2009). The ENSO–TC relationship here is identifiable prior to a cyclone season through the use of appropriate ENSO indices. This enabled Chand et al. (2010, hereafter CWC10) to develop suitable statistical forecast schemes to predict TC numbers in the FST region for the upcoming cyclone season.

However, while the schemes developed by CWC10 work well for years associated with El Niño and La Niña conditions, no substantial skill could be obtained for years associated with the ENSO-neutral conditions. The purpose of this study is to present a different methodological framework of seasonal cyclone modeling that uses a binary classification approach to predict the probability of seasonal TC activity in the FST region. As shown later, this new scheme has a substantial capability to make binary prediction of TC activity not only for El Niño and La Niña years but also for the years associated with ENSO-neutral conditions.

Seasonal prediction of TC activity using statistical methods is extensively used in various TC basins for over the past three decades [see, e.g., a review of Camargo et al. (2010)]. Pioneering work on the development of seasonal cyclone forecasting schemes was made by Nicholls (1979) for the Australian region and Gray (1984)
for the North Atlantic basin using regression-based linear statistical models. Subsequent studies led to development of prediction schemes for different cyclone basins with an improved methodological framework for statistical cyclone modeling. Vecchi et al. (2011), for example, developed a statistical-dynamical hurricane forecasting system for seasonal North Atlantic hurricane activity. Chan et al. (1998, 2001) and Liu and Chan (2003), for example, used the projection pursuit regression technique of Friedman and Stuetzle (1981) to develop seasonal forecasting schemes for the western North Pacific and the South China Sea. Elsner and Schmertmann (1993) considered a different approach to predict seasonal numbers of intense Atlantic hurricanes. They showed that a nonlinear Poisson model is superior to a linear statistical model in terms of improvements in the hindcast skill. Subsequent work—for example, Elsner and Jagger (2004, 2006), Chu and Zhao (2007), Flay and Nott (2007), and CWCl0—consequently used the Poisson regression models to predict TC counts in different basins. In addition, these studies have used a Bayesian approach to determine model parameters, as opposed to classical methods, such as least squares regression or maximum likelihood estimates. A Bayesian framework assumes that model parameters have a distribution rather than being fixed as in classical approaches. Inferences can therefore be made by computing the posterior probability density estimates of model parameters conditioned on the observed data. This has an obvious advantage over a linear or a classical method in that the uncertainties intrinsic to forecasts can be quantitatively expressed in probability statements.

However, while the Poisson regression models used in the aforementioned studies are effective for a series of rare events, they may be inappropriate for modeling binary classification problems, that is, whether a particular event occurs. This is clearly an issue for seasonal prediction of tropical cyclones. Therefore, in this study we construct a “probit” regression model using Bayesian fitting. A probit regression is considered a proper probability approach for describing binary response data (e.g., Albert and Chib 1993). Binary response data usually takes a value of one for the occurrence of an event (high TC activity in our case) and zero for the nonoccurrence (low TC activity). The probit regression approach is adopted here following its successful implementation in Chu et al. (2010) for forecasting regional and seasonal TC frequency in the western Pacific and in Chand and Walsh (2011) for forecasts of up to three days of TC formation in the Fiji region.

The scope of the present investigation is twofold. In the first part of this work we seek to develop an understanding of high and low TC activity over the FST region during the ENSO-neutral years based on the associated changes in preexisting large-scale environmental fields. This distinguishes the present work from some of our previous papers (e.g., Chand and Walsh 2009, 2011; CWCl0) in which we mainly examined the differences in environmental conditions for El Niño and La Niña years. In our previous work, very little attention was given toward understanding the large variation in TC activity during ENSO-neutral conditions. A further study is therefore deemed necessary to understand factors that account for high and low TC activity in the FST region during ENSO-neutral years.

In the second part of this work, we propose a new probit regression model that uses Bayesian fitting to predict TC activity in the FST region for an upcoming TC season, with the main emphasis being toward a skillful prediction for ENSO-neutral years.

This paper is structured as follows. Section 2 describes various datasets used in this study, while section 3 defines a proxy for TC activity in the FST region. Section 4 gives physical mechanisms accounting for the observed differences in high and low TC activity in the FST region. Section 5 outlines an approach for predictor selection. Sections 6 and 7 describe the model formulation and the model validation procedures, respectively. Finally, a summary is given in section 8.

2. Data

The TC data used in this paper are archived by the Joint Typhoon Warning Center (JTWC; available online at http://www.usno.navy.mil/NOOC/nmfc-ph/RSS/jtwc/best_tracks/) at 6-hourly intervals. Only data for the austral summer season (i.e., November–April) for the period beginning November 1985 and ending April 2009 are considered. Data prior to 1985 may have contained intensity biases (e.g., Harper et al. 2008) and are therefore not included in the analysis. Only TCs that reached at least a minimum of 17.5 m s⁻¹ sustained wind speed are included in the analysis. Extremely short-lived systems (i.e., systems that have a lifetime <1 day) are excluded. Altogether, 68 systems are considered in the analysis.

The monthly values of various ENSO indices are obtained from the Climate Prediction Center website (http://www.cpc.ncep.noaa.gov/data/indices/) for the period beginning November 1985 and ending April 2009. The SST data for the same period are obtained from the National Oceanic and Atmospheric Administration Extended Reconstructed SST (ERSST) version 2 dataset of Smith and Reynolds (2004), while the atmospheric variables are extracted from the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis 1 products (Kalnay et al. 1996).
3. Proxies for high TC activity

Different indices have been used in the past as proxies of TC activity. Emanuel (2005), for example, used a power dissipation index (PDI), defined as the cube of the maximum wind speed summed over the lifetime of a TC, to determine the potential annual threat that cyclones may cause to the North Atlantic and North Pacific regions. Webster et al. (2005) examined trends in the number of TCs that are in Category 4 and 5 of the Saffir–Simpson scale for different TC basins. Other studies (e.g., Camargo and Sobel 2005; Sobel and Camargo 2005; Chan 2007; Villarini and Vecchi 2012) have used a more conservative measure called the accumulated cyclone energy (ACE) (Bell et al. 2000). Here we also prefer ACE as a proxy of TC activity in the FST region for two reasons: (i) values of ACE, unlike the Saffir–Simpson scale, are not quantized and therefore any underestimate in maximum wind speeds will not result in different categorizations of TCs and (ii) even though the square of maximum wind speed may be biased toward more intense TCs, any error in the intensity estimates will be multiplied less strongly than in the PDI.

Annual values of ACE are obtained by calculating the sum of the squares of the estimated 6-hourly maximum sustained wind speed over the lifetime of a TC and then accumulating them for all TCs over the season under consideration, such that

\[
ACE = \sum_{i=1}^{N} \sum_{t_i} v(t)^2,
\]

where \(v(t)\) is the maximum wind speed obtained from the JTWC best-track data, in knots, of the TC at time \(t\) (defined here in terms of the number of 6-h blocks), \(t_{gi}\) and \(t_{di}\) are the genesis and dissipation time of the TC \(i\), and \(N\) is the total number of TCs in the season under consideration.

Overall, ACE in the FST region exhibits a large year-to-year variation (Fig. 1). Years of “high TC activity” are defined here as those years when ACE values exceeded the sample climatology (i.e., the 1985–2008 mean value). Similarly, years of “low TC activity” are those years when ACE values are equal to or less than the sample climatology. Because of our small sample size, one may argue how sensitive the choice of high and low ACE years is to statistical uncertainty in the sample climatology. To determine this, we have computed the 95% confidence intervals of the sample climatology. The lower and upper limits of ACE values at the 95% confidence interval are 1.50 × 10^5 and 1.71 × 10^5 kt^2, respectively, the sample climatology being 1.65 × 10^5 kt^2. Thus, years chosen for our “high ACE” and “low ACE” remain the same even with the choice of the 95% thresholds (not shown). This indicates that our result, based on the number of samples considered in the present investigation, is statistically not very sensitive to small variations in the chosen ACE threshold. Since this study focuses on the development of a binary classification scheme, years of high TC activity are given the binary label “1” and years of low TC activity are given the binary label “0.”

4. Large-scale conditions for high and low TC activity

We begin by examining various large-scale environmental conditions to account for the physical mechanisms that cause not only the overall observed changes in high and low TC activity in the FST region but also to understand changes in TC activity specific to ENSO-neutral years. The large-scale environmental conditions, which are widely known in the scientific literature for their role in affecting TC activity, include (i) the 850-hPa relative vorticity, (ii) environmental vertical wind shear between the 200-hPa and 850-hPa levels, (iii) the 200-hPa divergence, (iv) midtropospheric relative humidity, (v) sea surface temperature, and (vi) equivalent potential temperature gradient between the 1000-hPa and 500-hPa levels. The first three conditions are referred to as the dynamical conditions and the other three as the thermodynamical conditions. Note here that anomalously negative values of relative vorticity and environmental vertical wind shear, and anomalously positive values of other variables, indicate more favorable conditions than the overall climatology. The reverse is true otherwise.

In our present investigation, the November–April composite anomalies associated with both dynamical and thermodynamical fields are constructed for years corresponding to high and low TC activity relative to the overall 1985/86–2008/09 sample climatology (Fig. 2).
Fig. 2. Composites anomalies of various November–April large-scale environmental conditions for years associated with (left) high TC activity and (right) low TC activity. Stippling represents areas where anomalies are statistically significant at the 95% significance level as obtained using bootstrap sampling. The shading overlaid on relative vorticity indicates anomalous westerlies.
Results show that large-scale environments, particularly dynamical conditions in the region equatorward of ~15°S where TCs frequently form (Chand and Walsh 2009), are more favorable for TC formation during years of high TC activity than during years of low TC activity. Relative vorticity, for example, is more cyclonic (i.e., larger negative values) for years associated with high TC activity than low TC activity. Similarly, environmental vertical wind shear is anomalously low in high TC activity years rather than in low TC activity years. Note also the dominance of anomalous westerlies, which provide favorable conditions for TC development, in the 850-hPa windfield equatorward of ~15°S for years of high TC activity as opposed to years of low TC activity.

A major factor contributing to the observed difference in large-scale environmental conditions for the overall high and low TC activity is the El Niño and La Niña phenomena, respectively. About 75% of all El Niño years between the period 1985 and 2008 are associated with high TC activity in the FST region. Similarly, about 83% of all La Niña years during the period under consideration are associated with low TC activity. Years corresponding to ENSO events are obtained from the Climate Prediction Center website (at http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/ensostuff/ensoyears.shtml). The Climate Prediction Center (CPC) uses the 3-month running mean SST anomalies in the Niño-3.4 region (5°N–5°S, 120°–170°W) to determine the ENSO event: a year with five consecutive overlapping months with anomalies >0.5° (<−0.5°) is considered an El Niño (La Niña) year; otherwise it is considered an ENSO-neutral year. Composite anomalies of large-scale environmental conditions for El Niño and La Niña years (Fig. 3) are characteristically similar to those associated with high and low TC activity (Fig. 2), respectively. Note, for example, a strong dominance of cyclonic relative vorticity, favorable divergent atmosphere aloft, and significantly low environmental vertical wind shear (EVWS) in the main region of TC formation (i.e., equatorward of ~15°S) in these two figures. The thermodynamic conditions are also anomalously higher in the main region of TC formation for years of high TC activity, consistent with those for El Niño years. These results are consistent with Chand and Walsh (2009), who showed that large-scale environmental conditions, particularly dynamical conditions, are more favorable for TC formation in the FST region in El Niño years than in La Niña years, thus contributing to higher TC activity in El Niño years than in La Niña years.

An interesting result, which to some extent forms the basis of our subsequent investigation, arises after subdividing the ENSO-neutral years into those associated with high and low TC activity (Fig. 4). For the ENSO-neutral years associated with high TC activity, the three dynamical conditions, as well as relative humidity, are found to be more favorable for TC formation in the area north of the Fiji islands (~10°–15°S, 178°E) and in the Coral Sea region (west of ~175°E). Note that these areas are where TCs affecting the FST region during the ENSO-neutral years are frequently spawned [see, e.g., Fig. 2d of Chand and Walsh 2009]. Other thermodynamic fields such as the SST and the equivalent potential temperature gradient appear to play a lesser role in modulating the differences over the FST region during the ENSO-neutral years, as both are anomalously positive for years associated with high and low TC activity.

In summary, TC activity in the FST region is strongly modulated by large-scale environmental conditions. A major contribution to this modulation is the ENSO phenomenon. While TC activity in the FST region is usually high during El Niño years and low during La Niña years, variability during ENSO-neutral years is generally attributed to local variation in large-scale environmental conditions, particularly in the region north of the Fiji islands and the Coral Sea. If the variability of some preexisting large-scale environmental conditions is identified prior to a TC season, they can be used as predictors of TC activity in the FST region. In the following sections, we first examine the extent to which a season of high and low TC activity could be identified well in advance and then propose a Bayesian modeling framework for its prediction.

5. Potential predictors of TC activity

The selection of appropriate predictors is one of the most important steps for the development of a statistical forecast model. Here we select the predictors using the same approach suggested in CWC10.

The potential ENSO indices as candidate predictors of high TC activity in the FST region are identified using the following procedure. A correlation analysis is performed for the 1985/86–2008/09 seasons between the annual values of ACE for the November–April season and 3-monthly running mean values of various ENSO indices starting January–March preceding the TC season and ending October–December following the TC season. The statistical significance of correlation is determined using the Pearson correlation technique (e.g., Chu and Zhao 2007; CWC10). The critical value of the Pearson correlation coefficient for a sample size of 23, using a two-tailed test at the 95% significance level, is 0.41 (Sheskin 2007). Thus, any predictor with a correlation of magnitude below 0.41 is not considered. As in CWC10, it is evident here that the Niño-4 index gives the maximum preseason correlation during the May–July
FIG. 3. As in Fig. 2 but for the El Niño and La Niña years.
FIG. 4. As in Fig. 2 but for the ENSO-neutral years associated with high and low TC activity.
(MJJ) periods, with correlation coefficient $r = 0.55$ (Fig. 5). Therefore, the mean values of the May–July preseason Niño-4 index are retained for further analysis.

The large-scale environmental parameters used in CWC10 as potential predictors are also evaluated here (Table 1). As in CWC10, the annual values of ACE are correlated with these environmental variables for both the November–April tropical cyclone season (not shown) and the May–July preseason period (Fig. 6). We apply the procedure suggested in Chu and Zhao (2007) to determine the critical region for each candidate environmental parameter such that, for each of the environmental variables, a grid point with a Pearson correlation of magnitude above 0.41 is deemed significant at the 95% level and therefore selected as a critical region. However, questions may arise regarding the validity of such a parametric test due to our small sample size. Therefore, an additional nonparametric test is performed using a Monte Carlo simulation to determine the statistical significance of our correlation coefficients. Results obtained by the two methods (not shown) are very similar, so we only show statistics associated with the Pearson correlation.

To keep the results robust, a simple average within 5° square box centered over the critical region of the predictor variable is taken. Since favorable large-scale environmental conditions occur simultaneously with TC seasons (i.e., during the November–April season for the FST region), environmental variables for the November–April season cannot be used as predictors. Instead, the May–July preseason environmental variables are used, as they also give significantly high correlations (at the 95% significance level) with the November–April ACE values over certain critical regions. However, it is not very clear to us what physical mechanisms account for these statistically high correlations between the wintertime environmental variables and the summertime TC activity in the FST region. A similar pattern was also observed in past studies over the Australian region (e.g., Nicholls 1984; Ramsay et al. 2008; Werner and Holbrook 2011). Our working hypothesis is that it could be related to the atmospheric response of a developing ENSO event that reaches its peak phase in the austral summer season.

### 6. Model formulation

Since the prediction of high TC activity is evaluated as a binary classification problem, we construct a probit regression model, as it is considered a proper probability model for describing binary response data (e.g., Albert and Chib 1993). The probit model is formulated here using the procedure outlined in Chu et al. (2010) and also in Chand and Walsh (2011). A brief interpretation and formulation of the probit regression model in the context of predicting high and low TC activity in the FST region is given in the appendix. Using the overall probit model structure outlined in Eq. (A1), the predictor vector $X_i$ and the associated coefficient parameter vector $\beta$ for different predictor combinations over a total of $N$ years can be specified as follows:

$$X_i \in \{1, EVWS_i, VORT_i, DIV_i, Rhum_i, \Theta_{E_i}, \text{Niño-4}_i\}, \quad i = 1, 2, \ldots, N$$

$$\beta \in \{\beta_0, \beta_{EVWS}, \beta_{VORT}, \beta_{DIV}, \beta_{Rhum}, \beta_{\Theta_{E}}, \beta_{\text{Niño-4}}\}.$$ (2)

Values for the intercept ($\beta_0$) and parameter coefficients ($\beta$) are estimated via the Bayesian approach in which the Gibbs sampler (Gelfand and Smith 1990) is used to obtain the posterior distribution of each model parameter (see CWC10 for details of the Bayesian inference technique using the Gibbs sampler). A total of
15,000 simulations are performed here for different predictor combinations; the first 5000 iterations are discarded as a burn-in and the subsequent 10,000 iterations are used to obtain posterior distributions of model parameter coefficients. Burn-in up to the first 5000 iterations is necessary to ensure model convergence, which is diagnosed using the Gelman and Rubin (1992) diagnostic. This convergence diagnostic is a general approach for monitoring convergence of the Gibbs sampler in which various parallel chains are run with differing initial conditions. After convergence, posterior distributions of all chains reach a similar steady state.

An example of the posterior distribution of model parameters using all predictor variables is given in Fig. 7. Distributions are smoothed using a kernel density estimator so as to remove insignificant fluctuations. The autocorrelation values associated with each parameter reach zero fairly quickly. This indicates that the output of the Gibbs sampler is independently drawn from their joint posterior distribution. The posterior density distributions of parameters on either side of the zero line provide an insight into the relative contribution of each parameter to the regression model. Here, a large proportion of the distribution associated with relative vorticity...
(VORT), divergence (DIV), and relative humidity (Rhum) lies on one side of the zero line. This indicates that VORT, DIV, and Rhum as predictors for “TC activity” in the FST region play a key role in the prediction equation as opposed to other predictors. Indeed, as shown later, the model with VORT, DIV, and Rhum as predictors (the reduced-predictor model) appears to have a greater hindcast skill than the all-predictor combined model.

7. Model validation

Validation of a forecast model is necessary to assess its performance in practice. For better evaluation of the robustness of the predictor set, we use here two different validation techniques: (i) a leave-one-out cross-validation (LOOCV) (Elsner and Schmertmann 1994) technique and (ii) a fourfold cross-validation method (e.g., Werner and Holbrook 2011).

a. LOOCV technique

As in CWC10, we first use a LOOCV technique (e.g., Elsner and Schmertmann 1994) to assess the skill of the all-predictor combined and reduced-predictor models relative to each other and the overall climatology model. The climatology model contains only the regression constant (i.e., intercept) and the indicator variable.

The LOOCV technique works by successively omitting an observation from the dataset and repeating the modeling procedure to predict the omitted observation, with the resulting “prediction” often referred to as the “hindcast.” Because our annual ACE values exhibit no significant serial correlations (as per the autocorrelations associated with the climatology model shown in Fig. 7b), the use of the LOOCV procedure is justified, as the presence of serial correlation will introduce bias in the estimation of forecast skill. It is important to indicate here that in our cross-validation process we have completely removed all information for the year to be cross validated and reexamined the spatial correlation coefficients to ensure that our statistical models do not have biases (e.g., Elsner and Schmertmann 1994) or any artificial skill (e.g., DelSole and Shukla 2009).

The hindcast skills of different models are compared using the rms error (RMSE). The RMSE is the square root of the mean-squared error (MSE), which as defined by Elsner and Jagger (2006) is the squared difference between the posterior predicted probability and the observation, such that

\[
MSE = \frac{1}{N} \sum_{i=1}^{N} \sum_{k=0}^{1} p_i(k)[k - y_i]^2.
\]
Here $p(k)$ is the posterior predicted probability of a TC activity ($k = 0$ for a low activity and $k = 1$ for a high activity) and $y_i$ is a value equal to 1 or 0 depending whether the high or the low TC activity is observed over all years $i = 1, 2, \ldots, N$. The closer the hindcasts are to the observations, the lower the RMSE, and hence the better the model.

The RMSEs of different models are compared. Overall, both the all-predictor combined model (RMSE = 0.20) and the reduced-predictor model (RMSE = 0.17), on average, perform better than the climatology model (RMSE = 0.67). However, the reduced-predictor model is taken as our final choice, as it has the lower RMSE value and also on the basis that a model with fewer predictors is often preferred to reduce the risk of an overdetermined model (e.g., Mason and Baddour 2008).

Because the Bayesian approach provides a unified probabilistic framework for inferences, it is important to examine the predictive probability of observing TC activity for a particular year. For ease of comparison, we examine the model performance in different ENSO events obtained from the CPC website. The LOOCV hindcasts of TC activity for all El Niño and La Niña years are shown in Figs. 8 and 9, respectively. As is evident, our model has substantial skill to correctly predict TC activity during these ENSO events. With the exception of the 2007/08 season, the maximum probability coincides well with the observations. Similarly, the model also performs very well for the ENSO-neutral years (Fig. 10). With the exception of the 1993/94 season, the maximum probability coincides well with the observations. Note that predictions made here using May–July predictors are fairly robust, so the predictions

![Fig. 8](image-url)  
**Fig. 8.** Predictive distribution of the occurrence of TC activity for El Niño years obtained using LOOCV method. Binary number 1 indicates the probability of the occurrence of high TC activity and 0 indicates otherwise. Asterisks denote the observation.

![Fig. 9](image-url)  
**Fig. 9.** As in Fig. 8 but for La Niña years.
are not updated using subsequent October–December early cyclone season predictors as in CWC10.

b. Fourfold cross validation

To further evaluate the robustness of our predictor set for the reduced-predictor model, we applied a fourfold cross-validation method (e.g., Werner and Holbrook 2011). Here the data is first split into four consecutive 6-yr subsets. The model is then trained on three of the four subsets to hindcast the remaining 6-yr subset.

The predictive distribution of the occurrence of TC activity for each independent 6-yr subset (i.e., 1985/86–1990/91, 1991/92–1996/97, 1997/98–2002/03, and 2003/04–2008/09) is given in Fig. 11. It is found that the results of the fourfold cross-validation strongly compare with those of the LOOCV technique for all years except for the 1994/95 season (cf. respective years of Fig. 11 with those in Figs. 8–10). This additional validation method strongly suggests that our predictor set is very robust and the model skill is not artificial.

8. Summary

This paper presents a simple study on the development and evaluation of a statistical model to make probabilistic forecasts of tropical cyclone formation in the FST region using antecedent large-scale environmental conditions. It is a follow-up study to CWC10 primarily focused on the improvement of prediction skill of TC activity for the ENSO-neutral years.

Years of high TC activity in the FST region are usually associated with El Niño events when large-scale environmental conditions, particularly dynamical conditions such as cyclonic relative vorticity, environmental vertical wind shear, and divergence aloft, are more favorable for TC formation than those in La Niña years when TC activity is relatively low. For the ENSO-neutral years, high TC activity usually occurs in the FST region when environmental conditions, particularly low-level relative vorticity, environmental vertical wind shear, upper-level divergence, and midtropospheric relative humidity, are more favorable in the area north of the Fiji islands and in the Coral Sea region. Existence of different large-scale environmental conditions associated with high and low TC activity years is statistically identified up to the May–July preseason, thus enabling the possible development of a potential statistical prediction model to forecast the state of an upcoming TC season.

A probit regression scheme is developed using the Bayesian approach following the procedure described in Chu et al. (2010) and Chand and Walsh (2011). This scheme makes probabilistic prediction of high and low TC activity in the FST region using accumulated cyclone energy (ACE) as a proxy. Note that the probit regression is considered appropriate here given the binary nature of our TC classification problem. Its implementation using the Bayesian approach gives model parameters in terms of their posterior distribution, facilitating predictive inferences within a probabilistic framework.
Large-scale environmental conditions examined as candidate predictors include relative vorticity at the 850-hPa level, divergence at the 200-hPa level, environmental vertical wind shear between the 200-hPa and 850-hPa pressure levels, midlevel relative humidity, sea surface temperature, and the equivalent potential temperature gradient between 1000-hPa and 500-hPa levels. A number of prediction models based on different predictor combinations of mean May–July large-scale environmental conditions are evaluated. While all models examined for each set appear to have substantial skill, the reduced predictor model consisting of relative vorticity, divergence, and relative humidity is taken as our final choice based on our leave-one-out cross-validation assessment. Results based on hindcasts clearly suggest that our model predicts TC activity over the FST region with substantial skill during the May–July preseason for all years considered in the analysis, particularly for the ENSO-neutral years.

In summary, the present investigation has obvious benefits for the FST region. In addition to improved understanding of the physical mechanisms that account for years of high and low TC activity in the FST region, possible advance warning of the state of such a year using the prediction model developed here can have great socioeconomic benefits in terms of more effective and coordinated preparation for tropical cyclone events. Future work is recommended to update the model as more data become available.

**Acknowledgments.** We greatly appreciate constructive comments on our manuscript by Drs. Elizabeth Ebert and Andrew Dowdy of the Centre for Australian Weather and Climate Research. We also acknowledge our correspondence with Professor Johnny C. L. Chan of the Guy Carpenter Asia-Pacific Climate Impact Centre, City University of Hong Kong, on certain aspects of our work. We are grateful to all reviewers for their constructive comments on our manuscript.

**APPENDIX**

**Probit Regression Model by Truncated Normal Sampling**

Let \( y_i \) be an indicator variable, such that \( y_i = 1 \) when the TC activity is high and \( y_i = 0 \) when the activity is low for various years \( i = 1, \ldots, N \). If there exists an independent normally distributed latent variable vector \( Z_i \), such that
$Z_i = X_i \beta + \epsilon_i; \quad \epsilon_i \sim N(0,1); \quad i = 1, \ldots, N$. \hspace{1cm} (A1)

then for each year $i$, we define a binary class label $Y_i = 1$ if $Z_i \geq 0$ and $Y_i = 0$ if $Z_i < 0$, where $\beta$ represents the model coefficient vector associated with the predictor variables $X_i$ and $\epsilon_i$ is a noise vector assumed to be normally distributed with a mean of 0 and a variance of 1. The $Z_i$ is, of course, not known since the exact posterior distribution of parameter vector $\beta$ is not known. However, given the data $y_i$, the posterior distribution of regression parameter vector $\beta$ can be simulated via the Bayesian approach as discussed in, for example, CWC10 and Chu et al. (2010). Consequently, the values of $Z_i$, and hence the corresponding binary class label vector $Y_i$, can be obtained from a truncated normal distribution such that

$$Z_i \mid y_i, \beta \sim N(X_i^T \beta, 1)$$

truncated at the left by 0, if $y_i = 1$, and

$$Z_i \mid y_i, \beta \sim N(X_i^T \beta, 1)$$

truncated at the right by 0, if $y_i = 0$. \hspace{1cm} (A2)

The probability of “high TC activity” $P(y_i = 1)$ and the probability of “low TC activity” $P(y_i = 0)$ can therefore be expressed as

$$P(y_i = 1) = \frac{1}{L} \sum_{i=1}^{L} Y_i$$

and

$$P(y_i = 0) = 1 - P(y_i = 1).$$ \hspace{1cm} (A3)

where $L$ is the number of output of the Gibbs sampler ($L = 10000$ in our case).

REFERENCES


