Statistical prediction of sea breezes in Sydney Harbour

William T.M. Dunsmuir
University of New South Wales, Sydney, New South Wales, Australia

Elly Spark
Bureau of Meteorology, Sydney, New South Wales, Australia

Sung-Kwan Kim
University of New South Wales, Sydney, New South Wales, Australia

and

Shuling Chen
University of New South Wales, Sydney, New South Wales, Australia

(Manuscript received July 2002; revised March 2003)

This paper develops a logistic regression model for predicting the probability of a sea-breeze occurring at Fort Denison in Sydney Harbour during September, the period during which the Olympic Games were held in Sydney in 2000. Starting with approximately 140 potential predictor variables, a reduced set of nine derived predictor variables was obtained from which the final prediction model was derived. These variables appear to be meteorologically consistent and the final model performs very well on the training dataset and on a set of test data and gives acceptable false prediction error rates.

Introduction

This work was prompted by a requirement to forecast detailed wind information for Sydney Harbour during the sailing events of the Sydney 2000 Olympic Games. Sydney is located in a basin, which drains to the coast in the east. It is bounded in the west by the Blue Mountains, which rise to about 1000 m, and to the north and south by higher ground to about 400 m. These topographic features and the proximity to the ocean have major influences on the winds in the basin causing a number of local winds which cannot be captured by conventional synoptic scale numerical weather prediction (NWP) models. At the 1996 Olympics, two high resolution models at 8 km and 10 km were used to help predict sea-breezes (Rothfusz et al. 1998; Powell and Rinard 1998.) However, in 1996, when the Bureau of Meteorology commenced planning for the 2000 Olympics, there were no Australian operational models at this scale. High resolution versions were planned for its Limited Area Prediction System (LAPS) model (Puri et al. 1998) but the eventual performance was unknown. Moreover, as Sydney Harbour is only about 2 km wide, even a 5 km resolution model is not capable of capturing the wind variations which occur within the harbour. General features of the wind regimes in Sydney include a prevalent, persistent, light, overnight drainage flow from the west and a local daytime sea-breeze. During win-
ter, this sea-breeze is generally absent, and during summer it is frequent.

During the three months August through October, the sea-breeze occurs on about 40 per cent of days on Sydney Harbour. However, it is generally weak. Tracking of the sea-breeze on Doppler radar during September in 2000 showed that it can reach the western parts of Sydney, 60 kilometres inland, but may take over five hours to do so. On other days, it may penetrate only just beyond the coast. The forecasting of the winds is therefore a complex problem. Methodologies to forecast these winds could include high resolution numerical modelling, or a number of statistical techniques. Further details on features of winds in the region and a review of the methods being investigated for forecasting winds are provided in Connor et al. (2003).

An important part of predicting winds near a coastal location is to first predict the occurrence of a sea-breeze. Once that is determined the onset time and strength of the sea-breeze can each be modelled using alternative regression techniques as discussed in Spark et al. (1999, 2000). Operationally, during the Olympics, there was a requirement for the prediction of wind speed and direction. The methodology described in this paper was utilised as part of the ‘Regression Breeze’ model described in Connor et al. (2003). Verification statistics and comparisons with other prediction methods of wind speed and direction are presented for ‘Regression Breeze’ in Spark and Connor (2003). This paper describes in detail the development of a logistic regression model for predicting the probability of a sea-breeze occurring at Fort Denison in Sydney Harbour.

To date, there have not been any studies published concerning forecasting sea-breeze occurrence in Sydney Harbour using statistical models. Connor (1997) reports a similar study in Townsville, Queensland for the world ‘505’ class of single-hulled yachts during April 1996. He used statistical short term forecasting techniques to forecast wind speed and direction and temperature. He considered 198 potentially useful predictor variables, which were chosen on the basis of a physical or meteorological connection with the predictand. The predictor variables were gathered from three different predictor sets; potential observed predictors, potential forecast predictors and potential derived and combined predictors. Details can be found in Connor (1997). We follow a similar approach using analogous predictor variables. Connor derived a prediction of the occurrence or non-occurrence of a sea-breeze using least squares regression methods. In this paper we replace least squares with logistic regression which is a more appropriate statistical technique for predicting absence/presence responses.

In the next section we define sea-breeze events. In following sections we describe the steps taken to compile a large database of potential forecast variables which are considered in the model, and then we describe the logistic regression model that we use and explain how this was derived. Finally, the model is evaluated on the sample of data used to fit it, in addition to an evaluation on an independent dataset.

Overall, the model performs very well and has an operationally useful degree of accuracy for identifying sea-breeze events.

Defining sea-breeze events

The sea-breeze is a local wind circulation set up by the difference in temperatures over the land and sea (Crowder 1995; Zhang and Takle 1992). As the land heats during the day, the air above it becomes less dense and is replaced by denser air from over the sea. Above this surface flow, a reverse circulation takes place. The height of this reverse circulation can vary significantly. It is affected by atmospheric stability and the insolation. The sea-breeze depth is typically much greater in summer than in winter. In Sydney during September, in early spring, the surface sea-breeze flow remains below 900 hPa, although the reverse circulation can occur near this level. At the ground, and under very light synoptic gradient wind conditions, the surface component of the breeze will start perpendicular to the coast (approximately from the east in Sydney) and then gradually back to the northeast under the influence of the Coriolis force. The gradient wind is the wind that balances the pressure gradient force, the Coriolis force, and the centripetal acceleration due to the curvature of the isobars. It should be free from the influence of friction, buoyancy and sea-breezes, and would be expected to occur above the boundary layer. In common usage, forecasters within the Australian Bureau of Meteorology refer to wind at the 2000 ft (950 hPa) or 3000 ft (900 hPa) levels as the gradient wind, although the boundary layer can extend well above these heights.

The closed circulation is a definitive element of the sea-breeze phenomenon. Therefore, we distinguish between two modes of an onshore surface wind. Firstly, a classic sea-breeze where, after sufficient insolation, the surface air is onshore and an upper return flow is distinguishable. Secondly, an onshore synoptic wind from the sea, where the surface air and all the boundary layer winds have an easterly wind component with respect to the shoreline. We will use the term sea-breeze only to refer to the former. When the gradient wind is not negligible, the development
of the sea-breeze will be enhanced or limited, dependent on the speed and direction of the gradient wind (Houghton 1992; Spark and Connor 2003).

The data for this study were assembled from surface measurements at Fort Denison within Sydney Harbour, upper air data at Sydney Airport located approximately ten kilometres to the south, maximum and minimum temperature data at Richmond, located approximately 60 km inland, and sea surface temperature data at the Offshore Reference Station buoy, located about 3 km offshore south of Sydney Harbour. The data used were from 1991 to 1999 inclusive and a three-month window centred on September is used to increase the size of the dataset. Figure 1 is a frequency plot of the surface wind directions of all Fort Denison data for September for the years 1991 to 1998. This plot shows the surface wind direction frequency variation with time of day. Similar plots were used in describing the wind climatology at Savannah, Georgia, United States for the 1996 Olympic Games (Powell and Rinard 1998). The overnight flow is predominantly westerly. In the afternoon, in broad terms, there are four wind regimes: westertlies, southerlies, and two types of easterlies. Some easterlies continue as easterlies, and some become northeasterly as the day progresses. It is this last group of winds, the classic sea-breeze, which we wish to predict. Their direction is from 20° through to 150°. Figure 1 indicates that winds from 150° to 200° are most likely to be part of the ‘southerly’ group of winds.

Figure 2, a plot at 1500 LST (Local Standard Time which is UTC + 10) of the 900 hPa wind direction versus the Fort Denison surface wind directions for every half hour of the day, indicates there are essentially three different regions. These are the gradient wind offshore with the surface wind sea-breeze between 20° and 150° for Fort Denison (region A in Fig. 2), the gradient wind and the surface wind both onshore (region B in Fig. 2) and the remainder (region C in Fig. 2).

For predicting surface wind speed and direction individual linear regression equations must be developed separately for each of these three regions (Spark et al. 1999). Separation of region B from regions A and C is straightforward using gradient wind alone. However, to separate the data in region A (sea-breeze events) from region C of Fig. 2, a predictive model is required.

The direction of the dominant shoreline in the Sydney Region was taken to be 20° from north. A gradient wind is considered to be offshore if it has direction between 200° and 20° and onshore for directions from 20° to 200°. Hence there are two factors that we considered in arriving at our operational definition of sea-breeze events. The first is whether the gradient wind is offshore. The second is whether the Fort Denison surface wind direction is from 20° to 150° in the afternoon. Accordingly, we define sea-breeze events in terms of surface wind observations conditional upon gradient winds being offshore.

Several aspects need to be considered including the time at which we measure the gradient wind and the height at which we measure it. Sydney Airport data were available at heights of 950 hPa, 900 hPa and 850 hPa and at times of 0300 LST, 0900 LST and 1500 LST. The 0300 LST and 0900 LST data were
Initially rejected because of the possibility of synoptic changes occurring between the morning and afternoon. The 950 hPa winds were rejected as falling within the surface onshore flow of the sea-breeze circulation. Under light wind conditions occasionally onshore winds were observed at 850 hPa, but a sea-breeze circulation nevertheless developed with a return flow near 900 hPa. In our definition of the sea-breeze we therefore used the 900 hPa 1500 LST winds as representative of the gradient wind.

Various objective determinations of the sea-breeze occurrence can be constructed. The definition adopted for the final model presented here is that the 900 hPa wind at 1500 LST was offshore and the Fort Denison surface-wind direction at 1430 LST, 1500 LST and 1530 LST was consistently between 20° and 150°. For this definition there was a 40 per cent incidence of sea-breezes in the August, September, October period.

Development of the predictor data

Data sources

In this section we describe the process of selection of predictor variables that were considered as potentially useful for predicting the occurrence of sea breezes at Fort Denison. Following the development in Connor (1997), suitable observable meteorological variables were identified from which a number of other variables, thought to influence sea-breeze formation, were derived. Since not all relevant variables were available at a single location the following four sources of available data or forecasts were used.

1. Sydney Airport radiosonde data.
2. Richmond surface recordings.
3. Observatory Hill surface recordings.
4. Offshore Reference Station buoy data.

Upper air data were obtained at several times throughout the day (0300, 0600, 0900, 1200 and 1500 LST) and at several pressure levels (surface, 950 hPa, 900 hPa, 950 hPa and 700 hPa) which varied with location or source of data. In developing the prediction model, in addition to the directly observed predictor variables the model used variables that needed to be forecast prior to being used in the model and variables that were derived or combined from the observed or forecast predictors. The sea-breeze forecasts needed to be made prior to 0900 LST. Therefore, data input at 0900 LST had to be forecast in the operational implementation of the prediction model. For any variables required to be forecast, we used a perfect prognosis approach (Klein et al. 1959) which assumed that we were able to predict these variables perfectly. We discuss this further below.

Predictors likely to be useful for surface-wind prediction are pressure gradient, temperature, vertical velocity, relative vorticity, divergence, temperature advection, vorticity advection, moisture advection and stability indices. Similar to Connor (1997) we also investigated nonlinear effects based on transformation and combination of basic predictors in developing the forecast models. The derived or combined variables were selected using meteorological considerations as was done in Connor (1997). This process led to a possible 144 potential predictor variables being obtained for further consideration. Using a variety of graphical techniques, this large number of potential predictors was substantially thinned prior to logistic model fitting. Further details are provided in the next section. Some of the potential predictors were available early enough to be used directly in the model for the prediction of sea-breeze occurrence while others, as noted above, required forecasts.

Rectangular coordinates of wind

To account for the importance of land/sea thermal contrasts (Connor 1997) and to avoid modelling problems with using polar coordinates of winds, especially the direction, we transformed the original wind speed, s, and direction measurements, θ, into rectangular coordinates resolved into components parallel and orthogonal to the coastline. By approximating the angle of the regional coastline by a straight line at 20° east of north we derive onshore winds at time t and pressure p as

\[ U(t,p) = s(t,p) \sin(\theta(t,p)-20) \] and parallel to shore winds as

\[ V(t,p) = s(t,p) \cos(\theta(t,p)-20) \]

Averaged and differenced weather fields

In order to smooth out local variations or to overcome missing data problems in some of the observed fields, we applied temporal or vertical averaging. For example, average onshore boundary layer winds were formed at 0300 LST and 0900 LST as follows

\[ U(t) = (U(t,9500) + U(t,9000) + U(t,8500))/3 \]

where the average was formed over the available fields, in case any pressure level had missing data. At 1500 LST the averages used the 900 and 850 hPa fields only in order to avoid any influence of the surface onshore sea-breeze flow as this is frequently observed to intrude beyond the 950 hPa level but is unlikely to reach the 900 hPa level in September. The speed and directions were also calculated using these average winds.
Morning and afternoon average cloud cover variables were formed for both low cloud and total cloud by averaging the 0900 LST and 1200 LST readings and the 1200 LST and 1500 LST readings. If cloud cover is increased then there is less solar radiation reaching the surface and hence less mixing is expected within the boundary layer. As low cloud is generally more dense than upper cloud in Sydney, low cloud should be more influential than higher cloud in this process.

Derived fields
In addition to the above directly observable variables, and averages or differences of them, a number of derived variables of potential meteorological relevance were defined. Connor (1997), in his modelling for Townsville winds, found that important variables included upper wind components, lapse rates, thermal parameters such as land/sea temperature differences, 0900 LST surface-wind components, temperature advection terms and return flow indices. Detailed definitions of the derived variables included in the final logistic regression model are given below.

Developing the predictive model

The logistic regression model
Prediction of sea-breeze occurrence requires prediction of a binary valued outcome variable for which logistic or probit regression are appropriate techniques. These two methods usually give very similar predictive models and both will give valid predictions of probabilities between 0 and 1. An alternative that is sometimes used is a linear least squares regression model. This can be unsuitable since predictions of probabilities are not guaranteed to be between 0 and 1.

To describe the logistic regression model for predicting sea-breeze occurrence we define, on day $d$, the binary random variable $Y_d$ which takes the value 0 on a day where no sea-breeze occurred and the value 1 on days when a sea-breeze occurred. Here, a sea-breeze is defined using the operational definition given above. Given a set of predictors $x'_{d}$ (a row vector) on day $d$ we model the probability of a sea-breeze, $p_d=P(Y_d=1)$, occurring as

$$p_d = \frac{\exp(x'_{d}\beta)}{1 + \exp(x'_{d}\beta)} \quad \text{...4}$$

Note that the log odds of $p_d$ is given by

$$\logit(p_d) = \log\left(\frac{p_d}{1-p_d}\right) = x'_{d}\beta \quad \text{...5}$$

which is a linear function of the predictors using regression parameters $\beta$ which are estimated in the logistic regression fitting process. In our analysis the first element of $x_d$ is set to 1 to allow an intercept term in the linear function for the logit. Details about this model can be found in McCullagh and Nelder (1989). Maximum likelihood methods of fitting and assessing the model using generalised linear model fitting software are readily available in a variety of statistical packages. We used the S-PLUS statistical analysis package.

Preliminary screening of the variables
In all, there were initially over 140 potential predictor variables that might have been included in the model. Of these many were highly correlated with each other because of their proximity in space, time or pressure level. Clearly it would not be reasonable to include all of these in the logistic regression fitting process. Accordingly, steps were taken to cull these variables to a reasonable number with which to start the modelling process. In selecting these variables, several aspects were considered.

Our first step was to identify variables, or groups of highly collinear variables, that were unlikely to be useful for prediction of sea-breeze occurrence. Boxplots of each predictor variable were stratified into days with and without a sea-breeze and were then examined. This is a simple but powerful technique, which is illustrated in Fig. 3. This shows two examples of such box plots: Fig. 3(a), where the variable clearly discriminates between occurrence and non-occurrence; and, Fig. 3(b), where it does not. Any predictor variable that showed good separation with respect to the response variable, $Y_{d}$, was retained as a potentially important predictor for the next stage of the analysis. As a result of this preliminary screening, approximately 45 variables were dropped from further consideration leaving 98 potentially useful predictor variables.

Of these 98 predictor variables a number were highly collinear as could be expected. The use of highly collinear variables in logistic regression has the potential to lead to problems with the model fit and selection of variables. Measurements obtained at two or three different pressure heights, or that were proximate in time, were often highly correlated. For such groups of variables we chose to select a representative pressure height or, in some cases, to overcome missing data problems, averaged the measurements across the pressure heights. Averaging had the additional effect of smoothing out local variability in the predictor variables. Additionally, scatter plots showed some of the derived variables were highly collinear with the original measured variables. Any such derived variables were dropped from the potential predictor set.
Fig. 3  Split boxplots for two potential predictor variables. (a) 900 hPa wind speed at 1500 LST, (b) 700 hPa wind direction at 0900 LST. The boxplots of the predictor variables are stratified according to the non-occurrence (0) and occurrence (1) of a sea-breeze.

(a)

(b)

Finally, some variables that did not show clear separation on the split box plots were included because they were thought to be of meteorological significance. The cloud cover variables are an example of such potential predictor variables. Because of their discrete measurement scale, the use of box plots was not very informative.

Initial logistic regression model
After final screening by this process there were eighteen predictor variables remaining to test for the capacity to predict sea-breeze occurrence. This was considered to be a reasonable data set on which to perform the logistic regression fitting. It contained only a moderate number of variables, and these variables appeared to be logical on meteorological grounds and on previous research (Connor 1997). They were largely free of colinearity problems and many of them showed good separation of their distributions when stratified by the occurrence of a sea-breeze.

The logistic regression model using these variables was initially fitted with data from August, September and October for the period 1991 to 1997 inclusive (Kim 1999). In this model development, the sea-breeze definition required that the average speed at these three times was at least five knots, in addition to the directions at 1430 LST, 1500 LST and 1530 LST being consistently between 20° and 150°.

The $r$-statistics for individual coefficients were used to remove variables which were not significant at the five per cent level from the logistic regression model. This left eight variables. A partial deviance test (McCullagh and Nelder 1989) on the omitted variables was not significant at the five per cent level confirming that they could be removed from the model with no reduction in explanatory power. This model was used operationally during the 1999 Olympic Test Events with prediction performance results reported in Table 1.

Final model fit
While the initial model proved useful to forecasters, the time needed to input the required data, including forecast winds at several levels and times, was operationally unacceptable. In order to simplify its operational use for the Olympic Games, the final model excluded the use of averages of vertical wind speeds. Averages were replaced by values obtained at the single pressure level of 900 hPa. The model was fit using a total of 405 days taken from the months of August, September and October for the period 1991 to 1998 inclusive and for which there were no missing values in the predictors. This resulted in the use of nine predictor variables. Table 2 summarises the estimates and $r$-values of the variables obtained. The null deviance was 561.45 on 404 degrees of freedom and the residual deviance was 228.76 on 395 degrees of freedom.

The predictor variables in Table 2 are defined as follows.
1. Wind speed at 0900 LST and 900 hPa, $u(0900,900)$. This variable appears in the model with a negative sign which is as expected. When the gradient wind speed increases under offshore conditions it will eventually become so large that it will inhibit the gradient wind.
2. Dew-point temperature at 0900 LST, $T_d(0900)$. This has a positive sign in the model.
3. The atmospheric (900 hPa) wind components onshore, $U(1500,900)$, and parallel to shore, $V(1500,900)$, in the afternoon at 1500 LST. These
Table 1. Performance assessment of initial sea-breeze prediction model fit using data from August, September, October 1991 to 1997.

<table>
<thead>
<tr>
<th></th>
<th>SB Occurred</th>
<th>POD</th>
<th>No SB Occurred</th>
<th>POFD</th>
<th>KSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data: Logistic Regression</td>
<td>SB Predicted</td>
<td>152</td>
<td>87%</td>
<td>24</td>
<td>13%</td>
</tr>
<tr>
<td></td>
<td>No SB Predicted</td>
<td>22</td>
<td>155</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1998 Test Data: Logistic Regression</td>
<td>SB Predicted</td>
<td>20</td>
<td>80%</td>
<td>5</td>
<td>22%</td>
</tr>
<tr>
<td></td>
<td>No SB Predicted</td>
<td>5</td>
<td>18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1999 Olympic Test Event</td>
<td>No SB</td>
<td>1</td>
<td>100%</td>
<td>2</td>
<td>29%</td>
</tr>
<tr>
<td></td>
<td>No SB Predicted</td>
<td>0</td>
<td>5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Final Model Variables. Refer to text for definitions.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.5179</td>
<td>-3.17</td>
</tr>
<tr>
<td>s(0900,900)</td>
<td>-0.1590</td>
<td>-3.44</td>
</tr>
<tr>
<td>$T_d$(0900)</td>
<td>0.3032</td>
<td>5.81</td>
</tr>
<tr>
<td>$U$(1500,900)</td>
<td>0.2438</td>
<td>4.34</td>
</tr>
<tr>
<td>$V$(1500,900)</td>
<td>0.2923</td>
<td>5.24</td>
</tr>
<tr>
<td>$L$</td>
<td>0.4622</td>
<td>5.70</td>
</tr>
<tr>
<td>TT</td>
<td>-0.0326</td>
<td>-1.67</td>
</tr>
<tr>
<td>$L_{p,U}$</td>
<td>0.0148</td>
<td>3.43</td>
</tr>
<tr>
<td>$L_{p,V}$</td>
<td>0.0071</td>
<td>1.82</td>
</tr>
<tr>
<td>$C_{low}$</td>
<td>-0.2166</td>
<td>-1.82</td>
</tr>
</tbody>
</table>

Lapse rates are used to measure atmospheric stability. This particular lapse rate is a measure of afternoon instability within the boundary layer and is derived using the 0600 LST upper atmosphere as a reference. A large positive value indicates unstable conditions which should promote overturning and thereby help initiate the sea-breeze (Houghton 1992).

5. The Total-Totals index is another simple stability index

$$TT = T(0600,850) + T_d(0600,850) - 2T(0600,500) \ldots 7$$

where $T_d(0600,850)$ is the 850 hPa dew-point temperature at 0600 LST and $T(0600,500)$ is the 500 hPa temperature at 0600 LST. This index is commonly used as an indicator of thunderstorm potential and in the Sydney region values over 50 imply an increased probability of them occurring. It is a measure of the instability above the boundary layer. When this variable is large we would expect deep convection. This is likely to result in mixing down to the surface of winds at about 750 hPa or higher. The stronger westerly wind component at this pressure height could inhibit sea-breeze formation. This would suggest that an increased value of this index will lead to a decreased chance of sea-breeze formation as indicated by the negative sign on the coefficient for this term in the model fit.

4. Lapse rates indicate temperature differences at different pressure heights. A number of lapse rates were used. The lapse rate that we retained in the final model is defined as

$$L = T_{max,R} - T(0600,850) \ldots 6$$

where $T_{max,R}$ is the daily maximum temperature at Richmond and $T(0600,850)$ is the 850 hPa temperature at Sydney Airport measured at 0600 LST.

6. Wind variable products combine temperature contrast or lapse rates with wind variables and provide a measure of the tendency for mixing in the lower atmosphere. The final model used the following variables:

$$L_{p,U} = [T_d(0600) - T(0600,850)] \times U(0900,850) \ldots 8$$

$$L_{p,V} = [T_d(0600) - T(0600,850)] \times V(0900,850) \ldots 9$$

where $T_d(0600)$ is the 0600 LST surface temperature at Observatory Hill. It is not clear from meteorological considerations which sign should be attached to the coefficients of these variables.
7. Morning low cloud \( C_{\text{low}} \) defined to be the average of the low cloud readings at 0900 LST and 1200 LST at Observatory Hill. The sign on the model coefficient is negative as expected because with increased morning cloud there is less heating over land reducing the thermal contrast required for sea-breeze formation. Additionally, with increased cloud there is less vertical mixing in the boundary layer which is also expected to inhibit sea-breeze formation.

Note that one might have expected the model to directly include a thermal contrast variable (defined as the maximum temperature at Richmond minus the off-shore sea surface temperature). Since the sea-surface temperature varies only by a few degrees in the study period we suggest that lapse rate, \( L \), and morning low cloud, \( C_{\text{low}} \), are together acting as a substitute for thermal contrast in the model, because both of these are dependent on the Richmond maximum temperature.

**Evaluation of the forecasts**

The final fitted logistic regression model was used to forecast the occurrence of a sea-breeze whenever the estimated probability

\[
\hat{P}_d = \frac{\exp(x'_d \hat{B})}{1 + \exp(x'_d \hat{B})}
\]

is at least 0.5. The actual sea-breeze occurrences were cross-tabulated against the prediction of sea-breeze occurrences using the logistic regression model for two periods of data.

1. Training data (August, September and October in 1991 to 1998) based on 405 cases.
2. Test data (August, September and October in 1999) based on 49 cases for which all relevant data were available.

The performance of the logistic regression model was not assessed separately during the actual Sydney 2000 Olympic Games. However as mentioned in the Introduction, Spark and Connor (2003) do present such an assessment for the ‘Regression Breeze’ model of which this logistic regression is the first step in the overall forecast process.

The results are summarised in Table 3. We consider the results obtained using the logistic model to be very good for the two sets of data on which it was evaluated. As could be expected, the results for the training dataset are slightly better than for the test data and both outperform simple persistence forecasts. Note that this table includes only those days for which the 900 hPa wind at 1500 LST was offshore. From this table it can be seen that the number of sea-breeze days, as a percentage of the days when the gradient wind was offshore, was approximately 50 per cent. This compares with a 40 per cent occurrence of sea-breezes for all August, September, October days.

We also report the Hansen and Kuiper skill score (KSS) in Table 3. To define this, consider the simple 2 x 2 contingency table of the form:

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Not Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forecast</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td>Not Forecast</td>
<td>c</td>
<td>d</td>
</tr>
</tbody>
</table>

Using this notation, various equivalent definitions of the skill score can be defined (Wilks 1995; Seaman et al. 1996). One such definition is

\[ KSS = POD - POFD \]

where the probability of detection is

\[ POD = \frac{a}{a + c} \]

and the probability of false detection is

\[ POFD = \frac{b}{b + d} \]

This skill score can also be interpreted as a means for comparing the forecasts with random unbiased forecasts. It is equitable, in that both successfully forecast hits as well as successful forecast misses contribute to the overall score. For a perfect forecast \( KSS = 1 \). If there is an even distribution and there is no discrimination (a,b,c,d all equal) \( KSS = 0 \). Table 3 reports the values of POD, POFD and the KSS skill score, and demonstrates satisfactory skill for the logistic regression model.

**Discussion**

The final model relies on the perfect prognosis of a number of variables. The most significant of these are the forecast afternoon maximum temperature at Richmond and the afternoon wind speed and direction at 900 hPa. Although the 900 hPa level is above the surface flow of the sea-breeze circulation in September, Fig. 4 shows that this level can be within the return flow. In using a perfect prog approach, forecasts of this parameter derived from synoptic-scale models are not influenced by any potential sea breeze. However, in developing the equations, the return flow may have influenced this parameter to some extent.
Table 3. Performance assessment of final sea-breeze prediction model.

<table>
<thead>
<tr>
<th></th>
<th>SB Occurred Number</th>
<th>POD</th>
<th>No SB Occurred Number</th>
<th>POFD</th>
<th>KSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training Data:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>SB Predicted</td>
<td>183</td>
<td>88%</td>
<td>31</td>
<td>15%</td>
</tr>
<tr>
<td></td>
<td>No SB Predicted</td>
<td>24</td>
<td></td>
<td>167</td>
<td></td>
</tr>
<tr>
<td>Training Data:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persistence</td>
<td>SB Predicted</td>
<td>100</td>
<td>63%</td>
<td>59</td>
<td>39%</td>
</tr>
<tr>
<td></td>
<td>No SB Predicted</td>
<td>60</td>
<td></td>
<td>93</td>
<td></td>
</tr>
<tr>
<td>1999 Test Data:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>SB Predicted</td>
<td>29</td>
<td>91%</td>
<td>4</td>
<td>24%</td>
</tr>
<tr>
<td></td>
<td>No SB Predicted</td>
<td>3</td>
<td></td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>1999 Test Data:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Persistence</td>
<td>SB Predicted</td>
<td>177</td>
<td>61%</td>
<td>115</td>
<td>39%</td>
</tr>
<tr>
<td></td>
<td>No SB Predicted</td>
<td>114</td>
<td></td>
<td>184</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4  Time series of upper-wind components at Sydney Airport for the period 19-22 September 2000. The onshore components were compiled from pilot balloon data at 1000 ft intervals and displayed in m s$^{-1}$. The winds at 0000, 1200 and 1800 LST were interpolated from adjacent data, and onshore components are rendered as cool colours (blue) and offshore components as warm colours (red). The average gradient winds at 0900 LST and 2100 LST for each day are also shown as part of the time coordinate.

**Vertical Cross-section Sydney 19-22 September 2000**

We suggest that in future work, the 900 hPa winds be used at 0900 LST or 0300 LST to avoid this problem. This will also have the advantage to permit generalisation of the method to a whole year as the height of the sea-breeze circulation will not be an issue in developing the equations. Since the 900 hPa winds at 0900 LST and 1500 LST are highly correlated and either could have been selected from the initial list of potential predictor variables, this should make little difference to the final performance. Also, all the variables used at 0900 LST require prediction, although only at a very short time-scale. Use of the 0300 LST
winds would have the advantage of having the 1500 LST Richmond maximum temperature as the only forecast parameter in the model.

**Summary**

We have derived a model for predicting the occurrence of sea-breeze days, based on logistic regression relating the predictors to the probability of a sea-breeze occurring. Predictors were selected from a number of sources of relevant meteorological data. Meteorological knowledge, based on earlier work by Connor (1997), was used to derive new predictor variables likely to be important in forecasting sea-breeze occurrence. Graphical techniques were used to cull the large number of potential predictors to a manageable number of nineteen predictors prior to logistic regression fitting. Analysis of the significance of the coefficients led to a further reduction in the number of predictors to a final list of nine. The final model had coefficients for these variables that were sensible in terms of known influential factors for a sea-breeze. The predictive performance of the model within the training dataset and the test dataset were very good, in that they gave acceptable false prediction error rates and compared well with persistence.

**Acknowledgments**

The authors thank Greg Connor of the Australian Bureau of Meteorology, Townsville, for numerous valuable discussions and for permission to use Fig. 4. Funding for this research was provided by an Australian Research Council SPIRT grant in collaboration with the Australian Bureau of Meteorology.

**References**


