CHAPTER 3
METHODOLOGY

3.1 Observation data

The study obtained hourly data from 2001-2006 at meteorological stations of the Pollution Control Department (PCD-M2 and PCD-31), in the study area (Figures 3.1, 3.2 and Table A1). The two stations are 1.9 km apart, and about 6 km from the coastline. The key meteorological variables considered were wind speed and direction, global solar radiation, near-surface air temperature and sea surface temperature (RTG-SST data, from NCEP, U.S, Gemmill et al., 2007). The winds recorded at a height of 100-m at the PCD-M2 tower were selected for modeling to minimize the effects of interference from nearby obstacles or objects, such as trees and buildings. The global solar radiation data used was recorded at the nearby PCD-31 air quality monitoring station since sheltering of the pyranometer was absent when compared to PCD-M2. The wind speed (S), global solar radiation (GR), vertical temperature difference (VTD, here, 2-m air temperature minus 75-m air temperature), and land-sea temperature difference ($LSTD$, here, 2-m air temperature minus sea-surface temperature, Phan and Manomaiphiboon, 2012) were selected for a multivariate model. The selection of the variables is done in order to check the strength of the relationship as to which factors are most likely to influence wind speed. For a VAR model, in this particular study, arising from the relationship between various atmospheric variables such as wind speed, global solar radiation, and the temperature (within the surface heat flux) as shown in the log-law wind profile (Equation 2.1), it would be of interest to consider incorporating these variables for the task of wind speed forecasting. Initial plotting of the data was done to check for consistency, upon which quality checking for missing values, unusual patterns, and outliers was performed.

In this study, the input dataset was prepared for each month by concatenating all data of a particular month from all years. This is done since the meteorological variables in the different months have unique dynamics inherent in them depending on the season, for which different processes would be needed to represent them, as was done in previous studies (Torres et al., 2005). For each month, the data from the first four years were used as a training set to identify an optimal time-series model (i.e. to estimate model orders) as

outlined in the following sections, and the remaining data used for forecasting and performance evaluation (Figure 3.3). The trend in a time series is a slow, gradual change in some property of the series (e.g. mean) over the whole period of analysis. In case the data exhibits an increasing trend over the long-term, the autocorrelations are modified, since the trend will add in correlations that are not accounted for in time-series models. The application of the time-series models has the assumption that the time series is stationary (i.e. the process remains stable along a constant mean). Thus, detrending is done to satisfy this assumption. Simple linear trend in mean can be removed by subtracting a least-squares-fit straight line from each data point. In every fitting and forecasting, input time series is first detrended using ordinary linear regression.

(Intentionally left blank)

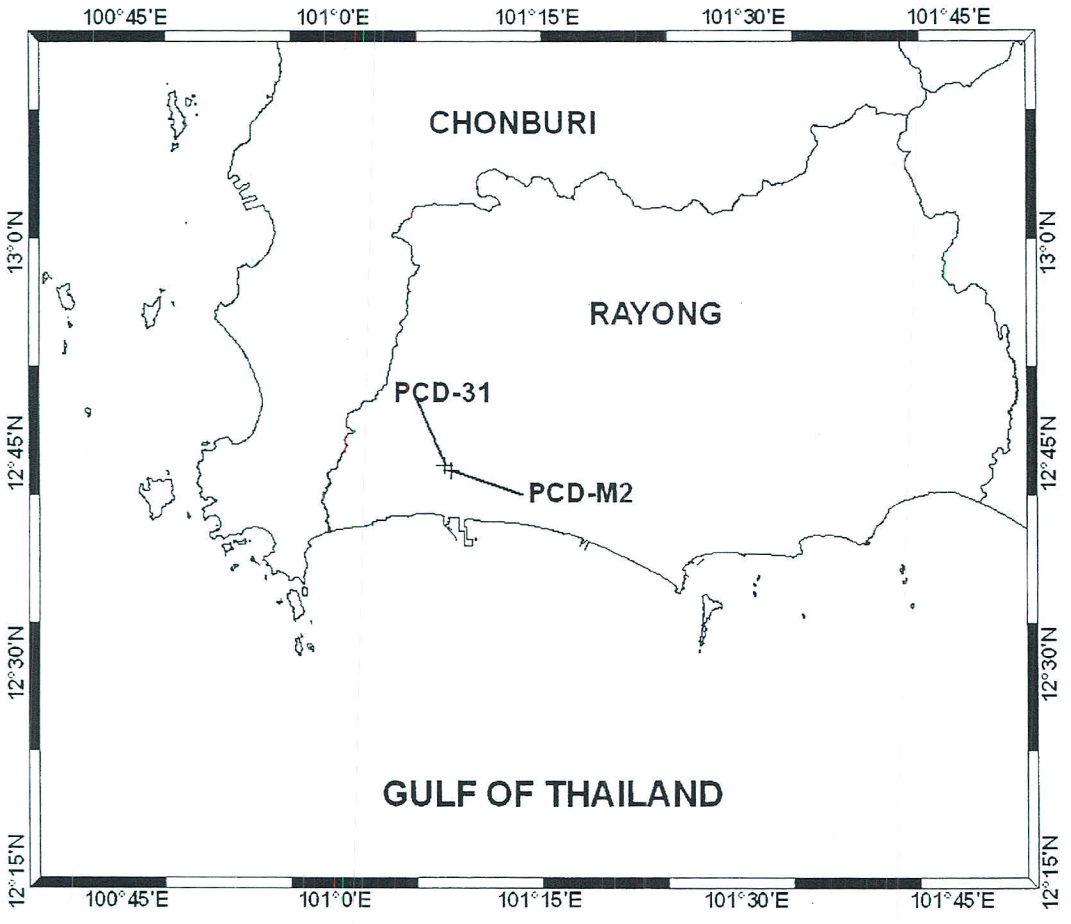


Figure 3.1: Meteorological stations PCD-M2 and PCD-31

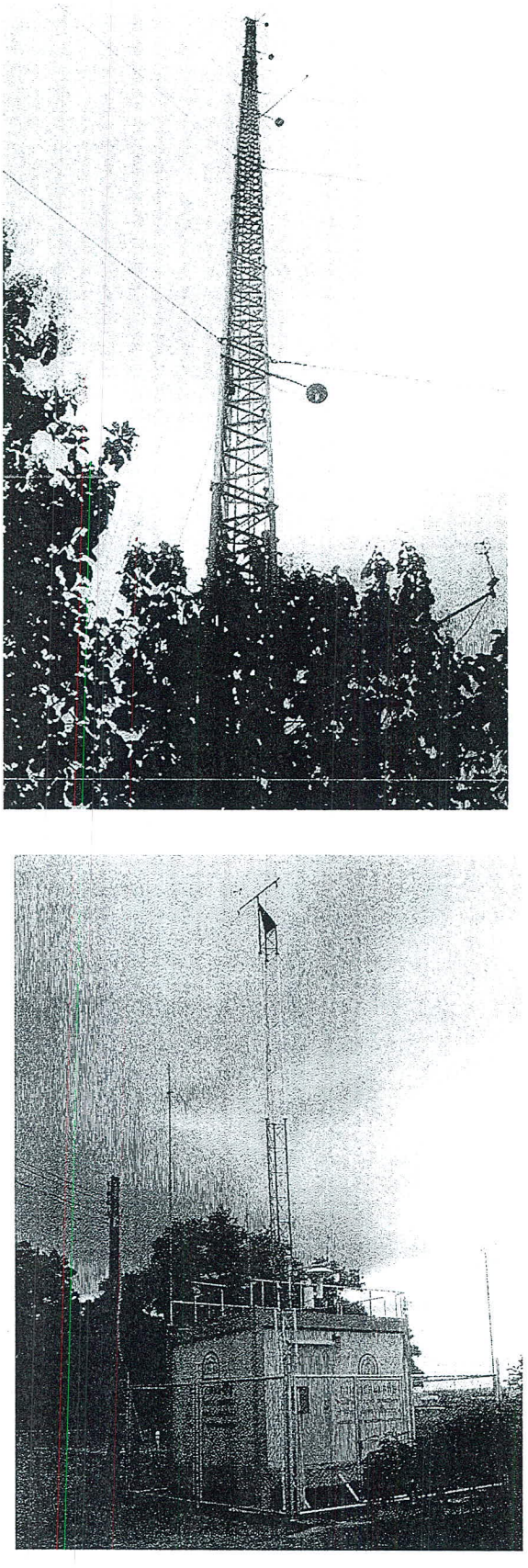


Figure 3.2: PCD-M2 (top) and PCD-31(bottom)

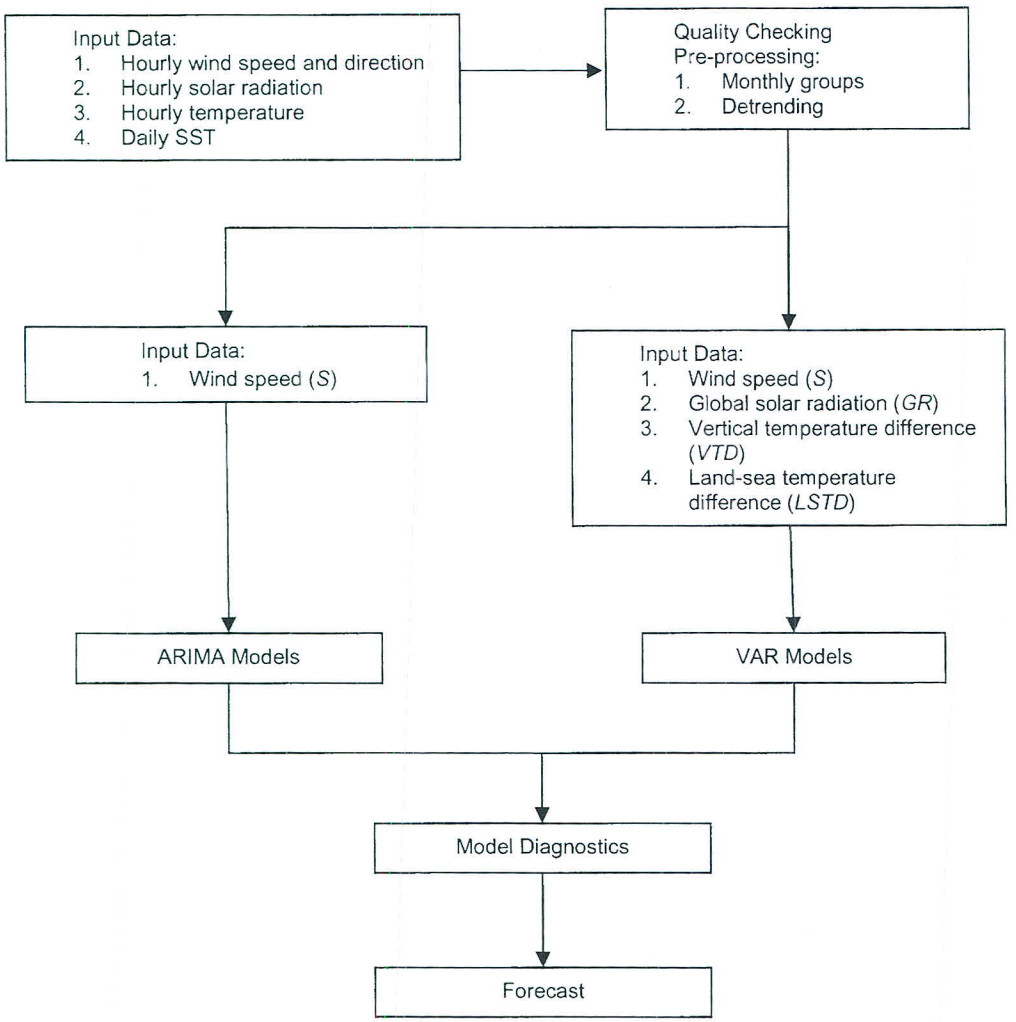


Figure 3.3: Flow chart for data pre-processing and time-series modeling

3.2 Observation data characteristics

The meteorological variables used for this study were surface wind, global solar radiation, near-surface temperature and sea surface temperature. Prior to modeling, it was needed to understand the behavior of these variables, such as the relation between any two variables and between the variable and its past values. This was done by comparative plots and the plots of the auto-correlation and cross-correlation of the variables.

3.2.1 Wind

Basic monthly statistics for wind speed at the site, based on the 6-year period, are shown in Table 3.1, and the corresponding monthly wind roses are given in Figure 3.4. As seen from the table, the wind is strongest in June-August (5.87-5.97 m s^{-1}), which is in the middle of the southwest monsoon. Missing values exist in the data but their amount is less than 30% for every month (except for 32.1% in December), which is considered sufficient for use. The wind roses also show the expected wind direction patterns according to the northeast monsoon (mid-October-January) and the southwest monsoon (May-September). Further, the relatively large portion of onshore directions (southwesterly-southeasterly) in January-April is seen, which is partially attributed to the presence of sea breeze and gulf winds. Transitional periods between the seasons are also indicated by the wind roses whereby for January, the shift from the northeast monsoon to the summer period is shown by the petals in the wind rose from mainly the northeast, south, and southwest. Another transition is between September and October, from the southwest monsoon season to the winter season.

As for the wind speed distribution, Figure 3.5 shows the annual mean probability density function plot for the period under consideration. It is seen that for most cases, the distributions show a positively skewed and long tailed pattern. The monthly probability density functions are shown in Figure 3.6, with some months having similar distributions depending on the season. For February-April, in each case there are two distributions superimposed on each other, which could be attributed to sea and land breezes, and gulf winds. For these months, there is relatively higher global solar radiation (Figure 3.7 (b)), and thus there is a clear distinction between the sea-breeze and the land-breeze.

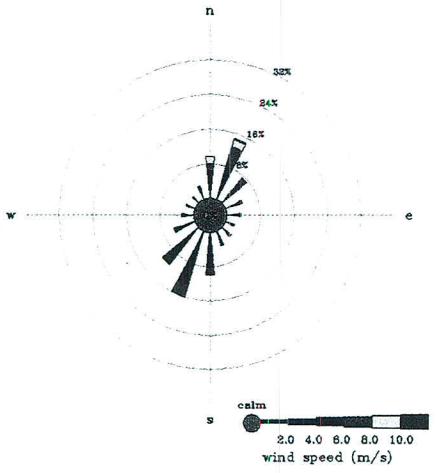
According to the distributions, sea-breeze has its peak at 6 m s^{-1}, whereas for land-breeze, it is much lower. Land breezes had lower speed because of the higher roughness length over land, whereas sea breezes on average had higher speeds owing to lower roughness length over the sea surface. For May-September, the distributions are slightly symmetrical, showing the steady weather of the southwest monsoon. For October, the peak shifts indicating the arrival of the northeasterly winds over land having lower speeds. The same pattern is seen for the distribution in November. In the month of December, there is more land-breeze because the winds come from the north and north-easterly directions. Figure 3.7 (a) also shows the monthly variation of wind speed and the respective standard deviation over 2001-2006. Figure 3.9 (a) indicates the diurnal variation of wind speed, with relatively higher magnitude experienced from 1200-1800 local time (LT). This corresponds to the sea-breeze start and ending time as identified by Phan and Manomaiphiboon (2012). Considering the distinct patterns and variations in the different months of the year depending on the season, it will be necessary to deal with monthly grouped data before the application of the time-series models.

(Intentionally left blank)

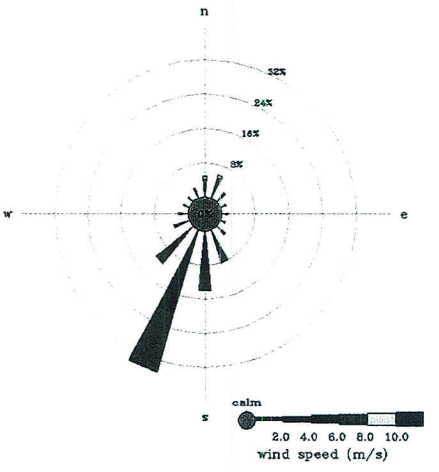
Table 3.1: Monthly statistics for 100-m wind speed (m s^{-1}) observed at PCD-M2 over 2001-2006

Month	Average	Standard Deviation	Minimum	Maximum	Missing (%)
Jan.	4.46	1.87	0.33	11.14	0.6
Feb.	5.10	1.68	0.60	10.53	1.0
Mar.	5.29	1.67	0.74	12.92	24.2
Apr.	4.99	1.60	0.50	10.81	14.5
May	5.51	2.03	0.82	12.12	24.5
Jun.	5.70	1.99	0.95	12.51	22.9
Jul.	5.87	2.00	0.80	12.66	5.5
Aug.	5.97	1.81	0.90	18.62	4.4
Sep.	4.94	1.84	1.02	14.51	21.4
Oct.	4.73	1.93	1.01	14.60	5.7
Nov.	5.25	2.27	0.86	12.42	10.9
Dec.	5.35	2.40	0.58	12.88	32.1

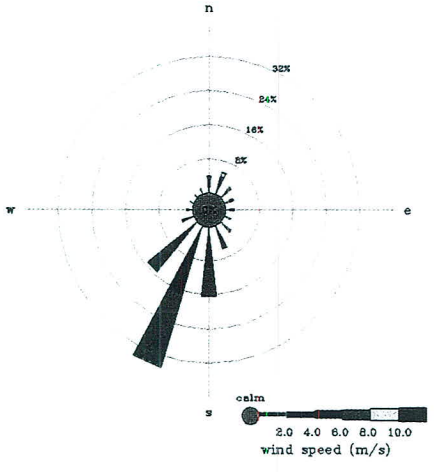
a) Jan.



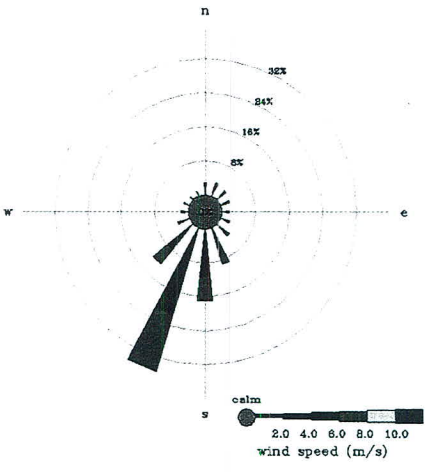
b) Feb.



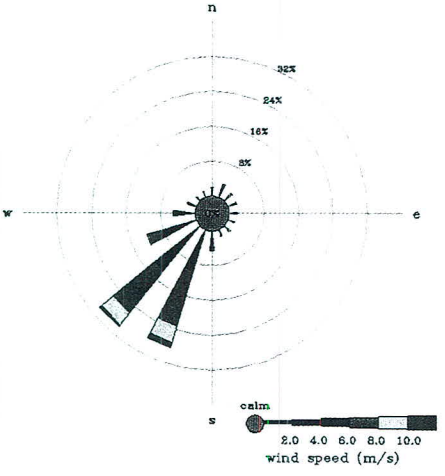
c) Mar.



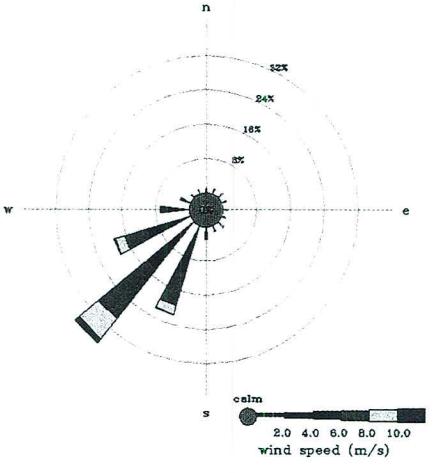
d) Apr.



e) May



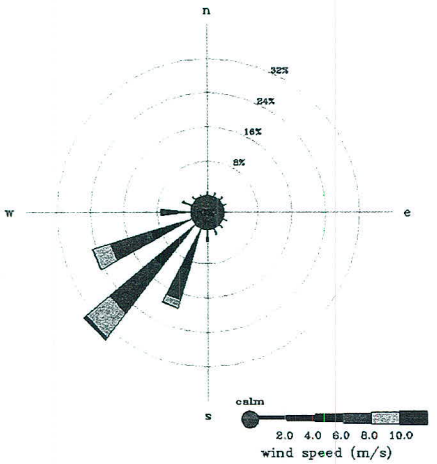
f) Jun.



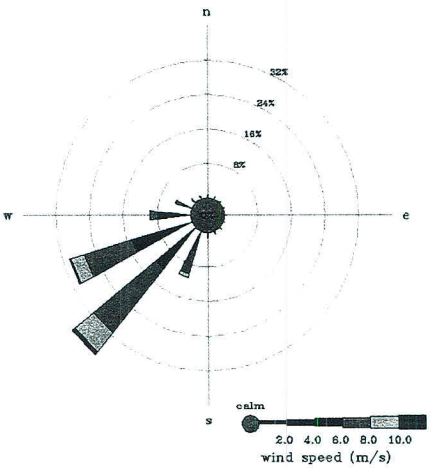
(continued on the next page)

Figure 3.4: Monthly 100-m wind roses at PCD-M2 over 2001-2006

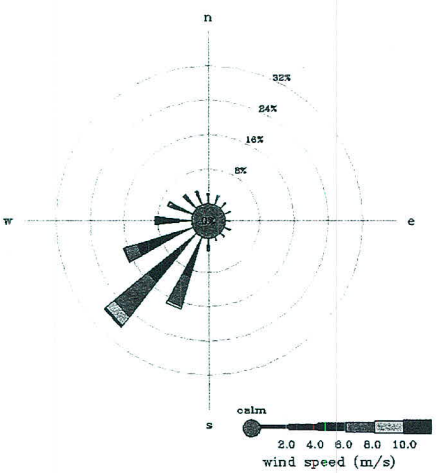
g) Jul.



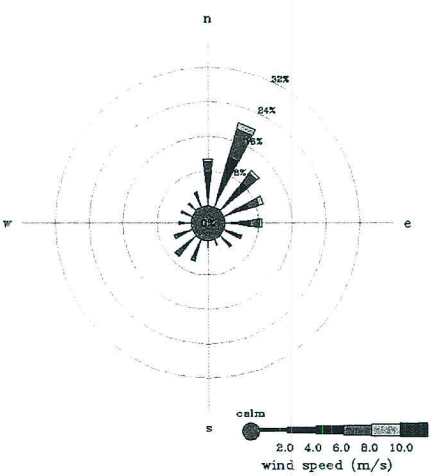
h) Aug.



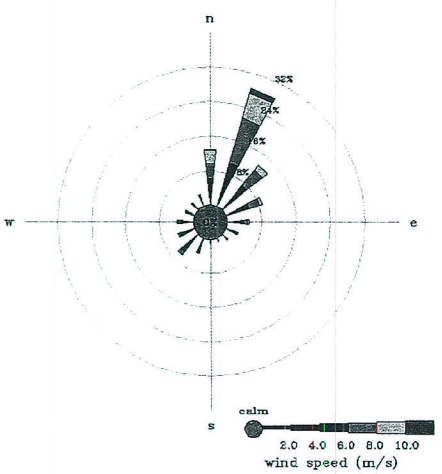
i) Sep.



j) Oct.



k) Nov.



l) Dec.

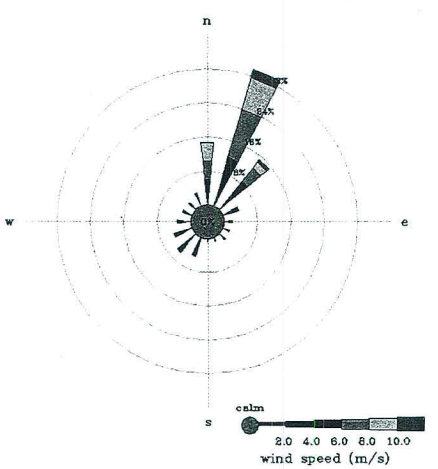


Figure 3.4: (Continued)

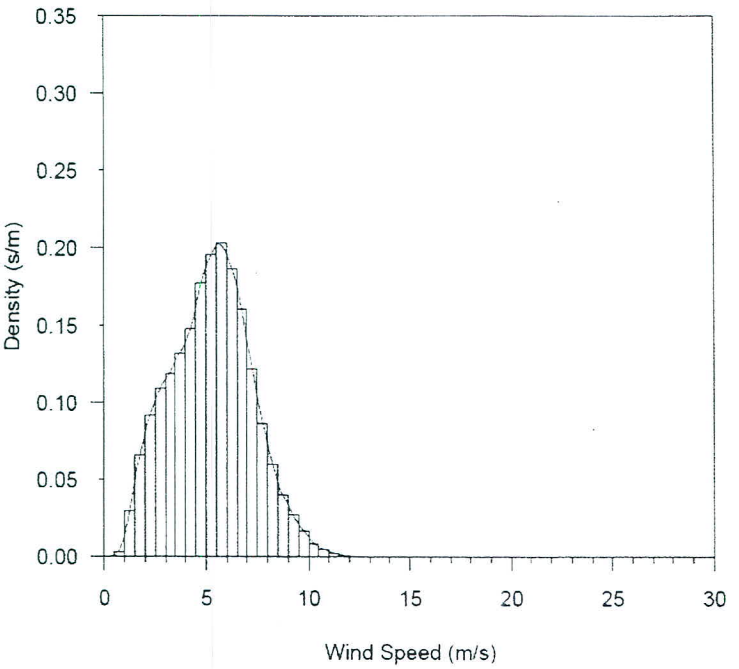
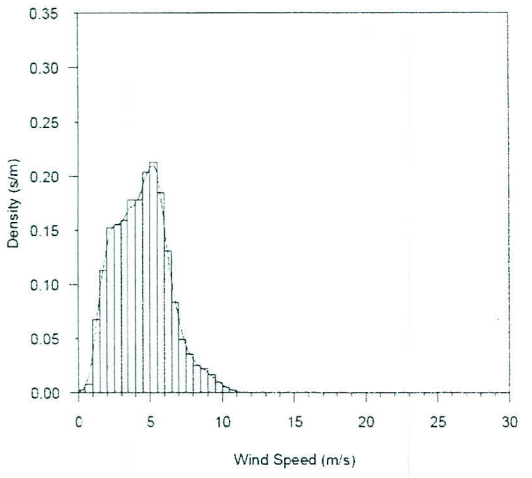
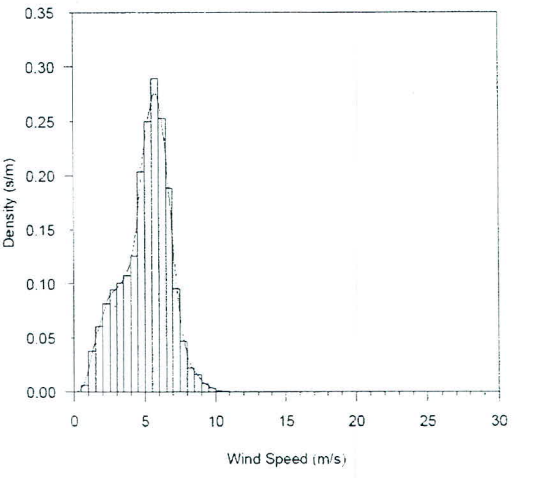


Figure 3.5: Probability density function of 100-m wind speed at PCD-M2 over 2001-2006

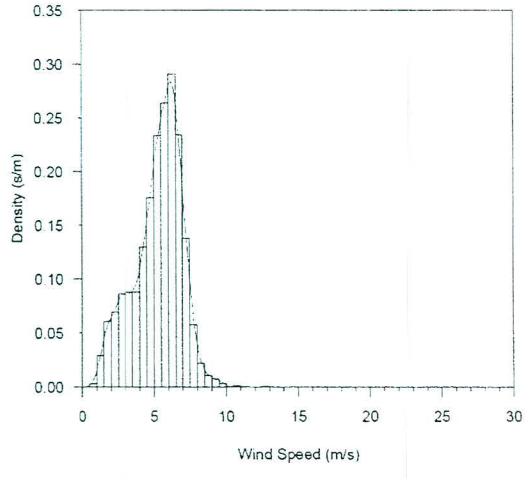
a) Jan.



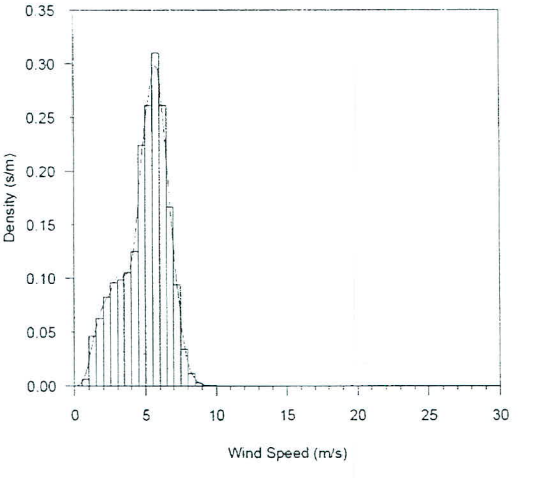
b) Feb.



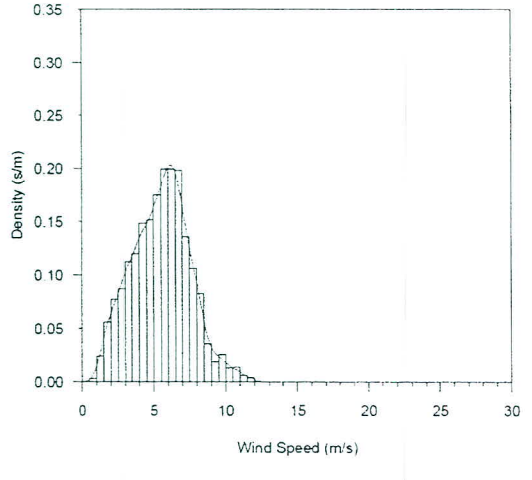
c) Mar.



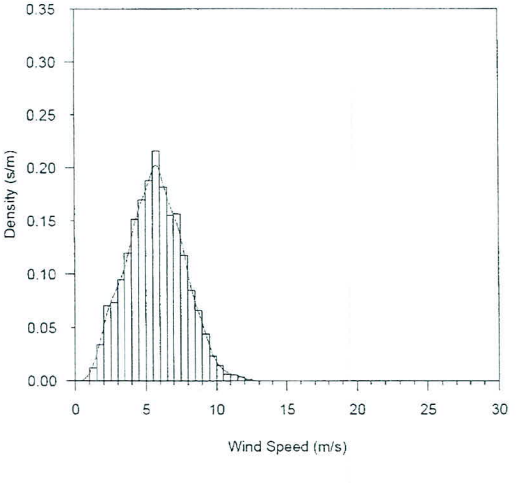
d) Apr.



e) May



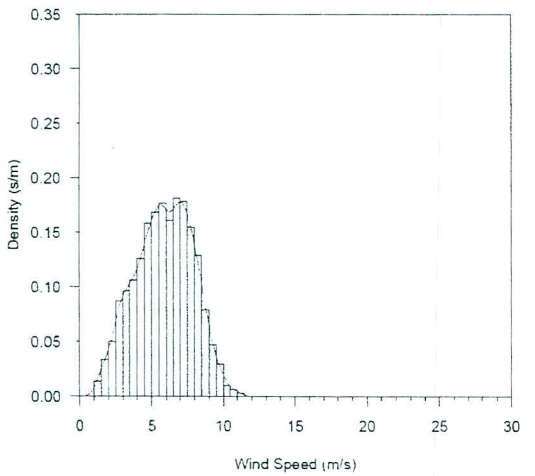
f) Jun.



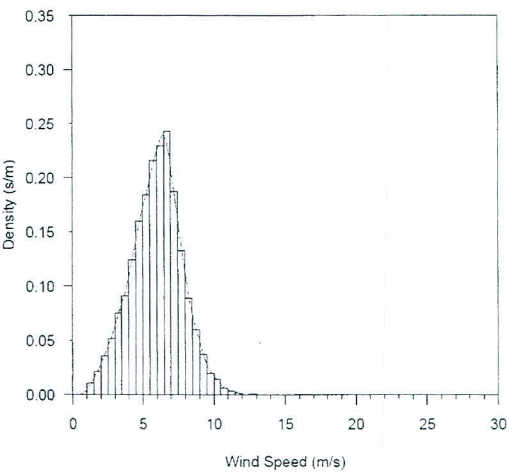
(continued on the next page)

Figure 3.6: Monthly probability density functions of 100-m wind speed at PCD-M2 over 2001-2006

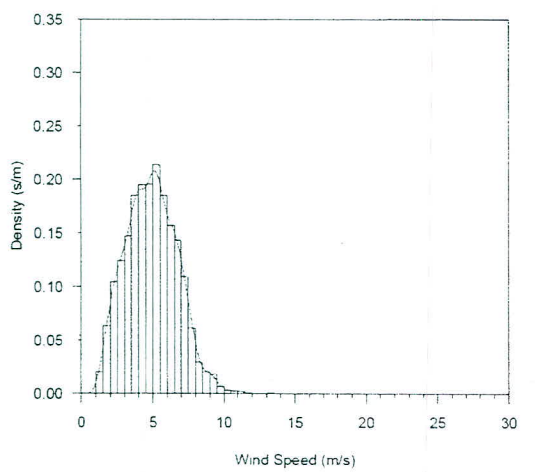
g) Jul.



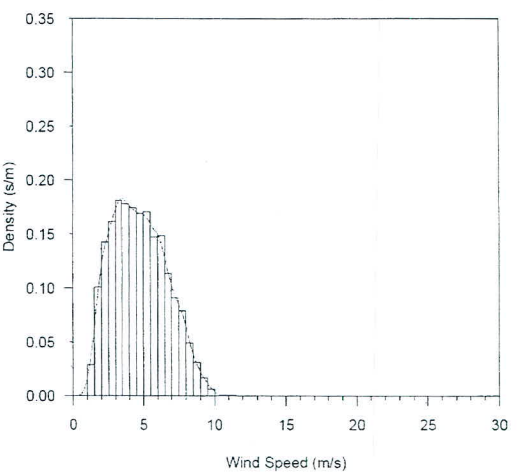
h) Aug.



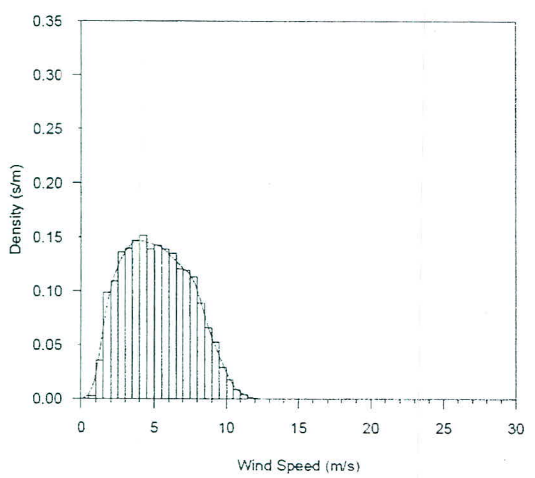
i) Sep.



j) Oct.



k) Nov.



l) Dec.

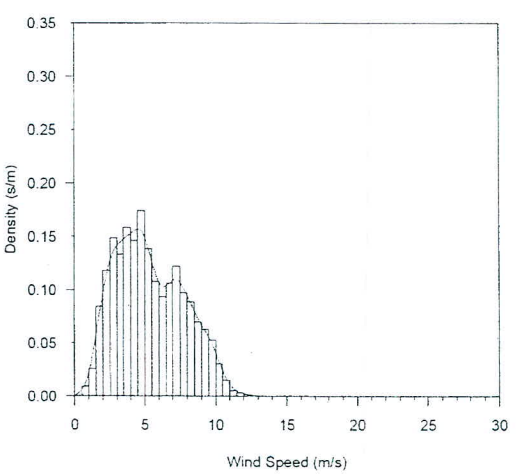


Figure 3.6: (Continued)

3.2.2 Global solar radiation

Considering the monthly daytime (0700-1800 LT) global solar radiation for the area, spanning the period 2001-2006, the results are shown in Figure 3.7 (b) and Table A2. The transition period from January-April shows an increasing trend in the global solar radiation, with the month of April showing the highest daytime average at 511 Wm^{-2}, which is in the summer period and in transition from the vernal equinox. Within the wet season there is a decrease in global solar radiation due to the presence of cloud cover, with values of 403-406 W m^{-2} in August-October. From November, a slight increase could be attributed to the shift to the dry northeast monsoon. Figure 3.9 (b) indicates the diurnal variation of the average global solar radiation, with a peak of 744 Wm^{-2} at 1300 LT. The hourly variation of the global solar radiation is incorporated in the log-law wind profile as a proxy in the Monin-Obhukov length for the kinematic surface heat flux.

3.2.3 Temperature

Descriptive information on the 2-m air temperature, 75-m air temperature, and sea-surface temperature is shown in Tables A3-A5, for which missing values were found to be lower than 30% in all months. A minimum of 10.2°C was recorded in November, 2004, for the near-surface temperature, and a maximum of 37.6°C in April, 2001. Figure 3.8 shows the monthly variation of 2-m air temperature, 75-m air temperature and sea-surface temperature over the Rayong site. The month of April shows the highest average temperature for the 2-m and 75-m air temperature, at 29°C and 28.5°C respectively, coinciding with the summer period. The highest average SST (~30.6°C) occurs in the month of May, which could be accounted for by the higher heat capacity of water, resulting in the retention of surface heat from the previous summer months. As discussed earlier, temperature is one of the proxies in the stability of the boundary layer in the log-law wind profile (Equation 2.1), and therefore an important variable to consider for the multivariate case. Figure 3.10 shows the diurnal variation of vertical and land-sea temperature difference with negative values being seen for night and early morning hours. The land-sea temperature difference is considered as one of the important factors in sea-breeze circulation (Phan and Manomaiphiboon, 2012). Positive values of the land-sea temperature difference are seen from 11 am-5 pm, which could lead to the development of sea-breeze.

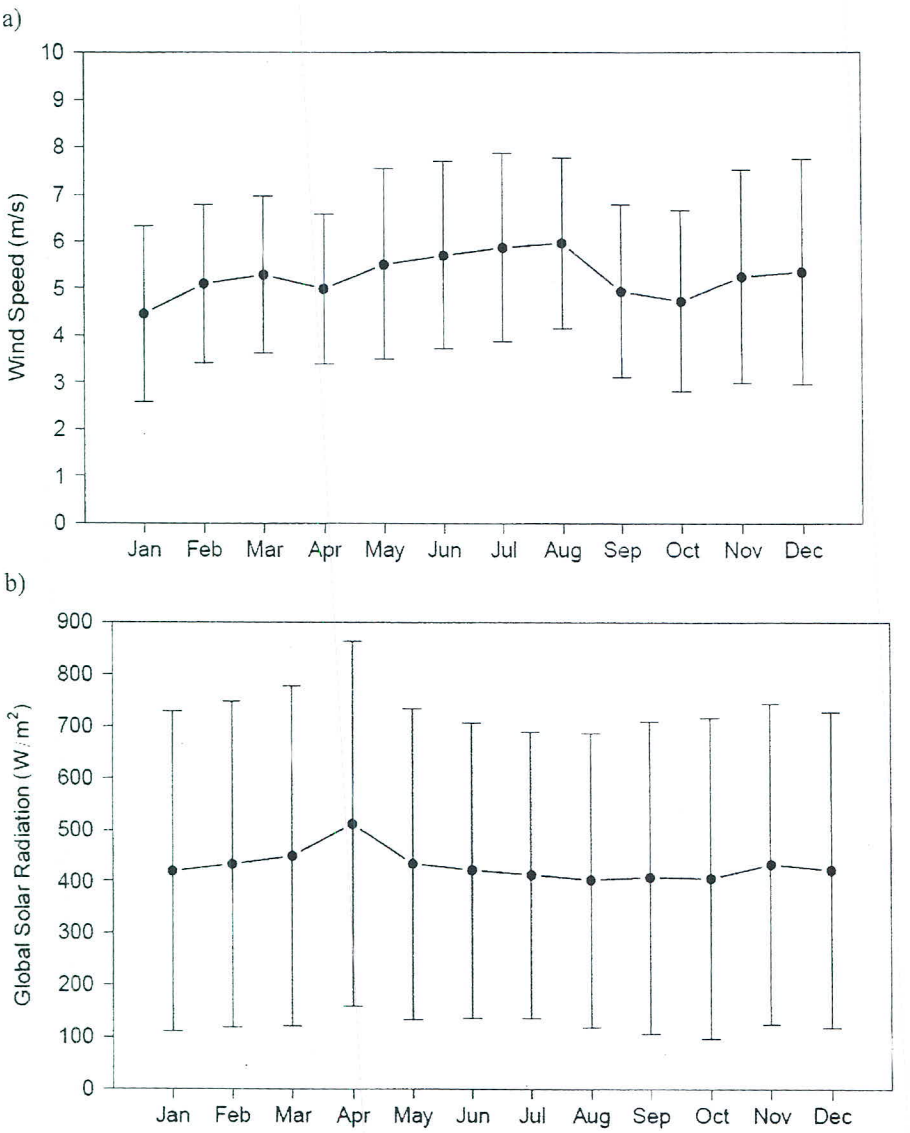


Figure 3.7: Monthly variation: (a) 100-m wind speed, and (b) daytime global solar radiation over 2001-2006. The vertical bars represent the standard deviation for each month

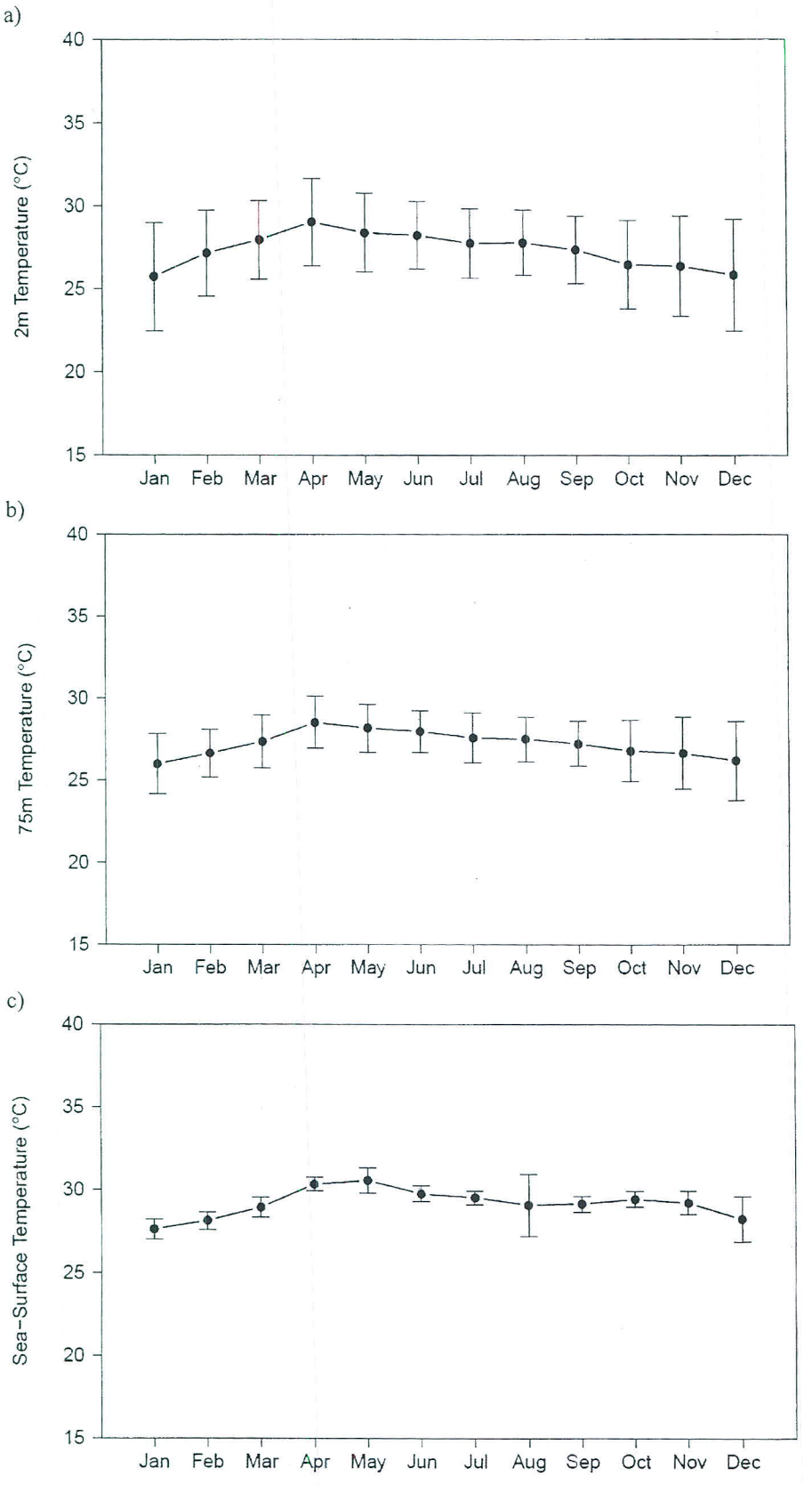


Figure 3.8: Monthly variation: (a) 2-m air temperature, (b) 75-m air temperature, and (c) sea-surface temperature over 2001-2006. The vertical bars represent the standard deviation for each month

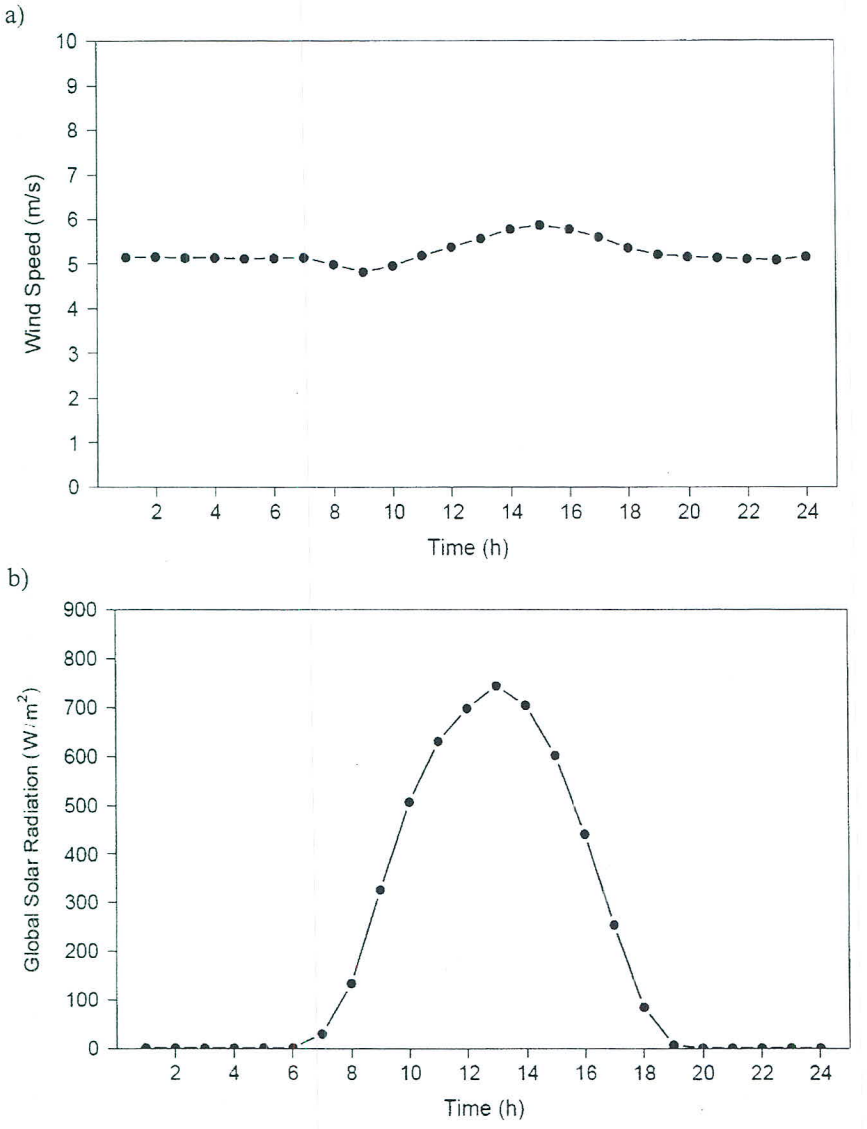


Figure 3.9: Diurnal variation: (a) 100-m wind speed, and (b) global solar radiation over 2001-2006

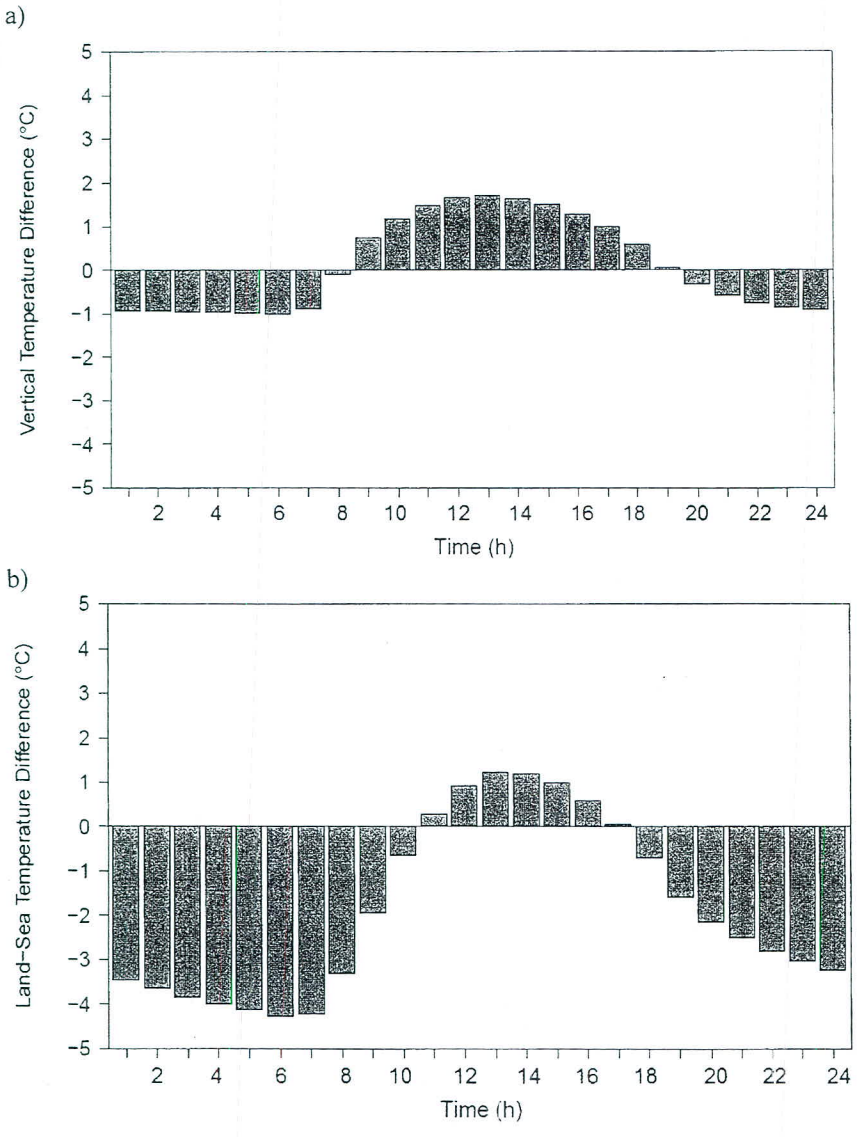


Figure 3.10: Diurnal variation: (a) vertical temperature difference (*VTD*), and (b) land-sea temperature difference (*LSTD*) over 2001-2006

3.3 Time-series modeling

3.3.1 Persistence models

Persistence models are typically used for wind speed forecasting to identify the overall merit of the development of sophisticated models (Monteiro et al., 2009; Wilks, 2006). A persistent model has a simple set-up. For example, the future values (up to the 24-hour forecast horizon) are assumed to be equal to the current observed value (Milligan et al., 2003), or equal to the average of the most recent 24 values. Alternatively, one may use the value at a future hour (of day) to equal that at the same hour (of day) of the most recent day. Here, the first definition was only included for comparison and will be referred to as PS.

3.3.2 ARMA and ARIMA models

As for time-series models, an ARMA(p, q) model represents a combination of two processes, AR (autoregressive) and MA (moving average). AR(p) relates current value x_t as a function of p past values, while MA(q) relates current value as a linear function of current random term and q most recent past random error terms. The parameters (p, q) are the ARMA model's order. As an extension to ARMA, ARIMA(p, d, q) incorporates time-series differencing in the modeling, and d denotes the order of differencing. Its mathematical expression is as follows:

$$\phi(B)(1-B)^d x_t = \theta(B)\omega_t, \tag{3.1}$$

where x_t is the observed value at time t, ω_t is a white noise process with zero mean and constant variance σ^2, B denotes the backshift operator, $B^k x_t = x_{t-k}$. $\phi(B)$ and $\theta(B)$ are the AR and MA operators, which have the following polynomial forms:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p, \text{ and} \tag{3.2}$$

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q, \tag{3.3}$$

where ϕ_i and θ_i are the model coefficients. When analyzing a time series, it is necessary to check the sample autocorrelation function r_k, calculated from the sample autocovariance function, c_k given by:

$$c_k = \frac{1}{n}\sum_{t=1}^{n-k}\left(x_t - \bar{x}\right)\left(x_{t+k} - \bar{x}\right), \text{ and} \tag{3.4}$$

$$r_k = \frac{c_k}{c_0}, \tag{3.5}$$

where n is the sample size, k denotes the time interval (lag, hereafter this term indicates hourly time intervals) and c_0 is the lag 0 autocovariance. The autocorrelation shows the linear dependence of the variable x_t with itself at k time intervals apart. The partial autocorrelation (PACF) at lag k is the value of the k^{th} coefficient of an AR(k) model. (Cowpertwait and Metcalfe, 2009).

It should be noted that the sample autocorrelation function (ACF) of the detrended data in each month exhibit strong diurnal variations (i.e. cycles over a lag of 24 hours) suggesting that removing such diurnality before the modeling may be useful. To do so, a detrended time-series of each month was subtracted with the average values of data corresponding to each individual hour (Figure 3.11). We shortly refer to this practice as dARIMA (where the prefix "d" denotes "using de-diurnalized data"). For SARIMA(p, d, q)×(P, D, Q)$_S$, it is in fact an extension of ARIMA with incorporation of seasonality. S is the seasonal period, (p, d, q) are the orders of the non-seasonal component of the model while (P, D, Q) are those of the seasonal component. In this study, S is 24. The mathematical expression of SARIMA is as follows:

$$\Phi_P\left(B^S\right)\phi(B)\left(1-B^S\right)^D(1-B)^d x_t = \Theta_Q\left(B^S\right)\theta(B)\omega_t, \tag{3.6}$$

where $\Phi_P(.)$ and $\Theta_Q(.)$ are the seasonal AR and MA operators of orders P and Q, respectively, and have the same polynomial form as those of $\phi(.)$ and $\theta(.)$, respectively (Chatfield, 1996).

For ARMA models, a time-series could be identified using the Box-Jenkins methodology (Figure 3.12). This involves visually examining the sample ACF and sample PACF plots of the series. However, doing so may be suitable when a time-series can be represented by a simple AR or MA process alone. For complex processes, a practical approach is to seek for a set of candidate models, vary the values of their model orders, fit the models with the data, and then use a criterion to select the most appropriate (i.e. optimal) model. In general practice, the Akaike information criterion (*AIC*) (Hyndman and Khandakar, 2008; R Development Core Team, 2011) is often recommended for use:

$$AIC = -2\ln(\text{maximum likelihood}) + 2m, \tag{3.7}$$

where the natural logarithm of the maximum likelihood function for the estimated model is calculated and m is the number of model coefficients or parameters. In selection, a model with a lower *AIC* value is preferred. After the fitting, it is always recommended to perform residual diagnostics to assess model adequacy since a model is typically formulated under a set of assumptions. Here, for a fitted model, the plot of residuals against time is to check zero mean and constant variance of the residuals, and the Q-Q plot of residuals is to check the assumption of normally distributed patterns. The Ljung-Box-Pierce statistic (Shumway and Stoffer, 2000) is to check for any statistical significance of grouped residuals autocorrelation up to lag K:

$$Q = n(n+2)\sum_{k=1}^{K}\frac{\hat{r}^2(k)}{n-k}, \tag{3.8}$$

where n is the sample size, $\hat{r}$ is the residual autocorrelation coefficient, and K the specified maximum lag value. Any model will be considered a good fit if the Q value falls below a specified critical value in a $\chi 2$ distribution with $K-p-q$ degrees of freedom (Wilks, 2006).

3.3.3 VAR models

As discussed earlier, the AR(p) model relates the current value of a variable to its most recent values. VAR (vector autoregressive), models are an extension of autoregressive models for multivariate data and can be expressed as:

$$Z_t = \sum_{i=1}^{p} \Phi_i Z_{t-i} + \omega_t, \tag{3.9}$$

where Z_t is a $(M \times 1)$ vector of M number of variables, $Z_{1t}, \ldots, Z_{Mt}$, Φ_i is a matrix holding the autoregressive coefficients with dimensions $(M \times M)$ and $\omega = (u_{1t}, \ldots, u_{Mt})$ is a white noise process of dimension $(M \times 1)$(Chatfield, 1996; Solari and van Gelder, 2011). As discussed earlier for univariate time series, the autocorrelations show the linear time dependence in a time series. As for the multivariate time series Z_t, there are cross-correlations between all possible pairs of variables in addition to the individual autocorrelations which indicate the level of inter-dependencies between the variables. The sample cross-covariance function and cross-correlation function (CCF) between variables x and y, $c_k(x, y)$ and $r_k(x, y)$ are given by:

$$c_k(x,y) = \frac{1}{n}\sum_{t=1}^{n-k}\left(x_t - \bar{x}\right)\left(y_{t+k} - \bar{y}\right), \text{ and} \tag{3.10}$$

$$r_k(x,y) = \frac{c_k(x,y)}{\sqrt{c_0(x,x)\,c_0(y,y)}}, \tag{3.11}$$

where c_0 is the lag 0 autocovariance of the variable (Cowpertwait and Metcalfe, 2009). The CCF will show the lead-lag relationship between the variables x and y. Considering the two time series (x_t and y_t), the series x_t may be related to past lags of the y-series. The CCF is helpful for identifying lags of the y-variable that might be useful predictors of x_t. The sample CCF is a set of sample correlations between y_{t+k} and x_t for $k = 0, \pm1, \pm2, \pm3, \ldots$, where a negative value for k is a correlation between the y-variable at a time before t and the x-variable at time t. When we have significant correlations with k negative, then it is said that y leads x. For this particular study, we will examine the instances where the y-variable(s) is the leading variable of the x-variable because the aim is to use values of the y-variable (here, e.g. global solar radiation, vertical temperature difference, and land-sea temperature contrast) to predict future values of x (here, wind speed).

In order to select an appropriate model corresponding to each monthly data, the method here was done following steps outlined for the univariate ARIMA models. The autocorrelations and cross-correlations of the variables are taken into consideration in the VAR model. De-diurnalized and detrended data for the variables considered are used for the model. The parameters of the VAR model are estimated through ordinary least squares regression and the number of lags equivalent to the VAR order p selected by the lag giving the minimum value of selection criteria *AIC* (Pfaff, 2008). From here, the model goodness-of-fit is tested through residual diagnostics as outlined in the previous section, and forecasts are then carried out.

(Intentionally left blank)

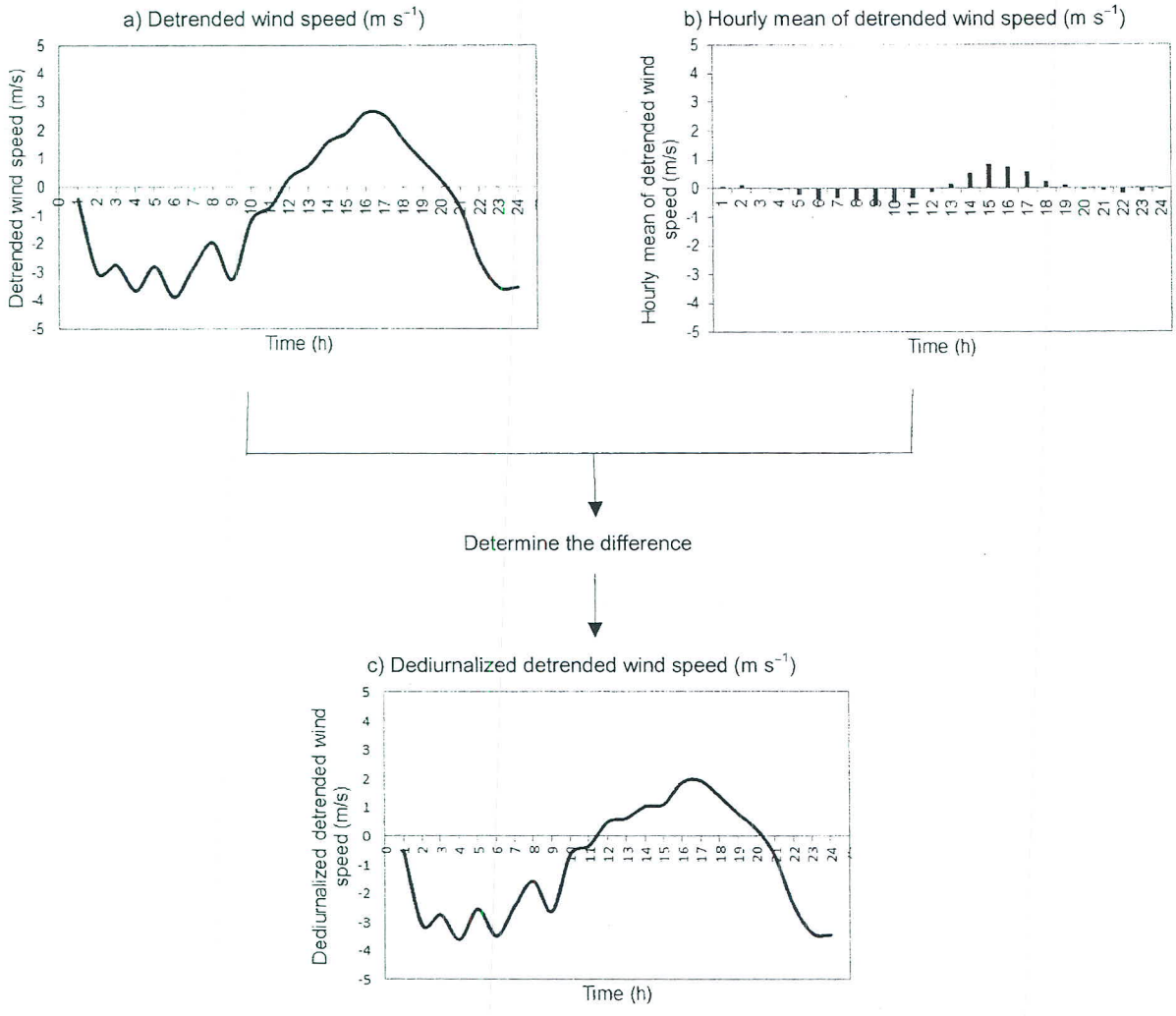


Figure 3.11: Dediurnalization for detrended hourly wind speed for March 23, 2001

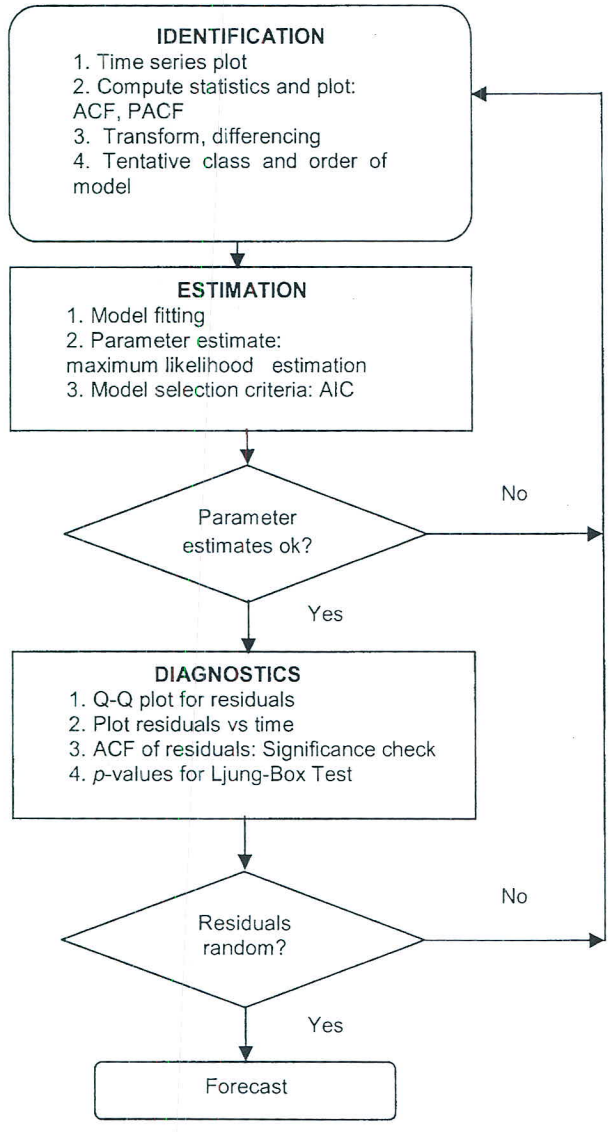


Figure 3.12: Box-Jenkins methodology for ARMA models

3.3.4 Forecast strategy

To perform forecast operation in the current study, a forecast episode is set to start at 0800 (local) hour and end at 0700 hour of the next day. As the operation continues, from the first episode (here, January 1, 2005) and the final episode (here, December 30, 2006), the historical data increases with time because the data used in past episodes are reused and added into the historical data pool for fitting a selected model (i.e. dynamic training). The point forecasts are done by expanding the equations given in the previous section so that x_t is on the left hand side and all the other terms are on the right. The equation is then rewritten by replacing t by $T+h$. Then, on the right hand-side, the future observations are replaced by their respective forecasts, future random errors by zero, and past errors by the corresponding residuals. Starting at h = 1 (0800 hour), these steps are then done for h = 2, 3...., 24 (0700 hour of the next day).

Taking Equations (3.1), (3.2) and (3.3) for the ARIMA(p, d, q), for d = 1 to indicate one level of differencing, expanding and rearranging the terms;

$$x_t = (1+\phi_1)\, x_{t-1} - (\phi_1 - \phi_2)\, x_{t-2} - \ldots - \phi_p x_{t-1-p} + \omega_t + \theta_1 \omega_{t-1} + \theta_2 \omega_{t-2} + \ldots + \theta_q \omega_{t-q}, \qquad (3.12)$$

replacing t by $T+1$ in the above equation;

$$x_{T+1} = (1+\phi_1)\, x_T - (\phi_1 - \phi_2)\, x_{T-1} - \ldots - \phi_p x_{T-p} + \omega_{T+1} + \theta_1 \omega_T + \theta_2 \omega_{T-1} + \ldots + \theta_q \omega_{T+1-q}. \qquad (3.13)$$

Given that we have observed values upto time T, all values on the right are known except for ω_{T+1} ... ω_{T+1-q} which are replaced by zero, and ω_{T-1} and ω_T for which the respective values of last observed residuals $\hat{\omega}_T$ and $\hat{\omega}_{T-1}$ are substituted (Cryer and Chan, 2008). For the one hour ahead forecast, we therefore have

$$x_{T+1|T} = (1+\phi_1)\, x_T - (\phi_1 - \phi_2)\, x_{T-1} - \ldots - \phi_p x_{T-p} + \theta_1\, \hat{\omega}_T + \theta_2\, \hat{\omega}_{T-1}. \tag{3.14}$$

The process continues in this manner for all the time-steps in the 24-hour forecast horizon. If any treatments on the original series before time-series modeling (e.g. detrending and de-diurnalization) are applied, predictions given by a model are converted back to the original form.

3.3.5 Forecast evaluation

To make relative judgment of how well a time-series model performs in terms of wind speed prediction, three statistical metrics were used: mean error (ME), Pearson's correlation coefficient (shortly, correlation or ρ) and skill score (SS). Their mathematical definitions are given as follows:

$$ME = \frac{1}{N}\sum_{i=1}^{N} |P_i - O_i|, \tag{3.15}$$

$$\rho = \frac{\frac{1}{N}\sum_{i=1}^{N} (P_i - \overline{P})\,(O_i - \overline{O})}{\sigma_P \sigma_O}, \text{ and} \tag{3.16}$$

$$SS = 1 - \frac{ME_{\text{Model}}}{ME_{\text{Reference}}}, \tag{3.17}$$

where P_i and O_i are the prediction and observation at hour i, respectively, and N is the total number of sample pairs from all forecast episodes combined. σ_P and σ_O are the standard deviations of predictions and observations, respectively, and $\overline{P}$ and $\overline{O}$ are the mean values of the predictions and observations. Mean error (ME) shows the average magnitude of the errors over an entire dataset and has positive values. ME is chosen as an accuracy measure since it gives the absolute magnitude of the prediction error, as opposed to the root mean square error which tends to be more sensitive to large errors. The skill

score is a relative measure for forecasts from a sophisticated model, compared to those from a reference model. The control forecasts are given by the persistence model as earlier outlined. SS is useful in equalizing the influence of inherently more or less challenging forecasting situations for different models. Positive values indicate better performance of the model, while negative values show that the model has no merit over the persistence model (Wilks, 2006; Landberg, 1999).

Taylor diagrams will be used to give a visual evaluation of different models in one plot by comparing to the observed data through the use of correlation coefficient, centered root-mean-square error (CRMSE) and standard deviation (Taylor, 2001; Wilks, 2006). In order to understand the geometric relationship among the statistics used in the construction of the Taylor diagram, the equation for the CRMSE is used:

$$CRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\left(P_i - \overline{P}\right) - \left(O_i - \overline{O}\right)\right)^2} \,. \tag{3.18}$$

Further expanding (3.18), together with the law of cosines, gives

$$CRMSE^2 = \sigma_P^2 + \sigma_O^2 - 2\sigma_P\sigma_O\rho \,. \tag{3.19}$$

Figure 3.13 and 3.14 shows this relationship between the statistical metrics and the predictions from a model and observations as represented in the Taylor diagram.

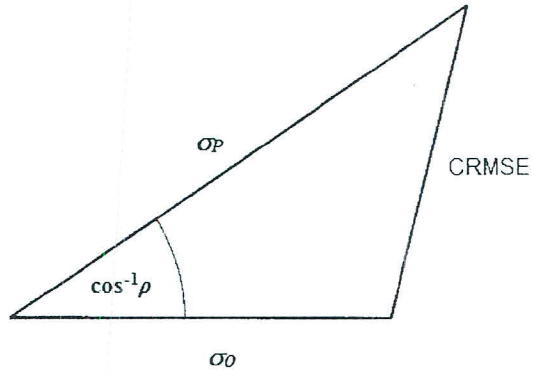


Figure 3.13: Relationship between correlation coefficient (ρ), CRMSE, and standard deviations σ_P and σ_O of predictions and observations

(Source: Taylor, 2001)

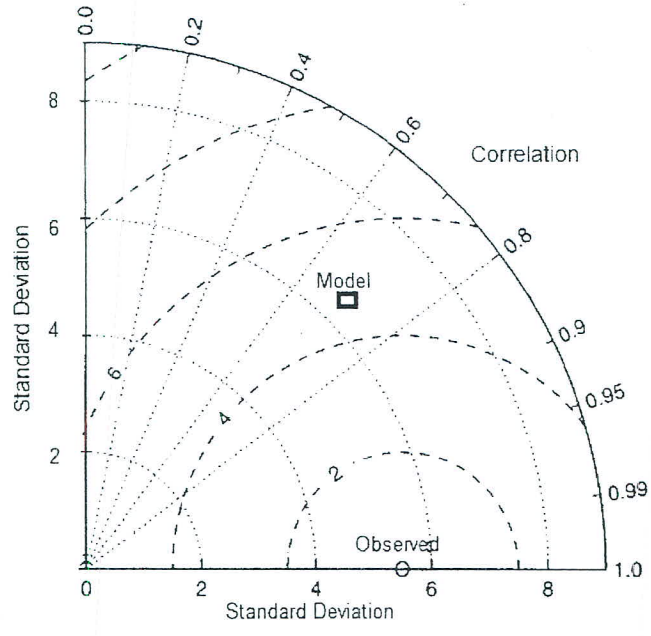


Figure 3.14: Taylor diagram for comparison of model performance

(Adapted from Taylor, 2001)