A hybrid ARIMA and artificial neural networks model to forecast particulate matter in urban areas: The case of Temuco, Chile

Luis A. Díaz-Robles a,*, Juan C. Ortega a, Joshua S. Fu b, Gregory D. Reed b, Judith C. Chow c, John G. Watson c, Juan A. Moncada-Herrera a

a School of Environmental Engineering, Catholic University of Temuco, Manuel Montt # 056, Casilla 15-D, Temuco, Araucania, Chile
b Civil and Environmental Engineering Department, University of Tennessee, Knoxville, TN, USA
c Division of Atmospheric Sciences, Desert Research Institute, Reno, NV, USA

Article info
Article history:
Received 8 January 2008
Received in revised form 16 July 2008
Accepted 18 July 2008

Keywords:
Particulate matter forecasting
Hybrid
ARIMA
Neural networks
Temuco

Abstract
Air quality time series consists of complex linear and non-linear patterns and are difficult to forecast. Box–Jenkins Time Series (ARIMA) and multilinear regression (MLR) models have been applied to air quality forecasting in urban areas, but they have limited accuracy owing to their inability to predict extreme events. Artificial neural networks (ANN) can recognize non-linear patterns that include extremes. A novel hybrid model combining ARIMA and ANN to improve forecast accuracy for an area with limited air quality and meteorological data was applied to Temuco, Chile, where residential wood burning is a major pollution source during cold winters, using surface meteorological and PM 10 measurements. Experimental results indicated that the hybrid model can be an effective tool to improve the PM 10 forecasting accuracy obtained by either of the models used separately, and compared with a deterministic MLR. The hybrid model was able to capture 100% and 80% of alert and pre-emergency episodes, respectively. This approach demonstrates the potential to be applied to air quality forecasting in other cities and countries.

1. Introduction
High levels or episodes of ambient particulate matter (PM) concentrations are major concerns for health effects and visibility impairment (Chow et al., 2002; Watson, 2002). Increased mortality and morbidity in communities with elevated PM concentrations have been reported by a variety of epidemiological studies (Sanhueza et al., 2005; Chow et al., 2006; Pope and Dockery, 2006). According to the Chilean regulations, the air quality index (índice de calidad del aire por PM, ICAP, in Spanish) is defined to assign episode types of atmospheric pollution for PM 10 (PM with aerodynamic diameter ≤ 10 micrometers). The ICAP is a simple standard scale of 24-h PM 10 average concentrations, see Table 1.

In Temuco, Chile, a few epidemiological studies have been conducted to establish the link between air quality and health. One recent study found a strong relationship between PM 10 and daily mortality cases (1997–2002) among subjects over 65 years old (Sanhueza et al., 2005). These authors found that the relative risk (RR) of a 100-μg m⁻³ increment in PM 10 was 1.236 (confidence interval, (CI), 95%, 1.004–1.522) for respiratory mortality, and 1.176 (CI 95%, 1.006–1.374) for cardiovascular mortality. These values were 15–17% higher than those reported for Santiago, Chile (Sanhueza et al., 1999). The authors used the same methodology for this city, where RRs were 1.061 (CI 95%, 1.017–1.106) for respiratory mortality and 1.025 (CI 95%, 1.005–1.046) for cardiovascular mortality. The authors suggested that the chemical composition and particle size differences for each city could explain the
disparity among their RRs for mortality (Sanhueza et al., 2005). Some studies have found higher concentrations of Polycyclic Aromatic Hydrocarbons (PAHs) in Temuco than Santiago, with concentrations that were 197 and 6.6 times higher, respectively, than the maximum limits allowed in the European Union (1 ng m$^{-3}$) (Tsapakis et al., 2002). Other studies have found that the PM$_{2.5}$ constitutes the 80–90% of the total PM$_{10}$ in Temuco, compared with 30–60% in Santiago (Sanhueza et al., 2005).

These studies document the need for better air quality management to reduce air pollution levels. An accurate air quality-forecasting model is needed to alert the population at large and to initiate preventative pollution control actions. This paper introduces a hybrid air quality-forecasting model for Chile that could be applied to other cases with similar terrain, emission sources, and databases.

### 1.1. Air quality forecasting models

Autoregressive Integrated Moving Average (ARIMA) (Box–Jenkins Time Series) (Box and Jenkins, 1970) and the multilinear regression (MLR) models have been widely used for air quality forecasting in urban areas, but they are of variable accuracy owing to their linear representation of non-linear systems (Goyal et al., 2006). Artificial neural networks (ANN) have been developed as a non-linear tool for pollution forecasting, principally using multilayer perceptron (MLP) architecture (Pérez and Reyes, 2002; Pérez et al., 2004; Pérez and Reyes, 2006; Schlink et al., 2006; Slini et al., 2006; Sofuoglu et al., 2006; Sousa et al., 2006; Thomas and Jacko, 2007). The comparison among these models is presented in Table 2.

These models have been compared to evaluate their robustness in air quality forecasting performance. Forecasting daily maximum ozone (O$_3$) concentration at Houston, a study showed that an ANN model was more accurate than either the ARIMA or MLR models, mainly because the data presented clearer non-linear patterns than linear ones (Prybutok et al., 2000). Goyal et al. (2006) pointed out those linear models such as MLR and ARIMA fail to predict extreme concentrations (episodes). These authors found that the combined ARIMA and MLR models predicted PM$_{10}$ levels more accurately than either model used independently for Delhi and Hong Kong. The ARIMA and ANN models have also been used for sales forecasting in Brazil, with the non-linear ANN model showing higher accuracy (Ansuj et al., 1996). Recent studies provide good descriptions of the hybrid ARIMA–ANN models (Aburto and Weber, 2007; Aslanargun et al., 2007; Gutiérrez-Estrada et al., 2007; Pulido-Calvo and Portela, 2007; Sallehuddin et al., 2007; Gutiérrez-Estrada et al., 2008). For economic time series forecasting, a study combined a seasonal ARIMA model with a back propagation ANN model (Tseng et al., 2002), showing that the hybrid performed better than ARIMA or ANN alone. Zhang (2003) tested a hybrid ARIMA and ANN model over three kinds of time series, and concluded that the linear and non-linear time series patterns in the combined model improved forecasting more than either of the models used independently. This hybrid ARIMA–ANN approach has recently been used for tourist arrival forecasting (Aslanargun et al., 2007), hydrology (Jain and Kumar, 2007), supply chain management (Aburto and Weber, 2007), freshwater phytoplankton dynamics (Jeong et al., 2008), watersheds (Pulido-Calvo and Portela, 2007), fish catch (Gutiérrez-Estrada et al., 2007) production values of the machinery industry (Chen and Wang, 2007), fish community diversity (Gutiérrez-Estrada et al., 2008), and in other areas (Tseng et al., 2002; Zhang, 2003; Zhang and Qi, 2005), but all of them are not related to air quality.

So far, only two hybrid models have been applied to air quality forecasting (Chelani and Devotta, 2006; Wang and Lu, 2006). Wang and Lu (2006) used MLP trained with a particle swarm optimization algorithm (MLP–PSO) and a hybrid Monte Carlo (HMC) method. This was applied for ground level O$_3$ forecasting in Hong Kong during 2000–2002 over 2 typical monitoring sites, of a total of 14, with different O$_3$ formation patterns in order to fully examine the feasibility and generality of the proposed predictive models. One was the Tsuen Wan (TW) site, which is located in the urban area and is surrounded by mountains, in which O$_3$ dynamic is mainly influenced by the high level of primary pollutants emitted from local traffic. The other was the Tung Chung (TC) site (Tung Chung Health Centre, located at the north of Lantau Island about 3 km southeast), which is a suburban residential area, where the annual average O$_3$ level is usually higher than that in TW site. The authors suggested that the O$_3$ pollution at the TC site is partially subjected to a regional influence of the Pearl River Delta pollution shifting. Their results indicated that the hybrid model produced good predictions of the maximum O$_3$ level at both sites. However, the model did not perform satisfactorily during an episode at the TC site, which is influenced by both local and regional emissions; in other words, long-range transport of O$_3$ precursors and two power plants emission of Hong Kong have more significant influence on the TC than the TW site. Chelani and Devotta (2006) included ARIMA with a non-linear dynamic technique to forecast nitrogen dioxide (NO$_2$) at a site in Delhi,
emissions originate from residential wood combustion. A statistical outlier analysis was performed on these data. The maximum 24-h PM$_{10}$ moving average concentrations in PM$_{10}$ moving averages per day (MaxPM$_{10}$ma) values were determined for each of the variables, where the main transformation functions were natural logarithms. Other variables related to daily PM$_{10}$ were also created; the maximum hourly PM$_{10}$ of the previous day (L1PM$_{10}$, $\mu$g m$^{-3}$), maximum 6-h PM$_{10}$ moving average concentration of the previous day (L6PM$_{10}$, $\mu$g m$^{-3}$), maximum 12-h PM$_{10}$ moving average concentration of the previous day (L12PM$_{10}$, $\mu$g m$^{-3}$), and maximum 24-h PM$_{10}$ moving average concentration of the previous day (L24PM$_{10}$, $\mu$g m$^{-3}$). Next, the data was standardized to obtain the identical scale along each axis in g-dimensional input space and to get constant variance for each variable. This analysis was performed for all models, as suggested for some studies (de Menezes and Nikolaev, 2006; Piotrowski et al., 2006; Gutiérrez-Estrada et al., 2008).

2.2. Variable selection and models construction

As shown in Table 4, the data set was partitioned into two sets; 92% for training and 8% for validation data, based on the ANN model requirements for training.

2.2.1. Linear approach: multiple linear regression (MLR) model

The multiple regression procedure was utilized over the training data set to estimate the significant regression coefficients $b_0$, $b_1$, ..., $b_q$ of the linear equation:

$$ y = b_0 + b_1 x_1 + \cdots + b_q x_q $$

(1)

where the regression coefficients $b_0$, $b_1$, ..., $b_q$ represent the independent contributions of each independent variable $x_1$, ..., $x_q$ to the prediction of the dependent variable $y$. To examine the independency of the variables, a multi-collinearity analysis was performed on the data set using the variation inflation factor (VIF) (DeLurgio, 1998). The global statistical significance of the relationship between $y$ and the independent variables was analyzed by means of analysis of variance (ANOVA, $\alpha$ level = 0.05) to ensure the validity of the model. A stepwise procedure of the JMP 6.0.2 (SAS Institute Inc., U.S.) tool was used for the MLR calibration. The final MLR model included the following normalized, independent, and standardized predictor variables that were statistically significant: previous day maximum hourly PM$_{10}$ (L1PM$_{10}$), WS, Tmin, and Tmax. These variables were also used as predictor variables in to the ARIMA($p,d,q$)X, ANN, and hybrid models.

2.2.2. Linear approach: Box–Jenkins ARIMA model

ARIMA linear models have dominated many areas of time series forecasting. As the application of these models...
is very common, it is described here briefly. The linear function is based upon three parametric linear components: autoregression (AR), integration (I), and moving average (MA) method (Box and Jenkins, 1970; Delurgio, 1998). The autoregressive or ARIMA($p$,0,0) method is represented as follows:

$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t$$  \hspace{1cm} (2)$$

where $p$ is the number of the autoregressive terms, $Y_t$ is the forecasted output, $Y_{t-p}$ is the observation at time $t-p$, and $\phi_1, \phi_2, \ldots, \phi_p$ is a finite set of parameters. The $\phi$ terms are determined by linear regression. The $\theta_0$ term is the

![Fig. 1. Las Encinas monitoring station in Temuco city, Chile.](image1)

![Fig. 2. Maximum 24-h PM$_{10}$ moving average by beta attenuation monitor (BAM) time series at the Las Encinas site in Temuco, Chile (from 07/21/2000 to 09/30/2006).](image2)
variables that composed the MLR model. The ARIMA model represents the independent external variables. In this case, an ARIMA($p,d,q$) or ARMA($p,q$) is a model for a time series that depends only on past values of itself and on past random terms $e_t$. This method has the form of Eq. (4).

\[ Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \mu - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} + e_t \]  

(3)

Finally, an ARIMA($p,d,q$) is a ARIMA($p,0,q$) model for a time series that has been differenced $d$ times.

The ARIMA models also have the capability to include external independent or predictor variables. In this case, the model is a multivariate model called MARIMA or ARIMAX. This model is represented as ARIMA($p,d,q$)$X$, where $X$ represents the independent external variables. In this study, the meteorological variables were the independent variables that composed the MLR model. The ARIMA model was obtained using the Time Series Forecasting System tool of the SAS 9.1 software (SAS Institute Inc., U.S.).

2.2.4. Linear and non-linear approach: hybrid ARIMA–ANN model

The combination of the ARIMA and ANN models was performed to use each model capability to capture different patterns in the air quality data. The methodology consisted of two steps: (1) in the first step, an ARIMAX model was developed to forecast MaxPM$_{10}$ma; and (2) in the second step, an ANN model was developed to describe the residuals from the ARIMAX model. In this study, MLP architectures with the LM training algorithm and different activation functions were used. The hybrid model was built using the Enterprise Miner tool of the SAS 9.1 software.

2.3. Measures of accuracy applied in the models performance

To assess the performance of the models during the training and validation phases several measures of accuracy were applied (Eqs. (6–12)), as there is not a unique and more suitable unbiased estimators employed to see how far the model is able to explain the total variance of the data (DeLurgio, 1998; Gutiérrez-Estrada et al., 2007; Pulido-Calvo and Portela, 2007). The proportion of the total variance in the observed data that can be explained by the model was described by the coefficient of determination ($R^2$). Other applied measures of variance were the coefficient of efficiency ($E^2$) (Nash and Sutcliffe, 1970; Kitanidis and Bras, 1980), the average relative variance (AVR) (Griñó, 1992), and the percent standard error of prediction (SEP) (Ventura et al., 1995). The $E^2$ and AVR were used to see how the models explain the total variance of the data and represent the proportion of variation of the observed data.
considered for air quality forecasting modeling. The SEP allows the comparison of the forecast from different models and different problems because of its dimensionless. For a perfect performance, the values of $R^2$ and $E^2$ should be close to one and these of SEP and ARV close to zero. The estimators to quantify the errors in the same units of the variance were the square root of the mean square error (RMSE), and the mean absolute error (MAE). The optimal model is selected when RMSE and MAE are minimized. The above estimators are given by:

$$E^2 = 1.0 - \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|^2}{\sum_{i=1}^{n}|y_i - \bar{y}|^2} \quad (6)$$

$$\text{ARV} = 1.0 - E^2 \quad (7)$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}} \quad (8)$$

$$\text{MAE} = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (9)$$

$$\text{SEP} = \frac{100}{\bar{y}} \text{RMSE} \quad (10)$$

where $y_i$ is the observed value, $\hat{y}_i$ is the forecasted value to $y_i$, $\bar{y}$ is the mean value of the series $y_i$, and $n$ is the number of the observations of the validation set.

In addition, the persistence index (PI) was used for the models performance evaluation (Kitanidis and Bras, 1980).

$$\text{PI} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y_{i-1})^2} \quad (11)$$

where $y_{i-1}$ is the observed MaxPM10ma at the time $i-1$, since only 1-day ahead forecasts were performed. A PI value of one indicates a perfect adjustment between forecasted and observed values, and a value of zero is equivalent to a model that always gives as prediction the previous observation. A negative value of PI reflects that the model is degrading the original information, denoting a worse performance than the model that always gives as prediction the previous observation (Anctil and Rat, 2005).

Other index used to identify the best model was the Bayesian Information Criterion (BIC) (Schwarz, 1978; Qi and Zhang, 2001).

$$\text{BIC} = n \log(\text{SSE}) + m \log(n) \quad (12)$$

where $m$ is the number of parameters of the model and SSE is the sum of the squared errors. In this equation, the first term measures the goodness-of-fit of the model, while the second term penalizes the number of the model parameters. The number of the ANN model parameters was considered as the number of weights (Chen and Hare, 2006; Gutiérrez-Estrada et al., 2007; Pulido-Calvo and Portela, 2007). The optimal model is selected when the BIC is the lowest.

3. Results and discussions

During the outlier analysis, just one high PM$_{10}$ concentration on March 9th 2001 ($284.2$ μg m$^{-3}$) was found and eliminated from the data. This high concentration is not common in the city in the late summer. This phenomenon was due mainly to an agricultural fire that occurred on the same day close to the city. Next, the stepwise method found a MLR model for the normalized and standardized variables (Eq. 13 and Table 5), where all the parameters had a significant $p$ value at a confidence level of 95%.

$$\text{MaxPM}_{10} \text{ma} = 0.80845(\text{L1PM}_{10}) - 0.13283(\text{WS}) - 0.15332(\text{Tmin}) + 0.05605(\text{Tmax}) \quad (13)$$

Table 5 suggests that the maximum hourly PM$_{10}$ concentration of the previous day, minimum temperature, and wind speed are more significant than maximum temperature on March 9th 2001 (284.2 μg m$^{-3}$) was found and eliminated from the data. This high concentration is not common in the city in the late summer. This phenomenon was due mainly to an agricultural fire that occurred the same day close to the city. Next, the stepwise method found a MLR model for the normalized and standardized variables (Eq. 13 and Table 5), where all the parameters had a significant $p$ value at a confidence level of 95%.

$$\text{MaxPM}_{10} \text{ma} = 0.80845(\text{L1PM}_{10}) - 0.13283(\text{WS}) - 0.15332(\text{Tmin}) + 0.05605(\text{Tmax}) \quad (13)$$

Table 5 suggests that the maximum hourly PM$_{10}$ concentration of the previous day, minimum temperature, and wind speed are more significant than maximum temperature on March 9th 2001 (284.2 μg m$^{-3}$) was found and eliminated from the data. This high concentration is not common in the city in the late summer. This phenomenon was due mainly to an agricultural fire that occurred the same day close to the city. Next, the stepwise method found a MLR model for the normalized and standardized variables (Eq. 13 and Table 5), where all the parameters had a significant $p$ value at a confidence level of 95%.

$$\text{MaxPM}_{10} \text{ma} = 0.80845(\text{L1PM}_{10}) - 0.13283(\text{WS}) - 0.15332(\text{Tmin}) + 0.05605(\text{Tmax}) \quad (13)$$

Table 5 suggests that the maximum hourly PM$_{10}$ concentration of the previous day, minimum temperature, and wind speed are more significant than maximum temperature to predict the maximum 24-hr PM$_{10}$ moving average concentrations at Temuco. This behavior of particulate matter pollution is similar than other similar woodsmoke polluted cities (Schreuder et al., 2006; Cavanagh et al., 2007; Larson et al., 2007; Naeher et al., 2007). The wind direction was not a significant variable, suggesting that the sources of PM were more local rather than transported from other regions.

These external variables were the same as those used in the ARIMAX model (Eq. 14 and Table 6).

$$\text{MaxPM}_{10} \text{ma} = \text{ARIMA}(1,0,1) + 0.60363(\text{L1PM}_{10}) - 0.15999(\text{WS}) - 0.18140(\text{Tmin}) + 0.09197(\text{Tmax}) \quad (14)$$

where ARIMA(1,0,1) is an ARIMA model with autocorrelation of order 1, without integration, and with a moving average.
average of order 1, represented by Eq. 10 and using the parameter estimated from Table 6.

\[ Y_t = 0.98122 Y_{t-1} + \mu - 0.87057 e_{t-1} + e_t \]  

(15)

Table 6 suggests that the autoregressive component of the ARIMA model is more significant than the moving average component, the maximum hourly PM\(_{10}\) concentration of the previous day, and the meteorological variables to predict the maximum 24-h PM\(_{10}\) moving average concentrations at Temuco.

The ANN model was built with the selected variables found through the MLR model and trained with the Levenberg–Marquardt algorithm and MLP architecture with one hidden layer, three neurons, and with a square activation function. This model is a black box for the model equations and coefficients, a disadvantage compared to the MLR or ARIMAX models.

To build the hybrid ARIMA–ANN model, two output variables of the ARIMAX model (Eq. 14) and the selected external meteorological variables from the MLR model were used as inputs to the new ANN model; the forecasted ARIMAX MaxPM\(_{10,ma}\), the errors associated with the ARIMAX model (\(e_{\text{arimax}}\)), L1PM\(_{10}\), WS, Tmin, and Tmax. The hybrid model was trained with the Levenberg–Marquardt algorithm and MLP architecture with one hidden layer, three neurons, and a square activation function. The model performances are reported in Fig. 3 and Table 7.

To evaluate the performance of these different models, specifically, three groups to measure the accuracy were...
considered: predictive capability ($R^2$, $E^2$, and ARV), precision (RMSE, MAE, and SEP), and goodness-of-fit (PI and BIC). To estimate these coefficients, all the modeled and observed variables and their residuals were re-calculated at their original units, in other words, non-normalized and non-standardized. For the predictive capability, the hybrid model beat the other three models at least in 10.4%. The $R^2$ between observed and estimated maximum 24-h PM$_{10}$ moving average concentrations in this validation phase indicated that a 98.28% of the explained variance was captured by the hybrid model, value better than the other models, Table 7. Similar conclusions were obtained in forecasting different kinds of time series (Zhang, 2003; Gutiérrez-Estrada et al., 2007; Pulido-Calvo and Portela, 2007). The biggest difference was detected for the $E^2$ coefficient (close to 22%), as that obtained for other study (Gutiérrez-Estrada et al., 2007). Regarding the accuracy, the hybrid model is consistently far the most accurate among the four models. The SEP coefficient appears as the more “relaxed”, while MAE as the most demanding among the three. In this case, this level of explained variance implied a SEP of 9.20%, a RMSE of 8.8 $\mu$g m$^{-3}$, and a MAE of 6.74 $\mu$g m$^{-3}$. Finally, the hybrid model appears as the model that better fit the forecasting also. In this regard, the observed differences in the BIC coefficient are distinguished (Table 7), because the parameter is one of the most efficient in the study of model comparison. As the determination coefficient has received much criticism on forecasting because it is not related to the difference between predicted and observed values, the PI was used complementarily (Goyal et al., 2006; Gutiérrez-Estrada et al., 2007). The PI value of the hybrid model was as high as 0.9676. This suggests that the hybrid model had better forecasting performance than those other models.

The performance in ARIMAX and ANN models is not robust when the data meet certain behaviors, such as linear and non-linear patterns, that usually are found in air quality time series. In this case, the 24-h PM$_{10}$ moving average concentrations time series presents those two patterns. This problem is solved with the hybrid ARIMA-ANN model, which is able to capture almost all peaks in the validation data set, compared with the other models, Fig. 3. Similar conclusions were obtained for other time series (Qi and Zhang, 2001; Zhang, 2003; Zhang and Qi, 2005; Aburto and Weber, 2007; Gutiérrez-Estrada et al., 2007; Pulido-Calvo and Portela, 2007). This detail is important to forecast PM episodes several hours in advance. With an $E^2$ of 0.9770 (Fig. 3d and Table 7), the hybrid model forecasted the maximum 24-h PM$_{10}$ moving average concentrations at the Las Encinas monitoring station of Temuco.

The results are also presented in the form of tables of contingency (Pérez and Reyes, 2006), Tables 8–11, which show a summary of correct forecasts for episode types and models. The columns show the number of the days forecast to be on a given type against the type of the observed day. A good model should capture high percentages of alert and pre-emergency episodes. Row %P represents the percentage of forecast days by type that were verified to occur. The percentage of false positives is represented by 100-%P. A good forecasting model should perform few false positives of episode types. Bold diagonal numbers are the successful forecasts by episode. The MLR and ARIMAX models had a poor performance to capture alerts and pre-emergency episodes, mainly pre-emergencies. This behavior can be expected, since these models are not good for non-linear patterns (Goyal et al., 2006). On the other hand, the non-linear ANN model had a better performance on the most dangerous episodes; getting 60% of successes in the pre-emergencies, but for alerts none days were

### Table 8

<table>
<thead>
<tr>
<th>Obs.</th>
<th>None</th>
<th>Alert</th>
<th>Pre-emergency</th>
<th>Emergency</th>
<th>Tot.</th>
<th>%P</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>169</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>171</td>
<td>99</td>
</tr>
<tr>
<td>Alert</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td>Pre-emergency</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Emergency</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Tot.</td>
<td>174</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>183</td>
<td>94</td>
</tr>
<tr>
<td>%P</td>
<td>97</td>
<td>33</td>
<td>33</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*None Alert Pre-emergency Emergency Tot. %P* 

### Table 9

<table>
<thead>
<tr>
<th>Obs.</th>
<th>None</th>
<th>Alert</th>
<th>Pre-emergency</th>
<th>Emergency</th>
<th>Tot.</th>
<th>%P</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>170</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>171</td>
<td>99</td>
</tr>
<tr>
<td>Alert</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>29</td>
</tr>
<tr>
<td>Pre-emergency</td>
<td>2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Emergency</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Tot.</td>
<td>176</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>183</td>
<td>94</td>
</tr>
<tr>
<td>%P</td>
<td>97</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 10

<table>
<thead>
<tr>
<th>Obs.</th>
<th>None</th>
<th>Alert</th>
<th>Pre-emergency</th>
<th>Emergency</th>
<th>Tot.</th>
<th>%P</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>170</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>171</td>
<td>99</td>
</tr>
<tr>
<td>Alert</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Pre-emergency</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td>Emergency</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Tot.</td>
<td>175</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>183</td>
<td>95</td>
</tr>
<tr>
<td>%P</td>
<td>97</td>
<td>0</td>
<td>43</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 11

<table>
<thead>
<tr>
<th>Obs.</th>
<th>None</th>
<th>Alert</th>
<th>Pre-emergency</th>
<th>Emergency</th>
<th>Tot.</th>
<th>%P</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>170</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>171</td>
<td>99</td>
</tr>
<tr>
<td>Alert</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>100</td>
</tr>
<tr>
<td>Pre-emergency</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>80</td>
</tr>
<tr>
<td>Emergency</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Tot.</td>
<td>170</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>183</td>
<td>99</td>
</tr>
<tr>
<td>%P</td>
<td>100</td>
<td>88</td>
<td>100</td>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*None Alert Pre-emergency Emergency Tot. %P*
patterns have not been considered. In fact, for 2006 model still has some problems to capture alerts and precaptures, the authors found a whole percentage of success on ANN and hybrid models, a large number of meteorological and air quality data. To have more accurate results on ANN and hybrid models, a large number of observations have to be considered to allow the enough training over those models. In this study, 2080 training observations were used.

The strong performance achieved with the novel hybrid ARIMA–ANN model to forecast PM10 at Temuco could be associated with a typical and almost invariant temporal profile of the hourly PM10 pollution mainly in winter season (May, June, and July), Fig. 4. Since Temuco PM10 is mainly attributed to RWC, this temporal profile and behavior could be similar to other woodsmoked cities, like Christchurch, New Zealand (Wang et al., 2006; Naehler et al., 2007; Titov et al., 2007), but different in other cities with PM10 originating from a mixture of sources, including long-range transport. This may generate more chaos over the air pollution time series, especially on linear and non-linear patterns, and the model selected must be able to predict PM successfully. A hybrid ARIMA–ANN model could meet those requirements. Follow-up papers will demonstrate model performance over several cities worldwide, which include different sources, meteorology, and geography.

4. Conclusions

A novel hybrid ARIMA–ANN model is proposed that is capable of exploiting the strengths of traditional time series approaches for air quality forecasting. Experimental results with meteorological and PM10 data sets indicated that the hybrid model can be an effective tool to improve the forecasting accuracy obtained by either of the models used separately, and compared with a statistical MLR model. The hybrid models took advantage of the unique capabilities of ARIMAX and ANN in linear and non-linear modeling over an air quality time series. The hybrid ARIMA–ANN model errored just in one forecast of pre-emergency episode type over the validation data set, whose concentration was in the border between the alert and pre-emergency episode classification. This hybrid methodology is able to process the air quality forecasting not only one month or a season, but also the whole year. The designation of Temuco as a non-attainment area for PM10 requires reliable air quality forecasting models that allow more accurate alerts for population exposure to the critical pollution episodes and to formulate control measures. To run the hybrid model, the authority needs just the SAS statistical software and meteorological and air quality observed data for the previous day.

Acknowledgments

The authors acknowledge the Comisión Nacional del Medio Ambiente (CONAMA) from the Araucania Region and the Secretaría Regional de Salud Araucanía for permission to use the meteorological and the PM10 data from Las Encinas monitoring station of Temuco, Chile.

References


Fig. 4. Mean of the hourly temporal profile for PM10 concentration in winter season (May, June, and July, 2000–2006) at the Las Encinas site in Temuco, Chile.


Pérez, P., Gramsch, E., 2007. PM10 Forecasting System in Santiago, Chile. In: A&WMA. 100th Annual Conference and Exhibition of the Air and Waste Management Association Pittsburgh, PA, US.


