Atmospheric Environment 42 (2008) 8331–8340

Contents lists available at ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

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 $\leq$ 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-$\mu g\, m^{-3}$ increment in PM_{10} was 1.236 (confidence interval, (CI), 95%, 1.004–1.522) for respiratory mortality, and 1.176 (CI95%, 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 (CI95%, 1.017–1.106) for respiratory mortality and 1.025 (CI95%, 1.005–1.046) for cardiovascular mortality. The authors suggested that the chemical composition and particle size differences for each city could explain the

* Corresponding author. Tel.: +56 45 205483; fax: +56 45 205430.
E-mail address: ldiaz@uct.cl (L.A. Díaz-Robles).

1352-2310/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2008.07.020

Table 1
Episode levels for air quality index (ICAP), Chile

Air quality levels for PM_{10} [$\mu g\,m^{-3}$]	ICAP	Ventilation condition	Episode type
0–194	0–200	Good to regular	None
195–239	201–300	Bad	Alert
240–329	301–500	Critical	Pre-emergency
>330	>501	Dangerous	Emergency

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,

Table 2
Comparison among the ARIMA, ANN, and MLR forecasting models (Zhang, 2003)

ARIMA	ANN	MLR
Better to capture the linear pattern of a time series	Better to capture the non-linear patterns of a time series	Better to capture the linear pattern of a time series
Great versatility	Great versatility	Few versatility
Better for seasonal patterns	Better to capture noise and extreme values (episodes)	Needs correction factors to capture extreme values
Needs historical data continuity	Does not need historical data continuity	Does not need historical data continuity

India, during 1999–2003. They found that the hybrid model had a better performance than ARIMA and the non-linear prediction used separately.

2. Methodology

2.1. Study area and available data

Hourly and daily time series of PM_{10} and meteorological data during 2000–2006 at the Las Encinas monitoring station in Temuco was used. Temuco, the capital of the Araucanía region of Chile (Latitude 38°45′S; Longitude 72°40′W; 100 m.a.s.l) with a population of approximately 300 000 inhabitants, is one of the most polluted cities in South America. The city is located in the center-south of Chile, equidistant between the Pacific Ocean and the Andes. The city placement corresponds to Cautín River-originated fluvial landmasses that developed in a crushed form between two hills, Ñielol (350 m) and Conunhueno (360 m). The region has a temperate rainy climate with Mediterranean influence. Precipitation occurs during all months of the year, with highest precipitation during winter and a dryer period during summer. Temuco has had a fast urban expansion; great economic growth; increased woodstove and industrial source emissions; and increasing vehicle exhaust.

Temuco was declared as non-attainment for PM_{10} in 2005. Due to the abundant wood supplies from local forests, almost 70% of the population uses wood for cooking or heating in winter. It is estimated that 87% of PM_{10} winter emissions originate from residential wood combustion (RWC) (Sanhueza et al., 2005). During 2006, 15 days were reported with PM_{10} concentrations exceeding the Chilean daily PM_{10} standard of 150 $\mu g\,m^{-3}$, among more than 60 days of exceedances in 2001–2006.

Temuco has a unique air quality monitoring station. Las Encinas station is located in the center of a vacant lot in a new residential area, 2.7 km west of the city center, Fig. 1. From this site, hourly temperature, relative humidity, wind direction, and wind speed are acquired. A beta gauge monitor (BAM, Environment S.A., Poissy, France) measures hourly PM_{10}. Daily precipitation, atmospheric pressure, and radiation measurements were obtained from the meteorological station at the Catholic University of Temuco, located ~2 km north east of the site, and 1 km north west of the city center (Fig. 1). To build the models, the available meteorological variables from those two stations were: minimum temperature (Tmin, °C); maximum temperature (Tmax, °C); relative humidity (RH, %); wind speed (WS, $m\,s^{-1}$); solar radiation (R, $Wh\,m^{-2}$day); atmospheric pressure (hPa, P); and precipitation (mm, PP). The maximum PM_{10} moving averages per day ($MaxPM_{10}ma$) values were calculated from the hourly PM_{10} measurements, and then a statistical outlier analysis was performed on these data using the Cook's distance, and DFFITS and DFBETAS coefficients (Belsley et al., 1980; Rawlings, 1988). Fig. 2 shows the maximum 24-h PM_{10} moving average concentrations in Temuco from July 2000 to September 2006. The statistics of the air quality and meteorological data are presented in Table 3. This table includes 1-h maximum PM_{10}, 24-h average PM_{10}, and maximum 24-h PM_{10} moving average concentrations as well, where the maximum concentrations were as high as 1164.0, 321.3, and 353.4 $\mu g\,m^{-3}$, respectively.

The dependent, $MaxPM_{10}ma$ variable, is log-normally distributed and has one marked seasonal comportment, and according to the United States Environmental Protection Agency (USEPA) and other studies (USEPA, 2003; Stadlober et al., 2008), a log transformation can be used to normalize the data to improve the modeling performance. After the outlier study, a transformation analysis was performed to obtain normal distributions for each of the variables, where the main transformation functions were natural logarithms (USEPA, 2003; Stadlober et al., 2008). 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, base 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 $b_0, b_1, \ldots, b_q$ of the linear equation:

$$y = b_0 + b_1 x_1 + \cdots + b_q x_q \tag{1}$$

where the regression coefficients $b_0, b_1, \ldots, b_q$ represent the independent contributions of each independent variable $x_1, \ldots, x_q$ 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, α 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

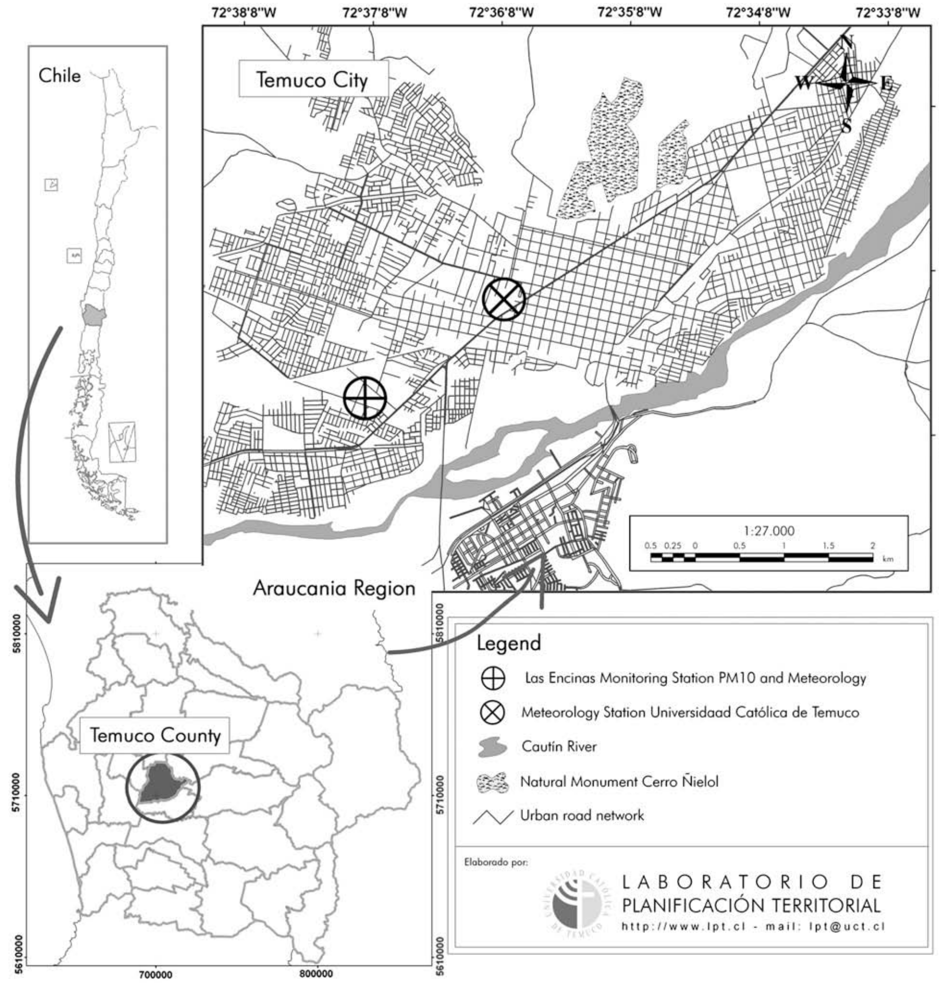

Fig. 1. Las Encinas monitoring station in Temuco city, Chile.

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:

$$Y_t = \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \tag{2}$$

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 ϕ terms are determined by linear regression. The θ_0 term is the

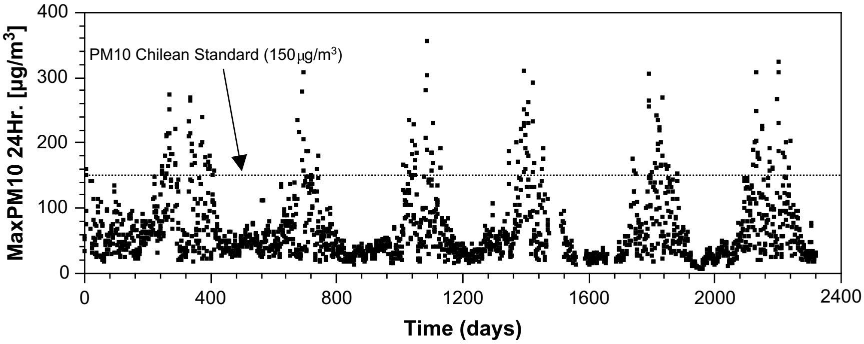

Fig. 2. Maximum 24-h PM_{10} moving average by beta attenuation monitor (BAM) time series at the Las Encinas site in Temuco, Chile (from 07/21/2000 to 09/30/2006).

Table 3
Statistical description of the meteorological and air quality data acquired from Temuco/Las Encinas station from 07/21/2000 to 09/30/2006

Variables	Mean	Standard deviation	Maximum	Minimum
Max 1-h PM_{10}, $\mu g\,m^{-3}$	166.5	168.5	1,164.0	6.0
24-h PM_{10}, $\mu g\,m^{-3}$	48.5	39.6	321.3	5.3
Max 24-h PM_{10} moving average, $\mu g\,m^{-3}$	63.0	50.2	353.4	4.1
Minimum temperature, °C	7.4	3.7	16.6	−4.9
Maximum temperature, °C	17.4	5.3	37.4	2.8
Wind speed, $m\,s^{-1}$	1.9	0.9	6.9	0.3
Precipitation, mm	3.3	7.6	78.6	0.0
Relative humidity, %	80.5	8.7	100.0	52.0
Solar radiation, $W\,m^{-2}$	3,755	2,438	9,185	135
Pressure, mbar	1,003	5	1,018	986

intercept and e_t is the error associated with the regression. This time series depends only on p past values of itself and a random term e_t. The moving average or ARIMA(0,0,q) method is represented as

$$Y_t = \mu - \theta_1 e_{t-1} - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} + e_t \quad (3)$$

where q is the number of the moving average terms, θ_1, θ_2, …, θ_q are the finite weights or parameters set, and μ is the mean of the series. This time series depends only on q past random terms and a present random term e_t. As a particular case, an ARIMA(p,0,q) or ARMA(p,q) is a model for a time series that depends on p past values of itself and on q past random terms e_t. This method has the form of Eq. (4).

$$\begin{aligned} Y_t = {} & \theta_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \mu - \theta_1 e_{t-1} \\ & - \theta_2 e_{t-2} - \cdots - \theta_q e_{t-q} + e_t \end{aligned} \quad (4)$$

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 *Times Series Forecasting System* tool of the SAS 9.1 software (SAS Institute Inc., U.S.).

2.2.3. Non-linear approach: artificial neural networks (ANN) model

On the other hand, according to Allende et al. (2002), an ANN model is a massively parallel-distributed processor that has a natural propensity for storing experiential knowledge and making it available for later use (Allende et al., 2002). It resembles the brain in two respects. The ANN models can recognize trends, patterns, and learn from their interactions with the environment. The most extensively studied and used ANN models are the multilayer feed forward networks (Rumelhart et al., 1986), which allow information transfer only from an earlier layer to the next consecutive layers. Each neuron j receives incoming signals from external variables or every neuron i in the previous layer and there is a synaptic weight (W_{ji}) associated with each incoming signal (x_i). According to Eq. (5), the effective incoming signal (N_j) to the node j is the weighted sum of all the incoming signals.

$$N_j = \sum_{i=1}^{m} x_i W_{ji} \quad (5)$$

The N_j passes through an activation function to produce the output signal (y_j) of the neuron j. The most widely studied activation functions are the logistic sigmoid, hyperbolic tangent sigmoid, and squared functions (Coulibaly et al., 1999).

For this model, and like the ARIMA model, the meteorological variables are the additional independent variables to define the patterns of the air quality time series that composed the best MLR model. A multilayer perceptron (MLP)-type ANN model architecture with a Levenberg–Marquardt (LM) (Shepherd, 1997) training algorithm was used to develop time series models of the non-linear type. The model was built using the *Enterprise Miner* 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 asses 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

Table 4
Training and validation data sets acquired from Temuco/Las Encinas station from 07/21/2000 to 09/30/2006

Data	Number of observations	From	To
Training	2080	07/21/2000	03/31/2006
Validation	183	04/01/2006	09/30/2006

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:

$$E^2 = 1.0 - \frac{\sum_i^n \left|y_i - \widehat{y}_i\right|^2}{\sum_i^n |y_i - \overline{y}|^2} \tag{6}$$

$$\mathrm{ARV} = 1.0 - E^2 \tag{7}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^n \left(y_i - \widehat{y}_i\right)^2}{n}} \tag{8}$$

$$\mathrm{MAE} = \frac{\sum_{i=1}^n \left|y_i - \widehat{y}_i\right|}{n} \tag{9}$$

$$\mathrm{SEP} = \frac{100}{\overline{y}}\mathrm{RMSE} \tag{10}$$

where y_i is the observed value, $\widehat{y}_i$ is the forecasted value to y_i, $\overline{y}$ 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).

$$\mathrm{PI} = 1 - \frac{\sum_{i=1}^n \left(y_i - \widehat{y}_i\right)^2}{\sum_{i=1}^n (y_i - y_{i-1})^2} \tag{11}$$

where y_{i-1} is the observed MaxPM$_{10}$ma 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).

$$\mathrm{BIC} = n\log(\mathrm{SSE}) + m\log(n) \tag{12}$$

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.

Table 5
Parameter estimates for the MLR model for the training data set for PM$_{10}$ samples acquired from Temuco/Las Encinas station from 07/21/2000 to 03/31/2006

Model parameter	Estimate	Standard error	t Ratio	Prob > $\|t\|$
L1PM$_{10}$	0.80845	0.010785	74.96	0.0000
WS	−0.13283	0.011363	−11.69	<0.0001
Tmin	−0.15332	0.012692	−12.08	<0.0001
Tmax	0.05605	0.011683	4.80	<0.0001

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 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%.

$$\mathrm{MaxPM_{10}ma} = 0.80845(\mathrm{L1PM_{10}}) - 0.13283(\mathrm{WS}) - 0.15332(T\mathrm{min}) + 0.05605(T\mathrm{max}) \tag{13}$$

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).

$$\begin{aligned}\mathrm{MaxPM_{10}ma} = {} & \mathrm{ARIMA}(1{,}0{,}1) + 0.60363(\mathrm{L1PM_{10}}) \\ & - 0.15999(\mathrm{WS}) - 0.18140(T\mathrm{min}) \\ & + 0.09197(T\mathrm{max})\end{aligned} \tag{14}$$

where ARIMA(1,0,1) is an ARIMA model with autocorrelation of order 1, without integration, and with a moving

Table 6
Parameter estimates for the ARIMA(1,0,1)-X model for the training data set for PM$_{10}$ samples acquired from Temuco/Las Encinas station from 07/21/2000 to 03/31/2006

Model parameter	Estimate	Standard error	t Ratio	Prob > $\|t\|$
Moving average. Lag 1	0.87057	0.0174	50.1508	<0.0001
Autoregressive. Lag 1	0.98122	0.0063	155.9363	<0.0001
Tmin	−0.18140	0.0136	−13.3511	<0.0001
Tmax	0.09197	0.0181	5.0762	<0.0001
WS	−0.15999	0.0127	−12.6129	<0.0001
L1PM$_{10}$	0.60363	0.0155	38.9242	<0.0001

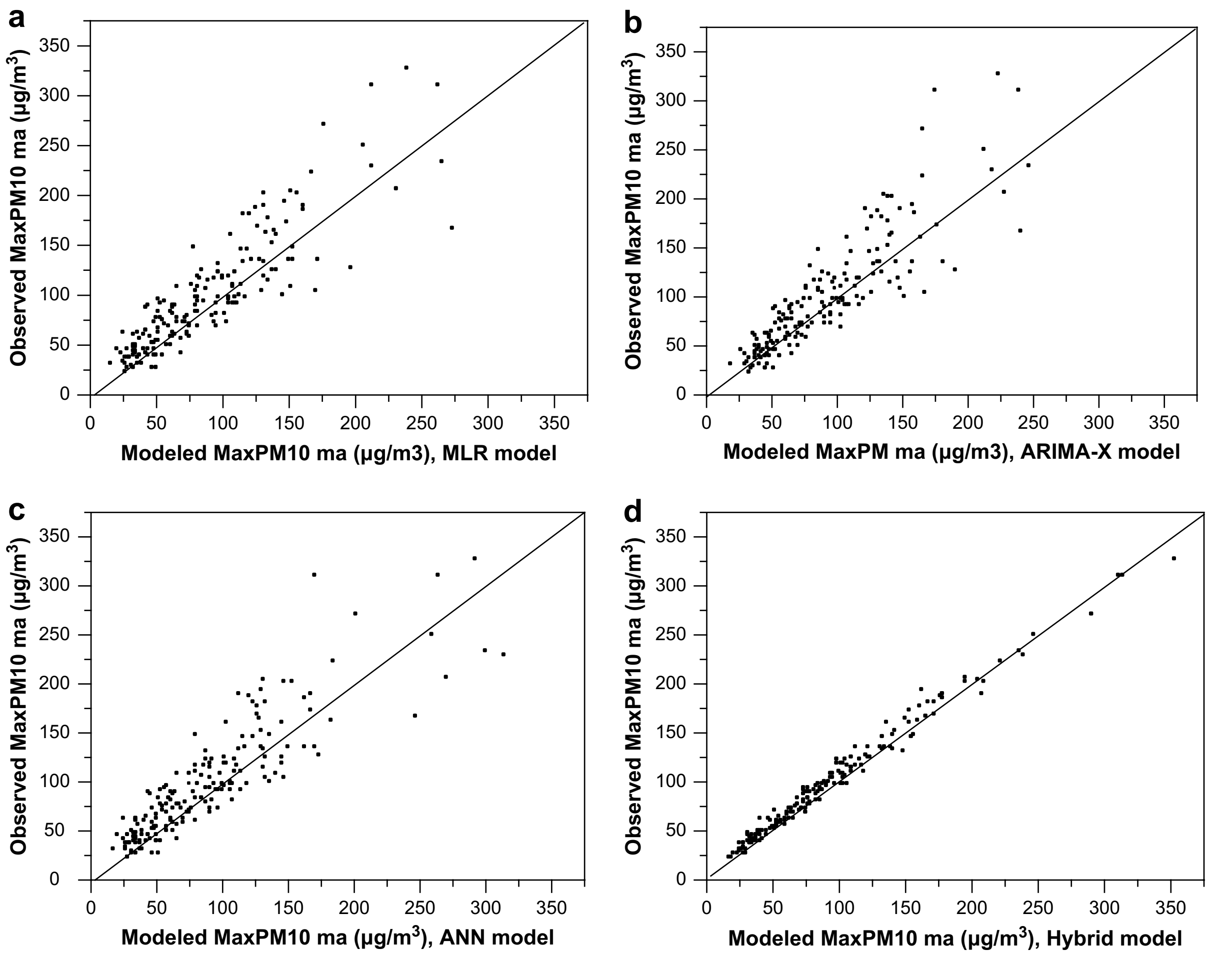

Fig. 3. Models performance using the validation data set for maximum 24-h PM_{10} moving average and meteorological measurements taken at Temuca/Las Encinas, Chile, from 04/01/2006 to 09/30/2006. The observed 24-h PM_{10} moving average concentration was compared with model performance from: a) MLR, b) ARIMA-X, c) ANN, and d) hybrid models.

average of order 1, represented by Eq. 10 and using the parameter estimated from Table 6.

$$Y_t = 0.98122 Y_{t-1} + \mu - 0.87057 e_{t-1} + e_t \tag{15}$$

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_{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

Table 7
Model performance statistics for the validation data set for PM_{10} samples acquired from Temuco/Las Encinas station from 04/01/2006 to 09/30/2006

Estimator	MLR	ARIMAX	ANN	Hybrid
R^2	0.7786	0.7691	0.7770	0.9828
E^2	0.7599	0.7588	0.7569	0.9770
ARV	0.2401	0.2412	0.2431	0.0230
RMSE, $\mu g\, m^{-3}$	28.39	28.46	28.57	8.80
MAE, $\mu g\, m^{-3}$	20.83	19.87	20.65	6.74
SEP	29.70	29.76	29.88	9.20
PI	0.6626	0.6611	0.6585	0.9676
BIC	2198.9	2210.1	2216.7	1801.2

Table 8
Contingency table for the MLR model of the validation data set, Temuco/Las Encinas station from 04/01/2006 to 09/30/2006

Obs.	Forecast					
	None	Alert	Pre-emergency	Emergency	Tot.	%O[a]
None	**169**	1	1	0	171	99
Alert	4	**2**	1	0	7	29
Pre-emergency	1	3	**1**	0	5	20
Emergency	–	–	–	–	–	
Tot.	174	6	3	0	183	94
%P	97	33	33			

[a] Percentage of observed days by type of episodes that were forecast to be in the type.

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 far 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.

Table 10
Contingency table for the ANN model of the validation data set, Temuco/Las Encinas station from 04/01/2006 to 09/30/2006

Obs.	Forecast					
	None	Alert	Pre-emergency	Emergency	Tot.	%O
None	**170**	0	1	0	171	99
Alert	4	**0**	3	0	7	0
Pre-emergency	1	1	**3**	0	5	60
Emergency	–	–	–	–	–	
Tot.	175	1	7	0	183	95
%P	97	0	43	0		

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 9
Contingency table for the ARIMA(1,0,1)-X model of the validation data set, Temuco/Las Encinas station from 04/01/2006 to 09/30/2006

Obs.	Forecast					
	None	Alert	Pre-emergency	Emergency	Tot.	%O
None	**170**	0	1	0	171	99
Alert	4	**2**	1	0	7	29
Pre-emergency	2	3	**0**	0	5	0
Emergency	–	–	–	–	–	
Tot.	176	5	2	0	183	94
%P	97	40	0	0		

Table 11
Contingency table for the hybrid model of the validation data set, Temuco/Las Encinas station from 04/01/2006 to 09/30/2006

Obs.	Forecast					
	None	Alert	Pre-emergency	Emergency	Tot.	%O
None	**170**	1	0	0	171	99
Alert	0	**7**	0	0	7	100
Pre-emergency	0	0	**4**	1	5	80
Emergency	–	–	–	–	–	
Tot.	170	8	4	1	183	99
%P	100	88	100	0		

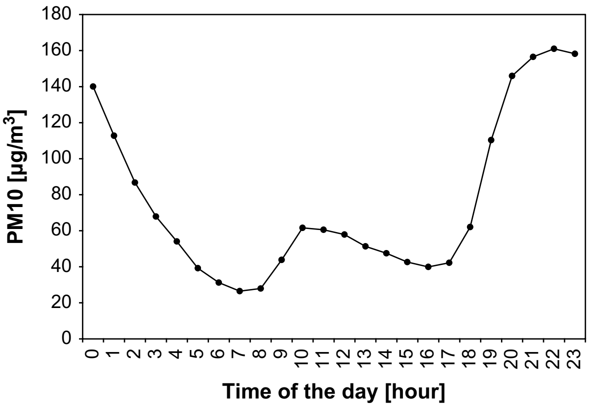

Fig. 4. Mean of the hourly temporal profile for PM_{10} concentration in winter season (May, June, and July, 2000–2006) at the Las Encinas site in Temuco, Chile.

captured. Finally, the hybrid model was able to capture all days and 80% of the alerts and pre-emergency episodes, respectively, Table 11. This advantage is extremely important to be applied in urban air quality forecasting tool. Recently, an ANN model developed for Santiago, Chile, has been effective to forecast those critical episodes, because it is based on complex algorithms that consider the global interactions among 8 monitoring stations (Pérez and Reyes, 2006; Pérez and Gramsch, 2007). However, this model still has some problems to capture alerts and pre-emergency episodes, because the linear time series patterns have not been considered. In fact, for 2006 forecasting of the maximum average 24-h PM_{10} concentrations, the authors found a whole percentage of success of 87%, nevertheless, detection of alert and pre-emergencies as low as 33% and 0%, respectively (Pérez and Gramsch, 2007).

In order to use this model for air quality forecasting and decision-making at Temuco, the hybrid model has to be run every day to forecast the next day using the 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 these models. In this study, 2080 training observations were used.

The strong performance achieved with the novel hybrid ARIMA–ANN model to forecast PM_{10} at Temuco could be associated with a typical and almost invariant temporal profile of the hourly PM_{10} pollution mainly in winter season (May, June, and July), Fig. 4. Since Temuco PM_{10} 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; Naeher et al., 2007; Titov et al., 2007), but different in other cities with PM_{10} 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 PM_{10} 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 erred 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 PM_{10} 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 Araucania for permission to use the meteorological and the PM_{10} data from Las Encinas monitoring station of Temuco, Chile.

References

Aburto, L., Weber, R., 2007. Improved supply chain management based on hybrid demand forecasts. Applied Soft Computing 7, 136–144.
Allende, H., Moraga, C., Salas, R., 2002. Artificial neural networks in time series forecasting: a comparative analysis. Kybernetik 38, 685–707.
Anctil, F., Rat, A., 2005. Evaluation of neural network streamflow forecasting on 47 watersheds. Journal of Hydrologic Engineering 10, 85–88.
Ansuj, A.P., Camargo, M.E., Radharamanan, R., Petry, D.G., 1996. Sales forecasting using time series and neural networks. Computers and Industrial Engineering 31, 421–424.
Aslanargun, A., Mammadov, M., Yazici, B., Yolacan, S., 2007. Comparison of ARIMA, neural networks and hybrid models in time series: tourist arrival forecasting. Journal of Statistical Computation and Simulation 77, 29–53.
Belsley, D., Kuh, E., Welsh, R., 1980. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley & Sons., New York.
Box, G.E.P., Jenkins, G.M., 1970. Time Series Analysis, Forecasting and Control San Francisco, CA.
Cavanagh, J.A.E., Brown, L., Trought, K., Kingham, S., Epton, M.J., 2007. Elevated concentrations of 1-hydroxypyrene in schoolchildren during winter in Christchurch, New Zealand. Science of The Total Environment 374, 51–59.
Chelani, A.B., Devotta, S., 2006. Air quality forecasting using a hybrid autoregressive and nonlinear model. Atmospheric Environment 40, 1774–1780.
Chen, D.G., Hare, S.R., 2006. Neural network and fuzzy logic models for pacific halibut recruitment analysis. Ecological Modelling 195, 11–19.
Chen, K.Y., Wang, C.H., 2007. A hybrid SARIMA and support vector machines in forecasting the production values of the machinery industry in Taiwan. Expert Systems with Applications 32, 254–264.

Chow, J.C., Bachmann, J.D., Wierman, S.S.G., Mathai, C.V., Malm, W.C., White, W.H., Mueller, P.K., Kumar, N., Watson, J.G., 2002. Visibility: science and regulation – discussion. Journal of the Air & Waste Management Association 52, 973–999.

Chow, J.C., Watson, J.G., Mauderly, J.L., Costa, D.L., Wyzga, R.E., Vedal, S., Hidy, G.M., Altshuler, S.L., Marrack, D., Heuss, J.M., Wolff, G.T., Pope, C. A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: lines that connect. Journal of the Air & Waste Management Association 56, 1368–1380.

Coulibaly, P., Anctil, F., Bobee, B., 1999. Hydrological forecasting with artificial neural networks: the state of the art. Canadian Journal of Civil Engineering 26, 293–304.

DeLurgio, S.A., 1998. Forecasting Principles and Applications. Tom Casson, New York.

Goyal, P., Chan, A.T., Jaiswal, N., 2006. Statistical models for the prediction of respirable suspended particulate matter in urban cities. Atmospheric Environment 40, 2068–2077.

Griñó, R., 1992. Neural networks for univariate time series forecasting and their application to water demand prediction. Neural Network World 2, 437–450.

Gutiérrez-Estrada, J.C., Silva, C., Yáñez, E., Rodríguez, N., Pulido-Calvo, I., 2007. Monthly catch forecasting of anchovy *Engraulis ringens* in the north area of Chile: non-linear univariate approach. Fisheries Research 86, 188–200.

Gutiérrez-Estrada, J.C., Vasconcelos, R., Costa, M.J., 2008. Estimating fish community diversity from environmental features in the Tagus estuary (Portugal): multiple linear regression and artificial neural network approaches. Journal of Applied Ichthyology 24, 150–162.

Jain, A., Kumar, A., 2007. Hybrid neural network models for hydrologic time series forecasting. Applied Soft Computing 7, 585–592.

Jeong, K.S., Kim, D.K., Jung, J.M., Kim, M.C., Joo, G.J., 2008. Non-linear autoregressive modelling by temporal recurrent neural networks for the prediction of freshwater phytoplankton dynamics. Ecological Modelling 211, 292–300.

Kitanidis, P.K., Bras, R.L., 1980. Real time forecasting with a conceptual hydrological model. 2: applications and results. Water Resources Research 16, 1034–1044.

Larson, T., Su, J., Baribeau, A.M., Buzzelli, M., Setton, E., Brauer, M., 2007. A spatial model of urban winter woodsmoke concentrations. Environmental Science & Technology 41, 2429–2436.

de Menezes, L.M., Nikolaev, N.Y., 2006. Forecasting with genetically programmed polynomial neural networks. International Journal of Forecasting 22, 249–265.

Naeher, L.P., Brauer, M., Lipsett, M., Zelikoff, J.T., Simpson, C.D., Koenig, J.Q., Smith, K.R., 2007. Woodsmoke health effects: a review. Inhalation Toxicology 19, 67–106.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. I: a discussion of principles. Journal of Hydrology 10, 282–290.

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.

Pérez, P., Palacios, R., Castillo, A., 2004. Carbon monoxide concentration forecasting in Santiago, Chile. Journal of the Air & Waste Management Association 54, 908–913.

Pérez, P., Reyes, J., 2002. Prediction of maximum of 24-h average of PM10 concentrations 30-h in advance in Santiago, Chile. Atmospheric Environment 36, 4555–4561.

Pérez, P., Reyes, J., 2006. An integrated neural network model for PM10 forecasting. Atmospheric Environment 40, 2845–2851.

Piotrowski, A., Napiorkowski, J.J., Rowinski, P.M., 2006. Flash-flood forecasting by means of neural networks and nearest neighbour approach – a comparative study. Nonlinear Processes in Geophysics 13, 443–448.

Pope, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: lines that connect. Journal of the Air & Waste Management Association 56, 709–742.

Prybutok, V.R., Yi, J.S., Mitchell, D., 2000. Comparison of neural network models with ARIMA and regression models for prediction of Houston's daily maximum ozone concentrations. European Journal of Operational Research 122, 31–40.

Pulido-Calvo, I., Portela, M.M., 2007. Application of neural approaches to one-step daily flow forecasting in Portuguese watersheds. Journal of Hydrology 332, 1–15.

Qi, M., Zhang, G.P., 2001. An investigation of model selection criteria for neural network time series forecasting. European Journal of Operational Research 132, 666–680.

Rawlings, J., 1988. Applied Regression Analysis. A Research Tool. Wadsworth & Brooks, Pacific Grove, California.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. "Learning" representations by back propagation errors. Nature 323, 533–536.

Sallehuddin, R., Shamsuddin, S.M.H., Hashim, S.Z.M., Abraham, A., 2007. Forecasting time series data using hybrid grey relational artificial neural network and auto regressive integrated moving average model. Neural Network World 17, 573–605.

Sanhueza, P., Vargas, C., Jiménez, J., 1999. Daily mortality in Santiago and its relationship with air pollution. Revista Medica De Chile 127, 235–242.

Sanhueza, P., Vargas, C., Mellado, P., 2005. Impact of air pollution by fine particulate matter (PM10) on daily mortality in Temuco, Chile. Revista Medica De Chile 134, 754–761.

Schlink, U., Herbarth, O., Richter, M., Dorling, S., Nunnari, G., Cawley, G., Pelikan, E., 2006. Statistical models to assess the health effects and to forecast ground-level ozone. Environmental Modelling & Software 21, 547–558.

Schreuder, A.B., Larson, T.V., Sheppard, L., Claiborn, C.S., 2006. Ambient woodsmoke and associated respiratory emergency department visits in Spokane, Washington. International Journal of Occupational and Environmental Health 12, 147–153.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Shepherd, A.J., 1997. Second-Order Methods for Neural Networks New York.

Slini, T., Kaprara, A., Karatzas, K., Moussiopoulos, N., 2006. PM10 forecasting for Thessaloniki, Greece. Environmental Modelling & Software 21, 559–565.

Sofuoglu, S.C., Sofuoglu, A., Birgili, S., Tayfur, G., 2006. Forecasting ambient air SO2 concentrations using artificial neural networks. Energy Sources Part B-Economics Planning and Policy 1, 127–136.

Sousa, S.I.V., Martins, F.G., Pereira, M.C., Alvim-Ferraz, M.C.M., 2006. Prediction of ozone concentrations in Oporto city with statistical approaches. Chemosphere 64, 1141–1149.

Stadlober, E., Hörmann, S., Pfeiler, B., 2008. Quality and performance of a PM10 daily forecasting model. Atmospheric Environment 42, 1098–1109.

Thomas, S., Jacko, R.B., 2007. Model for forecasting expressway fine particulate matter and carbon monoxide concentration: application of regression and neural network models. Journal of the Air & Waste Management Association 57, 480–488.

Titov, M., Sturman, A.P., Zawar-Reza, P., 2007. Application of MM5 and CAMx4 to local scale dispersion of particulate matter for the city of Christchurch, New Zealand. Atmospheric Environment 41, 327–338.

Tsapakis, M., Lagoudaki, E., Stephanou, E.G., Kavouras, I.G., Koutrakis, P., Oyola, P., von Baer, D., 2002. The composition and sources of PM2.5 organic aerosol in two urban areas of Chile. Atmospheric Environment 36, 3851–3863.

Tseng, F.M., Yub, H.C., Tzeng, G.H., 2002. Combining neural network model with seasonal time series ARIMA model. Technological Forecasting & Social Change 69, 71–87.

USEPAU.S., Environmental Protection Agency, 2003. Guidelines for Developing an Air Quality (Ozone and PM2.5) Forecasting Program, EPA-456/R-03–002. Office of Air Quality Planning and Standards.

Ventura, S., Silva, M., Pérez-Bendito, D., Hervás, C., 1995. Artificial neural networks for estimation of kinetic analytical parameters. Analytical Chemistry 67, 1521–1525.

Wang, D., Lu, W.Z., 2006. Ground-level ozone prediction using multilayer perceptron trained with an innovative hybrid approach. Ecological Modelling 198, 332–340.

Wang, H.B., Wamura, K., Shooter, D., 2006. Wintertime organic aerosols in Christchurch and Auckland, New Zealand: contributions of residential wood and coal burning and petroleum utilization. Environmental Science & Technology 40, 5257–5262.

Watson, J.G., 2002. Visibility: science and regulation. Journal of the Air & Waste Management Association 52, 628–713.

Zhang, G.P., 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing 50, 159–175.

Zhang, G.P., Qi, M., 2005. Neural network forecasting for seasonal and trend time series. European Journal of Operational Research 160, 501–514.