A hybrid model for online prediction of PM$_{2.5}$ concentration: A case study

Yasaman S. Sadabadi$^a$, Majid Salari$^{a,*}$, Reza Esmaili$^b$

$^a$Department of Industrial Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
$^b$Administration Manager of environmental pollution monitoring center Mashhad, Deputy of Services and the urban environment, municipality of Mashhad

$^*$Corresponding Author: Department of Industrial Engineering, Ferdowsi University of Mashhad, Mashhad, Iran. TEL: +985138805113. E-mail address: msalari@um.ac.ir
Abstract

In this paper, we aim at developing a model to predict the daily average concentration of particulate matters with a diameter of less than 2.5 micrometers (PM$_{2.5}$). In the introduced model, we incorporate Weather Research and Forecasting (WRF) meteorological model, Monte Carlo simulation, wavelet transform, and multilayer perceptron (MLP) neural networks. In particular, the MLP and wavelet transformation are combined for prediction. In order to predict the model’s input parameters, including wind speed, wind direction, temperature, rainfall, and temperature inversion, the WRF meteorological model is used. Finally, according to the available uncertainty in the input data and in order to achieve a more accurate prediction, the Monte Carlo simulation is utilized. In order to assess the effectiveness of the model in the real world, it has been conducted in an online mode for 35 days. Numerical results give an acceptable accuracy in terms of some widely used measures. In particular, taking into account the $R$ measurements, it is equal to 0.831 over the set of test instances.

Keywords

PM$_{2.5}$, Prediction, Neural networks, Wavelet transformation, Monte Carlo simulation, WRF model.

1. Introduction

Several big cities in developing countries are increasingly faced with the high levels of air pollution that affect their life quality and overall health [1]. Located in the northeast of Iran, Mashhad is one of these cities that is suffering from air pollution. The dominant pollutant in this city is particulate matters with a diameter of less than 2.5 micrometers (PM$_{2.5}$). This pollutant increases the risk of some diseases, including heart and lung diseases. Reliable prediction of this pollutant provides efficient activities in order to prevent air pollution crisis or at least reduce its adverse effects [2]. However, due to a large number of resources, complexity and variety of physical and chemical processes involved in the formation and transformation of pollutants; accurate prediction is very difficult. To this end, different types of approaches are presented which can basically be divided into three classes, including: empirical, deterministic and statistical approaches [3].

Empirical approaches, which are based on simple principles, are classified into persistence, climatology, and empiricism categories [3]. In the persistence methods, it is assumed that today’s observed pollutant level determines the predicted value of tomorrow [4]. In fact, in this technique, sudden changes in the air quality are not considered. Therefore, due to these sudden changes, this technique usually leads to low accuracy. The climatology method uses the mean value of pollutant values accumulated over 2-5 years, for prediction [3]. In other words, this method is effective only if there exist no changes in the data pattern for the time period considered. In case of changes in the pattern of data, this technique cannot provide an acceptable prediction. Finally, the empiricism method (criteria or rules of thumb) is based on the assumption that the thresholds (i.e., criteria) of meteorological variables or air quality variables can lead to future pollutant concentrations [3,4].

In the deterministic approaches, physical and chemical relations associated with the production and dispersion of pollutants are used to obtain a model for simulation purposes. This type of approach is formed without need to large volumes of measuring. However, it requires precise and complete information about pollutant productive resources, the way of emissions, and physical and chemical processes [5]. Although this type of approaches is even able to predict pollutants in places where there do not exist measurement stations, the knowledge needed for this class of methods is inadequate. Thus, in case that there is not enough knowledge, some approximations and simplifications are employed leading to some errors [6]. In many cases, the use of this approach is computationally costly [3,7].

Statistical approaches are created based on the existing relationship between the concentration of pollutants and effective parameters of air pollution. This type of approach often requires large amounts of data measured under different weather conditions. In this way, the model describes the relationship among variables through the data [5]. The major drawback of this approach is that it is suitable only for a specific area from which the data are taken and cannot be generalized to other regions with different meteorological conditions [5]. However, the statistical
approaches are generally more suitable for a specified area to explore the complex relationship between the concentration of pollutants and possible parameters [6]. In general, statistical approaches provide more accurate results than deterministic methods in reproducing local measured concentrations over short term periods [8]. Furthermore, statistical models are easier, quicker and economical tools to predict the pollutant concentration [9]. The most common statistical approaches are regression models [10,11], artificial neural networks (ANN) [12,13] and support vector machine (SVM) [14,15].

Recently, we have witnessed an increasing trend in the use of ANNs, as a nonlinear tool, for air pollution modeling with satisfactory results [5,16]. The first study in the field of pollution concentration prediction using neural networks was presented by Boznar et al. [17]. In this paper, the goal was to provide hourly prediction of sulfur dioxide (SO$_2$) for the region around the biggest Slovenian thermal power plant at Sostanj. In a paper by Perez and Reyes [18], the authors proposed a multi-layer neural network model and a linear model to predict the average of PM$_{10}$ concentration for up to 24-hour in advance. It has been shown that although ANN models lead to relatively more accurate results than those achieved by linear models, but the choice of input parameters are more important than the nature of the models (i.e., linear or nonlinear). Multi-layer perceptron (MLP) models have been used for the daily average prediction of air pollution index in [19]. The authors propose that the MLP model together with early stopping criterion provides good results. A two-stage neural network for daily prediction of ozone concentration has been proposed in [20]. In particular, at the first stage the meteorological conditions are clustered into meteorological regimes by using self-organizing map neural networks. Upon this step, at the second stage the MLP model is used to approximate the nonlinear relation of ozone-meteorology in each of the meteorological regimes. Hrust et al. [6] proposed a MLP neural network to hourly predict the concentration of CO, O$_3$, NO$_2$, and PM$_{10}$. In this paper, a new approach based on a group of univariate regression models is used for the selection of averaging intervals for input variables. Paschalidou et al. [21] presented several models for hourly prediction of PM$_{10}$ concentration in Larnaca, Limassol, Nicosia, and Paphos, Cyprus, using ANNs, including MLP and Radial Basis Function (RBF), as well as a model based on principal component regression analysis (PCRA). Their evaluation shows that the MLP performs better than the other models. They believe that the poor performance of the RBF is related to the fact that this model is more appropriate for clustering and pattern recognition problems. A hybrid model based on ANN and Taylor expansion has been also proposed in [22] to daily prediction of PM$_{10}$ and SO$_2$ concentrations. The authors claim that their model is appropriate in case that the existing input parameters are not complete. The results of the proposed model are superior to those of that achieved by the MLP and SVM. Perez and Gramsch [7] developed a model for hourly forecasting of PM$_{2.5}$ concentration via neural network in Santiago, Chile. The input data for this model are collected until 7 PM on the day of prediction. Moreover, the forecasting is done up to 15 hours in advance, starting at 8 PM of the present day. Gao et al. [23] investigated the feasibility of using the ANN model with 7 input variables to predict average ozone concentration in daytime (9:00 am - 6:00 pm) in the urban area of Jinan, a metropolis in Northern China. In this model, the input variables are selected with the forward selection procedure. Finally, uncertainty and sensitivity analysis are performed based on Monte Carlo simulations. The uncertainty analysis showed that the ANN model could properly predict the ozone level, while a few of the observed extreme high values fell outside of the 95% confidence interval. Furthermore, the Monte Carlo simulation techniques also used to investigate the sensitivity of the output ozone concentration to the meteorological and temporal input variables. Maximum temperature, atmospheric pressure, sunshine duration and maximum wind speed are identified as the input variables that significantly influence the range of predicted ozone concentration.

The wavelet transform is a tool that has been integrated with some other methods to predict the pollutant concentration. In particular, Siwek and Osowski [24] used the wavelet transformation in combination with a group of neural networks for daily prediction of PM$_{10}$ concentration. In this approach, the outputs of seven types of neural networks are utilized as inputs for another neural network in order to make a final prediction. The results confirm the effectiveness of the wavelet transform in combination with the neural network for air quality prediction. Feng et al. [5] have presented a hybrid model based on the analysis of air mass trajectory, wavelet transformation, and the MLP neural network to predict PM$_{2.5}$ concentration for two days in advance. A hybrid model, including stationary wavelet transformation (SWT) and back propagation neural network (BPNN) was proposed in [25] in order to daily
prediction of PM$_{10}$, SO$_2$, and NO$_2$ concentrations in Nan'an District of Chongqing, Chile. This model is based on the following procedure. First, the SWT is applied to decompose the time series of the pollutant concentrations into different scales, in which the information represents wavelet coefficients of pollutant concentration. Next, the wavelet coefficients together with other inputs are used to train an ANN in each scale. Finally, the estimated coefficients using the BPNN are employed to reconstruct the forecasting results through the inverse SWT. The results of the proposed model are superior to that obtained by the BPNN.

In this paper, we develop a hybrid model for daily prediction of PM$_{2.5}$ pollutant concentration. The introduced model incorporates WRF meteorological model, Monte Carlo simulation, wavelet transform, and multilayer perceptron (MLP) neural networks. In particular, the MLP and wavelet transformation are combined for prediction. In order to predict a set of model’s input parameters, including wind speed, wind direction, temperature, rainfall, and temperature inversion the WRF meteorological model is used. Finally, according to the available uncertainty in the input data and in order to achieve a more accurate prediction, the Monte Carlo simulation is utilized.

The paper is organized as follows: A brief explanation of the area under study, the existing data, and the variety of methods which are used in this paper, are introduced in Section 2. Section 3 provides the computational results. Finally, concluding remarks are reported in Section 4.

2. Data and methods

In this section, description of the study area, existing data, and the characteristics of the forecasting methods which are used to predict the PM$_{2.5}$ concentration, are reported.

2.1. Study area and available data

Mashhad, in the northeastern part of Iran, is geographically located in the range of longitude 59° 15' to 60° 36' and latitude 35° 43' to 37° 8'. In this city, there exist 11 monitoring stations to hourly measure air pollutants, including: CO, PM$_{10}$, PM$_{2.5}$, SO$_2$, and NO$_2$.

The first category of effective parameters is meteorological data. The effects of climate and meteorological factors on urban problems such as air pollution has become important in the recent years [26]. In particular, the meteorological parameters contain wind direction, wind speed, temperature, pressure, rainfall, and humidity, which are taken from the Meteorological Organization of Mashhad, Iran [27]. The meteorological standard for wind direction is as follows: 0 deg = wind from the north, 90 deg = wind from the east, 180 deg = wind from the south, 270 deg = wind from the west. The data are collected 8 times per day (every 3 hours). The wind speed and the wind direction are integrated by utilizing the following relations [5,13,24]. Preliminary results that this choice provides better prediction results.

\[
\text{wind}_x = w \cos(\phi) \\
\text{wind}_y = w \sin(\phi) 
\]

(1)

In these relations $w$ and $\phi$ are the wind speed and the wind direction, respectively. It is worth mentioning that for each involved parameter, the data have been collected from March 2014 to September 2015.

Another category of parameters is the traffic volume. Motor vehicles are one of the most important sources of air pollution in several cities due to the high level of pollution emission. The Sydney Coordinated Adaptive Traffic System (SCATS) data are used to assess the traffic volume in the studied area. In particular, SCATS uses sensors at each traffic signal to detect vehicle presence in each lane [28]. Since the main goal of this paper is to present an appropriate model which can be used in reality, it is recommended to use those input parameters that are predictable.
In particular, we processed the traffic data collected from 10% of the crossroads in Mashhad, Iran, from March 2014 to September 2015. The results show that there is a significant difference between the traffic volume during working days and weekends. Therefore, in our developed model the parameter corresponding to the traffic volume is considered as a binary parameter. Essentially, it takes 1 and 0 for weekends and working days, respectively.

There could be a situation in which a layer of cool air at the surface is overlain by a layer of warmer air and this phenomenon is called temperature inversion. Temperature inversion is an important factor which has a fundamental role in air pollution. In general, the temperature inversion is divided into two types, namely, subsidence and radiation. Information about this parameter is taken from the University of Wyoming in Laramie, Wyoming [29]. The information is then analyzed using RAwinsonde OBservation (RAOB) software in terms of the existence and type of the inversion. RAOB automatically decodes data from over 100 different formats and plots data on 12 interactive displays including skew-Ts, hodographs, and cross-sections. Produces displays of over 200 atmospheric parameters including inversions, icing, turbulence, wind shear, clouds, and more [30]. Airport station was Shahid Hashemi Nejad international airport, which is located in Mashhad. This input parameter is included in the model in a way that it takes 1 if the inversion type of radiation or subsidence happens at altitude below 1200 meters above the sea level, otherwise, the input value is equal to 0.

Time-dependent variables are also added into the input data set to reflect the relation between the pollutant concentration and seasonal cycles. To this end, the values 1 to 4 are associated with spring to winter, respectively. Finally, as it is shown in Figure 1, there exists a high correlation between the pollutant concentration in the previous day of prediction and the actual day. As a result, the pollutant concentration with lag 1 is considered as an input of the model. PM$_{2.5}$ concentration data for the required period is taken from Environment Pollution Monitoring Center (EPMC) of Mashhad [31].

Table 2 gives the correlation coefficients of the environmental parameters. The results indicate that there is a high correlation between pressure and humidity, temperature and pressure as well as temperature and humidity parameters. This fact is also well demonstrated in Figure 2 presenting the distribution of available data corresponding to the pressure and humidity (a), temperature and pressure (b), and temperature and humidity (c) parameters.

Table 2 gives the correlation coefficients of the environmental parameters. The results indicate that there is a high correlation between pressure and humidity, temperature and pressure as well as temperature and humidity parameters. This fact is also well demonstrated in Figure 2 presenting the distribution of available data corresponding to the pressure and humidity (a), temperature and pressure (b), and temperature and humidity (c) parameters.

In fact, in case that there is a correlation between two parameters, only one parameter can be considered as an input. Therefore, the following choices are possible: removal of pressure and humidity, temperature and pressure, temperature and humidity. In order to select the best scenario, three MLP neural networks are trained. In each network, one of the three mentioned combination of parameters is removed from the inputs of the model. In addition, 20% of the data are used to test the three neural networks. The results of the MAE measure show that the removal of the pressure and humidity is the best scenario by generating the minimum average MAE value. Hence, in the following, pressure and humidity are removed from the input parameters.

2.2. Solving method

In this study, we have proposed two models, namely offline and online models, to predict the level of PM$_{2.5}$ concentration for one day ahead. In the offline model we utilize the exact values of PM$_{2.5}$ and meteorological data
for prediction while in the online model we use the predicted values for the uncertain data. The details are provided in the following.

The goal of the developed offline model is to provide a tool for predicting the average value of PM$_{2.5}$ concentration for one day ahead. To do so, we have collected a data set including the PM$_{2.5}$ concentration, temporal variables, and meteorological data including temperature, rainfall, wind, and temperature inversion for the day of prediction (see, section 2.1). In the developed model, at the first step, a model based on the combination of the wavelet transform and the MLP neural network is created. To this end, the time series of the pollutant concentration with high variability is decomposed into several sub-series with low variability by the use of discrete wavelet transformation (see, section 2.2.2). Following this step, the MLP is applied in each of decomposition levels and the results at each level are summed up to achieve the final time series of pollutant concentration (see, section 2.2.2).

In the online model, at first the WRF meteorological model is used to predict the meteorological input parameters (see, section 2.2.3). Uncertainty in the model could be due to the input data of the model which is correlated with its accuracy and quality. Among of the eight input parameters, the type of day, season, and pollutant concentration in the previous day of prediction are not subject to uncertainty. However, other five parameters (temperature, wind in two directions, rainfall and temperature inversion) cannot be predicted with certainty. Therefore, uncertainty in input data cannot be eliminated. As a result, to reduce the model's uncertainty, due to the predicted input parameters, the Monte Carlo simulation is utilized (see, section 2.2.4). Using trained neural networks in each of decomposition levels, the prediction is achieved and ultimately the results of all predictions at all levels are summed up and the PM$_{2.5}$ concentration is predicted to one day ahead.

The general overview of the proposed model in online mode is depicted in Figure 3. In the following, we provide more explanation about each of the involved tools corresponding to the developed algorithm.

2.2.1. Neural network

We use the multi-layer perceptron neural network in the proposed model. The activation functions in hidden and output layers are logistic and linear function, respectively. We utilize the Levenberge-Marquardt (LM) learning algorithm for training the network. Moreover, the early stopping method is used to avoid over fitting the model with the training data. In the early stopping technique, the error of the validation data set during the training process is used to monitor the training procedure. The training procedure stops when the validation error for a certain number of iterations increases. The weights and biases are adjusted once the error is minimized. In order to determine the optimal topology of the neural network, i.e. number of hidden layers and neurons in each layer, we test all network topologies using the validation data and extract the best scenario. In particular, we consider to have a maximum of two hidden layers and a maximum of 20 neurons in each hidden layer.

2.2.2. Discrete wavelet transformation

Due to high variability of the time series of PM$_{2.5}$ concentration, an accurate prediction is difficult. The solution is to decompose the time series with high variability into sub-series with lower variability. Following this step, the prediction strategy can be applied to each level with lower variability. Finally, the results on all levels are summed up. In this paper, discrete wavelet transformation is used for this purpose.

Insert Figure 3 around here

In the composition process, the original time series is decomposed into detailed coefficients $D_j$ at different levels ($j = 1, 2, ..., J$) and a residual approximated coefficient $A_j$ through highpass and lowpass filters [32, 33]. In this approach, the number of points is halved in each level, such that the number of points in level $j$ is half of that in
level \((j - 1)\). Therefore, for reconstruction of the primary time series, a reconstruction algorithm is employed [32,33]. The objective is that the number of points in each level is equal to the original numbers. Figure 4 shows the approximations and details in \(J\) levels. The original signal is reconstructed by the approximation and details according to Equation 2.

We use Symlets sym6 wavelet as it provides the smallest variability at each decomposition level. To determine the appropriate number of levels, all possible scenarios are considered. Since in the decomposition approach the number of points in each level is halved, the maximum number of levels is \(\log_2 N\) in which \(N\) is the number of data. According to 469 available data, the maximum number of levels is 8. Using the MLP neural network, all possible scenarios are tested by utilizing a set of test data for each possible level \((i.e., J = 1, ..., 8)\). The results of the MAE index show that the 5 level scenario is the best choice. Figure 5 presents the results of 5-level wavelet decomposition of the data for PM\(_{2.5}\) concentrations in which \(S\) gives the original time series corresponding to the available date.

Insert Figure 4 around here

\[
S(n) = D_1(n) + D_2(n) + ... + D_J(n) + A_J(n) \tag{2}
\]

Insert Figure 5 around here

In converting the original time series to the wavelet representation, we have applied the strategy of using the whole data with moving windows by use of 364 previous items and the next day value. In order to apply the wavelet transformation for the decomposition of PM\(_{2.5}\) concentration on the previous day of prediction in the online mode, we take 364 previous days and adding the actual day at the end of the series.

2.2.3. WRF meteorological model

As already described in the previous sections, the goal of this paper is to develop a model for one day ahead prediction of PM\(_{2.5}\) concentration. To this end, it is initially required a set of input parameters, including wind, in two directions, temperature, rainfall, temperature inversion, type of day, and season of the year, for the day of prediction and PM\(_{2.5}\) concentration for the previous day. It is obvious that the actual value of some parameters, including wind speed in two directions, temperature, rainfall, and temperature inversion is not available for the day of prediction. Therefore, the predicted values of these parameters have to be used. To this end, the WRF meteorological model is utilized to predict these parameters for one day ahead. WRF is a state-of-the-art atmospheric modeling system designed for both meteorological research and numerical weather prediction. It offers a host of options for atmospheric processes and can run on a variety of computing platforms. The WRF model configuration is shown in Figure 6. Initially, terrestrial data (such as terrain, landuse, and soil types) and gridded data (such as GFS data) is required to run the WRF model. The WRF preprocessing system (WPS) is a set of three programs whose collective role is to prepare input to the real program for real-data simulation. ARW solver is the key component of the modeling system, which is composed of several initialization programs for idealized, and real-data simulations, and the numerical integration program. Several programs are supported for post processing i.e. NCAR Graphics Command Language (NCL). Graphic Tools facilitate visualization of the model output and we have used NCL in the present work. The details of WRF model are described in the user guide [34]. In this study, to extract the required meteorological data over Mashhad domain, WRF model is configured with two nested domains with horizontal resolutions of 5 and 15 km, and with a 100×99 and 85×67 grid points, respectively. The domains were run together efficiently using 2-way grid nesting in WRF. The vertical structure of the atmosphere is resolved with 30 vertical levels extending up to 5000 Pa. The Global Forecast System (GFS) data of 0.25°×0.25° and a vertical resolution of 27 pressure levels were used to define the initial and the boundary conditions. After running WRF, the NCL post-processing tool was used in order to translate meteorological data from WRF output to the required format. Finally, activated physical schemes in the WRF model are as follows: Microphysics, long wave radiation,
shortwave radiation, surface layer, land surface, planetary boundary layer, cumulus parameterization, fluxes, cloud effect and the number of soil layers in the land surface model.

Insert Figure 6 around here

2.2.4. Uncertainty analysis and Monte Carlo simulations

The model uncertainty is due to ambiguity in the input data. Among the eight input parameters of the model, type of the day, and season of the year are not subject to uncertainty and variability. However, the other six input parameters, i.e., wind in two directions, temperature, rainfall, PM$_{2.5}$ concentration and temperature inversion, cannot be predicted with absolute certainty. Therefore, uncertainty in input data cannot be eliminated from the model. To achieve a more accurate prediction, the Monte Carlo simulation is used. We do not consider the uncertainty in measuring the PM$_{2.5}$ concentration, because it does not improve the model results. The mechanism of the Monte Carlo method is based on the idea of taking a sample random population and estimating the desired outputs from this sample [35]. The expected value of a function $g$ of the uncertain variables of $x$ with a probability density function of $f_x(x)$ is given by Equation 3.

$$E(g(x)) = \int_{x \in X} g(x) f_x(x) dx$$  \hspace{1cm} (3)

In particular, $x$ represents the input parameters, including the temperature, wind in two directions, or the volume of the rainfall and $g(x)$ represents the predicted value of pollutant concentration (PM$_{2.5}$) resulted from the proposed model. By taking $n$ samples of $x$, i.e.($x_1, x_2, ..., x_n$), the Monte Carlo estimation can be achieved by utilizing Equation 4.

$$\bar{g}_n(x) = \frac{1}{n} \sum_{i=1}^{n} g(x_i)$$  \hspace{1cm} (4)

We assume each uncertain parameter to have a uniform distribution over a given interval. In particular, for each uncertain parameter, the lower and upper bounds of this interval are set by utilizing, respectively, the minimum and maximum values achieved from the WRF model and two websites, namely www.wunderground.com and www.timeanddate.com. The algorithm uses an iterated algorithm to set the appropriate number of samples ($n$). At the first step, the model is run 10 times and the average values of output is considered as the predicted PM$_{2.5}$ concentration. Essentially, in a run of the model for each uncertain parameter, the algorithm starts by taking a random sample from the corresponding interval to be used as the input parameter. Following this step, the number of samples is doubled (i.e., 2*10) and the algorithm is run again. This process continues as long as the mean difference of the PM$_{2.5}$ concentration for two consecutive iterations is greater than 0.01.

3. Results and discussion

In this section we evaluate the effectiveness of the proposed model. As mentioned earlier, the MLP neural network has to be applied at each decomposition level. To this end, 20% of the data (see, section 2.1) are selected, randomly, to be used as the test set. Furthermore, the test data are uniformly taken from different seasons of the year. The remaining data are considered as the training and validation sets. The MLP model is run 100 times and in each run all the training and validation data are chosen, randomly. Since the initial weights of the MLP are set randomly, the network is trained 10 times per run to select the best network. Finally, the network with the least mean square error over the validation data will be selected as the working one. By using the early stopping criterion [36] in the neural network, it is guaranteed to obtain similar results for both training and testing data.

To evaluate the results, two following approaches are utilized:
We use different measures, including the mean absolute error ($MAE$), the root mean square error ($RMSE$), the correlation coefficient ($R$), and the index of agreement ($IA$) which are reported in (5)-(8), respectively.

\[
MAE = \frac{1}{n} \left( \sum_{i=1}^{n} |p_i - p_a| \right)
\]  

(5) 

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (a_i - p_i)^2}
\]  

(6) 

\[
R = \frac{R_{pa}}{std(p)\cdot std(a)}
\]  

(7) 

\[
IA = 1 - \frac{\sum_{i=1}^{n} (a_i - p_i)^2}{\sum_{i=1}^{n} (|a_i - \bar{a}| + |p_i - \bar{a}|)^2}
\]  

(8) 

In these relations $n$ gives the number of instances, $p_i$ and $a_i$ represent, respectively, the predicted and observed values. In addition, $\bar{a}$ is the average of observed data, $std$ denotes the standard deviation, and $R_{pa}$ is the covariance value between the predicted and observed data.

The error is usually defined in terms of the band error. This criterion gives the difference between the observed and predicted intervals where the observed and predicted values fall. The range of pollution values is normally divided into five equal intervals. In this paper, the bands for PM$_{2.5}$ concentration are: [0-22], [23-44], [45-66], [67-88], [89-110]. According to the data, the number of PM$_{2.5}$ concentration in each band is 238, 234, 28, 6, and 1, respectively. As an example, the band error in measured and predicted pair (21, 24) is reported as +1, since 21 and 24 fall into the first and second intervals, respectively. The same measure is used in [5,37,38].

The proposed model is evaluated by applying the $MAE$, $RMSE$, $R$, $IA$ and band error criteria, for which the results are reported in Tables 3 and 4, respectively. The results in Table 3 show that the combination of wavelet transformation with MLP (MLP+W) provides a significant improvement in the performance of that achieved by the MLP.

Table 4 reports the output of the band error criterion. In this table, the numbers in parentheses give the error percentage while the numbers before the parentheses indicate the number of samples. The results show the superiority of the MLP+W model by obtaining a lower error than the MLP model. In particular, the error for MLP and MLP+W models are 30.85% and 27.66%, respectively, and the error is only ±1 band.

The scatter diagram of the results obtained by the MLP+W model compared to the measured data is plotted in Figure 7. In addition, Figure 8 illustrates the results of forecasting the test data obtained by the MLP+W model in a graphical way. Comparing MLP+W and MLP models show that some of the high peaks missed by the MLP model are almost anticipated by the hybrid model.

To find the distribution of the prediction errors, the histogram of the errors for the results in the MLP+W model is shown in Figure 9. In this diagram, the horizontal and vertical axes give, respectively, the error value and the number of samples.
The results of the proposed model in the online mode for one day ahead prediction of 19 days in May and 16 days in Nov 2016 are shown in Tables 5 and 6. The results in Table 5 indicate the high accuracy of the proposed model in the online mode. Table 6 shows that only the prediction of 13 days (from 35 days) is not classified in the desired category. However, in the 22 remaining days, the predicted values are in accordance with reality.

The scatter diagram of the prediction data resulted from the proposed model in the online mode and the measured data is shown in Figure 10.

4. Conclusions

In this paper, we proposed a hybrid model for the one day ahead prediction of the daily average concentration of particulate matter smaller than 2.5 micrometers in online mode. This model is based on the WRF meteorological model, Monte Carlo simulation, wavelet transform, and MLP neural network. In this model, the decomposition of time series of concentration pollutant into several levels with lower variability increases the accuracy of the final forecast. Moreover, the proposed model doesn’t need comprehensive information about air pollutants and pollution sources. The proposed model is also able to consider the nonlinear relation among various parameters. One of the main advantages of the proposed model is its proper implementation in real-world as it uses predictable input parameters.

To evaluate the performance of the model in real-world, the proposed model has been conducted in online mode for 35 days. In this regard, the values of parameters including temperature, rainfall, wind in two directions, and temperature inversion are obtained using the WRF meteorological model for one day ahead. Since these parameters cannot be predicted with specific certainty, for the improvement of the accuracy of the prediction, Monte Carlo simulation is used. The results of the prediction of the period under study in the online mode, show the good performance of the proposed model.

References

[1] Molina, M.J. and Molina, L.T. “Megacities and atmospheric pollution” Air. Waste. Manag. Assoc., 54, pp. 644-680 (2004).
[2] Kolehmainen, M., Martikainen, H. and Ruuskanen, J. “Neural networks and periodic components used in air quality forecasting”, Atmos. Environ., 35, pp. 815-825 (2001).
[3] Zhang, Y., Bocquet, M., Mallet, V. and et al. “Real-time air quality forecasting, part I: History, techniques, and current status”, Atmos. Environ., 60, pp. 632-655 (2012).
[4] U.S. EPA. “Guidelines for developing an air quality (Ozone and PM_{2.5}) forecasting program”, U.S. Environmental Protection Agency, Office of Air Quality Planning and Standards, Research Triangle Park, North Carolina. EPA-456/R-03-002 (2003).
[5] Feng, X., Li, Q., Zhu, Y. and et al. “Artificial neural networks forecasting of PM_{2.5} pollution using air mass trajectory based geographic model and wavelet transformation” Atmos. Environ., 107, pp. 118-128 (2015).
[6] Hrust, L., Klać, Z.B., Križan, J. and et al. “Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations”, Atmos. Environ., 43, pp. 5588-5596 (2009).
[7] Perez, P. and Gramsch, E. “Forecasting hourly PM_{2.5} in Santiago de Chile with emphasis on night episodes”, Atmos. Environ., 124, pp. 22-27 (2016).
[8] Donelly, A., Naughton, O., Broderick, B. and et al. “Short-Term Forecasting of Nitrogen Dioxide (NO_{2}) Levels Using a Hybrid Statistical and Air Mass History Modelling Approach”, Environ. Model. Assess., 22, pp. 231-241 (2016).
[9] Fernando, H.J.S., Mammarella, M.C., Grandoni, G. and et al. “Forecasting PM_{10} in metropolitan areas: efficacy of neural networks”, Environ. Pollut., 163, pp. 62-67 (2012).
Yasaman S. Sadabadi received the B.Sc. and M.Sc. degrees in Industrial Engineering from the Sadjad University of Technology and the Ferdowsi University, Mashhad, Iran in 2013 and 2016, respectively. Her research interests include prediction of pollutant concentration.
Majid Salari received the BSc and MSc degrees in Applied Mathematics from the Ferdowsi University of Mashhad, Iran in 2004 and 2006, respectively. In addition, he has PhD in Automations and Operations Research from the University of Bologna, Italy in 2010. In 2010, he joined the Department of Industrial Engineering, Ferdowsi University of Mashhad, as an Assistant Professor, and became an Associate Professor in 2016. His current research interests include combinatorial optimization, applied operation research, evolutionary computation, heuristic search, logistics, distribution, and vehicle routing.

Reza Esmaili received the Ph.D. degree in Climatology. He is the administrative manager of environmental pollution monitoring center of Mashhad (second metropolis of Iran) since 2014. Reza had done several scientific researches on urban bio-climatology and tourism climate in Iran during his Ph.D program. At the moment, most of his recent investigations are focused on air pollution due to his executive work experiences. His current research interests include urban environment's pollutant, air pollution, modelling of air pollution, climate risks and urban climatology.

Figures Caption
Figure 1: The relationship between PM$_{2.5}$ concentration for the previous ($d-1$) and the actual day ($d$)
Figure 2: The relation between the humidity and pressure (a), the pressure and temperature (b), humidity and temperature (c)
Figure 3: The general overview of the proposed model
Figure 4: Details and approximations achieved from discrete wavelet transform in J levels
Figure 5: The wavelet decomposition of the original time series $S$ of PM$_{2.5}$ concentration at 5 levels
Figure 6: The main structure of the WRF model
Figure 7: The scatter diagram of PM$_{2.5}$ concentration obtained by MLP+W and measured data
Figure 8: The resulted of prediction of the test data obtained by MLP+W: the predicted (red line) and measured (blue line) values of PM$_{2.5}$ concentration
Figure 9: The histogram of errors for PM$_{2.5}$ concentration obtained by the MLP+W model
Figure 10: The scatter diagram of PM$_{2.5}$ concentration obtained by proposed model and the measured data in online mode

Tables Caption
Table 1: Statistical description of the collected input data from March 2014 to September 2015
Table 2: The correlation coefficients between different parameters used in prediction of PM$_{2.5}$
Table 3: The results ($MAE$, $RMSE$, $R$, $IA$) of models on the test data
Table 4: The results (band error) of models on the test data
Table 5: The results ($MAE$, $RMSE$, $R$, $IA$) of proposed model in online mode
Table 6: The results (band error) of proposed model in online mode

![Figure 1](image)
Figure 2
Figure 3

Pollutant concentration (\(d-1\))

Wavelet transformation

Type of day (\(d\))

Season (\(d\))

\(D_1\)

\(D_2\)

\(D_3\)

\(D_4\)

\(D_5\)

\(A_5\)

Run MLP for \(D_1\)

Run MLP for \(D_2\)

Run MLP for \(D_3\)

Run MLP for \(D_4\)

Run MLP for \(A_5\)

Predicted pollutant concentration (\(d\))

Take a random value from \([W_x^L, W_x^U]\)

Take a random value from \([W_y^L, W_y^U]\)

Take a random value from \([T^L, T^U]\)

Take a random value from \([R_L, R_U]\)

Run the WRF model (\(d\))

Extracting meteorological data from two websites (\(d\))

NCL

Temperature inversion (\(d\))

Interval of wind\(x\) (\(d\))

\([W_x^L, W_x^U]\)

Interval of wind\(y\) (\(d\))

\([W_y^L, W_y^U]\)

Interval of temperature (\(d\))

\([T^L, T^U]\)

Interval of rainfall (\(d\))

\([R^L, R^U]\)
Figure 4

S

Figure 5

Figure 6
Table 1

| Variable             | Unit   | Range            | Mean  | St. Dev. |
|----------------------|--------|------------------|-------|----------|
| PM$_{2.5}$           | µg/m$^3$ | [4.83, 105.06] | 26.38 | 11.61    |
| Wind$_x$             | m/s    | [-8.16, 4.65]   | 1.03  | 1.56     |
| Wind$_y$             | m/s    | [-4.05, 6.35]   | 0.03  | 1.58     |
| Temperature          | ºC     | [-2.19, 34.56]  | 18.86 | 9.90     |
| Pressure             | hPa    | [991.91, 1039.28]| 1012.83 | 8.45    |
| Rainfall             | mm     | [0, 4]           | 0.07  | 0.28     |
| Humidity             | %      | [10.25, 99.5]   | 41.34 | 26.06    |
| Type of day          | ---    |                  | ---   | ---      |
| Season of year       | ---    |                  | ---   | ---      |
| Temperature Inversion| ---    |                  | ---   | ---      |

Table 2

| Wind$_x$ | Wind$_y$ | Temperature | Pressure | Rainfall | Humidity | PM$_{2.5}$ |
|----------|----------|-------------|----------|----------|----------|------------|
| 1        | -0.398   | 0.086       | 0.046    | -0.056   | -0.022   | -0.034     |
| -0.398   | 1        | 0.086       | 0.046    | -0.369   | -0.022   | -0.325     |
| 0.086    | 0.401    | 0.046       | -0.369   | -0.222   | -0.325   | 0.063      |
| 0.046    | -0.369   | -0.887      | 0.046    | -0.25    | -0.904   | 0.003      |
| -0.056   | -0.022   | -0.25       | -0.887   | 0.115    | 0.767    | -0.038     |
| -0.022   | -0.325   | -0.904      | -0.369   | 0.369    | 1        | 0.107      |
| -0.034   | 0.063    | -0.003      | 0.046    | 0.107    | 0.027    | 1          |

Table 3

|                  | MLP    | MLP+W  |
|------------------|--------|--------|
| MAE              | 6.090  | 5.397  |
| RMSE             | 8.216  | 7.060  |
| $R$              | 0.661  | 0.749  |
| LA               | 0.806  | 0.850  |
| ±1 band | MLP  | MLP+W |
|---------|------|-------|
| ±2 band | 0 (0.0) | 0 (0.0) |
| ±3 band | 0 (0.0) | 0 (0.0) |
| ±4 band | 0 (0.0) | 0 (0.0) |
| **Total** | **29 (30.85)** | **26 (27.66)** |

| **Online model** | **MAE** | **RMSE** | **R** | **IA** |
|------------------|---------|----------|-------|--------|
| **MLP**          | 5.323   | 6.783    | 0.831 | 0.894  |

| ±1 band | 13 (37.14) |
| ±2 band | 0 (0.0)    |
| ±3 band | 0 (0.0)    |
| ±4 band | 0 (0.0)    |
| **Total** | **13 (37.14)** |