Prediction of Airborne Nanoparticles at Roadside Location Using a Feed–Forward Artificial Neural Network

Article *in* Atmospheric Pollution Research · November 2016

DOI: 10.1016/j.apr.2016.11.004

Abdullah N. Al-Dabbous
Kuwait Institute for Scientific Research
13 PUBLICATIONS **55** CITATIONS

Prashant Kumar
University of Surrey
145 PUBLICATIONS **2,128** CITATIONS

Contents lists available at ScienceDirect

Atmospheric Pollution Research

journal homepage: www.elsevier.com

Prediction of airborne nanoparticles at roadside location using a feed–forward artificial neural network

Abdullah N. Al–Dabbous[a, *], Prashant Kumar[b, c], Abdul Rehman Khan[d]

[a] *Crises Decision Support Program, Environment and Life Sciences Research Center, Kuwait Institute for Scientific Research, P.O. Box 24885, 13109, Safat, Kuwait*
[b] *Department of Civil and Environmental Engineering, Faculty of Engineering and Physical Sciences (FEPS), University of Surrey, Guildford, GU2 7XH, United Kingdom*
[c] *Environmental Flow (EnFlo) Research Centre, FEPS, University of Surrey, Guildford, GU2 7XH, United Kingdom*
[d] *Environment and Life Sciences Research Center, Kuwait Institute for Scientific Research, P.O. Box 24885, 13109, Safat, Kuwait*

ARTICLE INFO

Article history:
Received 3 October 2016
Received in revised form 21 November 2016
Accepted 22 November 2016
Available online xxx

Keywords:
Artificial neural network
Prediction
Nanoparticles
Comparative evaluation
Multi–layer feed–forward network
Back–propagation training algorithm

ABSTRACT

Accurate prediction of nanoparticles is essential to provide adequate mitigation strategies for air quality management. On the contrary to PM_{10}, SO_2, O_3, NO_x and CO, nanoparticles are not routinely–monitored by environmental agencies as they are not regulated yet. Therefore, a prognostic supervised machine learning technique, namely feed–forward artificial neural network (ANN), has been used with a back–propagation algorithm, to stochastically predict PNCs in three size ranges ($N_{5–30}$, $N_{30–100}$ and $N_{100–300}$ nm). Seven models, covering a total of 525 simulations, were considered using different combinations of the routinely–measured meteorological and five pollutants variables as covariates. Each model included different numbers of hidden layers and neurons per layer in each simulation. Results of simulations were evaluated to achieve the optimum correspondence between the measured and predicted PNCs in each model (namely Models, M_1–M_7). The best prediction ability was provided by M_1 when all the covariate variables were used. The model, M_2, provided the lowest prediction performance since all the meteorological variables were omitted in this model. Models, M_3–M_7, that omitted one pollutant covariate, showed prediction ability similar to M_1. The results were within a factor of 2 from the measured values, and provided adequate solutions to PNCs' prognostic demands. These models are useful, particularly for the studied site where no nanoparticles measurement equipment exist, for determining the levels of particles in various size ranges. The model could be further used for other locations in Kuwait and elsewhere after adequate long–term measurements and training based on the routinely–monitored environmental data.

1. Introduction

Regulatory bodies worldwide have not reached a consensus regarding a legal threshold to control particle number concentrations (PNCs) in the ambient air (Kumar et al., 2014). The air quality standard for particles is based on mass concentrations of particulate matter less than 10 μm (PM_{10}) and 2.5 μm ($PM_{2.5}$). However, these standards do not regulate PNCs due to their negligible mass compared with PM_{10} and $PM_{2.5}$ (Heal et al., 2012). Consequently, most air quality monitors do not have nanoparticles monitors distributed in monitoring networks (Kumar et al., 2011b), which are highly expensive and very perceptive. Therefore, any relavent information about PNCs will be expedient using any consistent predictive model as a function of most commonly monitored pollutants in the ambient air.

The modelling of air pollutants usually fits into two modelling approaches: deterministic (i.e., dispersion models) and stochastic (i.e., statistical models) models that can be used in accurate modelling purposes (Mølgaard et al., 2012; Reggente et al., 2014). Artificial intelligence (AI), which was initially introduced by Robbins and Monro (1951), has wide applications in stochastic prediction models, where no equations are required to describe the physical processes in the model. The most commonly–used AI application in prediction is the artificial neural network (ANN). In these models, a set of training data is used to derive a statistical description (i.e., automatically developed by ANN) of the relation between inputs (covariates) and outputs (targets) that can make predictions of the output data from unseen (i.e., new) input data within the bounds of the training covariates range. This statistical description is considered as a black box, where unknown simultaneous computational process is applied to map the relation between covariates and targets, which is one of the drawbacks of the ANN approach.

The targetted concentrations (i.e., PNCs) collected at a receptor site are mainly from various known and unknown sources making varying contributions. Deterministic models require the knowledge of sources and concentrations of nanoparticles, and the associated transformation and dispersion processes (i.e., Gaussian and Eleurian models) that are yet not fully understood (Kumar et al., 2011a). Therefore, stochastic models are preferred to overcome the limitations of the deterministic models.

Unlike linear multivariate statistical methods (e.g., ordinary least squares method, and partial least squares), ANN is able to model complex non–linear relationships between given parameters, without any assistance from the user, and to easily deal with high–dimensional data (Svozil et al., 1997). ANN showed remarkable perfor-

Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control.

* Corresponding author.
Email address: adabbous@kisr.edu.kw (A.N. Al–Dabbous)

http://dx.doi.org/10.1016/j.apr.2016.11.004
1309-1042/© 2016 Published by Elsevier Ltd.

mance and accuracy in capturing the complex non–linear associations within data, compared to traditional statistical models. For example, Chelani et al. (2002) also showed superior prediction ability of ANN ($R^2 = 0.68$, 0.72 and 0.63 for industrial, commercial and residential sites, respectively) against the multivariate regression models ($R^2 = 0.57$, 0.52 and 0.48 for industrial, commercial and residential sites, respectively) for SO_2 daily concentrations at three sites in Delhi, India. Furthermore, Kukkonen et al. (2003) demonstrated that ANN ($R^2 = 0.71$) outmatched the linear statistical model ($R^2 = 0.47$) and the deterministic modelling system ($R^2 = 0.32$), when predicting NO_2 hourly concentrations at two monitoring stations in central Helsinki, Finland, from 1996 to 1999. Likewise, ANN has been shown to perform better ($R^2 = 0.65$) than multi–linear regression method ($R^2 = 0.60$) for predicting PM_{10} daily concentrations in Athens, Greece (Chaloulakou et al., 2003). In Athens again, ANN displayed better predictions for hourly PM_{10} concentrations than linear regression models at four urban and suburban locations since the R^2 for ANN were in the 0.80–0.89 range compared with 0.29–0.35 for linear regression models (Grivas and Chaloulakou, 2006). Furthermore, Paschalidou et al. (2011) showed better predictions by ANN ($R^2 = 0.65$–0.76) than those given by principal component regression analysis ($R^2 = 0.33$–0.38) for hourly PM_{10} concentrations in four urban locations in Cyprus. In Kocaeli (Turkey), Özdemir and Taner (2014) reported that predictions of PM_{10} hourly concentrations by ANN ($R^2 = 0.87$ and 0.49 for urban and industrial sites, respectively) outperformed multi–linear regression ($R^2 = 0.74$ and 0.36 for urban and industrial sites, respectively), highlighting the more efficient predictions by ANN. Among the aforementioned studies, ANN has shown the highest predictive accuracy in term of their appealing adaptive nature and ability of modelling complex non-linear high-dimensional data, and is thereby considered a better predictive modelling tool.

Many fields have utilised ANN successfully. Some of them include air pollution (Moustris et al., 2010), waste management (Antanasijević et al., 2013), medicine (Lo et al., 2013), ecology (Larsen et al., 2012) and chemistry (Svozil et al., 1997). In terms of air pollution, ANN has predicted successfully the concentrations of PM_{10} (Paschalidou et al., 2011), SO_2 (Moustris et al., 2010), O_3 (Kandya et al., 2013), NO_2 (Nagendra and Khare, 2006), NO_x (Perez and Trier, 2001), CO (Moustris et al., 2010) and H_2S (Baawain and Al-Serihi, 2014), but application of this approach to the PNC predictions remain very limited (Table 1).

The current analysis presented as a part of this work differs from previous PNC studies (Table 1) in the following unique ways. Firstly, the PNC measurements were recorded, at a sampling rate of 10 Hz with a time response ($T_{90–10\%}$) as low as 200 ms, using one of the

Table 1
Summary of ANN studies related to PNC predictions.

Author (year)	Location (type)	Method	Covariates	Targets	Time resolution	Notes
Hussein et al. (2006)	Helsinki, Finland (urban background)	Numerical fitting	wind speed, direction and temperature	$N_{10–100}$ and $N_{100–400}$	Hourly	The observed and predicted PNCs showed R^2 of 0.70 and 0.62 for $N_{10–100}$ and $N_{100–400}$, respectively.
Clifford et al. (2011)	Helsinki, Finland (urban background)	Generalised additive model and generalised linear model	wind speed and direction, temperature, relative humidity, rainfall, and solar insolation	$N_{10–100}$	Hourly	The generalised additive model outperforms the generalised linear model with R^2 value of 0.84 was found between the observed and predicted PNCs.
Mølgaard et al. (2012)	Helsinki, Finland (urban background)	Combination of regression model with an autoregressive model structure within a Bayesian framework	Wind speed, direction, temperature, relative humidity and traffic intensity	$N_{3–100}$ and $N_{100–950}$	3 h	The combined model showed R^2 of 0.67 and 0.57 between the observed and predicted $N_{3–100}$ and $N_{3–100}$, respectively.
Sabaliauskas et al. (2012)	Toronto, Canada (roadside)	Multiple linear regression model	$PM_{2.5}$, SO_2, O_3, NO, NO_2, CO, wind speed, temperature, relative humidity and solar radiation	$N_{8–10}$, $N_{10–20}$, $N_{20–30}$, $N_{30–40}$, $N_{40–50}$, $N_{50–60}$, $N_{60–70}$, $N_{70–100}$, $N_{100–200}$ and $N_{200–300}$	Daily	The comparison between the observed and predicted $N_{8–50}$, $N_{50–100}$ and $N_{100–300}$ showed R^2 of 0.52, 0.63 and 0.82, respectively.
Reggente et al. (2014)	Antwerp, Belgium (roadside)	Gaussian process regression and Bayesian linear model	O_3, NO, NO_2 and CO	$N_{25–300}$	5 and 30 min	NO and NO_2 provided more accurate predictions of PNCs than CO and O_3. Predictions of PNCs, using Gaussian process regression, with NO and NO_2 (up to $R^2 = 0.90$) outmatch the use of CO (up to $R^2 = 0.57$) and O_3 (up to $R^2 = 0.67$) as a covariates. Gaussian process regression showed better predictions than Bayesian linear model.
This study	Fahaheel, Kuwait (roadside)	ANN	PM_{10}, SO_2, O_3, NO_x, CO, wind speed and temperature	$N_{5–30}$, $N_{30–100}$ and $N_{100–300}$	5 min	Highest prediction ability occurred when all covariates were used, leading to $R^2 = 0.58$, 0.72 and 0.62 for $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$, respectively, between the measured and predicted PNCs. Relatively lower prediction ability occurred when meteorological variables were not incorporated as covariates, leading to $R^2 = 0.58$, 0.72 and 0.62 for $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$. Except for $N_{30–100}$ ($R^2 = 0.77$–0.82), a marginal decrease in the prediction ability was noticed for $N_{5–30}$ ($R^2 = 0.37$–0.52) and $N_{100–300}$ ($R^2 = 0.44$–0.50), when only the third quartile of the measured PNCs is used.

fastest available aerosol mobility size spectrometers, i.e., differential mobility spectrometer, DMS500 (Kumar et al., 2010). This has made a high resolution data available for the training and performance evaluation of our ANN models. Secondly, the sampling site is a representation of an urban location of an industrialised country of Arabian peninsula where petroleum and petrochemical products are the main source of revenue (Al-Dabbous and Kumar, 2015). Therefore, the model developed as a part of this work has a broad applicability after adequate training. Thirdly, unlike previous modelling efforts (Table 1), this is the first instance concerning the application of ANN for prediction of PNCs in the middle–east region. The novelty of the present ANN model is to relate simultaneously 3 targets, i.e., 5–30 nm ($N_{5–30}$; nucleation mode), 30–100 nm ($N_{30–100}$; Aitken mode), 100–300 nm ($N_{100–300}$; accumulation mode), to 5–7 (i.e., meteorological and pollutant) covariates in the best possible manner at a time resolution of 5 min, utilizing 525 simulations for evaluation. Lastly, other than the previous standard statistical modelling work of Reggente et al. (2014) and Sabaliauskas et al. (2012), most of studies have collected their data from an urban background locations (Table 1). Moreover, the work of Reggente et al. (2014) used a lower cut–off diameter of 25 nm; these are nucleation mode volatile particles and are cause of many uncertainties in nanoparticle models (Kumar et al., 2011a). Likewise, Sabaliauskas et al. (2012) considered a daily temporal resolution in their data analysis from urban locations. In this work, we use a lower cut–off diameter of 5 nm and an averaging time of 5 min, allowing to predict the nucleation mode particles that could contribute up to 77% of the total PNCs (Al-Dabbous and Kumar, 2014b; Kumar et al., 2009) and capture variability brought by nucleation mode particles to the total ambient PNCs, respectively.

In order to fill the above–noted research gaps, a supervised machine learning technique, namely multi–layer ANN, was applied to predict PNCs in three size ranges, i.e., $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$, using different combinations of seven routinely–measured meteorological (wind speed and temperature) and pollutant (PM_{10}, SO_2, O_3, NO_x and CO) variables as covariates.

2. Materials and methods

2.1. Multi–layer ANN

A multi–layer feed–forward ANN, trained with a supervised back–propagation training algorithm, is developed for the prediction of particles in three different sizes (Section 2.2). The general architecture of the network consists of input layer, hidden layers and output layer, as shown in Fig. 1. A single hidden layer is generally used in ANN prediction purposes (Hornik et al., 1989), however this practice is debated as more complex problems sometimes required more than one hidden layer (Chaloulakou et al., 2003). Therefore, networks with single, two and three hidden layers were assessed to choose optimum number of hidden layers than could yield acceptable prediction results. In these hidden layers, different numbers of hidden neurons were evaluated (described in Section 2.3). These layers are interconnected through a system of neurons by weights and output signals, which are originated from the neurons in input layer and fed forward towards the neurons in the following layer. The number of hidden neurons in the hidden layers was selected based on the commonly–used iterative approach (i.e., trial and error approach), and conditioned by $2^m \geq K$ rule in the case of a single hidden layer, where m and K are the number of hidden neurons and output variables, respectively (Chelani et al., 2002). On the other hand, the number of neurons in the input and output layers represent the covariate and the target variables, respectively. These neurons were interconnected in a feed–forward manner, where each neuron is linked to all neurons in the next immediate layer, and interacted by weighted connection. A Sigmoidal transfer function was used in the hidden layer, whereas a linear transfer function was used in the output layer as explained by Zhang et al. (1998). The best combination of the number of hidden layers and neurons that provided the minimum error was selected.

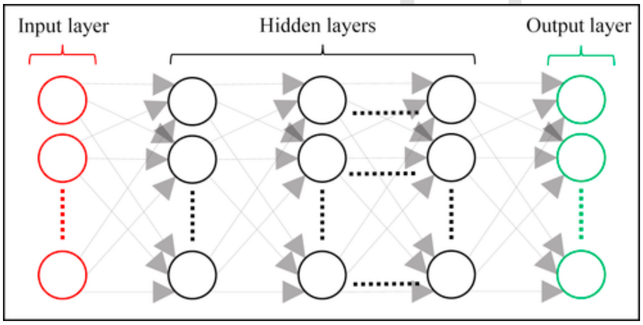

Fig. 1. General multi–layer feed–forward ANN architecture.

2.2. Data collection

High–resolution data of PNCs, ranging from 5 to 1000 nm, were collected at a near–road (15 m from the kerbside of Fahaheel highway) location in Fahaheel, Kuwait, during summer (27 May to 26 June 2013) by DMS500. However, only PNCs up to 300 nm were utilized in this study, as they represent the vast majority of total measured PNCs, i.e., >99% of PNCs (Al-Dabbous and Kumar, 2014a, b). Further measurements of PM_{10}, SO_2, O_3, NO_x, CO, wind speed and temperature were obtained from KEPA fixed monitoring station (Al-Dabbous et al., 2013), at a time resolution of 5 min. Wind speed and temperature were regarded as two of the major meteorological parameters that control airborne PNC dispersion (Al-Dabbous and Kumar, 2014b). PNC data, collected at 0.1 s time resolution, were averaged to 5 min to synchronize with the KEPA data. The PNC, pollutant and meteorological data sets were previously used and discussed in detail in Al-Dabbous and Kumar (2015, 2014b). The entire table consists of 8675 rows, representing the valid 5–minute observations over the 31–day of measurements, and 10 columns, representing all covariates and targets. Seven models were proposed to predict PNCs in the form of three different size ranges ($N_{5–30}$, $N_{30–100}$ and $N_{100–300}$) using different combinations of PM_{10}, SO_2, O_3, NO_x, CO, wind speed, and temperature as covariates (see Table 2). ANN model performance is controlled by a number of covariates, therefore, the maximum number of covariates (i.e., seven variables) is selected in

Table 2
Inputs and outputs variables proposed for each model.

Model ID	Model inputs	Model outputs
Model–1(Model, M_1)	PM_{10}, SO_2, O_3, NO_x, CO, wind speed and temperature	$N_{5–30}$, $N_{30–100}$ and $N_{100–300}$
Model–2 (M_2)	PM_{10}, SO_2, O_3, NO_x, CO	
Model–3 (M_3)	PM_{10}, SO_2, O_3, NO_x, wind speed and temperature	
Model–4 (M_4)	PM_{10}, SO_2, O_3, CO, wind speed and temperature	
Model–5 (M_5)	PM_{10}, SO_2, NO_x, CO, wind speed and temperature	
Model–6 (M_6)	SO_2, O_3, NO_x, CO, wind speed and temperature	
Model–7 (M_7)	PM_{10}, O_3, NO_x, CO, wind speed and temperature	

Model, M_1, which is considered as a reference combination for other models. In the other six models (Models, M_2–M_6), sensitivity of prediction on covariates has been examined.

In the first combination (model–1), all covariates related to pollutants and meteorological variables were included. In the second combination (model–2), only pollutants were utilised as covariates, and meteorological variables were excluded, in order to assess the influence of the meteorological variable in the prediction ability. Models 3–7 investigate the sensitivity of the prediction by removing one pollutant at a time from the covariates, while retaining the meteorological variables as covariates of the models. A total of 525 simulations (75 simulations for each of the seven models; described in Table 2) were formulated using different combinations of number of hidden layers and neurons. The 75 simulations were developed for single (25 simulations), two (25 simulations) and three (25 simulations) hidden layers, with the 25 simulations consisting of 2–50 hidden neurons (i.e., 2–20 neurons per layer with incremental factor of 1 neuron in each model, and 25–50 neurons per layer with incremental factor of 5 neuron in each simulation). The investigation of these different simulations contributes in obtaining an optimised model that provides the best prediction ability of PNCs in three size ranges ($N_{5–30}$, $N_{30–100}$ and $N_{100–300}$) for each model.

2.3. Description of ANN model

The ANN was created, trained and simulated with MATLAB (version: 8.3.0.532), using Neural Network Toolbox. The original dataset was divided into modelling and testing datasets, with ratios of 0.80 and 0.20 that were used to design the network and to test the network, respectively. As a first step of designing the network using the modelling dataset (80% of the original dataset), input and output variables were selected, and normalised into the range of −1 to 1 to improve computational performance. Thereafter, a multi–layer feed–forward back–propagation network was created using "newff" command, with hyperbolic tangent sigmoid transfer function "tansig" in the hidden layer. The transfer function used net input values to generate layers output values. Chaloulakou et al. (2003) found that the "tansig" transfer function lead to more accurate results than the Log–sigmoid transfer function "logsig". These transfer functions were defined as $(1+e^{-n})^{-1}$ for tansig and $2\left(1+e^{-2n}\right)^{-1}-1$ for logsig. The modelling dataset was further divided by default into three subsets: training (56% of the original dataset), validation (12% of the original dataset), and initial testing (12% of the original dataset) using "dividerand" function. Before training the network, weights and biases were set to initial values used by the feed–forward network using "init" command. Afterward, the created network was trained by Levenberg–Marquardt optimisation using the "trainlm" function, which was the fastest available back–propagation training function (Chaloulakou et al., 2003), with the mean square error performance function. The computational limitation of the "trainlm" function (i.e., high memory usage) was compensated by the use of fast processor and large memory workstation. In the network training process, the number of neurons varied from 2 to 20 neurons per layer with an incremental factor of 1 neuron in each simulation, and from 25 to 50 per layer with an incremental factor of 5 neurons in each simulation. In each respective simulation, the training process continued until the validation performance began to rise, whereupon the model was early–stopped. Thereafter, the trained and validated network was simulated in order to initially test the network using "sim" function by exposing the network to an unseen datum (i.e., testing subset; 12% of original dataset). The trained, validated and initially–tested network performance was checked by the MATLAB software using R^2 value. Modelling performance for the 525 simulations showed R^2 values of between 0.35 and 0.81, and the best–performing simulations in each studied model showed R^2 values of between 0.74 and 0.81 (Supplementary Information, SI Tables S1). The network was then exposed to another unseen datum (i.e., testing dataset; 20% of the original dataset) in order to generalise the model prediction ability. The output of the model was then compared with the actual values, and that was expressed by the following standard model evaluation statistics: R^2, root mean square error (*RMSE*), normalised *RMSE* (*NRMSE*) and index of agreement (*d*; sometimes referred to as IA) (Willmott, 1982), which were calculated using Microsoft Excel. These statistical performance indicators were proposed by Willmott (1982) and widely used in literature to assess the accuracy of the model predictions and to cross–compare these models similar to as reported by McKendry (2002) and Nagendra and Khare (2006). These criteria were defined as follows:

$$R^2 = \left(\frac{\sum_{i=1}^{n}\left(O_i - O_{avg}\right)\left(P_i - P_{avg}\right)}{\sqrt{\sum_{i=1}^{n}\left(O_i - O_{avg}\right)^2 \sum_{i=1}^{n}\left(P_i - P_{avg}\right)^2}} \right)^2$$

$$RMSE = \left(\frac{1}{n}\sum_{i=1}^{n}\left(P_i - O_i\right)^2 \right)^{\frac{1}{2}}$$

$$NRMSE = \frac{RMSE}{O_{max} - O_{min}}$$

$$d = 1 - \frac{\sum_{i=1}^{n}\left(P_i - O_i\right)^2}{\sum_{i=1}^{n}\left(\left|P_i - O_{avg}\right| + \left|O_i - O_{avg}\right| \right)^2}$$

where n is the number of samples considered; O_i and P_i are the measured and the predicted PNC values, respectively. O_{avg} and P_{avg} are the mean measured and mean predicted PNC values, respectively. O_{max} and O_{min} are the maximum and minimum measured PNC values, respectively. R^2 and d indicators are dimensionless descriptive statistical parameters ranging from 0 to 1 (perfect score = 1). *RMSE* and *NRMSE* values are in # cm^{-3} and %. The smaller the *RMSE* and *NRMSE* indicates a better model performance.

In what follows, these optimised models are simply referred as M_1–M_7, representing the best–performing simulation of models 1–7, and were selected for discussion. In each model, only the best simulation based on performance was selected for discussion.

3. Results and discussion

3.1. The architecture of the best–performing models

Following the iterative approach described in Section 2.1, the ANN architectures that yielded the best prediction ability in each of the proposed models are summarised in Table 3. Generally, the best–performing ANN simulations involved the use of three hidden layers, except for M_7, where two hidden layers showed the best performance (SI Tables S1). As a result of 75 simulations in each of the seven ANN models conducted to select the adequate number of neurons, a fully connected feed–forward ANN model with 45 or more

Table 3
Best architectures for ANN models, including the performance statistics of the independent testing data. R^2 and d indicators are dimensionless descriptive statistical parameters ranging from 0 to 1 (perfect score = 1). The smaller the normalised root mean square error (*NRMSE*) indicates a better model performance. *NRMSE* values are in %.

Model ID	Number of layers: Number of neurons[a]	Internal testing, R^2	$N_{5–30}$ ($O_{avg} = 7.75 \pm 5.36 \times 10^4$ cm^{-3})				$N_{30–100}$ ($O_{avg} = 1.33 \pm 1.25 \times 10^4$ cm^{-3})				$N_{100–300}$ ($O_{avg} = 6.09 \pm 2.13 \times 10^3$ cm^{-3})			
			P_{avg}(× 10^4# cm^{-3})	R^2	*NRMSE* (%)	d	P_{avg}(× 10^4# cm^{-3})	R^2	*NRMSE* (%)	d	P_{avg}(× 10^3# cm^{-3})	R^2	*NRMSE* (%)	d
M_1	3:45	0.81	7.62 ± 4.54	0.64^b	5.64	0.89	1.29 ± 1.16	0.79^b	3.70	0.94	6.00 ± 1.88	0.71^b	6.89	0.91
M_2	3:25	0.74	7.71 ± 4.33	0.58^b	6.11	0.86	1.34 ± 1.16	0.72^b	4.19	0.92	6.05 ± 1.78	0.62^b	7.90	0.88
M_3	3:45	0.81	7.90 ± 4.58	0.64^b	5.59	0.89	1.37 ± 1.12	0.76^b	4.00	0.93	6.15 ± 1.98	0.72^b	6.86	0.92
M_4	3:50	0.77	7.78 ± 4.55	0.63^b	5.75	0.89	1.40 ± 1.15	0.79^b	3.80	0.94	6.09 ± 1.94	0.71^b	6.96	0.91
M_5	3:45	0.79	7.72 ± 4.59	0.64^b	5.71	0.89	1.35 ± 1.16	0.77^b	3.95	0.93	6.10 ± 1.97	0.69^b	7.23	0.91
M_6	3:45	0.77	7.69 ± 4.54	0.66^b	5.54	0.89	1.34 ± 1.12	0.76^b	4.06	0.93	6.03 ± 1.83	0.63^b	7.80	0.88
M_7	2:50	0.79	7.53 ± 4.48	0.63^b	5.60	0.89	1.34 ± 1.17	0.75^b	4.07	0.93	5.98 ± 1.89	0.68^b	7.27	0.90

Note: [a] Number of neurons in each layer. [b] statistically significant (i.e., p–value < 0.05).

neurons in each hidden layer showed the best performance, except for M_2 where 25 neurons performed the best (SI Tables S1). The iterative approach proved to be more accurate and efficient than the previously used rule of thumb, described in details elsewhere (Nagendra and Khare, 2006), in optimising the number of hidden neurons in order to reach the best–performing ANN simulation. In what follows, ANN simulations (i.e., M_1–M_7) with the best prediction ability for each model were discussed.

The mean measured PNCs for the independent testing dataset (O_{avg}) was 7.75 ± 5.4 × 10^4 cm^{-3} for $N_{5–30}$, 1.33 ± 1.25 × 10^4 cm^{-3} for $N_{30–100}$ and 6.10 ± 2.14 × 10^3 cm^{-3} for $N_{100–300}$. These mean values were closely approached by the mean predicted PNCs (P_{avg}) in all the proposed ANN models within a range of around ±2.18 × 10^3, ±7.30 × 10^2 and ±1.15 × 10^2 cm^{-3} for $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$, respectively, as tabulated in SI Tables S2. On the whole, promising ANN simulations were developed with R^2 and d values of the training dataset reaching up to 0.79 and 0.94, respectively, and the least *NRMSE* value reaching 3.70 (Table 3), revealing a good agreement between the measured and predicted PNCs.

3.2. Comparative evaluation of the performance of best–performing models

Based on the inter– and intra–comparison among the 525 simulations (75 simulations for each of the seven models), the best–performing prediction model that used all variables as covariates (i.e., Model, M_1; see Table 3) yielded the best correspondence between the measured and predicted PNCs, according to the R^2, d and *NRMSE* performance indicators. The architecture of model M_1 consisted of three hidden layers with 45 neurons in each layer. The performance of this model gave relatively very good results, compared with other ANN models (i.e., M_2), according to R^2 (0.64, 0.79 and 0.71; p–value < 0.05), d (0.89, 0.94 and 0.91) and *NRMSE* (5.60, 3.70 and 6.89%) for $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$, respectively, as seen in Table 3. Thereafter, the sensitivity of the model to the input variables was tested by inter–comparing the best–performing prediction simulations in each model. The inter–comparison between the best–performing simulations (i.e., Models, M_1–M_7) revealed that ANN simulations not including meteorological parameters as covariates resulted in a relatively inadequate prediction ability (i.e., M_2; see Table 3), compared with other ANN simulations (i.e., Models, M_1, M_3–M_7), highlighting the importance of meteorological variables in enhancing the prediction ability. M_2 consisted of three hidden layers with 25 neurons in each layer and displayed R^2 (0.58, 0.72 and 0.62; p–value < 0.05), d (0.86, 0.92 and 0.88) and *NRMSE* (6.11, 4.19 and 7.90%) for $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$, respectively, as presented in Table 3. As for M_3–M_7, performance was nearly as good as in M_1 (see Fig. 2), according to the R^2, d, and *NRMSE* values reported in Table 3, reflecting the importance of only six variables without any specific preference. This would allow to predict $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$ in case of one of the covariates (i.e., pollutants) is missing due to failures of the measuring instrument.

3.3. Prediction of high PNCs

The essential quality of a prediction model is its ability to accurately predict high–pollution concentration, in the current case high PNCs (but within the bounds of the training PNC range). As stated earlier, there was no legal threshold for controlling PNCs in the ambient air; therefore the third quartile of the measured PNCs was considered as the threshold representing high PNCs. More precisely, 9.50 × 10^4, 1.44 × 10^4 and 7.20 × 10^3 cm^{-3}, corresponding to the 75th percentile values, were used as the threshold for $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$, respectively. The inter–comparison of the prediction ability among the best–performing simulations (i.e., M_1–M_7) for the high PNCs (Fig. 3) mimicked the same trend as measured when all PNCs were considered (Fig. 2). However, the results presented in Table 4 showed a marginal decrease in the prediction ability for $N_{5–30}$ (R^2 = 0.37–0.52) and $N_{100–300}$ (R^2 = 0.44–0.50), whereas $N_{30–100}$ maintained good correspondence between the measured and predicted high PNCs (R^2 = 0.77–0.82). Similarly, a corresponding decrease in d was observed for $N_{5–30}$ (d = 0.76–0.80) and $N_{100–300}$ (d = 0.44–0.82), while d remained above 0.90 for $N_{30–100}$. The *NRMSE* of the high PNCs predictions increased for $N_{5–30}$ (*NRMSE* up to 18.29%) and $N_{100–300}$ (*NRMSE* up to 16.93%), and markedly maintained for $N_{30–100}$ (*NRMSE* up to 7.82%). Models $M_{1–7}$ tend to under–predict the high PNCs for $N_{5–30}$ and $N_{100–300}$ (Fig. 3).

3.4. General discussion

In general, the prediction ability was better for $N_{30–100}$ than for $N_{5–30}$ and $N_{100–300}$, according to the aforementioned statistical performance indicators (i.e., R^2, d and *NRMSE*; see Tables 3 and 4). All statistical prediction simulations usually based on the previous history of relationships between covariates and outputs, and the prediction simulations for different size ranges $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$ have significantly unique relationships between covariates and outputs. For instance, particles in the nucleation mode ($N_{5–30}$) are more sensitive to transformation processes due to their volatility and unstable nature (Morawska et al., 2008). This leads to a very short lifetime in the atmosphere, therefore the relationships between the covariates and $N_{5–30}$ are not well established. Similarly, accumulation mode particles

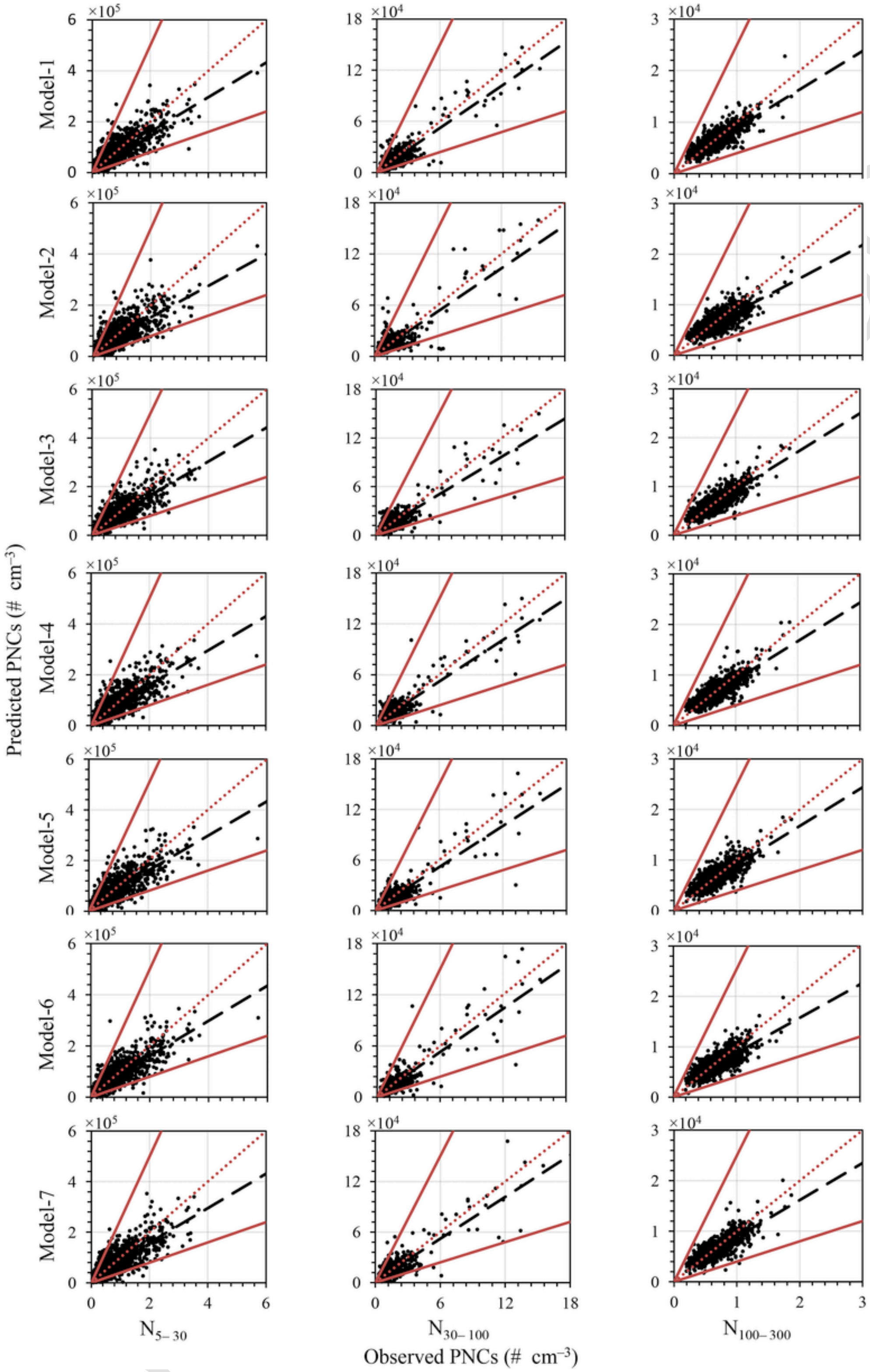

Fig. 2. Scatter plots of the predicted (y–axes) versus the measured (x–axes) PNCs in the three size ranges (5–30, 30–100 and 100–300 nm) for the best–performing simulation in each model. Each scatter plot shows the best–fit lines (dashed black line), the 1:1 lines (dotted red line), and the factor of two of measured PNCs (solid red lines).

($N_{100-300}$) have much longer lifetimes compared to smaller particles, causing them to be transported for larger distances (Laakso et al., 2003). Given that local pollutants and meteorological variables were used as covariates, mapping of the relationships between long–range transported accumulation mode particles and covariates is not well understood, leading to relatively lower prediction ability. The locally–produced Aitken mode particles (N_{30-100}) are less effectively removed by transformation processes (e.g., evaporation and coagula-

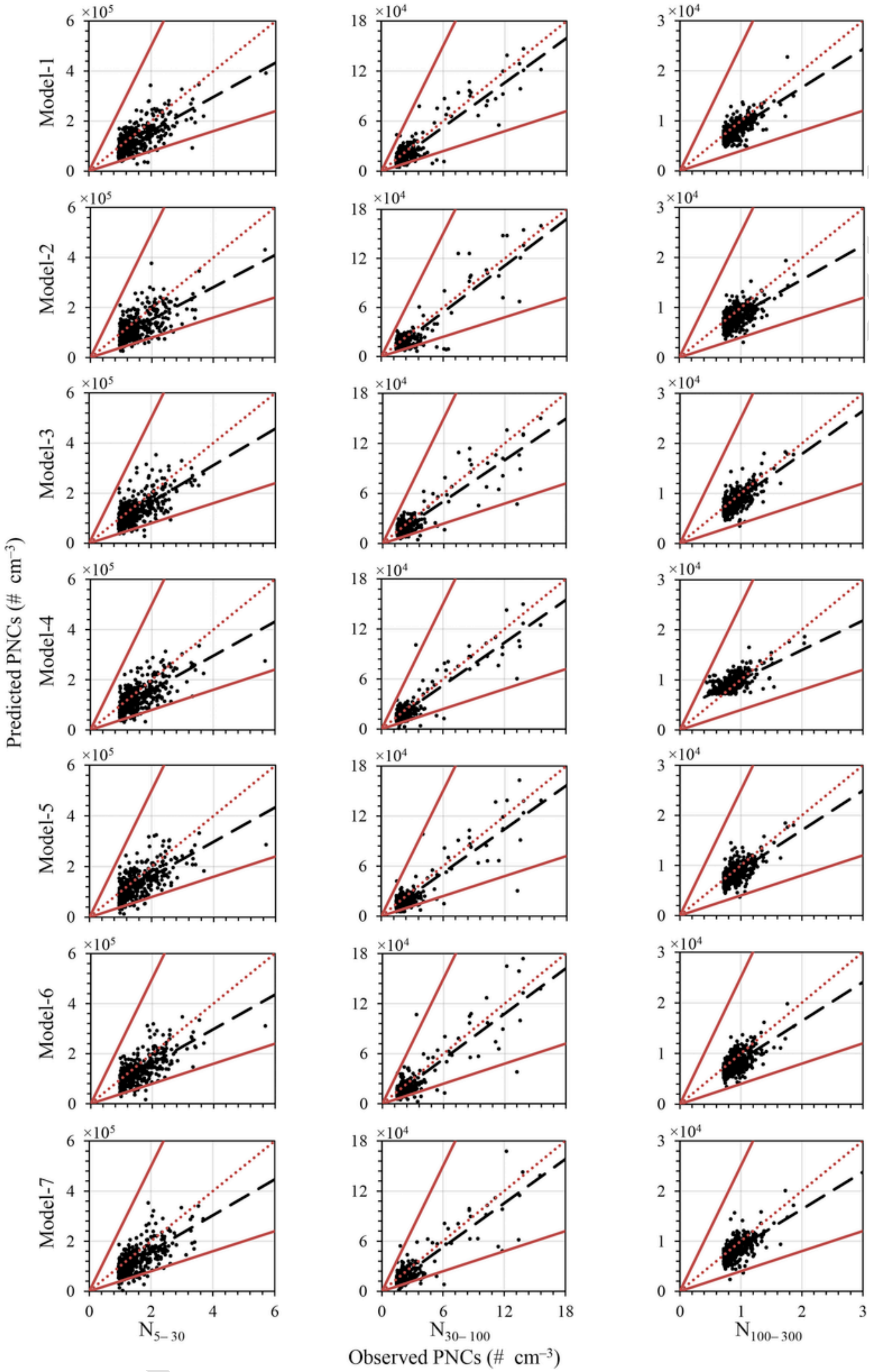

Fig. 3. Scatter plots of the predicted (y–axes) versus the measured (x–axes) PNCs in three size ranges (5–30, 30–100 and 100–300 nm) for the high PNCs (i.e., the third quartile of the measured PNCs) for the best–performing simulation in each model. Each scatter plot shows the best–fit lines (dashed black line), the 1:1 lines (dotted red line), and the factor of two of measured PNCs (solid red lines).

Table 4
Best architectures for ANN models for the high PNCs (i.e., the third quartile of the measured PNCs), including the performance statistics of the independent testing data. The coefficient of determination (R^2) and the index of agreement (d) are dimensionless descriptive statistical parameters ranging from 0 to 1 (perfect score = 1). The smaller the normalised root mean square error (*NRMSE*) indicates a better model performance. *NRMSE* values are in %.

Model ID	Number of layers: Number of neurons[a]	$N_{5–30}$ (O_{avg} = 14.90 ± 5.68 × 10^4 cm^{-3})				$N_{30–100}$ (O_{avg} = 2.44 ± 2.03 × 10^4 cm^{-3})				$N_{100–300}$ (O_{avg} = 9.09 ± 1.74 × 10^3 cm^{-3})			
		P_{avg}(× 10^4# cm^{-3})	R^2	*NRMSE* (%)	d	P_{avg}(× 10^4# cm^{-3})	R^2	*NRMSE* (%)	d	P_{avg}(× 10^3# cm^{-3})	R^2	*NRMSE* (%)	d
M_1	3:45	12.48 ± 5.49	0.50[b]	10.37	0.80	2.11 ± 2.00	0.81[b]	6.67	0.94	8.29 ± 1.92	0.49[b]	14.32	0.79
M_2	3:25	12.10 ± 5.49	0.44[b]	11.37	0.76	1.99 ± 2.08	0.80[b]	7.26	0.93	8.07 ± 1.98	0.37[b]	16.93	0.72
M_3	3:45	12.72 ± 5.61	0.48[b]	17.55	0.80	2.15 ± 1.92	0.80[b]	6.97	0.94	8.52 ± 2.08	0.52[b]	13.66	0.82
M_4	3:50	12.61 ± 5.53	0.48[b]	10.47	0.79	2.22 ± 1.97	0.82[b]	6.54	0.95	8.39 ± 2.10	0.50[b]	6.96	0.80
M_5	3:45	12.60 ± 5.66	0.47[b]	10.68	0.79	2.12 ± 2.05	0.79[b]	7.29	0.94	8.45 ± 2.01	0.47[b]	14.27	0.80
M_6	3:45	12.61 ± 5.51	0.50[b]	10.30	0.80	2.14 ± 2.16	0.77[b]	7.82	0.93	8.08 ± 2.02	0.44[b]	16.25	0.75
M_7	2:50	12.26 ± 5.52	0.48[b]	18.29	0.78	2.13 ± 2.03	0.78[b]	7.35	0.93	8.23 ± 2.00	0.42[b]	15.67	0.75

Note: [a] Number of neurons in each layer. [b] statistically significant (i.e., p–value < 0.05).

tion) from the atmosphere, compared with $N_{5–30}$, allowing the prediction models to better understand their relationships with the covariates. In summary, the deviations between the measured and predicted PNCs were not substantial. This observation is further supported by the fact that predicted PNCs were within a factor of two to the measured PNCs (Figs. 2 and 3), suggesting that the proposed prediction models can provide adequate solutions to PNCs prognostic demands.

4. Conclusions

A preliminary modelling effort was made in order to study the potential of predicting PNCs in the three size ranges ($N_{5–30}$, $N_{30–100}$ and $N_{100–300}$) as function of meteorological parameters and pollutants concentration during summertime in the urban area of Fahaheel, Kuwait, using multi–layer feed–forward ANN, trained by back–propagation according to Levenberg–Marquardt optimisation. Seven models, covering a total of 525 simulations, were investigated using different combinations of input variables (PM_{10}, SO_2, O_3, NO_x, CO, wind speed, and temperature). The input variables selected in each model were as follows: (i) model 1; all variables, (ii) model 2; all variables except wind speed and temperature, and (iii) models 3–7; all variables with the exception of a pollutant at each time. Every model was optimised by altering the number of hidden layers and number of neurons corresponding to each layer in order to achieve the best prediction performance. In each model, 75 simulations were tested, covering 1–3 hidden layers and 2–50 hidden neurons in each layer (i.e., from 2 to 20 neurons per layer with an incremental factor of 1 neuron in each simulation, and from 25 to 50 per layer with an incremental factor of 5 neurons in each simulation).

Out of the total 75 simulations in each of the seven studied models, the best–performing simulation for each model (i.e., M_1–M_7) was selected. These selected simulations provided a satisfactory prediction accuracy, based on R^2 and d values of up to 0.79 and 0.94, respectively, between the measured and the predicted PNCs. M_1, which uses all variables as covariates, provided the best correspondence between the measured and predicted PNCs, according to R^2 (0.64, 0.79 and 0.71; p–value < 0.05), d (0.89, 0.94 and 0.91) and *NRMSE* (5.60, 3.70 and 6.89%) for $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$, respectively. The improvement in prediction performance in M_1 is attributed to the use of more covariates than other M_2–M_7, allowing the ANN to accurately map the non–linear relation between the measured and predicted PNCs. Conversely, M_2, which did not include meteorological parameters (i.e., wind speed or temperature) as covariates, showed the least performance, compared with M_1, M_3–M_7, with R^2 (0.58, 0.72 and 0.62; p–value < 0.05), d (0.86, 0.92 and 0.88) and *NRMSE* (6.11, 4.19 and 7.90%) for $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$, respectively. The low performance of M_2, reflected the strong influence of meteorological variables on PNCs. M_3–M_7 showed almost similar performances to M_1, according to R^2, *d, NRMSE* values, allowing them to predict $N_{5–30}$, $N_{30–100}$ and $N_{100–300}$ in case one of the pollutants (i.e., PM_{10}, SO_2, O_3, NO_x and CO) is missing from input variables. It can be concluded that there is no significance of any pollutant on other and performance is similar considering four covariates instead of total five covariates. In an extended analysis to investigate the prediction ability of the proposed simulations for high PNCs, a marginal decrease in the prediction ability between the measured and predicted PNCs was noticed for $N_{5–30}$ (R^2 = 0.37–0.52) and $N_{100–300}$ (R^2 = 0.44–0.50), whereas a consistant model performance was observed for $N_{30–100}$ (R^2 = 0.77–0.82). In all studied ANN models, the prediction ability for $N_{30–100}$ showed better performance than $N_{5–30}$ and $N_{100–300}$, due to their lower sensitivity to transformation processes compared to $N_{5–30}$ and their local origin compared to the long–range transported $N_{100–300}$. In general, predicted PNCs in all ANN models trials were approximately within a factor of two of the measured PNCs. These simulations can be applied, particularly for limited measured nanoparticles data, for accurate predictions of various size ranges of nanoparticles as function of routinely–monitored criteria pollutants and meteorological data. These proposed simulations are of high importance for assessing the nanoparticles levels of different size ranges for sites where nanoparticles measurement devices are unavailable. The applicability of ANN models is a great advantage for the prediction of nanoparticles in various size ranges for other locations in Kuwait and elsewhere if trained based on long–term measurements.

Acknowledgements

The authors are grateful to the funding support received from the Kuwait Institute for Scientific Research for the experimental work. The authors also thank Kuwait Environment Public Authority for providing primary pollutants data, and Kuwait National Meteorological Network for providing the meteorological data.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.apr.2016.11.004.

References

Al-Dabbous, A., Khan, A., Al-Rashidi, M., Awadi, L., 2013. Carbon dioxide and volatile organic compounds levels in mosque in hot arid climate. Indoor Built Environ. 22, 456–464.

Al-Dabbous, A.N., Kumar, P., 2014a. The influence of roadside vegetation barriers on airborne nanoparticles and pedestrians exposure under varying wind conditions. Atmos. Environ. 90, 113–124.

Al-Dabbous, A.N., Kumar, P., 2014b. Number and size distribution of airborne nanoparticles during summertime in Kuwait: first observations from the Middle East. Environ. Sci. Technol. 48, 13634–13643.

Al-Dabbous, A.N., Kumar, P., 2015. Source apportionment of airborne nanoparticles in a Middle Eastern city using positive matrix factorization. Environ. Sci. Process. Impacts 17, 802–812.

Antanasijević, D., Pocajt, V., Popović, I., Redžić, N., Ristić, M., 2013. The forecasting of municipal waste generation using artificial neural networks and sustainability indicators. Sustain. Sci. 8, 37–46.

Baawain, M.S., Al-Serihi, A.S., 2014. Systematic approach for the prediction of ground-level air pollution (around an industrial port) using an artificial neural network. Aerosol Air Qual. Res. 14, 124–134.

Chaloulakou, A., Grivas, G., Spyrellis, N., 2003. Neural network and multiple regression models for PM10 prediction in Athens: a comparative assessment. J. Air Waste Manag. Assoc. 53, 1183–1190.

Chelani, A.B., Rao, C.C., Phadke, K., Hasan, M., 2002. Prediction of sulphur dioxide concentration using artificial neural networks. Environ. Model. Softw. 17, 159–166.

Clifford, S., Choy, S.L., Hussein, T., Mengersen, K., Morawska, L., 2011. Using the generalised additive model to model the particle number count of ultrafine particles. Atmos. Environ. 45, 5934–5945.

Grivas, G., Chaloulakou, A., 2006. Artificial neural network models for prediction of PM 10 hourly concentrations, in the Greater Area of Athens, Greece. Atmos. Environ. 40, 1216–1229.

Heal, M.R., Kumar, P., Harrison, R.M., 2012. Particles, air quality, policy and health. Chem. Soc. Rev. 41, 6606–6630.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Netw. 2, 359–366.

Hussein, T., Karppinen, A., Kukkonen, J., Härkönen, J., Aalto, P.P., Hämeri, K., Kerminen, V.-M., Kulmala, M., 2006. Meteorological dependence of size-fractionated number concentrations of urban aerosol particles. Atmos. Environ. 40, 1427–1440.

Kandya, A., SN, S.M., Tiwari, V.K., 2013, doi:10.4172/2165-784X.S1-006. Forecasting the tropospheric ozone using artificial neural network modelling approach: a case study of megacity Madras, India. J. Civ. Environ. Eng.

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., 2003. Extensive evaluation of neural network models for the prediction of NO 2 and PM 10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmos. Environ. 37, 4539–4550.

Kumar, P., Fennell, P.S., Hayhurst, A.N., Britter, R.E., 2009. Street versus rooftop level concentrations of fine particles in a Cambridge street canyon. Bound. Layer Meteorol. 131, 3–18.

Kumar, P., Ketzel, M., Vardoulakis, S., Pirjola, L., Britter, R., 2011a. Dynamics and dispersion modelling of nanoparticles from road traffic in the urban atmospheric environment—a review. J. Aerosol Sci. 42, 580–603.

Kumar, P., Morawska, L., Birmili, W., Paasonen, P., Hu, M., Kulmala, M., Harrison, R.M., Norford, L., Britter, R., 2014. Ultrafine particles in cities. Environ. Int. 66, 1–10.

Kumar, P., Robins, A., Vardoulakis, S., Britter, R., 2010. A review of the characteristics of nanoparticles in the urban atmosphere and the prospects for developing regulatory controls. Atmos. Environ. 44, 5035–5052.

Kumar, P., Robins, A., Vardoulakis, S., Quincey, P., 2011b. Technical challenges in tackling regulatory concerns for urban atmospheric nanoparticles. Particuology 9, 566–571.

Laakso, L., Hussein, T., Aarnio, P., Komppula, M., Hiltunen, V., Viisanen, Y., Kulmala, M., 2003. Diurnal and annual characteristics of particle mass and number concentrations in urban, rural and Arctic environments in Finland. Atmos. Environ. 37, 2629–2641.

Larsen, P.E., Field, D., Gilbert, J.A., 2012. Predicting bacterial community assemblages using an artificial neural network approach. Nat. methods 9, 621–625.

Lo, B.W., Macdonald, R.L., Baker, A., Levine, M.A., 2013. Clinical outcome prediction in aneurysmal subarachnoid hemorrhage using Bayesian neural networks with fuzzy logic inferences. Comput. Math. Methods Med. 2013.

McKendry, I.G., 2002. Evaluation of artificial neural networks for fine particulate pollution (PM10 and PM2. 5) forecasting. J. Air Waste Manag. Assoc. 52, 1096–1101.

Mølgaard, B., Hussein, T., Corander, J., Hämeri, K., 2012. Forecasting size-fractionated particle number concentrations in the urban atmosphere. Atmos. Environ. 46, 155–163.

Morawska, L., Ristovski, Z., Jayaratne, E., Keogh, D.U., Ling, X., 2008. Ambient nano and ultrafine particles from motor vehicle emissions: characteristics, ambient processing and implications on human exposure. Atmos. Environ. 42, 8113–8138.

Moustris, K.P., Ziomas, I.C., Paliatsos, A.G., 2010. 3-Day-ahead forecasting of regional pollution index for the pollutants NO_2, CO, SO_2, and O_3 using artificial neural networks in Athens, Greece. Water Air Soil Pollut. 209, 29–43.

Nagendra, S.S., Khare, M., 2006. Artificial neural network approach for modelling nitrogen dioxide dispersion from vehicular exhaust emissions. Ecol. Model. 190, 99–115.

Özdemir, U., Taner, S., 2014. Impacts of meteorological factors on PM10: artificial neural networks (ANN) and multiple linear regression (MLR) approaches. Environ. Forensics 15, 329–336.

Paschalidou, A.K., Karakitsios, S., Kleanthous, S., Kassomenos, P.A., 2011. Forecasting hourly PM10 concentration in Cyprus through artificial neural networks and multiple regression models: implications to local environmental management. Environ. Sci. Pollut. Res. 18, 316–327.

Perez, P., Trier, A., 2001. Prediction of NO and NO 2 concentrations near a street with heavy traffic in Santiago, Chile. Atmos. Environ. 35, 1783–1789.

Reggente, M., Peters, J., Theunis, J., Van Poppel, M., Rademaker, M., Kumar, P., De Baets, B., 2014. Prediction of ultrafine particle number concentrations in urban environments by means of Gaussian process regression based on measurements of oxides of nitrogen. Environ. Model. Softw. 61, 135–150.

Robbins, H., Monro, S., 1951. A stochastic approximation method. Ann. Math. Stat. 400–407.

Sabaliauskas, K., Jeong, C.-H., Yao, X., Jun, Y.-S., Jadidian, P., Evans, G.J., 2012. Five-year roadside measurements of ultrafine particles in a major Canadian city. Atmos. Environ. 49, 245–256.

Svozil, D., Kvasnicka, V., Pospichal, J., 1997. Introduction to multi-layer feed-forward neural networks. Chemom. Intell. Lab. Syst. 39, 43–62.

Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorological Soc. 63, 1309–1313.

Zhang, G., Patuwo, B.E., Hu, M.Y., 1998. Forecasting with artificial neural networks:: the state of the art. Int. J. Forecast. 14, 35–62.