Available online at www.sciencedirect.com
SCIENCE DIRECT®
Atmospheric Environment 39 (2005) 6524–6536
ATMOSPHERIC ENVIRONMENT
www.elsevier.com/locate/atmosenv

Evaluation of an integrated modelling system containing a multi-layer perceptron model and the numerical weather prediction model HIRLAM for the forecasting of urban airborne pollutant concentrations

Harri Niska[a,*], Minna Rantamäki[b], Teri Hiltunen[a], Ari Karppinen[b], Jaakko Kukkonen[b], Juhani Ruuskanen[a], Mikko Kolehmainen[a]

[a]*Department of Environmental Sciences, University of Kuopio, PO Box 1627, FIN-70211 Kuopio, Finland*
[b]*Finnish Meteorological Institute, Sahaajankatu 20 E, FIN-00880 Helsinki, Finland*

Received 28 February 2005; received in revised form 16 June 2005; accepted 13 July 2005

Abstract

In this paper, a multi-layer perceptron (MLP) model and the Finnish variant of the numerical weather prediction model HIRLAM (High Resolution Limited Area Model) were integrated and evaluated for the forecasting in time of urban pollutant concentrations. The forecasts of the combination of the MLP and HIRLAM models are compared with the corresponding forecasts of the MLP models that utilise meteorologically pre-processed input data. A novel input selection method based on the use of a multi-objective genetic algorithm (MOGA) is applied in conjunction with the sensitivity analysis to reduce the excessively large number of potential meteorological input variables; its use improves the performance of the MLP model. The computed air quality forecasts contain the sequential hourly time series of the concentrations of nitrogen dioxide (NO_2) and fine particulate matter ($PM_{2.5}$) from May 2000 to April 2003; the corresponding concentrations have also been measured at two urban air quality stations in Helsinki. The results obtained with the MLP models that use HIRLAM forecasts show fairly good overall agreement for both pollutants. The model performance is substantially better, when the HIRLAM forecasts are used, compared with those obtained both using either HIRLAM analysis data or meteorological pre-processor, for both pollutants. The performance of the currently widely used statistical forecasting methods (such as those based on neural networks) could therefore be significantly improved by using the forecasts of NWP models, instead of the conventionally utilised directly measured or meteorological pre-processed input data. However, the performance of all operational models considered is relatively worse in the course of air pollution episodes.

Keywords: Multi-layer perceptron; Numerical weather prediction; Genetic algorithms; Air quality forecasting; Model input selection

1. Introduction

Urban air pollution and, in particular, peak pollution episodes comprise an acute environmental problem in many densely populated and industrialised areas,

*Corresponding author. Fax: +358 17 163222.
E-mail address: harri.niska@uku.fi (H. Niska).

1352-2310/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2005.07.035

causing adverse health effects and even an increased mortality among susceptible population subgroups. Available information concerning European peak pollution episodes in 13 countries has been reviewed by Kukkonen (2001). The causes of air pollution episodes are complex and depend on various factors including emissions, meteorological parameters, topography, atmospheric chemical processes and solar radiation. The relative importance of such factors is dependent on the geographical region, the surrounding emission source areas and the related climatic characteristics, as well as the season of the year (e.g., Piringer and Kukkonen, 2002).

As episodic situations are typically caused by unfavourable meteorological conditions, the statistical forecasting of peak pollution levels has mostly been based on the regression analysis between airborne concentrations and meteorological parameters (e.g., Ziomas et al., 1995). In operational deterministic air pollution forecasting systems, the meteorological parameters can be derived from numerical weather prediction (NWP) models, and these are subsequently utilised as input for deterministic dispersion models. However, the NWP models have originally been designed for meteorological predictions in a synoptic and meso-scale; instead of local-scale predictions within the lowest atmospheric layers (e.g., Baklanov et al., 2002). The currently available NWP models have only moderate capabilities to predict the meteorological conditions characteristic of peak pollution episodes (Pielke and Uliasz, 1998; Pohjola et al., 2004).

Statistical nowcasting methods are currently used for air pollution forecasting by the local authorities worldwide. Artificial neural networks (NNs) and, in particular multi-layer perceptrons (MLP), have been shown to be useful and fairly accurate tools for air quality forecasting, as these can accommodate complex non-linear relationships between emissions, air quality, meteorological parameters and other factors (Comrie, 1997; Gardner and Dorling, 1999; Kolehmainen et al., 2001). An extensive evaluation and inter-comparison of five NN methods, a linear statistical model and a deterministic modelling system for the prediction of urban NO_2 and PM_{10} concentrations was presented by Kukkonen et al. (2003). However, the NN models established have been evaluated mostly using measured or pre-processed meteorological data, and only few studies have aimed at evaluating the NN models using actual forecasted meteorological data extracted from NWP models (Termonia and Quinet, 2004; Hooyberghs et al., 2005).

This paper focuses on the air quality forecasting using MLP models combined with one NWP model, and on the model evaluation for the measured NO_2 and $PM_{2.5}$ concentrations in an urban area. The main objective of this study was to compare the numerical performance of the forecasting model and the corresponding so-called nowcasting model; the latter uses only measured or pre-processed meteorological data as input. Additionally, the paper aims at presenting and applying novel method based on a multi-objective genetic algorithm for selecting optimal inputs of MLP model that is necessary for enhancing the accuracy of modelling.

2. Materials and methods

2.1. Experimental data

2.1.1. Concentration and meteorological data

The concentration and meteorological data used in this study were measured and gathered in Helsinki, Finland, during the period from 1 May 2000 to 30 April 2003. The concentration data comprised hourly ambient airborne pollutant concentrations for the following species: NO_x, NO_2, CO, O_3, PM_{10} and $PM_{2.5}$, monitored at two urban air quality stations (Vallila and Kallio) in central Helsinki and processed according to the QA/QC procedures of the Helsinki Metropolitan Area Council (YTV). The stations selected represent urban traffic (Vallila) and urban background (Kallio) environments; they were selected for the forecasting of NO_2 and $PM_{2.5}$, respectively.

The meteorological parameters were evaluated and pre-processed by the meteorological pre-processing model MPP-FMI (Karppinen et al., 1997, 2000). The pre-processed meteorological data computed for the location of central Helsinki was selected to be used in this study, as it is the best representative for the whole of the urban area (compared with, e.g. utilising data measured solely at the airport), and contains also relevant derived meteorological parameters, such as, e.g. the Monin–Obukhov length and the mixing height. The pre-processed meteorological data (MPP-FMI) is based on a combination of the data from the synoptic stations at Helsinki-Vantaa airport (about 15 km north of central Helsinki) and Helsinki-Isosaari (an island about 20 km south of central Helsinki). The mixing height of the atmospheric boundary layer was evaluated using the meteorological pre-processor, based on the sounding observations at Jokioinen (90 km northwest) and the routine meteorological observations (Karppinen et al., 1997). This pre-processor has also been adapted to better allow for urban meteorological conditions (Karppinen et al., 2000).

The quality of the data sets was examined, especially, in the context of missing values. As a result, fairly small fractions, ranging from 1% to 6%, of missing concentration data were detected. An imputed concentration data set was created, in which the missing values were replaced, in order to obtain a harmonised database that is well suited for the inter-comparison of models. The incomplete data was imputed using the method based on

the combination of self-organising map (SOM) and linear interpolation; such methods have been recently shown to be well applicable to air quality data sets (Junninen et al., 2004).

2.1.2. Numerical weather prediction data

The NWP data was derived from the Finnish variant of the HIRLAM model which is a limited-area weather forecasting model developed in cooperation between European national weather services. In this study, the HIRLAM version 4.6.2 (Eerola, 2001, 2002) was utilised. The operational set-up of this HIRLAM version contains two computation sites: the ATA which covers Europe and Northern Atlantic with the horizontal resolution of 44 km and the ENO which covers Northern Europe with the horizontal resolution of 22 km. These models are run in parallel so that the ATA creates boundaries and makes analysis for the ENO. Both the models have 190×140 horizontal grid points in transformed latitude–longitude grid, and 31 vertical levels reaching the altitude of approximately 30 km.

The grid point (the longitude of 24.81°E and latitude of 60.33°N) of the ENO that was nearest to the selected air quality monitoring stations was selected. For this point, all the forecasts made within 6 hourly intervals (00, 06, 12, 18 UTC) were employed. We utilised the data from the lowest model levels of 26–31 where the level 31 is closest to ground. The following meteorological variables were included: wind u- and v-components ($m\,s^{-1}$), temperature (K), kinetic energy of turbulence ($J\,kg^{-1}$), specific humidity ($kg\,kg^{-1}$), specific cloud condensate ($kg\,kg^{-1}$) and total cloud cover (fraction). Additionally, the following parameters at the ground surface (based on extrapolation analysis of HIRLAM data) were included: pressure at 2 m (Pa), temperature at 2 m (K) and accumulated fluxes including heat flux ($J\,m^{-2}$) and radiation ($J\,m^{-2}$).

2.2. Multi-layer perceptron models

The MLP is a system composed of simple interconnected neurons producing signals as a function of the sum of the neuron inputs modified by a non-linear transfer function; for a more profound introduction see e.g. Haykin (1999). The MLP provides a non-linear tool for tackling regression problems by approximating any smooth, differentiable function; this is particularly applicable when modelling complex processes. In the applications of atmospheric sciences, the MLP has been utilised for varying tasks, such as prediction, function approximation, or pattern classification (see Gardner and Dorling, 1999).

Recent investigations related to the prediction of hourly time series of air quality using MLP models (e.g. Kolehmainen et al., 2001; Kukkonen et al., 2003) have addressed solely the nowcasting, i.e. the air quality forecasting using numerical weather forecasting models has been outside the scope. In this study, we address the actual forecasting in time. Basically, the modelling system is based on a generic MLP model in which the objective is to model an unknown relationship between hourly airborne concentrations and meteorological parameters.

2.2.1. Integration of the NWP data and MLP models

The use of actual weather forecast data was avoided in training, as this can lead to learning non-physical relationships between concentrations and meteorological parameters. This is due to the uncertainties associated with the meteorological forecasts of NWP models. Therefore, the analysis of HIRLAM was found to comprise more firm basis for training. The analysis of HIRLAM specifies the initial conditions of the state variables at the beginning of a numerical simulation $T = 00$ h (denoted also as $T+00$). It includes only the mathematical construction (data assimilation) of the required meteorological three-dimensional fields, but no forecasting of the future states of the atmosphere by the model. The analysis state therefore theoretically includes the smallest possible computational inaccuracy. The actual forecast of HIRLAM was utilised for the evaluation of the modelling system.

The temporal resolution of the HIRLAM forecasts were limited to 6-hourly intervals, which was due to present operational practice where four forecasts are daily run. Data interpolation was therefore applied in order to achieve the same temporal resolution (1 h) in all the data sets. This was seen to be particularly necessary to ensure the consistent inter-comparison practice. The nearest-neighbour interpolation combined with the sliding mean method within a period of 4 h was applied for simplicity.

2.2.2. The set-up and training of the integrated modelling system

A forecasting period of 24 h was selected for practical regulatory reasons; shorter time forecasts are of minimal value for air quality management purposes. In this study, four modelling systems based on the MLP model and various meteorological input data sets were constructed for the daily forecasting of urban air quality; $T+24$ h. The input variables are presented in Table 1 for each model considered. Two of the models are based on the combination of the MLP model and the HIRLAM NWP outputs, using either the analysis ($T = 00$ h) of HIRLAM or the 24 hourly forecasts ($T+24$ h) of HIRLAM; these models are denoted here as MLP + NWP00 and MLP + NWP24, respectively.

The other two models are based on the MLP models that utilise solely pre-processed meteorological data produced by the MPP-FMI model; these models are

Table 1
The input variables.and their time lags for the MLP models considered

Input variables	Units	Time lags of input variables			
		MLP + MPP00	MLP + MPP24	MLP + NWP00	MLP + NWP24
Temporal data					
Sine and cosine of hour	—	$T+24$	$T+24$	$T+24$	$T+24$
Sine and cosine of year day	—	$T+24$	$T+24$	$T+24$	$T+24$
Sine and cosine of week day	—	$T+24$	$T+24$	$T+24$	$T+24$
Concentration data					
NO_x, NO_2 and O_3	$\mu g\,m^{-3}$	T	T	T	T
PM_{10} and $PM_{2.5}$	$\mu g\,m^{-3}$	T	T	T	T
Pre-processed meteorological data					
Pressure and temperature	Pa, K	T	$T+24$	—	—
Humidity	%	T	$T+24$	—	—
State of ground and albedo	—	T	$T+24$	—	—
Cloudiness	(0–8)/8	T	$T+24$	—	—
Dewpoint, wetbulb and temperature scale	K	T	$T+24$	—	—
Rain	mm	T	$T+24$	—	—
Height of low clouds	M	T	$T+24$	—	—
Sine and cosine of direction of flow	—	T	$T+24$	—	—
Wind speed	m/s	T	$T+24$	—	—
Sunshine duration and solar elevation	h, rad	T	$T+24$	—	—
Solar and net radiations	$W\,m^{-2}$	T	$T+24$	—	—
Moisture parameter	—	T	$T+24$	—	—
Monin–Obukhov length	m	T	$T+24$	—	—
Friction and convective velocities	$m\,s^{-1}$	T	$T+24$	—	—
Turbulence and latent heat flux	$W\,m^{-2}$	T	$T+24$	—	—
Mixing height	m	T	$T+24$	—	—
Gradient of potential temperature	$K\,m^{-1}$	T	$T+24$	—	—
Numerical weather prediction data (model surface levels 26–31)					
U- and V-components of wind	$m\,s^{-1}$	—	—	T	$T+24$
Kinetic energy of turbulence	$J\,kg^{-1}$	—	—	T	$T+24$
Temperature	K	—	—	T	$T+24$
Specific humidity and cloud condensate	$kg\,kg^{-1}$	—	—	T	$T+24$
Total cloud cover	%	—	—	T	$T+24$
Pressure and temperature at 2 m	Pa, K	—	—	T	$T+24$

The input time $T+24$ is the time for which the forecast applies ($+24$ h).
The variables of the HIRLAM model are utilised from the model surface levels of 26–31 where the level 31 is closest to ground.

denoted MLP + MPP00 and MLP + MPP24, respectively. The meteorological data were extracted at the forecasting time ($T+00$ h) and at the daily forecast time ($T+24$ h), for the models MLP + MPP00 and MLP + MPP24, respectively. These latter two models were applied in previous evaluation studies by Kolehmainen et al. (2001) and Kukkonen et al. (2003). These models are useful as numerical references for inter-comparing various modelling options. Clearly, the model MLP + MPP24 is not feasible for any actual forecasts, as it utilises meteorologically pre-processed data at the daily forecast time (i.e. it requires as input meteorological data that will be available only in the future). The predictions of this particular model should therefore only be considered as a numerical reference for inter-comparing various models.

All the MLP models were back-propagation (BP) trained using the learning algorithm of Levenberg–Marquardt and 1000 training epochs. For controlling the over-fitting, we kept to the early stopping method of training; regularisation techniques were not employed here. The training was stopped, when the error of validation set (20% of the training data) increased for 10 iterations and the weights and biases at the minimum of the validation error were utilised. The architectural issues of the MLP models were largely based on the previous study by Niska et al. (2004), in which one hidden layer with 20 hidden nodes, sigmoid transfer

functions for hidden units and linear transfer function for output were found to be sufficient.

2.3. Determination of model-specific inputs

In this study, the selection of optimal model inputs was particularly necessary due to the large number of possible meteorological input variables (ranging from 50 to 100); part of these may have a negligible effect or be totally irrelevant. The large number of meteorological variables was due to the utilisation of parameters on several vertical levels of HIRLAM. The input selection was employed here by a novel way, namely by posing it as the multi-objective optimisation problem where input subset dimensionality was minimised and model performance maximised. Basically, the input selection was performed through a wrapper approach (Kohavi and John, 1997) where input subsets are evaluated using the learning algorithm (MLP model) itself.

2.3.1. Implementation of multi-objective genetic algorithm

We used an evolutionary multi-objective optimisation strategy based on a genetic algorithm (MOGA) for searching feasible model inputs, because such approaches have been demonstrated to be well applicable for NN models associated with large input data dimensions (e.g. Emmanouilidis et al., 1999; Oliveira et al., 2003). The implementation of this strategy was based on the toolbox of genetic and evolutionary algorithms (GEATbx), in which the multi-objective optimisation scheme follows mainly Fonseca and Fleming (1993) and partly Srinivas and Deb's (1994) work. The overall structure of proposed selection scheme is presented in Fig. 1.

As the core of the MOGA, we used a structured genetic algorithm where populations were utilised within the unrestricted migration/island model (e.g. Cantú-Paz, 1995); the base population of 100 individuals was divided into five relatively isolated subpopulations each having 20 individuals. In this scheme, 10% of individuals were exchanged between subpopulations within 20 generation intervals employing the fitness-based migration, i.e. the best individuals were exchanged. Inside the subpopulations, different mutation rates, ranging in [0.5, 2.5] (values represent the average number of mutations per individual), were employed in order to maintain both rough and fine search capabilities. As a termination criteria we used 500 iterations.

The operators were selected based on the investigations made by Emmanouilidis et al. (1999, 2000), whereby the advantages of the random sampling tournament selection and the subset size-oriented common features (SSOCF) recombination in multi-objective input selection have been demonstrated. It has been, especially, shown that the SSOCF is capable of constructing useful building blocks and maintaining the distribution across the range of Pareto front. The mutations were employed within the framework of binary mutations since the problem (input selection) was represented (encoded) as a bit string.

2.3.2. Searching Pareto-optimal input subsets

As opposed to the single-objective optimisation, no unique optimum can be attained in multi-objective case, but instead a set of trade-off (non-dominate) solutions.

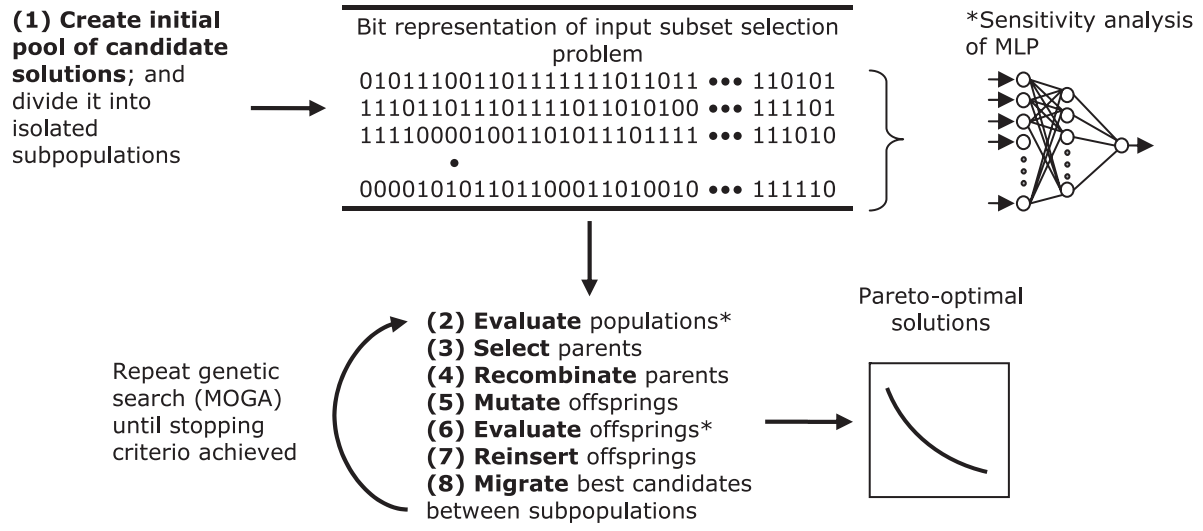

Fig. 1. The overall structure of the evolutionary multi-objective input selection strategy based on the use of MOGA and the sensitivity analysis of MLP model.

These solutions are known as the Pareto-optimal set (Goldberg, 1989) where no improvement in any objective is possible without sacrificing at least one of the other objectives. To attain well-distributed Pareto front (trade-off curve), we applied the modified Pareto ranking and goal attainment techniques provided by the GEATbx. For a more detailed description of these techniques, the reader is referred to the documentation of the GEATbx (www.geatbx.com) and to the above-mentioned references.

Two multi-objective optimisation criteria: (1) the number of inputs and (2) the model performance to an input subset were applied in order to find the most reliable complexity/accurate trade-off of MLP models. The first criterion was achieved simply as the number of selected inputs (bit = 1). The value of latter criterion was computed using the sensitivity of MLP model to estimate the relationship between the inputs and the final performance (Moody and Utans, 1991; Oliveira et al., 2003). In this scheme, the sensitivity of MLP model to an input subset was assessed by simulating the MLP model (trained on all the inputs) on a test data set where unselected inputs were replaced by their respective means computed on training data set.

Here, the sensitivity of input subset was defined as an absolute difference between the model performance achieved for an input subset and the model performance achieved for all input variables. Consequently, the input subsets having low sensitivity values (this indicates that the most relevant network input variables are selected) should be preferred in the selection strategy since they possibly contain relevant variables needed in the forecasting. This was simply achieved by minimising the sensitivity of input subsets. To measure the model performance, we used the well-known index of agreement (Willmott et al., 1985; Willmott, 1981) calculated as follows:

$$d = 1 - \left[\frac{\sum_{i=1}^{N}(P_i - O_i)^2}{\sum_{i=1}^{N}(|P_i - \overline{O}| + |O_i - \overline{O}|)^2}\right], \quad (1)$$

where N is the number of observations, O_i is the observed data point, P_i is the predicted data point and $\bar{O}$ is the mean of observed data. The index of agreement (d) varies from 0.0 (theoretical minimum) to 1.0 (perfect agreement between observed and predicted values).

2.4. Evaluation and inter-comparison of MLP models

The evaluation of MLP models was performed using the scheme that is illustrated in Fig. 2. The evaluation includes the following main phases: (1) the selection of model-specific inputs using the MOGA in conjunction with the sensitivity analysis of MLP, (2) the validation of achieved Pareto-optimal front and (3) the final evaluation of MLP using a Pareto-optimal input subset.

The data set contains the concentration and meteorological data for each hour during the period from 1 May 2000 to 30 April 2003; this constitutes 24 215 records of data. In phases 1 and 2, the data set was partitioned at random to three data sets: training data set (60% of all data records), test data set (20%) and

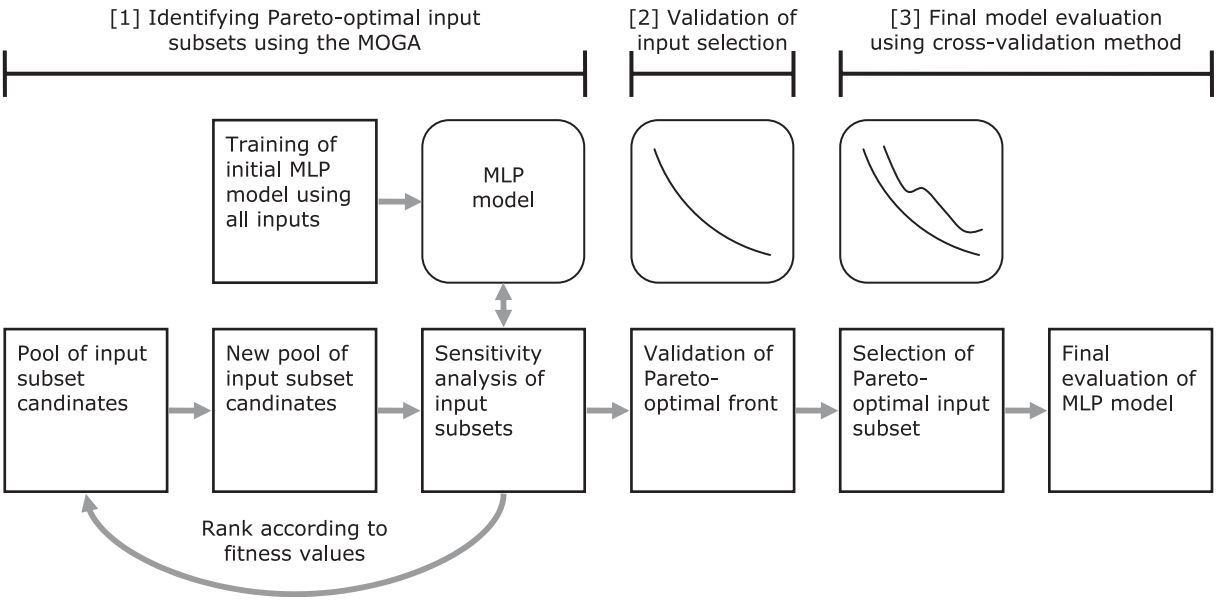

Fig. 2. The main phases of model evaluation based on the use of the MOGA in conjunction with the sensitivity analysis of MLP model.

validation data set (20%). The training data set was used for training the initial MLP, the test data set for performing feature selection and the validation data set for validating the achieved Pareto-optimal feature subsets. The validation step was carried out because it has been found that the Pareto-optimal feature subsets do not provide enough information per se, in order to select the best feature subset (Oliveira et al., 2003). This was caused by the fact that the Pareto-optimal feature subsets cannot provide a good generalisation on a different data; they provide good performance on the test data (used to compute the sensitivity of input subsets) only.

The best solution (containing the subset of input variables that yields the lowest sensitivity) of validated Pareto-optimal front was chosen to be used in the final evaluation. In the final step, we cross-validated the NN model sequentially on an annual basis. This means that the data corresponding to each of the above-mentioned years (defined as from 1 May 2000 to 30 April 2001, from 1 May 2001 to 30 April 2002 and from 1 May 2002 to 30 April 2003) are used in turn for the validation of models, and the data from the other two years is used for the training of models. This procedure has the advantage of yielding three separate sets of statistical model performance parameters. Comparing the results obtained during several years also provides confidence that the conclusions will not be dependent on the specific annual meteorological conditions.

For each validation period, several statistical measures were calculated: the index of agreement (Eq. (1)), mean absolute error (MAE), the squared correlation coefficient (R^2) and the fractional bias (FB):

$$\mathrm{FB} = \frac{\bar{P} - \bar{O}}{0.5(\bar{P} + \bar{O})}, \tag{2}$$

where $\bar{P}$ and $\bar{O}$ are the means of predicted and observed concentrations, respectively.

Additionally, the root mean square errors (RMSE) and its systematic and unsystematic components were calculated:

$$\mathrm{RMSE_s} = \left[\frac{1}{N}\sum_{i=1}^{N}(\hat{P}_i - O_i)^2\right]^{1/2}, \tag{3}$$

$$\mathrm{RMSE_u} = \left[\frac{1}{N}\sum_{i=1}^{N}(\hat{P}_i - P_i)^2\right]^{1/2}, \tag{4}$$

where $\hat{P}_i$ is a least-squares estimate of the predicted data point. The sum $\mathrm{RMSE}^2 = \mathrm{RMSE_s}^2 + \mathrm{RMSE_u}^2$ indicates the total error into systematic and unsystematic components.

When designing a modelling system for operational air quality forecasting, approaches that can forecast critical episodes must be preferred. We therefore also applied statistical measures for indicating whether an episodic threshold is exceeded or not (binary forecast), these measures are the fraction of false alarms (FA), the fraction of correctly predicted episodes (TA) and the episodic success index (SI). The SI is calculated from the true positive rate (TPR), representing the sensitivity of model (the fraction of correct alarms) and the false positive rate (FPR), representing the specificity of model (the relative fraction of FA).

$$\mathrm{TPR} = A/M, \quad 0 \leqslant \mathrm{TPR} \leqslant 1, \tag{5}$$

$$\mathrm{FPR} = (F - A)/(N - M), \quad 0 \leqslant \mathrm{FPR} \leqslant 1, \tag{6}$$

where A is the number of correctly predicted exceedances, M is the number of all observed exceedances, F is the number of all predicted exceedances and N is the total number of observations. Sensitivity and specificity are combined into the SI as follows:

$$\mathrm{SI} = \mathrm{TPR} - \mathrm{FPR}, \tag{7}$$

ranging in $[-1, 1]$. For a perfect model, $\mathrm{TPR} = 1.0$, $\mathrm{FPR} = 0.0$ and $\mathrm{SI} = 1.0$.

Threshold concentration values for episodic situations of NO_2 and $PM_{2.5}$ were defined according to the national guidelines or the data itself; the latter if guidelines were not available. The national guidelines for NO_2 are $70\,\mu g\,m^{-3}$ (daily mean) and $150\,\mu g\,m^{-3}$ (hourly mean) that are defined on a monthly basis, as the 99th percentile of the hourly values. During the period 2000–2003 at the station of Vallila, there have been only three exceedances of hourly guideline ($150\,\mu g\,m^{-3}$). Therefore, the daily mean value of $70\,\mu g\,m^{-3}$ was used as the threshold value of NO_2 in the computations. The examination of measurement data of NO_2 showed that there was overall 628 daily hours exceedances for this threshold, i.e. ~2.6% of all hours (97.4th percentile) in the data period. We defined also a threshold value for $PM_{2.5}$. However, there are no national or EU guidelines for the concentrations of $PM_{2.5}$. We therefore simply used the 97.4th percentile ($25\,\mu g\,m^{-3}$) of the measurement data (the station of Kallio) to determine a threshold.

3. Results and discussion

3.1. Examination of Pareto-optimal input subsets

The computed Pareto-optimal fronts regarding to the primary objectives: the model performance (the index of agreement) and the model complexity (the number of inputs), and their validation curves are presented in Fig. 3 for the four models considered. We have not presented the lists of selected variables as their detailed analysis is outside the scope of this study.

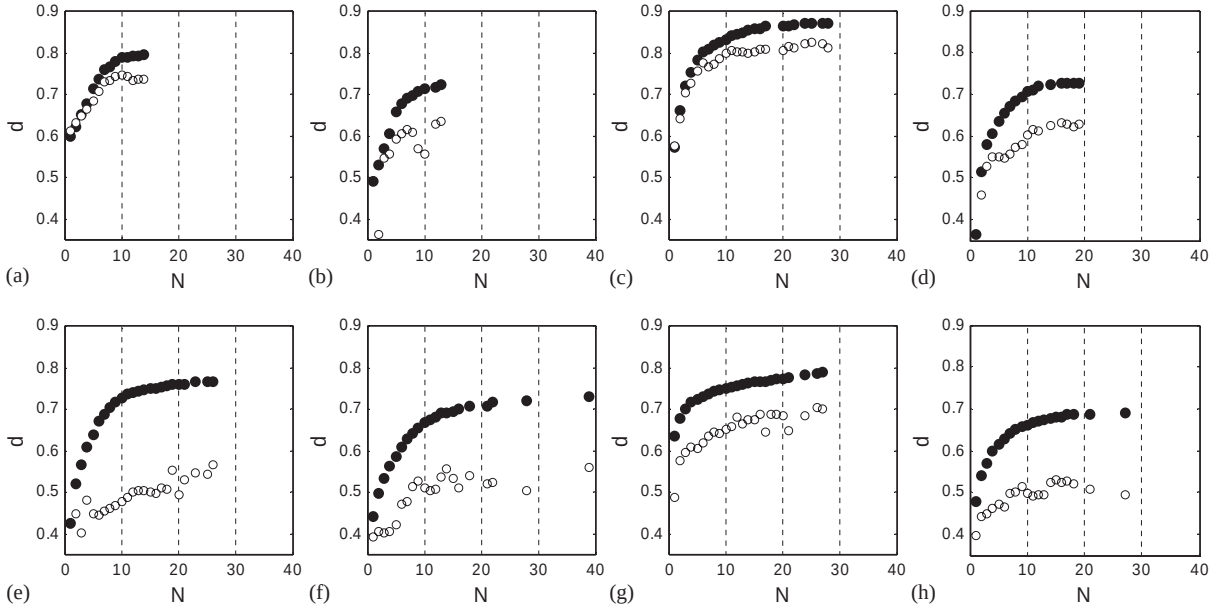

Fig. 3. Pareto-optimal fronts (points) and their respective validation curves (circles) for the period of 1 May 2001 to 30 April 2002 for the four models considered where d is the index of agreement achieved and N is the number of input variables utilised. The figure plots a–d present the forecasting of NO_2 and the plots e–h present the forecasting of $PM_{2.5}$ such that (a) and (e) correspond to MLP + NWP24, (b) and (f) correspond to MLP + NWP00, (c) and (g) correspond to MLP + MPP24 and (d) and (h) correspond to MLP + MPP00.

The results show that the total number of input variables needed for the forecasting ranged from 10 to 30, and from 20 to 40 in case of NO_2 and $PM_{2.5}$, respectively. The reason for these differences may be the basically different nature of the $PM_{2.5}$ concentrations compared with those of NO_2. In Helsinki, approximately half of the fine particulate matter concentrations in street level is originated from long-range transport (LRT), while more than 90% of the street-level concentrations of NO_2 is originated from local traffic (Karppinen et al., 2000). However, both meteorological data sets considered, HIRLAM and MPP-FMI data, represent local-scale meteorological conditions; the prevailing conditions during LRT are not taken into account. Both the nowcasting and forecasting of the fine particulate matter concentrations is therefore expected to be a substantially more demanding task, compared with that of nitrogen oxides. The above-mentioned, required larger number of input variables is therefore probably caused by the difficulties of the statistical methods in finding relevant meteorological parameters for all the cases.

In case of NO_2, using the HIRLAM data as input, a smaller number of variables was needed for achieving the optimum performance, compared with the corresponding cases using the meteorologically pre-processed data, both for the nowcasting ($T = 00$ h) and forecasting ($T = 24$ h) cases. The performances of validation curves obtained for the forecasting of $PM_{2.5}$ were remarkably worse, compared to those of the Pareto-optimal fronts. Therefore, the use of validation phase was found to be particularly important in case of $PM_{2.5}$ forecasting.

3.2. Results of the statistical evaluation of model performance

The final results of statistical model evaluation have been presented in Table 2 (NO_2) and Table 3 ($PM_{2.5}$). For both pollutants, the results have been presented as average values of the statistical indicators and their standard deviations over the cross-validation periods considered. The cross-validation periods were used for evaluating the year–year variation of the results. The evaluation of model performance in episodic conditions using the parameters FA, TA and SI was also included.

The overall agreement using daily HIRLAM weather forecasts (MLP + NWP24) can be considered to be fairly good for both pollutants: for instance, the d values ranged from 0.79 to 0.80 and from 0.63 to 0.81 for NO_2 and $PM_{2.5}$, respectively. As expected, the model performance substantially improves using the HIRLAM forecasts (MLP + NWP24), compared with the corresponding cases using HIRLAM analysis data as input (MLP + NWP00), for both pollutants.

In case of $PM_{2.5}$ forecasts, there was a substantial variation between the years considered. The pronounced

Table 2
The statistical model evaluation parameters of the forecasted and measured hourly time series of NO_2 concentrations at the station of Vallila, presented as average values and their standard deviations of a total of 10 repeated runs

Ind.	MLP + NWP24						MLP + NWP00					
	Year 1		Year 2		Year 3		Year 1		Year 2		Year 3	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
RMSE	12.18	0.11	12.63	0.21	13.96	0.09	13.51	0.17	14.12	0.18	16.26	0.12
$RMSE_s$	8.30	0.16	8.77	0.33	9.39	0.17	10.64	0.29	11.29	0.35	13.21	0.16
$RMSE_u$	8.91	0.09	9.09	0.07	10.33	0.24	8.33	0.13	8.47	0.19	9.48	0.33
d	0.80	0.01	0.80	0.01	0.79	0.00	0.72	0.01	0.71	0.01	0.64	0.01
R^2	0.47	0.01	0.48	0.02	0.43	0.00	0.35	0.02	0.35	0.02	0.23	0.01
MAE	8.82	0.13	9.18	0.18	10.04	0.10	10.09	0.14	10.34	0.10	11.98	0.07
FB(%)	−6.40	0.10	−5.50	0.30	−6.20	0.10	−7.30	0.20	−7.00	0.40	−8.30	0.10
FA(%)	55.95	19.07	17.86	10.49	57.46	21.81	100.00	0.00	11.54	15.11	93.67	13.12
TA(%)	0.60	0.49	5.00	2.55	3.22	1.46	0.00	0.00	3.02	2.56	0.24	0.60
SI	0.01	0.00	0.05	0.03	0.03	0.01	0.00	0.00	0.03	0.03	0.00	0.01
Ind.	MLP + MPP24						MLP + MPP00					
	Year 1		Year 2		Year 3		Year 1		Year 2		Year 3	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
RMSE	9.28	0.18	9.76	0.25	12.20	0.24	13.03	0.33	13.69	0.23	16.22	0.17
$RMSE_s$	5.00	0.18	5.45	0.34	6.08	0.37	10.06	0.59	10.76	0.35	12.96	0.23
$RMSE_u$	7.82	0.11	8.09	0.10	10.56	0.39	8.27	0.24	8.47	0.14	9.73	0.41
d	0.90	0.00	0.90	0.01	0.86	0.00	0.74	0.02	0.73	0.01	0.65	0.01
R^2	0.69	0.01	0.69	0.02	0.58	0.01	0.39	0.03	0.39	0.02	0.24	0.01
MAE	6.82	0.14	7.23	0.19	8.84	0.19	9.73	0.25	10.13	0.17	12.05	0.15
FB(%)	−3.20	0.20	−2.50	0.20	−3.70	0.20	−6.60	0.40	−6.30	0.30	−7.90	0.20
FA(%)	37.22	7.04	16.31	3.53	64.35	4.54	72.81	22.97	13.27	13.10	76.14	23.49
TA(%)	19.87	4.31	36.83	3.55	22.42	2.51	0.60	0.73	3.56	1.56	0.71	1.19
SI	0.20	0.04	0.37	0.04	0.21	0.02	0.01	0.01	0.04	0.02	0.01	0.01

The values are presented for the cross-validation periods: year 1 is defined as the period from 1 May 2000 to 30 April 2001, year 2 from 1 May 2001 to 30 April 2002 and year 3 from 1 May 2002 to 30 April 2003.

year–year variation was affected by the occurrence of severe episodes particularly during the third year (1 May 2002 to 30 April 2003). The scatter plots of observed versus predicted for the third year are presented in Figs. 4a–h; moreover, the corresponding histograms of prediction are presented in Figs. 5a–h. There are a few episodic data points of high observed concentrations that have not been correctly predicted in case of $PM_{2.5}$. Both meteorological data sets considered here represent local-scale meteorological conditions; the prevailing larger-scale meteorological conditions during LRT episodes for $PM_{2.5}$ are not therefore taken into account by the models.

The artificial alternative of using pre-processed meteorological data at the time $T = 24$ h (MLP + MPP24) was moderately better than the corresponding operative modelling option of using NWP forecasts. This result should only be considered as a numerical checking that using actual measurements at the time $T = 24$ h would provide a more accurate result than using the corresponding NWP model forecasts.

The corresponding performance of the models in case of the highest concentrations, as measured by the SI, FA and TA parameters, were more modest for all the models, irrespective of the pollutant. An exception was detected for the forecasting of $PM_{2.5}$ during the period from 1 May 2001 to 30 April 2002; this period did not include any severe episodes. As a general result, all the models under-estimated the highest concentrations, as has also been found in earlier investigations (Kolehmainen et al., 2001; Kukkonen et al., 2003). Similar to the overall model performance, the non-operational MLP + MPP24 model was somewhat better for forecasting episodes, compared with the corresponding model using NWP forecasts, MLP + NWP24.

Table 3
The statistical model evaluation parameters of the forecasted and measured hourly time series of $PM_{2.5}$ concentrations at the station of Kallio, presented as average values and their standard deviations of a total of 10 repeated runs

Ind.	MLP + NWP24						MLP + NWP00					
	Year 1		Year 2		Year 3		Year 1		Year 2		Year 3	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
RMSE	4.61	0.05	4.44	0.09	8.51	0.19	4.98	0.04	4.63	0.07	8.78	0.29
$RMSE_s$	3.12	0.11	2.90	0.20	6.24	0.27	3.61	0.14	3.02	0.13	6.38	0.14
$RMSE_u$	3.40	0.09	3.36	0.13	5.77	0.28	3.42	0.16	3.50	0.12	6.03	0.49
d	0.79	0.01	0.81	0.01	0.63	0.02	0.73	0.01	0.80	0.01	0.61	0.01
R^2	0.44	0.01	0.49	0.02	0.20	0.02	0.35	0.01	0.44	0.01	0.17	0.02
MAE	3.34	0.04	3.30	0.06	5.30	0.20	3.64	0.03	3.49	0.06	5.67	0.29
FB(%)	−5.58	0.46	−3.77	0.66	−6.89	0.52	−6.05	0.38	−3.11	0.42	−6.45	0.41
FA(%)	57.76	32.49	10.40	5.56	70.82	5.39	71.12	13.44	11.24	3.93	62.45	7.47
TA(%)	3.25	4.11	40.67	8.63	11.76	4.92	5.08	4.29	51.67	6.05	18.68	3.74
SI	0.03	0.04	0.41	0.09	0.10	0.05	0.05	0.04	0.52	0.06	0.17	0.03
Ind.	MLP + MPP24						MLP + MPP00					
	Year 1		Year 2		Year 3		Year 1		Year 2		Year 3	
	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
RMSE	4.33	0.08	4.01	0.12	7.56	0.11	5.08	0.11	4.63	0.14	8.44	0.14
$RMSE_s$	2.99	0.13	2.62	0.17	5.39	0.16	4.09	0.21	3.55	0.23	6.93	0.10
$RMSE_u$	3.13	0.04	3.03	0.03	5.30	0.18	3.00	0.11	2.96	0.06	4.82	0.19
d	0.81	0.01	0.85	0.01	0.72	0.01	0.68	0.03	0.76	0.03	0.57	0.01
R^2	0.50	0.02	0.57	0.02	0.33	0.02	0.31	0.03	0.43	0.04	0.17	0.02
MAE	3.11	0.06	2.99	0.08	4.74	0.10	3.70	0.08	3.43	0.09	5.33	0.15
FB(%)	−5.14	0.50	−2.78	0.27	−5.16	0.33	−7.17	0.43	−4.25	0.70	−7.21	0.30
FA(%)	55.71	18.63	9.74	3.22	51.27	6.99	88.82	12.86	13.60	4.43	57.59	6.32
TA(%)	6.59	6.29	56.17	3.60	26.89	3.33	0.63	0.82	36.92	12.10	11.43	2.32
SI	0.06	0.06	0.56	0.04	0.25	0.03	0.01	0.01	0.37	0.12	0.11	0.02

The values are presented for the cross-validation periods: year 1 is defined as the period from 1 May 2000 to 30 April 2001, year 2 from 1 May 2001 to 30 April 2002 and year 3 from 1 May 2002 to 30 April 2003.

4. Conclusions

In this paper, the MLP model and the data of HIRLAM NWP model were integrated and evaluated for the forecasting of NO_2 and $PM_{2.5}$ concentrations in central Helsinki. Additionally, a novel approach based on the MOGA and the sensitivity analysis of MLP model was used to select relevant model input variables. The utilisation of this method was required in order to reduce the excessively large number of potential meteorological input variables. Besides the evaluation of the developed operational MLP models that utilise the NWP data as input against measured concentrations, we also inter-compared these models with the corresponding MLP models that utilise the pre-processed meteorological data as input. The latter category of models (utilising pre-processed or directly measured meteorological data) constitute the statistical nowcasting methods that are currently widely used worldwide.

The results obtained with the operational MLP models that use HIRLAM forecasts showed fairly good agreement for both pollutants, as was found out using a variety of statistical model performance measures. The model performance was substantially better for both pollutants, when the HIRLAM forecasts ($T = 24$ h) were used, compared both with the case in which the HIRLAM analysis data ($T = 00$ h) was used, and the case in which meteorological pre-processor data ($T = 00$ h) was used as input. The HIRLAM forecasts therefore substantially improved the model performance, compared with the currently commonly used statistical prediction methods. However, the performances of all operational models considered deteriorated in episodic situations as has also been found in several other investigations (e.g. Kolehmainen et al., 2001; Kukkonen, et al., 2003).

A general conclusion is that the combination of NWP and MLP can be a useful tool for assessing NO_2 and

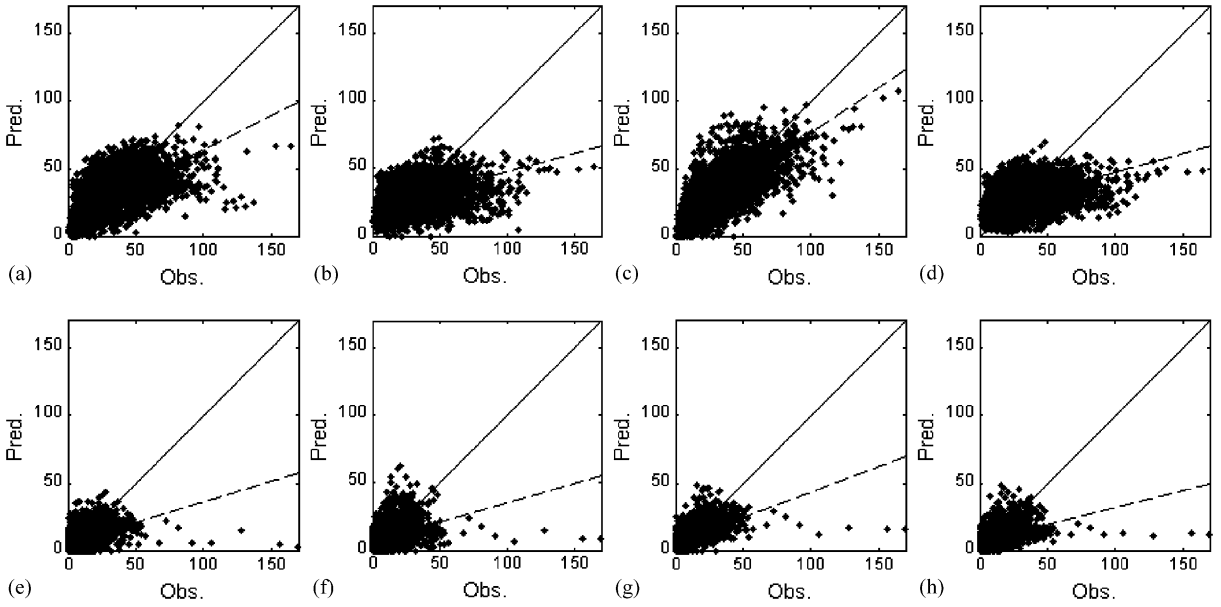

Fig. 4. The scatter plots of the measured and forecasted concentrations for the period of 1 May 2002 to 30 April 2003 for the four models considered. The figure plots a–d present the forecasting of NO_2 and the plots e–h present the forecasting of $PM_{2.5}$ such that (a) and (e) correspond to MLP + NWP24, (b) and (f) correspond to MLP + NWP00, (c) and (g) correspond to MLP + MPP24 and (d) and (h) correspond to MLP + MPP00.

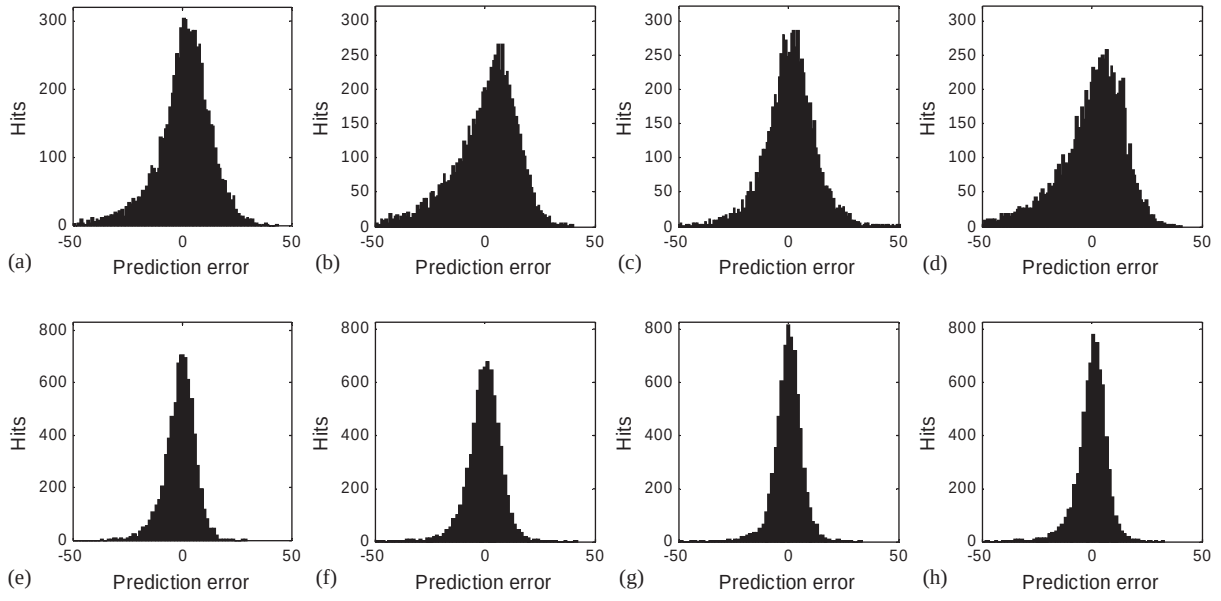

Fig. 5. The histograms of prediction error for the period of 1 May 2002 to 30 April 2003 for the four models considered. The figure plots a–d present the forecasting of NO_2 and the plots e–h present the forecasting of $PM_{2.5}$ such that (a) and (e) correspond to MLP + NWP24, (b) and (f) correspond to MLP + NWP00, (c) and (g) correspond to MLP + MPP24 and (d) and (h) correspond to MLP + MPP00.

$PM_{2.5}$ hourly concentrations in an urban area. However, the performance of the proposed system depends ultimately on site-specific conditions such as climatic and topographical factors, and its utilisation requires appropriate good quality site- and time-specific data for model training. In Northern Europe, the frequency of very stable atmospheric conditions with low wind speeds tends to be higher than that in lower latitudes; these are

commonly the most difficult meteorological conditions to model (Kukkonen et al., 2003; Rantamäki et al., 2005).

The capabilities of the presented modelling system could be improved by enhancements in the physical parameterisations and the spatial resolution of the HIRLAM model. Clearly, nested model computations utilising the HIRLAM and non-hydrostatic mesoscale meteorological model (such as, for instance, the MM5 model) would probably improve the accuracy of forecasting of the relevant meteorological variables. The MLP model could also be enhanced to better consider the temporal patterns of time series, for example by allowing for the selection of relevant time-lags of inputs using the proposed input selection technique or using recurrent neural networks.

Acknowledgements

The financial support of the Academy of Finland (FORECAST, project no. 49946) is gratefully acknowledged. We also wish to acknowledge the financial support of the European Comission for the FUMAPEX project and the Cluster of European Air Quality Research (CLEAR). The Helsinki Metropolitan Area Council (YTV) is thanked for providing the air quality monitoring data and Dr. S. Dorling (University of East Anglia, UK) for his valuable comments.

References

Baklanov, A., Rasmussen, A., Fay, B., Berge, E., Finardi, S., 2002. Potential and shortcomings of NWP models in providing meteorological data for UAP forecasting. Water, Air, and Soil Pollution 2, 43–60.

Cantú-Paz, E., 1995. A summary of research on parallel genetic algorithms. Technical Report IlliGAL No. 95007, University of Illinois at Urbana-Champaign.

Comrie, A.C., 1997. Comparing neural networks and regression models for ozone forecasting. Journal of Air and Waste Management Association 47, 653–663.

Eerola, K., 2001. The new operational HIRLAM at the Finnish Meteorological Institute. HIRLAM Newsletter 35, 22–28.

Eerola, K., 2002. The operational HIRLAM at the Finnish Meteorological Institute. HIRLAM Newsletter 41, 19–24.

Emmanouilidis, C., Hunter, A., MacIntyre, J., Cox, C., 1999. Selecting features in neurofuzzy modelling using multi-objective genetic algorithms. In: Willshaw, D., Murray, A. (Eds.), Proceedings of the Ninth International Conference on Artificial Neural Networks. IEE, London, UK, pp. 749–754.

Emmanouilidis, C., Hunter, A., MacIntyre, J., 2000. A multi-objective evolutionary setting for feature selection and a commonality-based crossover operator. In: Proceedings of CEC00, 2000 Congress on Evolutionary Computation. IEEE, Piscataway, NJ, USA, pp. 309–316.

Fonseca, C.M., Fleming, P.J., 1993. Genetic algorithms for multi-objective optimisation: formulation, discussion and generalisation. In: Forrest, S. (Ed.), Proceedings of the Fifth International Conference on Genetic Algorithms. Morgan Kaufmann Publishers, San Mateo, CA, USA, pp. 416–423.

Gardner, M.W., Dorling, S.R., 1999. Artificial neural networks (the multi-layer perceptron)—a review of applications in the atmospheric sciences. Atmospheric Environment 32, 2627–2636.

Goldberg, D.E., 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley, Reading, Massachusetts.

Haykin, S., 1999. Neural Networks: A Comprehensive Foundation, second ed. Prentice Hall, Upper Saddle River, NJ.

Hooyberghs, J., Mensink, C., Dumont, G., Fierens, F., Brasseur, O., 2005. A neural network forecast for daily average PM_{10} concentrations in Belgium. Atmospheric Environment in press.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M., 2004. Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38, 2895–2907.

Karppinen, A., Joffre, S., Vaajama, P., 1997. Boundary layer parametrization for Finnish regulatory dispersion models. International Journal of Environment and Pollution 8, 557–564.

Karppinen, A., Joffre, S.M., Kukkonen, J., 2000. The refinement of a meteorological preprocessor for the urban environment. International Journal of Environment and Pollution 14, 565–572.

Kohavi, R., John, G.H., 1997. Wrappers for feature subset selection. Artificial Intelligence 97, 273–324.

Kolehmainen, M., Martikainen, H., Ruuskanen, J., 2001. Neural networks and periodic components used in air quality forecasting. Atmospheric Environment 35, 815–825.

Kukkonen, J. (Ed.), 2001. Meteorology Applied to Urban Air Pollution Problems. COST action 715, Working Group 3, Status Report, Directorate-General for Research, Information and Communication Unit, European Commission, Brussels, pp. 73 (http://cost.fmi.fi/statusreportprinted.pdf).

Kukkonen, J., Partanen, L., Karppinen, A., Ruuskanen, J., Junninen, H., Kolehmainen, M., Niska, H., Dorling, S., Chatterton, T., Foxall, R., Cawley, G., 2003. Extensive evaluation of neural networks of NO_2 and PM_{10} concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmospheric Environment 37, 4539–4550.

Moody, J., Utans, J., 1991. Principled architecture selection for neural networks: application to corporate bond rating predictions. In: Moody, J., Hanson, S.J., Lippmann, R.P. (Eds.), Proceedings of Advances in Neural Information Processing Systems. Morgan Kaufmann, San Mateo, CA, pp. 683–690.

Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., Kolehmainen, M., 2004. Evolving the neural network model for forecasting air pollution time series. Engineering Applications of Artificial Intelligence 17, 159–167.

Oliveira, L.S., Sabourin, R., Bortolozzi, F., Suen, C.Y., 2003. A methodology for feature selection using multi-objective genetic algorithms for handwritten digit string recognition.

International Journal of Pattern Recognition and Artificial Intelligence 17 (6), 903–930.

Pielke, R.A., Uliasz, M., 1998. Use of meteorological models as input to regional and mesoscale air quality models—limitations and strenghts. Atmospheric Environment 32, 1455–1466.

Piringer, M., Kukkonen, J., 2002. Mixing height and inversions in urban areas. In: Piringer, M., Kukkonen, J. (Eds.), Proceedings of workshop, 3–4 October 2001, Toulouse, France. COST Action 715, EUR 20451, European Commission, Brussels, pp. 113.

Pohjola, M.A., Rantamäki, M., Kukkonen, J., Karppinen, A., Berge, E., 2004. Meteorological evaluation of a severe air pollution episode in Helsinki on 27–29 December 1995. Boreal Environmental Research 9, 75–87.

Rantamäki, M., Pohjola, M.A., Tisler, P., Bremer, P., Kukkonen, J., Karppinen, A., 2005. Evaluation of two versions of the HIRLAM numerical weather prediction model during an air pollution episode in southern Finland. In: Sokhi, R. (Ed.), Fourth International Conference on Urban Air Quality: Measurement, Modelling and Management 2003 (special issue); Atmospheric Environment 39/15, 2775–2786.

Srinivas, N., Deb, K., 1994. Multiobjective optimization using nondominated sorting in genetic algorithms. Evolutionary Computation 2 (3), 221–248.

Termonia, P., Quinet, A., 2004. A new transport index for predicting episodes of extreme air pollution. Journal of Applied Meteorology 43 (4), 631–640.

Willmott, C.J., 1981. On the validation of models. Physical Geography 2, 184–194.

Willmott, C.J., Ackleson, S., Davis, R., Feddema, J., Klink, K., Legates, D., O'Donnell, J., Rowe, C., 1985. Statistics for the evaluation and comparison of models. Journal of Geophysical Research 90 (C5), 8995–9005.

Ziomas, I.C., Melas, D., Zerefos, C.S., Bais, A.F., 1995. Forecasting peak pollutant levels from meteorological variables. Atmospheric Environment 29, 3703–3711.