J20.5 **ENVINNET: A NEURAL NETWORK FOR HINDCASTING PM_{10} IN URBAN PHOENIX**

M. C. Mammarella[1]*, G. Grandoni[1], P. Fedele[1], R. A. Di Marco[1], H.J.S. Fernando[2], R. Dimitrova[2], P. Hyde[2]
[1] ENEA, Italian National Agency for New Technologies, Energy and Environment
[2]Center for Environmental Fluid Dynamics, ASU, Tempe, USA

1. INTRODUCTION

The Phoenix metropolitan area is currently designated as 'serious' with regard to violation of the U.S. National Ambient Air Quality Standards (NAAQS) for particulate matter of aerodynamic diameter less than 10μm (PM_{10}). Most of the severe PM_{10} violations have been attributed to regional natural exceptional events or local exceptional (episodic) events associated with windblown dust emanating from area sources such as construction and agricultural sites, vacant lots and alluvial channels. During such events, the pollution concentration spikes for a short period of time, thus raising the 24-hour average of PM_{10} anomalously. This may lead to the excess of PM_{10} concentration above the currently set 24-averaged standards, 150μg/m^3, determined in the interest of protecting public health (primary standard) and the environment (secondary). For PM_{10}, both standards are the same. Even if the NAAQS are not exceeded during such an event, severe health repercussions can occur due to pulsed PM_{10} events. For example, thunderstorm-induced asthma epidemics that swamp hospital emergency rooms within 20 minutes of the onset of a storm have been attributed to suspension of micron-sized starch granules that originate in pollen (Venables et al. 1997). Unlike for industrial and transportation-networks associated sources, it is difficult to predict and control PM pollution arising from natural episodic events. In general, deterministic models have been used for such predictions, but unavailability of the up-to-date pollution inventories, the complexity of models and the fact that air pollution prediction models, such as CMAQ, do not have sound dust entrainment module have been the bane in developing operational forecasting tools based on deterministic models (Choi et al. 2006; Choi & Fernando 2007).

To this end, the Italian National Agency for New Technologies, Energy and Environment (ENEA) has been working with ASU to develop a stochastic model based on neural networks, which is called EnviNNet. The neural networks (NN) have already been used in air quality forecasting in Rome, Milan and Napoli with considerable success. Developing EnviNNet required careful selection of a subset of input variables, paying attention to site-specific exceptional events, including time lag effects. The data noise needed to be considered in order to satisfactorily adapt the non-linear dynamic interaction between meteorological and pollution related processes. EnviNNet employs 3-layer MLP network architecture with hidden nodes, full connection between layers and no connection between neurons in the same layer topology and exponential transfer function. In training the network, characteristics of high, low and episodic air pollution events at a particular data site were taken into account.

Since PM_{10} concentrations are heterogeneous throughout the year, to ensure that EnviNNet is robust and adaptive to the local climate, specific input data subsets were built by combining time section data from different years. Meteorological and pollution data were taken from selected four-to-six-month windows as well as time periods that show noteworthy patterns. For the Phoenix area, as discussed below, EnviNNet showed better performance compared to conventional deterministic models in predicting PM_{10} peaks. It should be noted that testing has been conducted only for a single station and a single pollutant in a specific geographical area, and hence further model evaluations are necessary before arriving at conclusions.

2. ENVINNET

Considering its wide usage in atmospheric applications (Gardner & Dorling 1998; Nishka et al. 2004), the kind of NN used was a Multi Layer Perceptron (MLP). The MLP structure consists of an interconnected system of nodes (neurons) within a hidden layer that employs non-linear continuous transfer functions that connect input and output vectors. Bearing in mind that the main purpose of EnviNNet is the prediction of PM_{10}, the following design choices were made:

- **Architecture**: three-layer MLP network, with the number of hidden nodes selected to reliably rebuild data from a test data-set.
- **Topology**: full connection between layers and no connection between neurons in the same layer;
- **Transfer function:** exponential, to ensure positive functions at the output node and hyperbolic tangent for hidden nodes. The latter is an excellent compromise for a non-linear function both globally and locally;
- **Information flow**: feed-forward;
- **Representation of input variables**: standardization of inputs to eliminate problems due to different measurement scales for different predictors;

**Corresponding author address*: M. C. Mammarella, ENEA L.Thaon di Revel 76, 00196 Rome, Italy;
e-mail: mariacristina.mammarella@sede.enea.it

- **Reconstruction of the output signal**: through linear combination and exponentialization of the outputs of hidden nodes.
- **Training method:** parameters are learned or estimated through the conjugate-gradient method. It is preferred over the back-propagation method as it exploits both first-order (gradient) and second-order (curvature) data during optimization of the objective function.

Mathematically, the three-layer NN has the following form:

$$y = f(\varphi(x, w))$$

where **x** represents input data, **w** the coefficients (parameters estimated by learning), **f** the activation functions from layers 2 to 3 and **φ** the activation functions from layers 1 to 2. The choice of **f** and **φ** determines the output; for example, if **φ** is a hyperbolic tangent and **f** is linear, an input of meteorological and pollution variables are transformed to an output with both negative and positive values, but if **f** is exponential, the output can have only positive values. As such, an exponential function for the output and a hyperbolic tangent activation function for the hidden neurons were selected.

The schematic chart of the neural network architecture is shown in Figure 1.

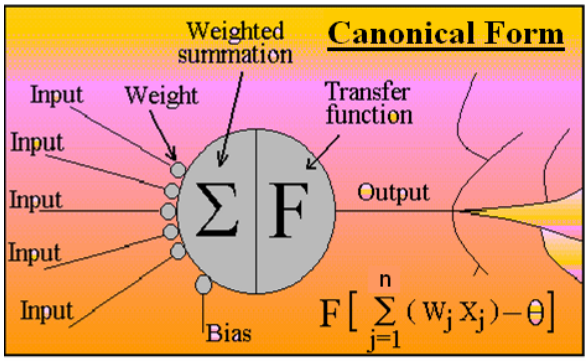

Figure 1. Neural network architecture

3. COMPARISON BETWEEN CMAQ AND ENVINNET – RESULTS AND DISCUSSION

Two different types of predictive systems – deterministic (CMAQ) and stochastic (EnviNNet) were evaluated against observations at one monitor in the Central Phoenix (CP) area. A one month 'design' period covering November 2005 was selected, considering that, in general, winter months exhibit the highest PM_{10} concentrations and hospital visits due to respiratory illnesses. The selected period has both high and low PM_{10} days including exceptional events with very high particle concentration due to storm comditions. The study domain with location of the available monitors is shown in Figure 2. The observations are taken only from CP in this study, but other sites for a future forecasting network are also shown in the figure.

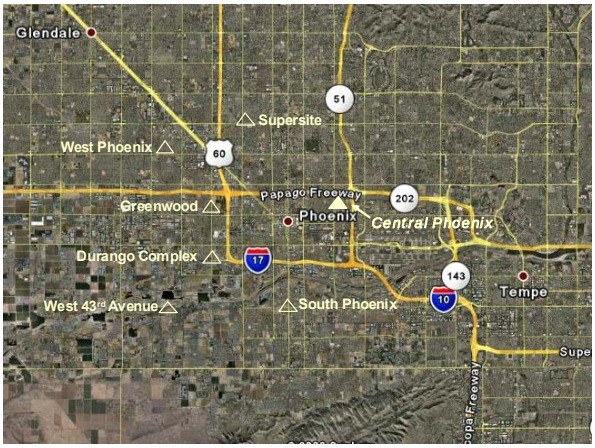

Figure 2. The locations of pollution and meteorological monitoring stations in Phoenix, Arizona.

The simulations were conducted using the regional air quality modeling deterministic system Models-3. It consists of mesoscale meteorological model MM5 (Grell&Dudia, 1994), Sparse Matrix Operator Kernel Emissions Processor SMOKE, and Chemical Transport Model CMAQ (Byun&Ching, 1999). Three nested domains with 36-, 12- and 4-km grid resolution were utilized to predict flow, emissions and air quality respectively. The outcomes from the finest domain were compared with the results from EnviNnet and observed data. The other domains were used to provide boundary and initial conditions for the inner domains only.

The meteorological and PM_{10} data from two years (2005, 2006) were used with the stochastic model EnviNNet to select the set of appropriate pre-processed historic time series at the Central Phoenix site (CP) to predict the concentration in the hindcasting mode. Statistical and descriptive indicators were found to represent behaviour of the CP neighbourhood via various classes of the input parameters. The indicators account for aggregated homogeneous spatio-temporal bands of microclimatic factors and air pollutant concentration. The latter is dependent on the wind transport, diffusion of PM_{10}, and emissions, which can be separately specified for working days and holidays.

The comparison of EnviNNet performance against the predictions of the deterministic modeling system is presented below. The regulatory PM_{10} enforcements are made based on 24-hour averages, which were used to estimate the relationship between pollution and asthma events (Dimitrova et al. 2008). Both models were evaluated against the observation as shown in Figure 3. The neural network satisfactorily predicts the PM_{10} peaks whereas CMAQ has problems of capturing the exceptional events satisfactorily.

Although the deterministic methods would have given ideal predictions under ideal initial, boundary and pollution inventory conditions with all scales of motion fully resolved, the present status of the model is far from this state. The model does not resolve processes with scales smaller than the grid size,

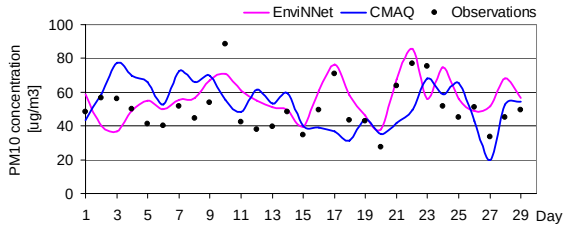

Figure 3. 24-hour PM_{10} concentrations predicted by CMAQ and EnviNNet in comparison with the observations.

which are parameterized. The model performance is based on the hourly data modeled by individual methods and the observed data are shown in Figures 4 and 5. The coeficients of determination (R2) are also shown in the plots. The R^2 is approximately two times higher for EnviNNet compared to CMAQ. The predicted hourly concentrations are closer to the observed values and crowd around the perfect coefficient of determination (R^2=1) while the scatter plot for CMAQ is more dispersed.

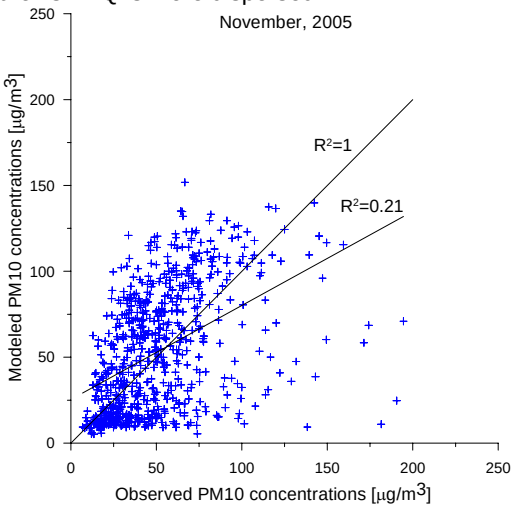

Figure 4. Scatter plots of PM_{10} concentration predicted by CMAQ in comparison with hourly observed data.

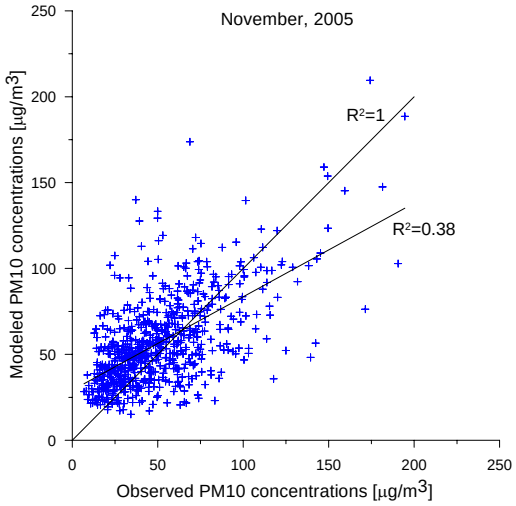

Figure 5. Scatter plots of PM_{10} concentration predicted by EnviNNet in comparison with hourly observed data.

In addition, different statistics were computed to estimate the model performance, and they show reasonable agreement between the calculated and observed values of the both models (Table 1).

Table 1. Summary of the statistical measures that compares the observations and model concentrations of PM_{10} for study periods.

Model	MAE	RMSE	IA	Period
CMAQ	26.12	34.4	0.68	November
EnviNNet	19.01	25.02	0.77	November
CMAQ	18.88	31.94	0.75	5-9 Nov.
EnviNNet	16.5	20.11	0.82	5-9 Nov
CMAQ	22.2	36.72	0.59	25-29 Nov.
EnviNNet	18.15	23.81	0.79	25-29 Nov.

The calculations were made for one month and for two five-day periods with good agreement and high disagreement for both models in comparison with the observations (see Appendix for the statistics definitions). The Index of Agreement (IA) is more than 0.6, which shows good correspondence between the calculated and observed data. Generally, EnviNNet gives better IA in comparison with CMAQ for all periods considered here. The Mean Absolute Errors (MAE) are less than 26 for CMAQ and less than 19 for EnviNNet. The Root Mean Square Errors (RMSE) are in the range of 20 – 37 for different periods. The stochastic model EnviNNet yields smaller errors than CMAQ.

The emissions inventory in its model-ready state has a critical bearing on the model performance. About 70% of ambient PM_{10} in Phoenix comes from fugitive emison including a large portion of soil dust issions. These dust emissions from activities such as traffic on unpaved roads, land clearance, and building and road construction are difficult to quantify well. Furthermore, emissions from many source categories cannot accurately be distributed in time and space. This questionable reliability of the emissions (an ever present stumbling block in most air pollution studies) adversely affects CMAQ performance.

The advantages of the model EnviNNet are that this system does not need the emission inventories and it dynamically checks its PM_{10} predictions against actual data so that the continued re-analysis of the data refines the learned network, thus improving its predictive capabilities.

4. CONCLUSIONS

As an alternative to the deterministic Models-3 system, the neural network is found to better predict moderate to high PM_{10} concentrations than CMAQ. The principal advantage of using EnviNNet in an asthma warning system is that it would not require an emission inventory or daily complicated computational efforts of the grid-based, deterministic modeling system. The neural network is much easier and quicker to use than CMAQ and could be partially automated for issuing health warnings. The shortcoming of EnviNNet is a limited geographical

coverage. Only one site was examined, therefore similar work at additional sites is needed.

Implementing EnviNNet on several stations in the Phoenix metropolitan can provide pollution distribution both spatialy and temporally, using interpolation methods (e.g., Kriging).

References:

Byun, D.W. and Ching, J.K.S., 1999: Science Algorithms the EPA Model-3 Community Multiscale Air Quality (CMAQ) Modeling System. *Washington, DC: U.S. Environmental Protection Agency, Office of Research and Development, EPA-600/R-99/030*.

Choi, Y-J., Hyde, P., Fernando, H.J.S., 2006: Modeling of episodic particular matter events using a 3-D air quality model with fine grid: Applications to a pair of cities in the U.S./Mexico border. *Atmospheric Environment*, 40, 5181-5201.

Choi, Y-J. and Fernando, H.J.S., 2008: Implementation of a Windblown Dust Parameterization into MODELS-3/CMAQ: Application to Episodic PM Events in the U.S./Mexico border. *Atmospheric Environment*, 42, 6039-6046.

Venables, K.M., Allitt, U., Collier, C.G., Emberlin, J., Grieg, J.B., Hardaker, P.J., Highham, J.H., Laing-Morton, T., Maynard, R.L., Murray, V., Strachan, D., Tee, R.D., 1997: Thunderstorm-related asthma - the epidemic of 24/25 June 1994. *Clinical Exper. Allergy*, 27, 725-736.

Gardner M. W. and Dorling S. R., 1998: Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32, 2627–2636.

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.

Appendix

Definitions of statistics:

The following indicators were used for performance evaluation. Here **P** is the predicted value, **O** the observed value, and $\overline{\mathbf{P}}$ and $\overline{\mathbf{O}}$ the mean values.

$$\mathbf{MAE} = \frac{1}{\mathbf{N}}\sum_{i=1}^{N}|\mathbf{P}_i - \mathbf{O}_i| \quad \text{(Mean Absolute Error)}$$

$$\mathbf{MB} = \frac{1}{\mathbf{N}}\sum_{i=1}^{N}(\mathbf{P}_i - \mathbf{O}_i) \quad \text{(Mean Bias)}$$

$$\mathbf{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(\mathbf{P}_i - \mathbf{O}_i)}{\mathbf{N}}} \quad \text{(Root Mean Square Error)}$$

$$\mathbf{IA} = 1 - \frac{\sum_{i=1}^{N}\left(\mathbf{P}_i - \mathbf{O}_i\right)^2}{\sum_{i=1}^{N}\left(\left|\mathbf{P}_i - \overline{\mathbf{O}}_i\right| + \left|\mathbf{O}_i - \overline{\mathbf{O}}_i\right|\right)^2} \quad \text{(IA Index of Agreement)}$$