Presented at the 96th AMS Annual Meeting, New Orleans, LA, Gen 10-14, 2016

4.3

OPTIMIZATION OF INPUT DATA FOR NEURAL NETWORK BY THE DEFINITION OF THE MATHEMATICAL ARGUMENT DIAGRAM: A CASE STUDY OF OZONE PREDICTION

A. Pelliccioni1,2 , S. E. Haupt 3,4

1 INAIL-DIMEILA, Via Fontana Candida 1, 00040 Monteporzio Catone, Roma (Italy)

2 DICEA, Università di Roma "La Sapienza", Via Eudossiana 18, 00184 Roma (Italy)

3 National Center for Atmospheric Research, Research Applications Laboratory

4 The Pennsylvania State University Meteorology Department

1. Introduction.

In most scientific approaches, the theory is reserved for explaining the phenomena observed in nature that are often revealed by empirical observation. Measurements and theory are intimately interconnected and science requires explanation of measurements by theory. This observation leads to a question regarding the consistency of scientific information and the best methodology to obtain theoretical results. Arguments about consistency of scientific information are relevant to application of neural networks (NN) to environmental problems. In fact, the evaluating the information contained in the input data is crucial to successful application of NNs for forecasting.

The goal of this work is to show an application of understanding experimental data for prediction of ozone from an urban dataset.

The relevance of input information during the training phase will be emphasized.

2. Deterministic and intelligent models for air pollution

It is often necessary to build models to predict air pollution based on historical data. Many models exist to simulate environmental conditions. The deterministic models are based on the knowledge of boundary conditions and of the physical equations. Lagrangian and Eulerian models (Finlayson-Pitts B at al (1999), Zhang J. et al. (2007)) are the two main physical approaches applied for air pollution predictions. Each deterministic model describes some conditions and simulates the pollutants under those conditions.

When deterministic models are compared with measurements from monitoring stations, sometimes the results are disappointing, especially for the simulation of the chemical reactions in urban areas. In these cases where deterministic models miss the prediction, an intelligent modeling approach (such as NN or support vector machine (SVM) can be a valid alternative. For the prediction of ozone some authors use NN (Comrie (1997), Gardner et al (2000), Ibarra-Berastegi et al. (2008)), while others the SVM (Feng et al. (2011), Ortiz-García et al (2010)). Here we have used the neural network model.

To reproduce observed ozone the main factors to consider are the following:

- Dispersion induced by meteorological conditions (Chen et al (2011)
- Thermal and mechanical turbulence in the boundary layer (McElroy et al. (1986))
- Emission factors characteristic of each source (Lighty et al. (2000))
- Simulation of chemical reactions during the transport time from emission the monitoring station

However, one must consider additional challenges to simulating air pollution in urban areas. While the relationship between cause and effect is often easy to determine for closed systems (i.e. wind-tunnels (Blocken et al (2007)) or water-channel simulations (Yee et al. (2006)), in the case of urban data this relationship is much more difficult to determine. In real situations, the complexity of the models is linked to the complexity of the boundary conditions. In such cases, one must consider unknown information as a primary issue to address before of any model application.

3. Definition of the mathematical arguments diagram (MAD)

To optimize input data to an NN we have introduced the definition of the Mathematical

[1] *Corresponding Author Address*:
Armando Pelliccioni, Inail. Dimeila, Rome, Italy.
Email: a.pelliccioni@inail.it

Arguments Diagram (MAD) concept. This concept consists of a graph that explores the relationship between input variables used to train the NN. The MAD graph shows the information levels involved in the physical process that we wish to simulate. Figure 1 shows a general scheme for MAD.

The MAD graph contains the following information:

- the independent variable (X_1, … Xn) or input for NN models;
- the inner parameters of NN models (K_1… K_p);
- the argument equation for the dependent variable (Y(X_1..X_n;[K_1..K_p]).

The graph is fundamental to defining all relevant variables connected with the physics.

The output argument equation (the black square box in Figure 1) is:

$$Y=Y(X_1, \dots Xn; [K_1,..,Kp]) \qquad (1)$$

There exists a strict correspondence between the MAD graph and the argument equation. For each MAD, an argument equation is given and vice versa.

The information level number related to the process is the number of square boxes used as input information to the argument equation (number of red boxes in the MAD).

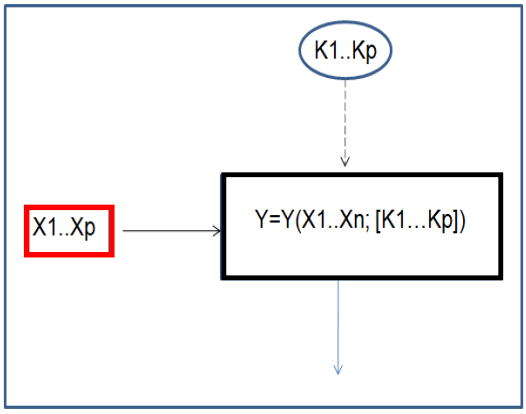

Figure 1: General MAD scheme

The MAD graph indicates the explicit relationships between the relevant input variables needed by model to reproduce the output variables (the black arrow that connects the boxes in Figure 1).

The MAD graph makes no explicit statement about the specific mathematical relationship between the variables (i.e., if the model is linear or nonlinear). The physics must be consider at this point. The system being studied must be synthesized to reduce all conservation equations to (1), where all variables and parameters and their connections are well identified.

This methodology provides for classification of all relevant parameters and variables for the mathematical model.

For each MAD a fixed number of information levels must be assigned. For the air pollution transport and dispersion mode the argument equation is quite intricate and the complete MAD involves up to seven information levels (not shown here).

4. MAD scheme for photochemical reaction of Ozone in atmosphere

For the photochemical production of NO2, the maximum information levels is assumed to be three. These reactions are activated by solar radiation, temperature and involve NO, O3 and NO2 pollutants

In urban areas, photochemical reactions involve the production of NO2 by the O3 and NO2 pollutants (Seinfeld et al. (2012)). The reaction usually takes place on sunny days with high air temperatures.

In Figure 2 the MAD scheme for the first information level are shown. This first information level (L1) corresponds to the minimum information level to simulate the production and depletion of O3 by NO and NO2.

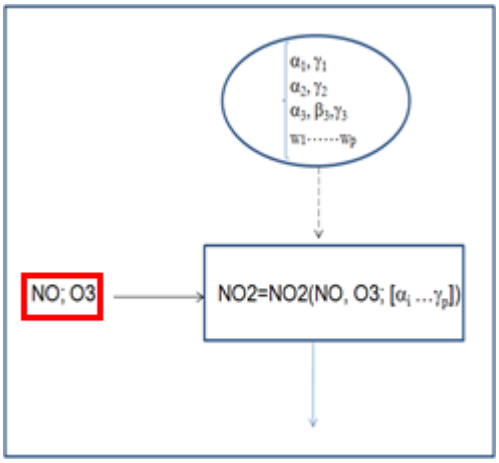

Figure 2: MAD scheme First-level MAD (L1)

At this first level, the MAD for O3 is only obtained by providing NO and NO2 concentrations. For the regression model, all parameters $\alpha_1 \dots \gamma_p$ are constant and fixed during training (i.e. consistent input dataset)

For the NN model, the $\alpha_1 \dots \gamma_p$ parameters are coincident with synaptic weights linked to hidden neurons in the NN models.

The argument equation for the first MAD scheme is:

$$O_3(t) = O_3(NO; NO_2) \qquad (2)$$

At the second level of information (L2), the O3, obtained by the NO and NO2 concentrations, is

filtered by the knowledge derived from photochemical reactions (Figure 3).

Such information is connected with K_1(T,SR,RH) and K_2(T,SR,RH), where K_1 and K_2 are the chemical rate connected to the photochemical reactions in atmosphere.

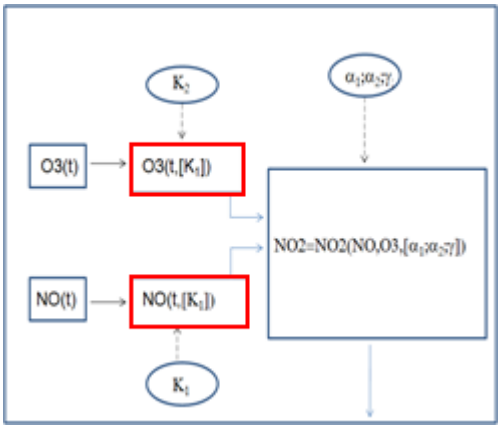

Figure 3: MAD scheme Second-level MAD (L2)

The values of K_1 and K_2 are unknown and difficult to obtain for an urban area. The O3 concentrations can be reproduced by the NN using NO and NO2 as inputs, together with a proxy variable for the reaction rate parameters K_1 and K_2. The latter are dependent on the observed air temperature, solar radiation and relative humidity.

The argument equation associated MAD (Figure 3) is:

$$O_3(t) = O_3(NO; NO_2; T, RH, SR) \quad (3)$$

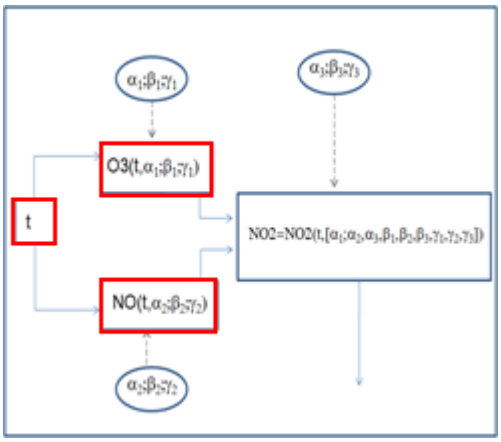

Figure 4: MAD scheme Third-level MAD (L3)

The third level of information (L3) is introduced by including the variable, time of day (h), which is associated with atmospheric stability conditions. It can be demonstrated that an averaged value of Monin Obukhov length L can be associated with each hour of the day (h) (Pelliccioni et al. (2012)). The MAD scheme (Figure 4) linking the variables is the equation:

$$O_3(t) = O_3(NO; NO_2; T, SR, RH, hh) \quad (4)$$

The analysis of the hour as an input variable to the NN can usually proceed in either of two ways. First, using a direct variable, hour of day, without any transformation (one dimensional analysis - 1D). Second, by employing a sine-cosine transformation of the hour variable (h) to account for diurnal-nocturnal cycles (two dimensional analysis - 2D). In this work, we operate with a 1-D analysis.

5. Dataset description

In Figure 5 shows the network of ozone monitoring stations in Rome. We use data from the station called Villa ADA (VAD), which is classified as a background monitoring station (yellow circle in Figure 5). The data are collected hourly and we use data from the full year period of 2006.

Pollutants concentrations measured at VAD include: Ozone (O3), Nitrogen oxide (NO) and dioxide (NO2), carbon oxide (CO) and total particulate matter and ultrafine particulates (PM10 and PM2.5 respectively). The meteorological variables measured include relative humidity (RH), solar radiation (SR), temperature (T), wind speed (WS) and direction (WD) and pressure (P). The total number of variables measured is 12. Adding the exogenous variable, hour of day (h), brings the total to 13. This number represents the maximum information or data available at the VAD station.

The three levels of the MAD scheme correspond with following percentage of input data, or information level (IL):

L1→ 2/13≅ 15% of IL (5)

L2→ 5/13≅ 38% of IL (6)

L3→ 6/13≅ 46% of IL (7)

Here, we have tested a fourth simulation (named ALL), where the Rain (R) and wind speed (WS) are added to L3 variables.

For this simulation, the percentage of data corresponding to the ALL simulation is equal to 8/13≅ 61%.

The final classification as suggested by the MAD analysis are the following:

L1 → O3(NO,NO2) (9)

L2 → O3(NO,NO2,T,RH,SR) (10)

L3 → O3(NO,NO2,T,RH,SR,h) (11)

ALL→ O3(NO,NO2,T,RH,SR,h, WS,R) (12)

The data was divided in two groups. The training dataset (comprised of 60% of randomly selected data) and the test dataset (comprised of the remaining 40% of total data). We use a multilayer perceptron (Gardner et al. (1998), Pelliccioni et al. (2006)) with 10 and 12 hidden neurons for the NN.

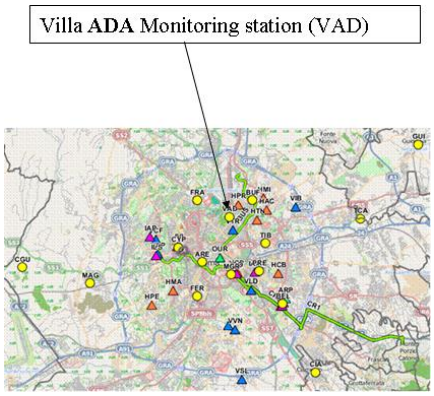

Figure 5: Rome monitoring stations

6.0 Results and discussion

We examine the results from the test dataset (see above). We describe the results for applying both linear regression and neural network models.
We explain the model performance by splitting the ozone concentrations in three ranges:

- Lower range: ozone observations under 50 mg/m3.
- Middle range: ozone observations between 50 and 100 μg/m^3.
- Upper range: ozone observations greater than 100 μg/m^3.

6.1 Linear regression results

Figure 6 compares the observed and modelled ozone for all input information (L1, L2, L3 and ALL) using the regression model.

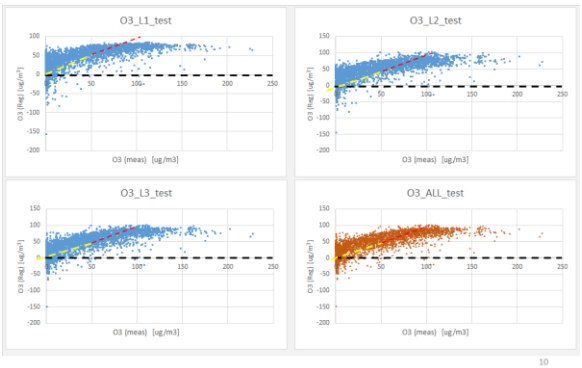

Figure 6: Modeled ozone against observed by L1, L2, L3 and ALL simulations: regression model

If we concentrate our attention in the middle range(red line in Figure 6), the L1 cases underestimate the observed (see Table 1).

The increase of information improves the observed ozone in the middle range when using the regression model (going from left to right in Table 1).
The L3 simulation is the best (average $O3_{lin}(L3)=67.5\pm16.9$ μg/m^3 as compared to the observed $O3_{meas}=74.4\pm14.8$ μg/m^3 - see Table 1).

	O3_meas	O3_L1	O3_L2	O3_L3	O3_ALL
Average	74.4	64.2	64.5	67.5	67.5
SD	14.8	12.7	15.2	16.9	17.0

Table 1: average measure ozone in the middle range. Comparison with regression model results for different information levels.

For the upper range of observations, all simulations show a worse behavior. The linear models doesn't succeed in forecasting higher ozone values ($O3_{lin}(L1\text{-}L3)= 70.9\div79.3$ μg/m^3 compared to the observed $O3_{meas}=122.5\pm21.0$ μg/m^3 - see Table 2).

	O3_meas	O3_L1	O3_L2	O3_L3	O3_ALL
Average	122.5	70.9	74.9	78.9	79.3
SD	21.0	13.3	16.3	14.9	15.1

Table 2: average measure ozone in the upper range. Comparison with regression model results for different information levels

For the lower concentration range (observed under 50 μg/m^3), the linear models produce an overestimation ($O3_{lin}(L1\text{-}L3)= 29\pm25$ μg/m^3 as compared to the observed $O3_{meas}=14.2\pm15.4$ μg/m^3 - see Table 3).
Furthermore, no significance values of the correlation between the model and observations is obtained (yellow dashed line in Figure 6).

	O3_meas	O3_L1	O3_L2	O3_L3	O3_ALL
Average	14.2	30.0	29.1	26.9	26.8
SD	15.4	25.4	25.1	24.9	24.9

Table 3: average measure ozone in the lower range. Comparison with regression model results for different information levels.

The ALL simulation does not improve the performance at all with respect to L1-L2 and L3. Table 4 indicates a flat behavior for R-squares starting from the L2 simulation.

	L1	L2	L3	ALL
R^2-TEST	0.47	0.58	0.58	0.58

Table 4: L1, L2, L3 and ALL simulations. R-squares values by regression model for overall dataset.

Table 5 reports the average negative values using the linear models (all points under black dashed line in Figure 6). The negative values decrease with increasing values of input

information (from the value of -17.0 μg/m^3 for L1 up to -15.4 μg/m^3 for the ALL simulation).

	L1_Reg	L2_Reg	L3_Reg	ALL_Reg
O3_avg(meas)	7,49	8,32	6,95	6,93
O3_Reg	-16,97	-16,76	-15,70	-15,40

Table 5: average regression modelling ozone for negative contribution. Comparison with observations and information levels.

6.2 Neural Network results

Figure 7 compares simulations trained using a neural network model. Improvement over the linear models is evident.

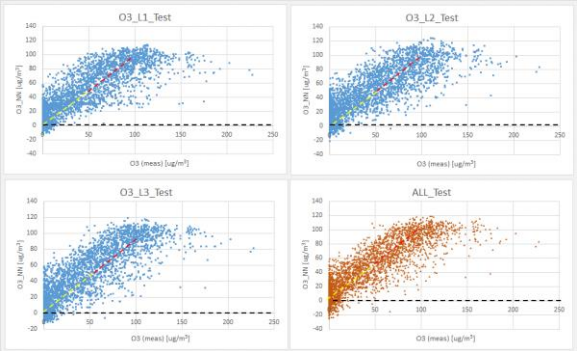

Figure 7: Modeled ozone against observed by by L1, L2, L3 and ALL simulations: neural network

Table 6 reports the average predicted values in the middle range. The modeled ozone values provide very good predictions. The observed ozone is $O3^{NN}_{meas}$=74.0±14.5 μg/m^3 when the NN model produces about 72 μg/m^3.

	O3_meas	O3_L1	O3_L2	O3_L3	O3_ALL
Average	74.0	72.6	71.5	72.5	72.5
SD	14.5	21.8	22.8	23.2	22.8

Table 6: average measured ozone in the middle range. Comparison with NN model results for different information levels.

The results for the upper range are reported in table 7. The observed ozone values ($O3^{NN}_{meas}$=121.7±21.1 μg/m^3) are slightly underestimated by the neural network ($O3^{NN}_{model}$ increase starting from 87.3 μg/m^3 up to 93.0 μg/m^3 for ALL).

It is worth noting that the L3 and ALL simulations predict quite similar levels. This fact indicates no improvement due to adding the new variables from MAD scheme. While the performance of the MAD scheme seems to be insensitive to the choice of input information for linear models, the neural network appears to be much more sensible to the input data defined by the MAD scheme.

The NN advantages are particularly evident in the lower range simulation (Table 8).

	O3_meas	O3_L1	O3_L2	O3_L3	O3_ALL
Average	121.7	87.3	89.7	92.1	93.0
SD	21.1	17.7	17.7	16.1	16.1

Table 7: average measured ozone in the upper range. Comparison with NN model results for different information levels.

	O3_meas	O3_L1	O3_L2	O3_L3	O3_ALL
Average	14.2	22.9	23.6	22.5	21.5
SD	15.4	22.4	23.3	22.5	22.2

Table 8: average measure ozone in the lower range. Comparison with NN model results for different information levels.

The negative values as computed by the NN model (Table 9) are much better than by the linear model (see Table 5). In fact, we have found an underestimation of ozone between -4.3 μg/m^{3}up to -6.9 μg/m^3. (against the measured values between 1.9 μg/m^3 up to 3.4 μg/m^3).

The ALL simulation produces higher negative values (about -7 μg/m^3) respect to L3 simulation.

	L1_NN	L2_NN	L3_NN	ALL_NN
O3_avg(meas)	1,90	2,55	2,98	3,41
O3_NN	-4,29	-4,23	-4,19	-6,90

Table 9: average NN modelling ozone for negative contribution. Comparison with observations and information levels

Significant results are also evident in the R-square coefficients for the different simulations (L1, L2, L3 and ALL-see Table 10).

	L1	L2	L3	ALL
R^2-TEST	0.69	0.68	0.71	0.73

Table 10: L1, L2, L3 and ALL simulations. R-squares values by NN model for overall dataset.

6.3 Discussion

We have confirmed that prediction of ozone for urban area is best reproduced by using non-linear models.

The application of the MAD scheme to select data is always efficient, independently by the assumed information. From the results of the ALL simulation, not all of the information added to a MAD scheme enhances the performance.

The selection of input data using equations (2), (3) and (4) can be considered a winning strategy for forecasting the behavior of the photochemical reaction in the urban area of Rome.

7. Conclusion

The MAD analysis provides a conceptual model for the optimization of input variables to reproduce target data (e.g., Ozone) by neural networks or regression models.

For modeling photochemical reactions of air pollution, the MAD analysis demonstrates the necessity classifying model inputs in information levels. The analysis of the MAD scheme for the photochemical production of ozone suggests three information levels (L1, L2 and L3) is appropriate.

We have also tested the inclusion of the variables external to photochemical smog (as in the ALL simulation) to verify the the MAD choice.

We find that introducing the exogenous time variable (named h) into the NN, provides useful information. The variable h can be linked with the Monin- Obukhov length. Here we use a one dimensional transformation for (h).

The new time variable improves the accuracy of the NN with respect to a MLR model as well as the performance with respect to the utilization of the typical reactants for ozone production (NO-NO2) and to the physical conditions (T,RH,SR).

The MAD analysis demonstrates, for the evaluation of photochemical reaction of air pollution, the necessity of classifying input information into three levels. The results show that model performance increases with increasing amounts of input information. Adding new information (as in the ALL formulation) does not produce any significant increase in performance for the L3 level input data.

The MAD analysis focuses on the optimal data selection. Our preliminary results demonstrate that model performance is linked to a correct choice of input data associated with the physics rather than with variables chosen by blind statistical selection.

REFERENCES

Blocken, B., Carmeliet, J., Stathopoulos, T. CFD evaluation of wind speed conditions in passages between parallel buildings-effect of wall-function roughness modifications for the atmospheric boundary layer flow. Journal of Wind Engineering and Industrial Aerodynamics, 95 (9-11), 2007, pp. 941-962.

Chen, F., Kusaka, H., Bornstein, R., Ching, J., Grimmond, C.S.B., Grossman-Clarke, S., Loridan, T., Manning, K.W., Martilli, A., Miao, S., Sailor, D., Salamanca, F.P., Taha, H., Tewari, M., Wang, X., Wyszogrodzki, A.A., Zhang, C. The integrated WRF/urban modelling system: Development, evaluation, and applications to urban environmental problems. International Journal of Climatology, 31 (2), 2011, pp. 273-288.

Comrie, A.C.. Comparing neural networks and regression models for ozone forecasting. Journal of the Air and Waste Management Association, 47 (6), 1997, pp. 653-663.

Feng, Y., Zhang, W., Sun, D., Zhang, L.. Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification. Atmospheric Environment, 45 (11), 2011, pp. 1979-1985.

Finlayson-Pitts, Barbara J., and James N. Pitts Jr. Chemistry of the upper and lower atmosphere: theory, experiments, and applications. Academic press, 1999.

Gardner, M.W., Dorling, S.R. Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences. Atmospheric Environment, 32 (14-15), 1998, pp. 2627-2636.

Gardner, M.W., Dorling, S.R.. Statistical surface ozone models: An improved methodology to account for non-linear behaviour. Atmospheric Environment, 34 (1), 2000, pp. 21-34.

Ibarra-Berastegi, G., Elias, A., Barona, A., Saenz, J., Ezcurra, A., Diaz de Argandoña, J.. From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao. Environmental Modelling and Software, 23 (5), 2008, pp. 622-637.

Lighty, J.S., Veranth, J.M., Sarofim, A.F. Combustion aerosols: Factors governing their size and composition and implications to human health. Journal of the Air and Waste Management Association, 50 (9), 2000, pp. 1565-1618.

McElroy, J.L., Smith, T.B. Vertical pollutant distributions and boundary layer structure observed by airborne lidar near the complex southern California coastline. Atmospheric Environment (1967), 20 (8), 1986. pp. 1555-1566.

Ortiz-García, E.G., Salcedo-Sanz, S., Pérez-Bellido, Á.M., Portilla-Figueras, J.A., Prieto, L.. Prediction of hourly O3 concentrations using support vector regression algorithms. Atmospheric Environment, 44 (35), 2010, pp. 4481-4488.

Pelliccioni, A., Monti, P., Gariazzo, C., Leuzzi, G. Some characteristics of the urban boundary layer above Rome, Italy, and applicability of Monin-Obukhov similarity. Environmental Fluid Mechanics, 12 (5), 2012, pp. 405-428.

Pelliccioni, A., Tirabassi, T. Air dispersion model and neural network: A new perspective for integrated models in the simulation of complex situations. Environmental Modelling and Software, 21 (4), 2006, pp. 539-546.

Seinfeld, J.H., and Spyros N. P. Atmospheric chemistry and physics: from air pollution to climate change. John Wiley & Sons, 2012.

Yee, E., Gailis, R.M., Hill, A., Hilderman, T., Kiel, D. Comparison of Wind-tunnel and Water-channel Simulations of Plume Dispersion through a Large Array of Obstacles with a Scaled Field Experiment, Boundary-Layer Meteorology, December, 121(3), 2006, pp 389-432.

Zhang, Z., Chen, Q., Comparison of the Eulerian and Lagrangian methods for predicting particle transport in enclosed spaces Atmospheric Environment, 41(25) 2007, pp 5236-5248.