Forecasting powder dispersion in a complex environment using Artificial Neural Networks

Pierre Lauret, Frederic Heymes, Serge Forestier, Laurent Aprin, Alexis Pey, Marcia Perrin

Forecasting powder dispersion in a complex environment using artificial neural networks

Pierre Lauret[a], Frederic Heymes[a], Serge Forestier[b], Laurent Aprin[a], Alexis Pey[b], Marcia Perrin[c]

[a]Institute of Risk Science (ISR), Ecole des mines d'Alès, Alès, France

[b]Tuev Sued, Mattenstrasse 24, CH-4002 Basel

[c]Novartis Pharma AG, Lichstrasse 35, CH-4002 Basel

Corresponding author: pierre.lauret@mines-ales.fr

Highlights

- Atmospheric powder dispersion is modeled in a complex urban area.
- 290 daily mean concentration measurements are recorded.
- Artificial Neural Network (ANN) model is trained and evaluated.
- The ANN model satisfies air quality model evaluation criteria.
- The ANN model computing time is nearly instantaneous (less than one second).

1. Abstract

Atmospheric dispersion prediction skill is required for any industry processing hazardous material. This is a sensitive task since many parameters are involved: source term, atmospheric conditions, and local configuration. Behavior of dust dispersion is difficult because of the diameter scattering, agglomeration, sedimentation, range of densities… Furthermore, production sites may be located inside a complex environment such as urban areas, where accuracy of classical dispersion models is low. This paper aims to evaluate the efficiency of an Artificial Neural Networks (ANN) model to predict dust dispersion in an urban area without prior knowledge of the source term. The experimental database consists of 290 daily mean concentration measurements on a site located 500m away from the emission source. The inputs are selected from meteorological data from a MeteoSwiss station located 4.5km south. The training phase is done through early stopping application. ANN model selection is performed on the best coefficient of determination value. Model performance is evaluated using classical air quality criteria and shows good results. Nevertheless, ANN model tends to underestimate high concentrations while overestimating low concentrations. Results are included within acceptable range. Improvements can be achieved by adding information of the source term as an input for the ANN model.

2. Keywords

Atmospheric dispersion modeling, Artificial Neural Networks, dust dispersion, Complex environment

3. Introduction

Atmospheric dispersion prediction ability is required for any industry processing hazardous materials. One can distinguish chronic pollution resulting from continuous release with low level of concentration and accidental pollution resulting from an accidental event (leakage, human error …). Industrial processes can generate such chronical concentrations in the atmosphere. The World Health Organization has recently published a warning report on air quality (WHO, 2016) that links particulate solids under 10 μm and 2.5 μm level and the increase of several diseases (stroke, heart disease, lung cancer, chronic and acute respiratory diseases, asthma). Each country institution set limits to material emission and impose studies of the behavior of substance in case of release. For example, the Swiss legislation for the protection against major accidents (OPAM) requires that the accidental release of a highly active powder should be studied before

giving the authorization of production. Annex 4 (332 and 333) of the order 814.012 refers to the study of different cases of dispersion and the consequences for both people and the environment. For most simple cases, classical models are perfectly indicated. In free field, Gaussian models are accurate. In more complex situations, Computational Fluid Dynamics models are required. However, there are situations where explicit modelling is ineffective. These are more particularly situations where the source term is unknown, as in the case of diffuse industrial emissions, uncontrolled discharges. There is currently no method for predicting the effects of such poorly known emissions. This paper focuses on the study of the behavior of an undefined release of a powder in a complex environment: urban, including the presence of a river. To model this situation, Artificial Neural Network method is implemented due to its efficiency in complex situations where direct modelling is difficult (Lauret et al., 2016a). To do so, there is a need to process a dust monitoring combined with registration of meteorological data near emission site to build, verify and validate a dispersion model. The developed model here is thus based on real life experiments as there is a monitoring database available with over 2 years of registration. It gives daily mean concentration of particles at a station 500 m away from the emission area. It is distinct from classical tools because of the intrinsic consideration of the emission area environment with nearly instantaneous results.

3.1. Modeling of dust dispersion in lack of knowledge of source term

Numerous parameters impact powder dispersion in the atmosphere and are required to accurately model the dispersion. Characterization of the source term gives the initialization of the model. Depending on the aim of the modeling, it can be fully defined, or estimated from qualitative or qualitative observations in terms of flow rate, velocity and physical properties. Meteorological parameters are directly linked to the dispersion due to the implication in the atmospheric flow equations. Once again, to model correctly, it is important to get information closer to the emission source to be as accurate as possible. Finally, direct environment from the emission source to the monitoring point define how the flow is influenced by the buildings in the vicinity, orography and surface roughness (road, grass, water…) (Hosker Jr, 1985).

Usual modeling techniques are Computational Fluid Dynamics or CFD models, integral models or Gaussian models. Each model presents advantages and limits. Gaussian models correspond to an analytical resolution of the advection equation, using standard deviations to calibrate turbulent diffusion coefficients. Their limitations are due to hypothesis required to solve the equation: atmospheric boundary layer characterization is evaluated according to several categories, specific obstacles are not considered, material is considered as passive, accuracy is better in far field. These coefficients were tested for urban environment using Indianapolis experiments (Hanna et al., 1999). Despite these limitations, these models are widespread because of their easiness and quickness of use and the regulator acceptability. Intermediate simplified CFD models exist, but increase both complexity and computing time. Diagnostic wind flow models like mass-consistent models are able to reconstruct a steady-state wind field from initial experimental data, while keeping predictions of orography effects (Castellani et al., 2015). They are based on simplified steady-state solutions of the Navier-Stokes equations. Eulerian models from Computational Fluid Dynamics solved these equations based on finite elements method. At first, the wind flow is determined. Turbulence is solved using closure equations of the system. These equations are transport equations of turbulent quantities. The turbulent diffusion coefficient is introduced in the advection-diffusion equation to model the dispersion. Dispersion can be computed using a Lagrangian method: behaviors of particles are followed and are proportionally linked to concentration, depending on initial conditions.

Some alternative models using statistical techniques attempt to speed up classical CFD techniques. Stavrakakis et al. (2011) develop an Artificial Neural Networks (ANN) model to yield relationships between air velocity and geometrical characteristics. Then, the near-optimal geometrical solution is computed using the CFD model, saving several CFD simulations, each

one representing approximately 10h CPU time. Vendel (2011) generates an important database of different possible wind fields around a specific site. The time spent to build this large database is spared when an event occurs: the trajectory simulation is computed using the interpolate velocity field according to the actual meteorological conditions recorded on the site. A Lagrangian approach is thus used. Another approach consists in creating a large database of CFD calculations in order to give knowledge of fluid mechanics equations to statistical tools like ANN (Lauret et al., 2016b). Again, computing time is reported in a learning phase while operating phase is nearly instantaneous. Moreover, Cao (2007) carried out a study to determine, by ANN method, the dispersion coefficients used in the Gaussian model. This study helps to improve Gaussian models by adding continuous standard deviation values, adapted to every specific configuration. ANN has already been used to forecast tracer concentrations at a given site using spatially distributed sensors (Podnar et al., 2002). A rudimentary comparison with traditional statistical methods revealed that the ANN performed better and showed fewer limitations as a tool for tracer modeling, especially for long-term prediction. Boznar et al., (1993, 2004) present a model using meteorological values (air temperature, global solar radiation, wind speed, wind direction, maximal air temperature) and previous pollutant concentrations (NO, NO2, NOx, CO, O3) to perform a 12-hours forecasting of Ozone concentrations. This model shows sufficient capabilities to inform citizens about possibilities of high and alarm concentrations

The study case developed here corresponds to the dispersion of highly active powder from a source located in a complex environment. Concentrations of the powder are registered at several stations away from the emission sources, scattered around the city. Daily mean concentrations from November the 13th 2013 to April the 1st 2015 are integrated to feed the database. Information about the source term is limited to the location. Meteorological data are retrieved from a MeteoSwiss station located at 4.5 km at the south of the emission source. The situation presented here is specific because of the complexity of the site (buildings, Rhin river, car circulation …), the difficulty to determine source term and the need for fast modeling. In this context, machine learning tools like Artificial Neural Networks (ANN) are of particular interest. The aim of this model is to forecast daily mean concentration at a specific location using ANN without knowledge of the emission source term.

3.2. Artificial Neural Networks

ANN are machine learning models based on the systemic paradigm and are able to identify a nonlinear behavior from a database without physical assumption acting as a black box. The Multilayer Perceptron ANN model used here has two essential properties. Universal approximation (Hornik et al., 1989) and parsimony (Barron, 1993) make it able to predict efficiently future behaviors on never encountered situations within the variables range of the examples database. A neuron is a nonlinear, parameterized, bounded function. Variables are assigned to the inputs of the neuron. The output of a neuron is the result of nonlinear combination of the inputs, weighted by the parameters and using an s-shaped function like a sigmoid. A neural network is the composition of several neurons. Parameters calibration is done through application of an algorithm using the training database and designed to decrease the model error. In this work, the Levenberg-Marquardt method is used (Hagan and Menhaj, 1994). The function realized by the ANN is continuously tested on a disjoined set of examples, namely the validation set. This set is employed to avoid overtraining using early stopping (Sjöberg et al., 1995). Lastly, performances of the model must be measured on another set, never used during training or stopping: the test set.

4. Material and method

4.1. Example database creation

Database consists of 290 daily mean concentrations measurements on a site located 500m away from the emission source recorded from November the 13th 2013 to April the 1st 2015. The daily monitored mean concentration dataset shows a large discrepancy in the values. Figure 2 represents both the concentration distribution and the cumulative sum of elements in the sample. More than 85 % of the concentrations are under 0.05 $\mu g.m^{-3}$. Meanwhile, most important concentrations to forecast are the high values of concentrations. Extreme values of the distribution, over 0.15 $\mu g.m^{-3}$ represent less than 1.4 % of the dataset. It is a difficult challenge to forecast such an unbalanced distribution, according the importance to the inputs selection and training procedure.

In order to optimize the training process of the ANN, it is first important to feed it with variables directly affecting the daily mean concentration at the station. As said before, data on the emission source term are not available. Thus, the only available data are linked to the meteorology. These data are collected from Basel MeteoSwiss station, located 4.5 km south from the emission.

Hourly data are provided during the period of interest and reported in table 1:

As the goal of this work is to provide the mean daily concentration at a specific location, data has to be modified on daily basis. Meteorological data was processed to fit day duration. In order to avoid the gap between 0° and 360° in the direction of the wind, values of both the cosinus and the sinus are used as inputs for the ANN. All temperature values were integrated in only two inputs parameters: mean and standard deviation of the whole set of daily max, min and mean temperature. The mean and standard deviation values were also used for the following parameters: relative humidity, sun radiation, wind direction, wind velocity, wind gust and atmospheric pressure. For the rain and the shinning sun duration, the total sum of the day was used. It results in 17 different inputs for the neural networks as seen in the box plot in figure 3:

As we can see on this figure, values are not comparable one to another without additional mathematical correction. To ensure that the training algorithm is not going to promote one variable instead of another, it is important to reduce the data in the same interval.

4.2. ANN Architecture

Architecture of the ANN is directly linked to his ability to forecast concentrations at the considered location. As the number of examples is low, several different architectures have to be tried and evaluated in order to select the best model. In this work, Levenberg-Marquardt algorithm has been used. In order to avoid an overtraining of the data set, early stopping is used. It consists of dividing the database in three parts:

- The training set represents 80% of the database.
- The stop set is used to avoid overtraining: when the mean squared error stops decreasing on it, the training phase is interrupted.
- The test set is used to assess the model quality.

Previous works (Lauret et al., 2016b) present the need to study the proper number of neurons in hidden layer. Moreover, several initializations are required to correctly train the ANN. Indeed, due to the relatively small size of the example database, a focus on the best model selection is made. The coefficient of determination is used to evaluate the performance of the training phase of the neural network (Kong A Siou et al., 2012). In this work, 1 to 30 neurons in hidden layer have been tried. Each ANN has been trained with 20 initializations to ensure the best fitting to the concentrations

4.3. Performance criteria

A forecasting model has to be evaluated against measured data. In air quality, several criteria are used to do so. Chang and Hanna (2004) proposed four different quality criteria. It is important to use all of them to avoid misunderstanding of the model performance. In this work, the factor of two (FAC2), the Normalized Mean Squared Error (NMSE), the Fractional Bias (FB) and the coefficient of determination (R^2) are used. The last one replaces the coefficient of correlation because of its use in neural networks application. Expressions of these criteria are detailed in the following table:

As previously mentioned, each one has an importance. The target values for these criteria are as following: R^2 and FAC2 is one; FB and NMSE is 0. FB measures the systematic errors which lead to always underestimate or overestimate the measured values. FB values range between -2 (extreme underprediction) to 2 (extreme overprediction). Therefore, matching perfect target FB value does not mean perfect modeling, because of possible cancelling errors. NMSE measures systematic and random errors. As detailed by Hanna and Chang (2012), acceptable values are :

- $|FB| \lesssim 0.67$, i.e., the relative mean bias less than a factor of ~ 2
- $NMSE \lesssim 6$, i.e., the random scatter $\lesssim$ 2.4 times the mean
- $FAC2 \gtrsim 0.3$, i.e., the fraction of y within a factor of two of y^p that exceeds 0.3

In the following section, these criteria are used to evaluate the model performance to forecast the concentrations at a specific site.

5. Results and discussion

5.1. Best model selection

The selection of the best ANN model is based on the evaluation of the coefficient of determination on the test set of the examples database. Figure 4 represents the evolution of the coefficient of determination with the increase of the number of neurons in hidden layer. Considering the median value of 20 initializations for each number of hidden neurons, one can see an improvement of the performance of the neural network up to 6 neurons in hidden layer. Above this value, the coefficient of determination decreases and reaches a chaotic zone from 9 to 26 neurons in hidden layer. Values above 26 show a stabilization of the R^2 value. If only the maximum value is considered, then the best performance is evaluated with 3 neurons in hidden layer. An increase in the maximum value is observed from 27 to 30 neurons in hidden layer.

Initialization plays an important role in the best ANN selection because of the training dependency. As shown in figure 5, the median R^2 value from 20 initializations corresponds to most of the ANN training. Four training show different behavior, three are out of acceptability range whereas one gives an improvement in the training phase.

The best selected model corresponds to the initialization #5 of the neural network with 3 neurons in the hidden layer. This model is evaluated in the results and discussion section.

5.2. ANN model evaluation

In the training phase three data sets were created. The test set is used to evaluate the performance of the ANN versus the measured data. It is composed of 29 examples representing both high and low concentrations. Results in term of air quality criteria are reported here:

According to the guidelines on air quality performance criteria, the ANN model gives good results. The value of the FAC2 is above 0.3. NMSE is under 1 while fractional bias criterion is close to 0. It indicates that no systematic errors are made, even if random errors may be present. From a global point of view, the model neither under nor over estimates the concentrations. The coefficient of determination indicates a value of 0.6 that is quite encouraging. To better

understand the results, figure 6a and 6b show the scatter of data observed and forecasted. In 6a, only the test set has been represented while all the concentrations are plotted in 6b.

In figure 6a, almost all data are included in the factor of two of the observed concentrations. There are overestimations of the low level concentrations, inferior to 0.04 $\mu g.m^{-3}$. Inversely, concentrations above 0.5 $\mu g.m^{-3}$ are underestimated, within a factor of two. These first conclusions are also present on figure 6b where all the concentrations of the dataset are considered. The concentrations of the validation set are lightly underestimated ($FB \simeq 0.88$). Considering most important values of forecasted concentration on the training set, the major part is underestimated while other examples reach the perfect match. Values around zero are both under or over estimated. In terms of computing time, results are obtained nearly instantaneously (less than one second), enhancing improvement compare to CFD models that require more than one hour. Knowledge of the behavior of the model is very important because of the trend to underestimate high values. While operating, this ANN model might be used to foresee precautions measure if needed or adaptive actions on the process.

6. Conclusions

The aim of this work is to prove that using ANN to forecast powder concentrations without knowledge of the emission source is possible. The presented results emphasize this possibility. The proposed dispersion model forecasts concentrations of a powder at a specific location. The emission source corresponds to a production site distant of 500 m. The configuration of the environment can be considered as complex because of the presence of numerous buildings, roads, parks and a river. An Artificial Neural Networks is used to perform this forecasting. The training of the model is achieved through the use of 290 daily mean concentrations recorded from November the 13th 2013 to April the 1st 2015. The measured concentrations were clustered in three set to process the training phase of the ANN. Each set has to correctly represent the whole set so stratified data sampling was realized. The validation of the model was based on the coefficient of determination, the factor of two, the fractional bias and the normalized mean square error. Acceptable values of these criteria are reached which confirms the possibility to use ANN even in case of poor knowledge of the source term even if the ANN model tends to underestimate high concentrations while low concentrations are overestimated. This behavior has to be known before using the model. Nevertheless, improvements can be done to the model by supplying additional information about the source term as inputs for the neural network. Likewise, the MeteoSwiss station is located at more than 4.5 km south and can possibly be not representative of the atmospheric flow at the emission source. Moreover, another argument for adding meteorological data acquisition system at the emission source is that urban flow of air is very specific and can largely vary with small shifting of the measure. Moreover, despite the number of examples that can be considered as low, modeling is satisfying. As the sampling is going on, database grows and better trainings can be realized. Authors suggested the implementation of a self-training algorithm, improved each day from data acquired the day before. The next goal of this study is to forecast concentrations for each measurement station and realize a mesh on the entire city.

7. References

Barron, A.R., 1993. Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inf. Theory 39, 930–945. doi:10.1109/18.256500

Boznar, M., Lesjak, M., Mlakar, P., 1993. A neural network-based method for short-term predictions of ambient SO2 concentrations in highly polluted industrial areas of complex terrain. Atmos. Environ. Part B. Urban Atmos. 27, 221–230. doi:10.1016/0957-1272(93)90007-S

Boznar, M.Z., Mlakar, P., Grasic, B., 2004. Neural networks based ozone forecasting, in: 9th International Conference on Harmonisation within Atmosphéric Dispersion Modelling for Regulatory Purposes. pp. 356–360.

Castellani, F., Astolfi, D., Burlando, M., Terzi, L., 2015. Numerical modelling for wind farm operational assessment in complex terrain. J. Wind Eng. Ind. Aerodyn. 147, 320–329. doi:10.1016/j.jweia.2015.07.016

Chang, J.C., Hanna, S.R., 2004. Air quality model performance evaluation. Meteorol. Atmos. Phys. 87, 167–196. doi:10.1007/s00703-003-0070-7

Hagan, M.T., Menhaj, M.B., 1994. Training Feedforward Networks with the Marquardt Algorithm. IEEE Transations Neural Networks 5, 989–993.

Hanna, S., Chang, J., 2012. Acceptance criteria for urban dispersion model evaluation. Meteorol. Atmos. Phys. 116, 133–146. doi:10.1007/s00703-011-0177-1

Hanna, S.R., Egan, B.A., Purdum, J., Wagler, J., 1999. Evaluation of the ADMS, AERMOD, and ISC3 dispersion models with the optex, duke forest, kincaid, indianapolis, and lovett field data sets. Int. J. Environ. Pollut. 3.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer Feedforward Networks are Universal Approximators. Neural Networks 2, 359–366.

Hosker Jr, R.P., 1985. Flow and Diffusion Near Obstacles, in: Atmospheric Science and Power Production. pp. 241–326.

Kong A Siou, L., Johannet, A., Valérie, B.E., Pistre, S., 2012. Optimization of the generalization capability for rainfall-runoff modeling by neural networks: The case of the Lez aquifer (southern France). Environ. Earth Sci. 65, 2365–2375. doi:10.1007/s12665-011-1450-9

Lauret, P., Perrin, M., Heymes, F., Aprin, L., Slangen, P., Pey, A., Steinkrauss, M., 2016a. Atmospheric powder dispersion in an urban area. Chem. Eng. Trans. 48, 91–96. doi:10.3303/CET1648016

Lauret, P., Heymes, F., Aprin, L., Johannet, A., 2016b. Atmospheric dispersion modeling using Artificial Neural Network based cellular automata. Environ. Model. Softw. 85, 56–69. doi:10.1016/j.envsoft.2016.08.001

Podnar, D., Koračin, D., Panorska, A., 2002. Application of artificial neural networks to modeling the transport and dispersion of tracers in complex terrain. Atmos. Environ. 36, 561–570. doi:10.1016/S1352-2310(01)00446-0

Sjöberg, J., Zhang, Q., Ljung, L., Benveniste, A., Deylon, B., Glorennec, P.Y., 1995. Nonlinear Black-Box Modeling in System Identification: a Unified Overview. Automatica 31, 1691–1724.

Stavrakakis, G.M., Karadimou, D.P., Zervas, P.L., Sarimveis, H., Markatos, N.C., 2011. Selection of window sizes for optimizing occupational comfort and hygiene based on computational fluid dynamics and neural networks. Build. Environ. 46, 298–314. doi:10.1016/j.buildenv.2010.07.021

Vendel, F., 2011. Modélisation de la dispersion atmosphérique en présence d'obstacles

complexes : application à l'étude de sites industriels. Ecole Centrale de Lyon.

WHO Media centre, 2016. Ambient (outdoor) air quality and health [WWW Document]. Fact sheet. URL http://www.who.int/mediacentre/factsheets/fs313/en/

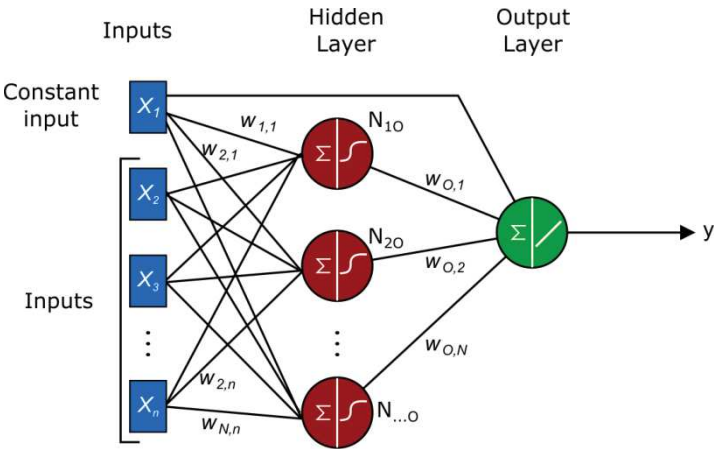

Figure 1: Multilayer perceptron scheme

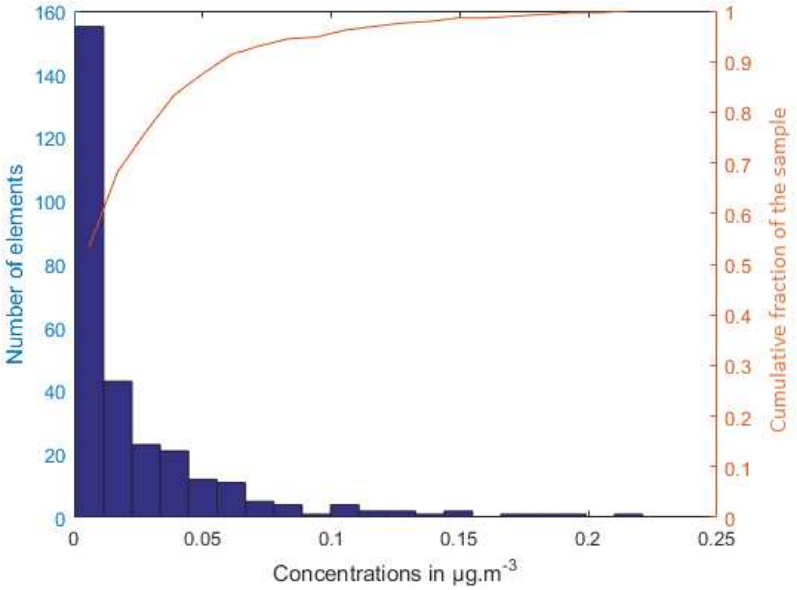

Figure 2: Concentrations distribution

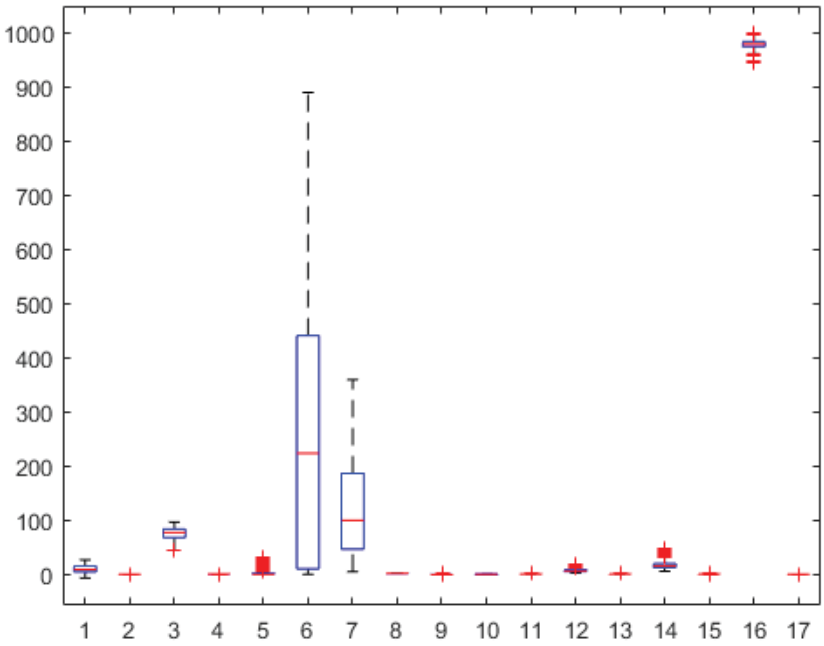

Figure 3: box plots of the 17 different inputs of the ANN

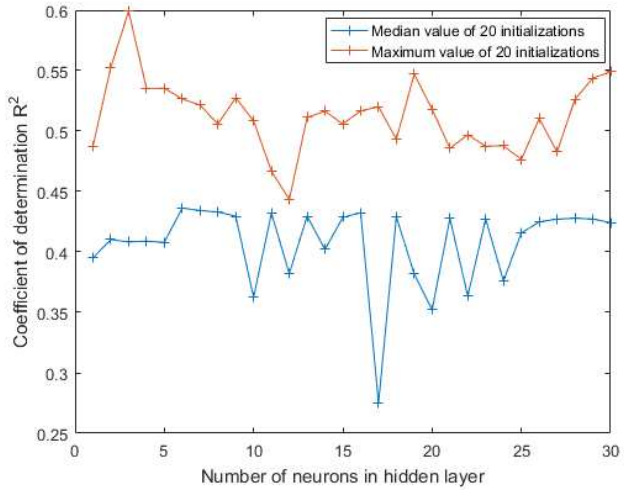

Figure 4: Coefficient of determination evolution with number of neurons in hidden layer

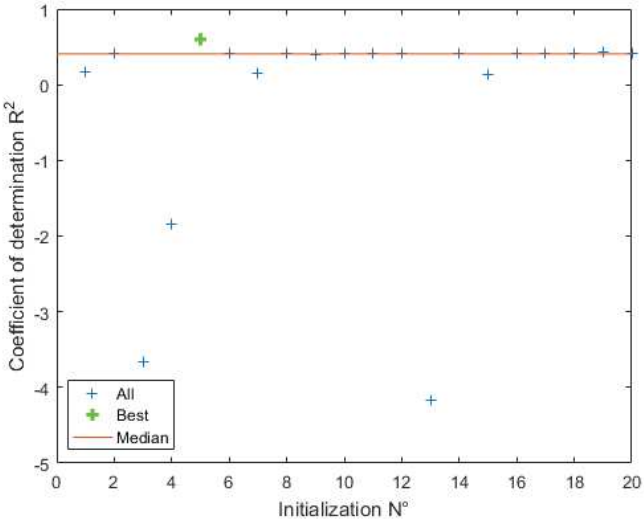

Figure 5: Coefficient of determination as a function of weights initialization

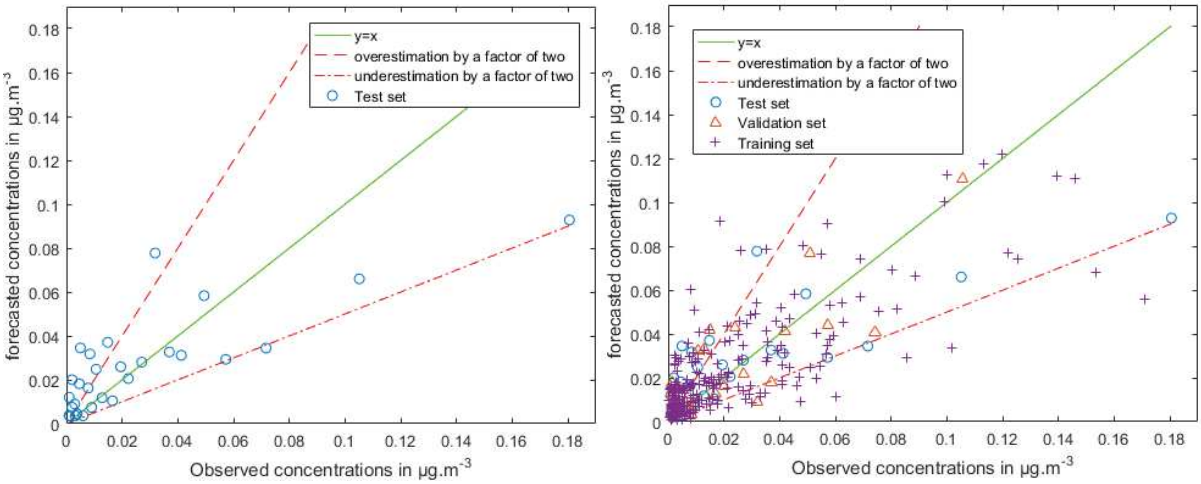

Figure 6: a) scatter plot of ANN model versus observed concentrations (test set only) – b) scatter plot of ANN model versus observed concentrations (all concentrations available).

Table 1: Hourly data from MeteoSwiss station

Variable	Temperature			Relative humidity	Total rain	Sun duration	Global radiation	Wind				Atmospheric pressure
	Mean	Min	Max					Direction	Mean	Max	Gust	
Unity	°C			%	mm	min	$W.m^{-2}$	°	$Km.h^{-1}$			hPa
Daily modification	Mean and standard deviation			Mean and standard deviation	Sum	Sum	Mean and standard deviation	Mean and standard deviation of both the cosinus and sinus	Mean and standard deviation			Mean and standard deviation

Table 2: Performance criteria used in the study - y^p represents observed value; y represents measured value.

Criteria	R2	FAC2	FB	NMSE
Expression	$R^2 = 1 - \frac{\sum_i (y^p - y)^2}{\sum_i (y^p - \overline{y^p})^2}$	Fraction where : $0{,}5 \leq \frac{y}{y^p} \leq 2$	$FB = 2\frac{(\overline{y^p} - \bar{y})}{(\overline{y^p} + \bar{y})}$	$NMSE = \frac{\overline{(y^p - y)^2}}{\overline{y^p}\,\bar{y}}$
Target value	1	1	0	0
Remarkable value	0: Model is equivalent of mean of possible value	-	Positive value: global under estimation Negative value: global over estimation	-

Table 3: Evaluation of performance criteria on the test set.

Performance criteria	R^2	FAC2	NMSE	FB
Value	0.6	0.55	0.7	-0.02