Computation of the airborne contaminant transport in urban area by the artificial neural network

A. Wawrzynczak[1,2][0000−0001−8292−6875] and
M. Berendt-Marchel[1][0000−0001−7204−7367]

[1] Siedlce University, Faculty of Exact and Natural Sciences, Institute of Computer Sciences, Poland
[2] National Centre for Nuclear Research, Swierk-Otwock, Poland
anna.wawrzynczak-szaban@uph.edu.pl

Abstract. Providing the real-time working system able to localize the dangerous contaminant source is one of the main challenges of the cities emergency response groups. Unfortunately, all proposed up to now frameworks capable of estimating the contamination source localization based on recorded by the sensors network the substance concentrations are not able to work in real-time. The reason is the significant computational time required by the applied dispersion models. In such reconstruction systems, the parameters of the given dispersion model are sampled to fit the model output to the registrations; thus, the dispersion model is run tens of thousands of times. In this paper, we test the possibility of training an artificial neural network (ANN) to effectively simulate the atmospheric toxin transport in the highly urbanized area. The use of a fast neural network in place of computationally costly dispersion models in systems localizing the source of contamination can enable its fast response time. As a training domain, we have chosen the center of London, as it was used in the DAPPLE field experiment. The training dataset is generated by the Quick Urban & Industrial Complex (QUIC) Dispersion Modeling System. To achieve the ANN capable of estimating the contaminant concentration, we tested various ANN structures, i.e., numbers of ANN layers, neurons, and activation functions. The performed tests confirm that trained ANN has the potential to replace the dispersion model in the contaminant source localization systems.

Keywords: machine learning · neural networks · airborne contaminant transport computation.

1 Work motivation

The main task of the emergency response groups existing in all cities is a quick reaction to any threats to people and the environment. The primary factor determining the success or failure of a given action is the response time. Nowadays, the chemicals are used in most areas of the industry, making the transport and

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_30

storage of the toxic materials pose a constant risk of releasing it into the atmosphere. In the cases when the source of the failure resulting in releasing the toxin into the atmosphere is known, the emergency responders can quickly undertake all necessary actions to minimize the consequences of such release. The more challenging are situations when the sensors, distributed over a city, report the non-zero concentration of the dangerous substance, which source is not known. In such cases, important is to have a system able to, in a real-time estimate the most probable location of the contamination source based solely on the concentration data reported from the sensors network. The algorithms that can cope with the task can be divided into two categories. First are based on the backward approach, but those are dedicated to the open areas or a continental-scale problem. Second, are based on the forward approach. In this case, the appropriate dispersion model parameters are sampled (among them source location) to chose the one giving the smallest distance measure between the model outputs and sensors measurement in considered spatial domain.

Such an inverse problem has no unique analytical solution but might be analyzed with probabilistic frameworks, as the Bayesian approach, where all searched quantities are modeled as random variables. Bayesian approach transforms the inverse mentioned above problem into searching for a posterior distribution based on the sampling of an ensemble of simulations using a priori knowledge and observed data. Stochastic reconstruction of the contamination source consists of two principal mechanisms. One is the dispersion model suitable for modeling of the airborne contaminant in considered terrain, and the second is the sampling algorithm able to find the optimal dispersion model parameters based on the model output comparison and the contaminant registrations. Regarding the efficiency of the applied parameter scanning algorithm, each reconstruction requires multiple runs of the dispersion model. The reconstruction in urban terrain, which is of interest in this paper, requires advanced dispersion models taking into account the turbulence of the wind field around the buildings. The most reliable and exact are the computational fluid dynamics models (CFD), but those are very computationally demanding. We must be aware of the fact that to find the most probable contamination source, the dispersion model has to be run tens of thousands of times. So, the applied dispersion model has to be fast to be applied in a real-time working emergency system.

The first reconstructions in urban scales using building models was reported in [1] and [2]. In [1], authors used an adjoint representation of the source-receptor relationship and applied a Bayesian inference methodology in conjunction with Markov Chain Monte Carlo sampling procedures. In [2] authors applied the methodology presented in [3] to the reconstruction of the flow around an isolated building and the flow during IOP3 and IOP9 of the Joint Urban 2003 Oklahoma City experiment. In this reconstruction, the FEM3MP [4] model was applied to predict the atmospheric dispersion of the released substance.

In [5] authors applied the approximate Bayesian computation algorithm (ABC) to localize the source of contamination in the highly urbanized terrain of the center of London utilizing the real field experiment data from DAPPLE experi-

DOI: 10.1007/978-3-030-50417-5_30

ment [6]. As the forward dispersion model, the Quick Urban Industrial Complex (QUIC) Dispersion Modeling System was applied [7]. The successful estimation of the release source required over 10000 runs of the dispersion model. Even though the QUIC model is able to simulate the airborne contaminant transport in the city relatively quickly, a single simulation over the $800m \times 800m$ domain takes as minimum $2\ minutes$. Thus, reconstruction on a single computer requires over 330 hours. Computation time can be shortened by using a distributed system, but it is still impossible to apply it in the real-time working localization system when the answer time is crucial. Moreover, the required simulation time will increase with the enlargement of the considered terrain, e.g., for the whole city.

Even though in [5] the fast convergence of the ABC algorithm was proven, the whole framework cannot be implemented in the real-time emergency system due to the long computational time required by the dispersion model in urbanized terrain. This conclusion was an inspiration for the study presented in this paper. The idea is to train the artificial neural network (ANN) to be efficient in the simulation of the airborne contaminant transport in the urbanized terrain. If it succeeds, the ANN might work as the forward model in the system localizing the contamination source in real-time. Of course, the ANN has to be trained on the fixed city topology using the real wind conditions. This process requires lots of simulations serving as the training data-sets for the ANN. The process of training the ANN is computationally expensive, but ones trained, the ANN would be a high-speed tool for estimation the point-concentrations for a given contamination source.

2 ANN training city domain

Training ANN requires a large representative, reliable set of data. In this case, it should be measurements of the contaminant being a result of various release rates under different meteorological conditions. In this paper, we decided to check the possibility to train the ANN to simulate the airborne toxin transport in the area of central London where the DAPPLE experiment [6] was conducted (the main crossroad is of Marylebone Road and Gloucester Place, 51.5218N 0.1597W). The ideal situation would be if we could train the ANN on the real data. Unfortunately, it is not possible to obtain a set of data from real gas releases in urban areas that will be large enough to be a reliable set to train the ANN. Even though the city domain considered in this paper was the place of carrying out the large real field experiment DAPPLE the data available from its Trials are very limited. From four Trials, we have concentrations at 15 receptor positions for 30 minutes with 3-minutes intervals. This gives us, in sum, about 600 point-concentrations for four various source positions and release rates. This number of data is not enough to properly train the ANN. The only solution is to use the verified and well-recognized dispersion model to generate the data-set utilized to train, test, and validate the ANN. For that reason, we have used the QUIC Dispersion Modeling System. QUIC is intended for applications where the dispersion of air

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_30

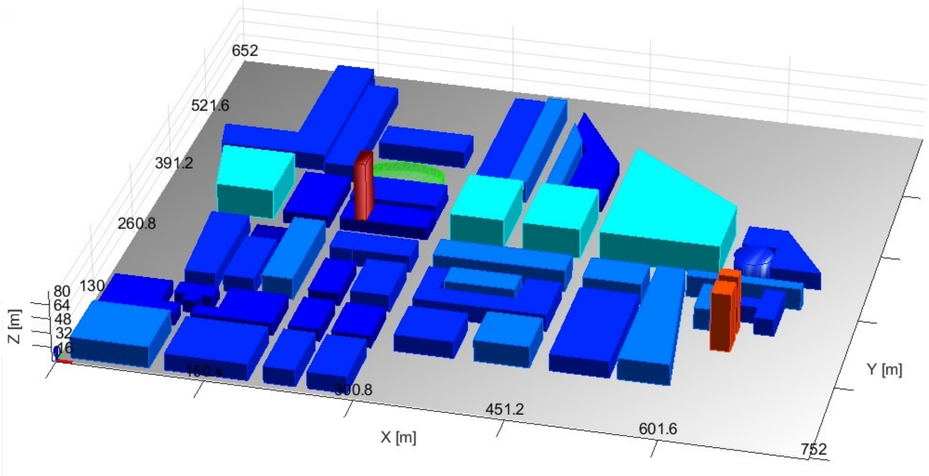

Fig. 1. The domain representing the area of central London assumed during the preparation of the ANN testing data-set (the main crossroad is of Marylebone Road and Gloucester Place, 51.5218N 0.1597W).

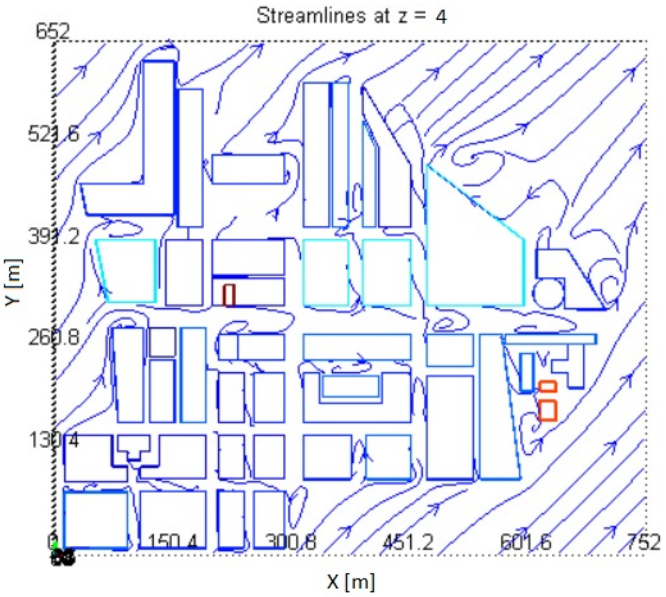

Fig. 2. The sample streamlines of the wind in the area of central London assumed during the preparation of the ANN testing data-set.

pollutants released near buildings must be computed relatively quickly [7]. The

effectiveness of the QUIC model as the forward model in the reconstruction of the contaminant source based on the field experiment DAPPLE data was proven in [5].
The QUIC system comprises of a wind model QUIC-URB, a dispersion model QUIC-PLUME, and a graphical user interface. The modeling strategy adopted in QUIC-URB was originally developed by Rockle [8] and uses a 3D mass-consistent wind model to combine properly resolved time-averaged wind fields around buildings [9]. The code has been tested for both idealized and real-world cases (e.g.,[7, 10]).

To test the possibility of applying the ANN to simulate the airborne contaminant dispersion in the urbanized terrain, we have prepared the domain of size $752\ m \times 652\ m \times 80\ m$ in which we have placed representations of the original buildings. The average building height in the area is $21.6m$ (range 10 to $64m$). The whole considered domain and the estimated by the QUIC-URB sample wind field around the buildings are presented in Figs. 1 and 2.

In this domain, we have set the simulations of an ideal gas continuous release and registered its concentration for thirty minutes. To reflect the real measurement conditions we have randomly drawn the 600 contamination source locations, release rate within the interval $Q \in < 10Mg; 500Mg >$ and its duration within interval $< 2min, 30min >$ and 100 registration points (representing the sensor locations) per single release. The registered concentrations were normalized and logarithmized with an added background Gaussian noise at the level of $10^{-5} g/m^3$. The sample simulated by the QUIC model propagation of the released 250 Mg of gas during the first 30 minutes within the domain is presented in Fig. 3.

3 The selected ANN topology

Artificial neural networks (ANN) are computational models that consist of interconnected elements called neurons. They are modeled on the construction of natural neurons and synapses connecting them [11]. ANN is capable of learning from training samples without knowing any laws or equations. There are several types of neural networks. In this paper, we used one of the simplest and widely used ANNs, the feed-forward neural network, e.g., [12]. This network has a one-way structure, i.e., the signal flows only in one direction from input nodes to output nodes. Feedforward neural networks were successfully used to predict the transport of pollutants in open areas, e.g., [13–15].

The ANN contaminant dispersion model is considered as a system that receives information from n distinct sets of inputs $X_i (i = 1, \dots, n)$, namely contaminant source parameters and sensor location, and produces a specific output, in our case the concentration of the gas in the passed as the input location. No prior knowledge about the relationship between input and output variables is assumed. The input variables should be independent of each other, and each one is represented by its own input neuron $i = 1, \dots, n$. Each neuron calculates a linear combination of the weighted inputs ω_{ij}, including a bias term b_i, from the

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_30

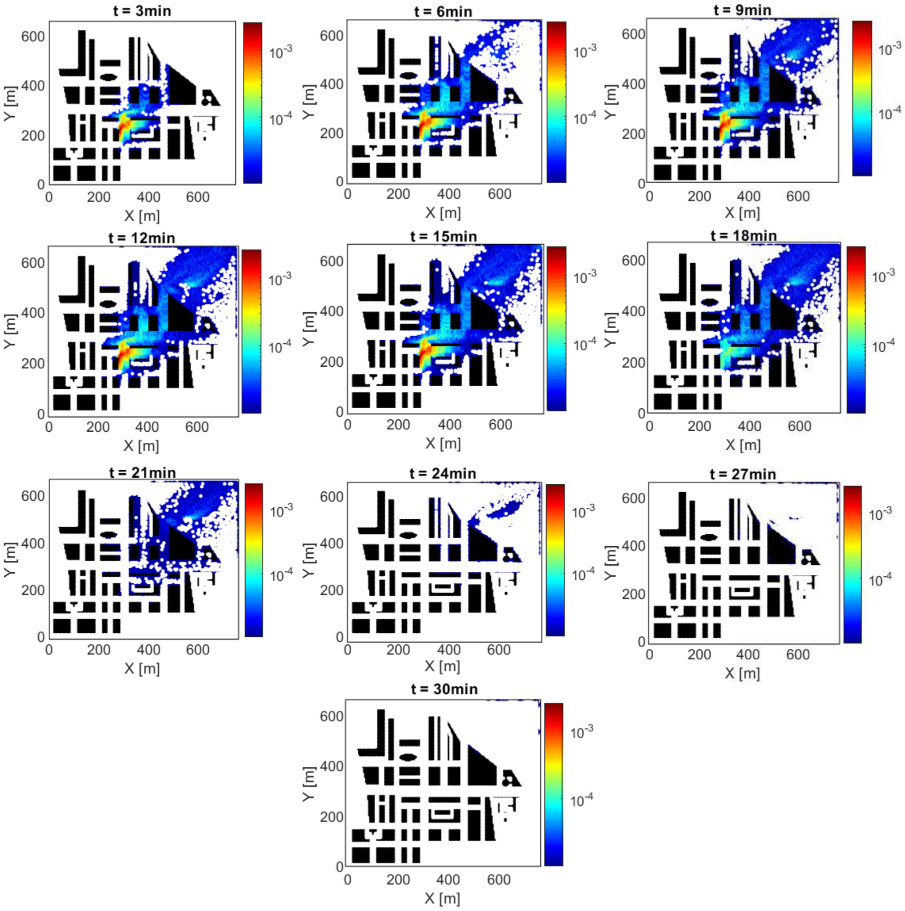

Fig. 3. The normalized concentration of the gas during thirty minutes after the 15 minute release of 250 Mg of gas from the source located at $x = 300$m, $y = 240$m, $z = 7$m within the considered domain as simulated by the QUIC model.

links feeding into it and the corresponding summed value $C_j = \sum_i \omega_{ij} X_i + b_i$ is transformed using a function f , either linear or non-linear for example log-sigmoid or hyperbolic tangent. The bias term is included in order to allow the activation functions to be offset from zero, and it can be set randomly or to the desired value. The output obtained is then passed as a new input $\tilde{X}_j = f(C_j)$ to other nodes in the following layer, usually named hidden layer. Though one is allowed to use several neurons in this hidden layer, it is generally advantageous to somehow minimize the number of hidden neurons, in order to improve the generalization capabilities of the model and also to avoid over-fitting. Having such a framework of input variables and sets of functions, the ANN has to be trained in order to obtain the best estimate for each weight ω. The weight values

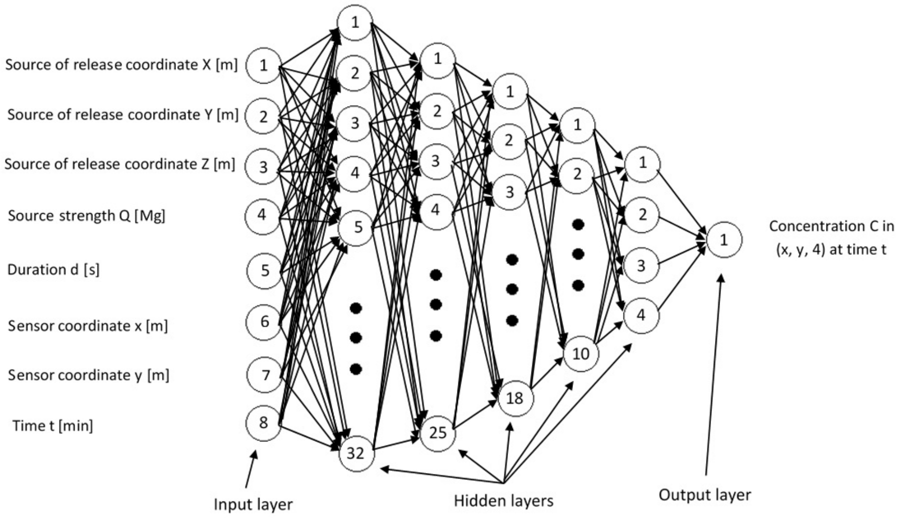

Fig. 4. The topology of the seven-layered feed-forward applied ANN with listed characteristics of input and output neurons.

are determined by an optimization procedure, the so-called learning algorithm [16]. The result of the neural network is compared with the target in order to calculate a predefined value of the error function. The error is sent over the network, and the algorithm adjusts the weights of each connection, respectively, to reduce the value of the error function. This repeated process corresponds to the number of training iterations that causes the network result to coincide with a state where the error between the output and the target is minimal.

During the ANN setup phase, numerous tests were performed with different combinations of the hidden layers, neuron number, learning rates, and activation functions. Fig. 4 illustrates the selected topology. In the input layer, we introduce eight neurons representing the coordinates of the contamination source, release rate, and its duration and the coordinates of the registration point (sensor) within the domain and registration time after the initiation of the release. In the output layer, the tracer concentration at a given point and time will be given. It appeared that the ANN performed the best when the five hidden layers were introduced, with 32, 25, 18, 10, and 4 neurons in subsequent layers. The Levenberg-Marquardt learning method was used, which minimizes an error function in "damped" procedures, i.e., select steps proportional to the gradient of the error function. We have tested various activation functions, and the best results were achieved using the hyperbolic tangent *tanh* function in all hidden layers, and linear function in the output layer. The crucial for ANN better performance occurred scaling of all input parameters to be in the interval $(0, 1)$. Scaling allowed escaping from the problem of different scales and model instability. Additionally, the output concentrations were logarithmized. Loga-

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_30

rithmization improved the ANN learning process because it allowed narrowing the range of concentration values. In consequence, to escape from the logarithm of the zero, the scaling was set to the interval (0,1), while hyperbolic tangent gives outputs in the range (-1,1). The input data set described in Section 2 was divided to training data set - 70%, validation and testing data-sets 15% each. The objective function describing the mean squared error between the real concentration and network output was set to reach $1e-04$ value.

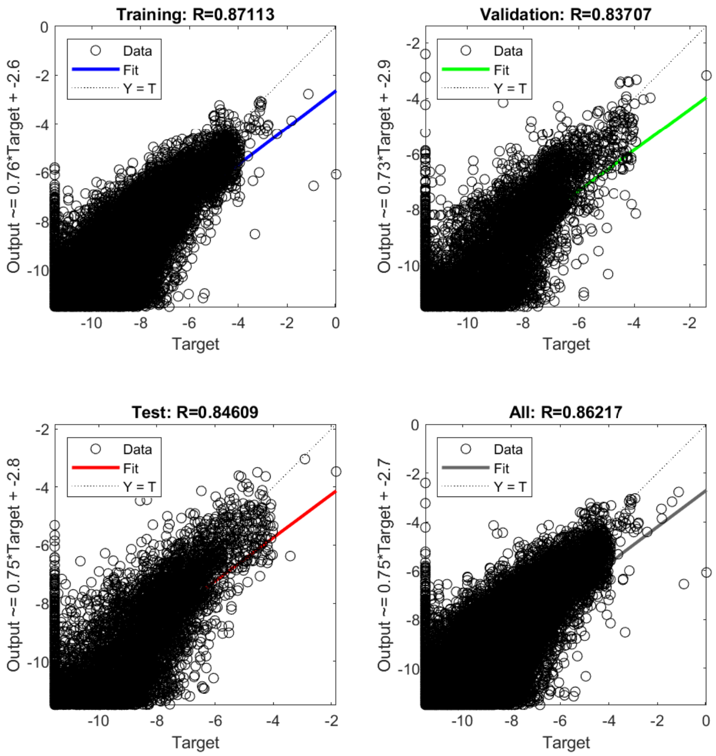

Fig. 5. The scatter plots representing results of training, testing, and validation process of the ANN. The dashed line represents the ideal fit.

4 Results

The results of the ANN training are presented in Fig. 5. Each point represents the single-point concentration as predicted by the trained ANN versus the input

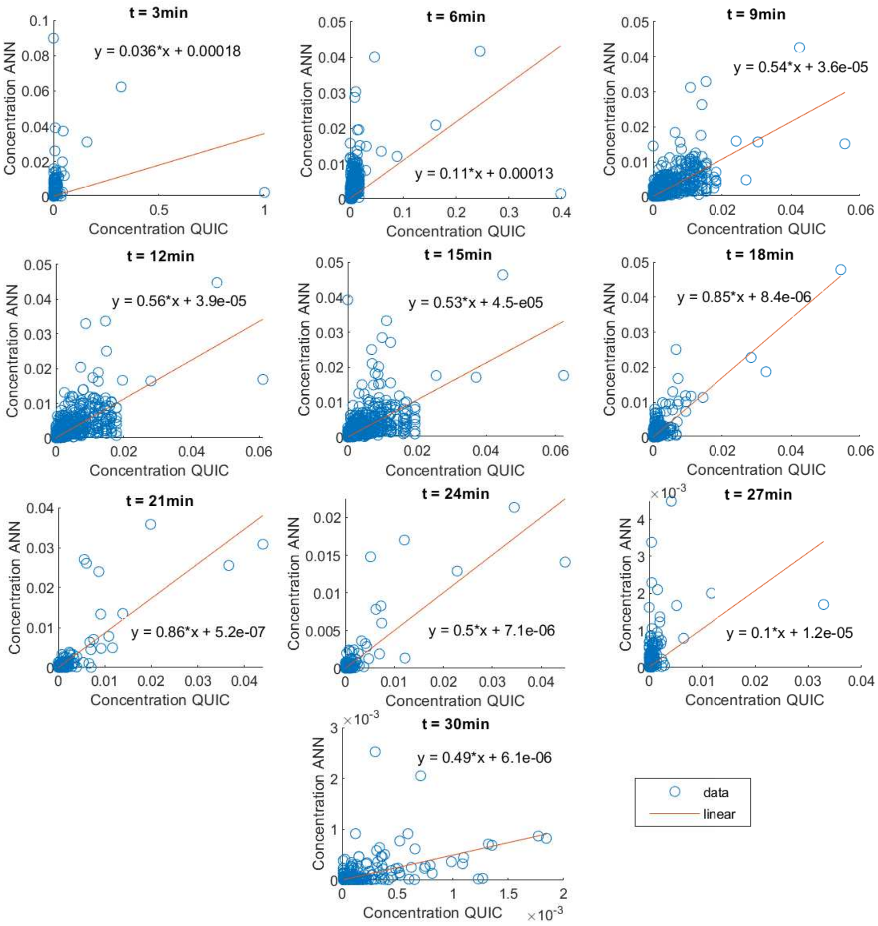

Fig. 6. The scatter plots representing results of training ANN versus the QUIC training-set data broken down by time. The line represents the linear fit.

concentration from the QUIC model included in the training, validation, and test datasets. The ANN was trained under the Deep Learning Toolbox of the Matlab software. Taking into account the complexity of the transport of the airborne contaminant in the turbulent wind around the buildings, the quality of the trained ANN is quite good. The R-value for training is 0.87, and together with test and validation data equals 0.86. These values indicate a significant relationship between the outputs and targets. The regression lines show that ANN slightly underestimates the higher concentrations. A more detailed analysis of the ANN performance is shown in Fig. 6. Fig. 6 presents the comparison of the concentrations predicted by the trained ANN versus the concentrations from the QUIC model, taking into account the time dependence of concentration at the given registration point. The concentrations were sampled every three-minutes.

ICCS Camera Ready Version 2020
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-50417-5_30

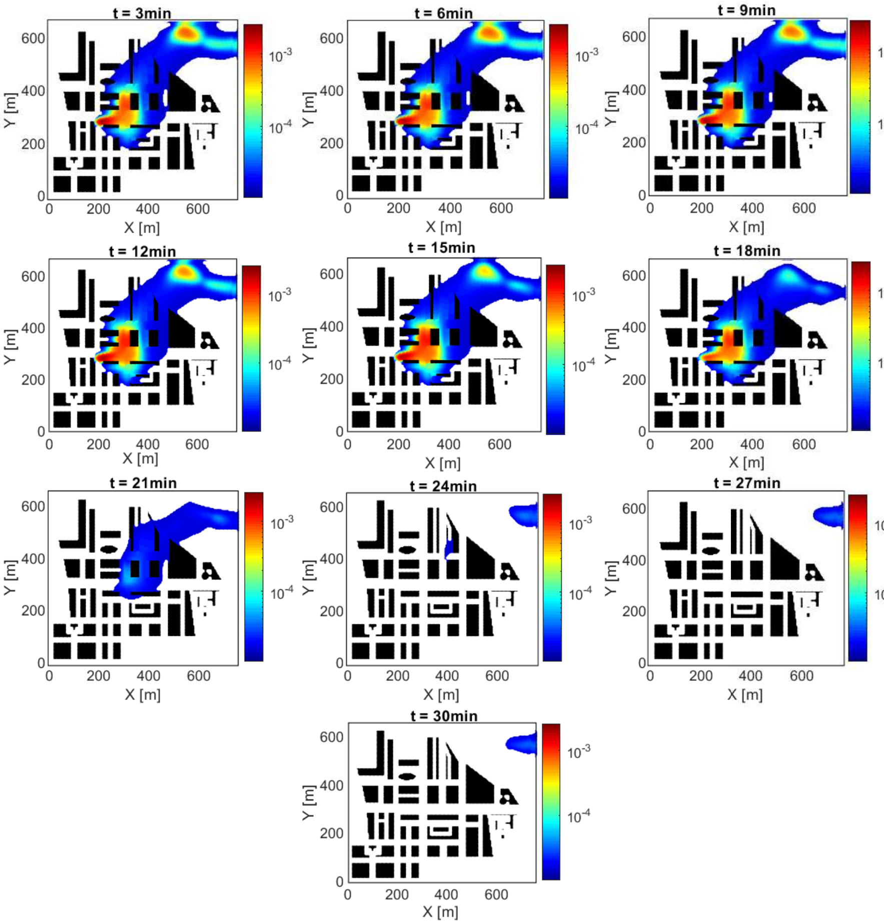

Fig. 7. The dispersion of the contaminant simulated by the ANN during consecutive thirty minutes after the 15 minute release of 250 Mg of gas from the source located at $x = 300$m, $y = 240$m, $z = 7$m within the considered domain.

The scatter plots display that the best agreement is achieved after eighteen minutes from starting of the release. With time the ANN starts to underpredicts the concentrations. After nine minutes, the correlation coefficient increases from 0.54 up to 0.86 in the 21st minute. Moreover, it is visible that ANN has a tendency to underpredicts the smaller concentrations, while for higher concentrations level of agreement increases. A possible reason is that in the training dataset, more scenarios are leading to smaller concentrations and fewer favorable to increased concentrations. Nevertheless, the crucial question is, does the ANN learned the physics standing behind the gas dispersion over the highly urbanized area? Figs. 7 and 8 presents the simulated by the ANN contaminant transport for thirty minutes after two release scenarios. In the simulation of the

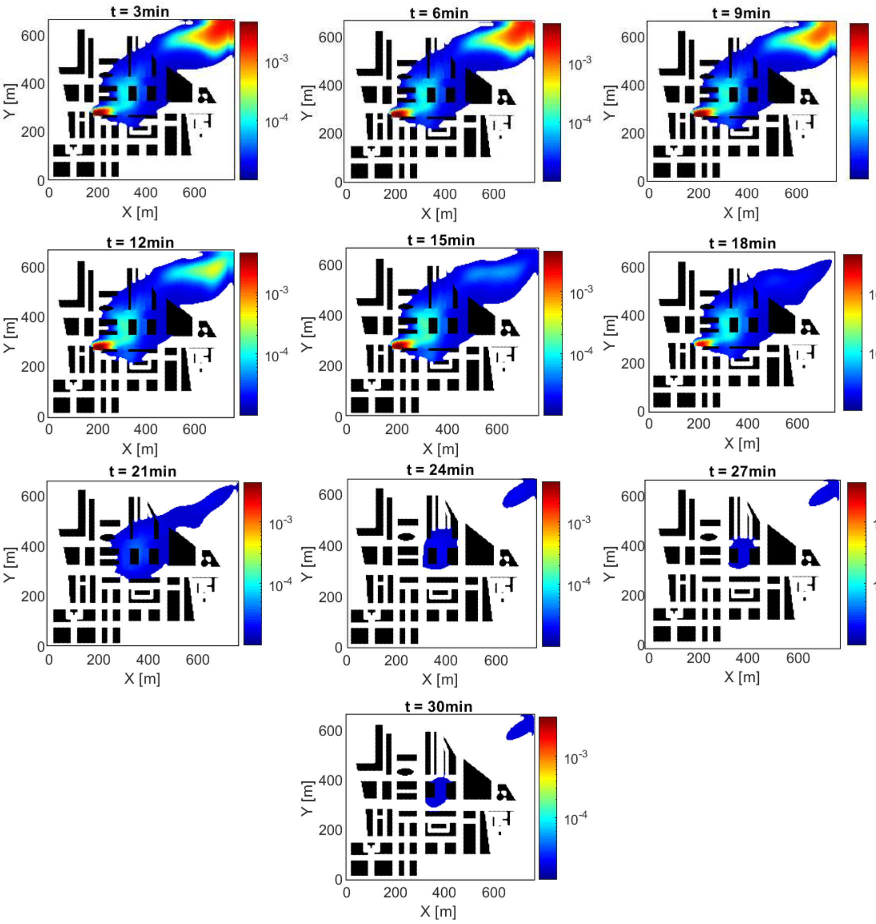

Fig. 8. The dispersion of the contaminant simulated by the ANN during consecutive thirty minutes after the 15 minute release of 400 Mg of gas from the source located at $x = 300$m, $y = 240$m, $z = 7$m within the considered domain.

gas by the trained ANN we have assumed the source location at the position with coordinates x = 300m, y = 240m, z = 7m within the considered domain. The concentrations predicted by the ANN were sampled homogenously every 4 meters within the domain with 3 minutes time span. One can see that the simulated by the ANN dispersion of the contaminant gas agrees with the QUIC simulation presented in Fig. 3. The gas is spread in the wind direction set to 225^o. As a result, the concentration of gas slowly decreases. Comparison of Fig. 7 where the released mass equals 250 Mg and Fig. 8 for released mass equal to 400 Mg confirms the correct estimation of gas concentrations by ANN. The concentrations predicted by ANN for release of 400 Mg of gas are greater than in

DOI: 10.1007/978-3-030-50417-5_30

the case of 250 Mg release. The regions of higher concentrations stay longer in the vicinity of buildings. We can conclude that simulated by the trained ANN transport of contaminants in the vicinity of the center of London agrees with the simulations performed by the QUIC model (Fig. 3).

5 Summary and future work

Presented results confirm that trained ANN can sufficiently simulate the turbulent transport of airborne toxins in the highly urbanized area. Such a result has not been published before. Even though we do not obtain the one-to-one agreement between the QUIC and ANN model concentrations, the trajectory of gas particles and gradient of concentrations predicted by ANN agree with the expectations. Obtained results suggest that the trained ANN can be successfully used in the contaminant source localization system as the forward dispersion model. In such systems, the contaminant source is estimated based on sampling the dispersion model set guided by minimizing the distance measure between the real concentrations from the sensors network and concentrations expected from the forward dispersion model. Therefore, more crucial is that ANN should correctly estimate the concentration gradients than its exact values. The main aim of the application of the trained ANN in such a localization system was to enable its operation in real-time. The time required by the presented in this paper ANN to estimate thirty-minute gas concentrations in a 196 000 sensor-points, as required by the simulations presented in Fig. 7 was equal to 3 s, while for the QUIC model it is estimated as at least 300 s, this gives us 100 times speed up. Taking this into account the reconstruction time in the real accidental situation can be short, resulting in the fast localization of the contaminant source.

The continuation of the presented research results will be the use of a trained neural network in place of the dispersion model for reconstruction based on real data from the DAPPLE field experiment, as it was presented in [5].

6 Acknowledgement

We thank Michael Brown and Los Alamos National Laboratory for the possibility to use the Quick Urban & Industrial Complex Dispersion Modeling System. Special thanks to Piotr Kopka for help in preparing the QUIC simulation. The authors acknowledge the partial financial support by the Polish National Science Centre, grant no. 2016/22/E/HS5/00406.

References

1. Keats, A., Yee, E., Lien, F.-S.: Bayesian inference for source determination with applications to a complex urban environment. Atmospheric environment, **41**(3), 465–479 (2007)

2. Chow, F.-K., Kosovia, B., Chan, S.: Source inversion for contaminant plume dispersion in urban environments using building-resolving simulations. Journal of applied meteorology and climatology **47**(6), 1553–1572 (2008)
3. Johannesson, G., Hanley, B., Nitao, J.: Dynamic bayesian models via monte carlo-an introduction with examples (No. UCRL-TR-207173). Lawrence Livermore National Lab., Livermore, CA (US) (2004)
4. Chan, S.-T., Leach, M.-J.: A validation of FEM3MP with Joint Urban 2003 data. Journal of applied meteorology and climatology **46**(12), 2127–2146 (2007)
5. Kopka, P., Wawrzynczak, A.: Framework for stochastic identification of atmospheric contamination source in an urban area. Atmospheric environment **195**, 63–77 (2018)
6. Wood, C.-R., Arnold, S.-J., Balogun, A.-A., Barlow, J.-F., Belcher, S.-E., Britter, R.-E., ... Neophytou, M.-K.: Dispersion experiments in central London: the 2007 DAPPLE project. Bulletin of the American Meteorological Society **90**(7), 955–970 (2009)
7. Williams, M.-D., Brown, M.-J., Singh, B., Boswell, D.: QUIC-PLUME theory guide. Los Alamos National Laboratory **43** (2004)
8. Röckle, R.: Bestimmung der Strömungsverhltnisse im Bereich komplexer Bebauungsstrukturen. na. (1990)
9. Sherman, C.-A.: A mass-consistent model for wind fields over complex terrain. Journal of applied meteorology **17**(3), 312–319 (1978)
10. Gowardhan, A.-A., Brown, M.-J., Pardyjak, E.-R.: Evaluation of a fast response pressure solver for flow around an isolated cube. Environmental fluid mechanics **10**(3), 311–328 (2010)
11. Bishop, C.-M.: Neural networks for pattern recognition. Oxford university press (1995)
12. Hecht-Nielsen, R.: Theory of the backpropagation neural network. In: Neural networks for perception, pp. 65–93. Academic Press., ISBN 9780127412528 (1992)
13. Hossain, K.: Predictive Ability of Improved Neural Network Models to Simulate Pollutant Dispersion. International Journal of Atmospheric Sciences (2014)
14. Ma, D., Zhang, Z.: Contaminant dispersion prediction and source estimation with integrated Gaussian-machine learning network model for point source emission in atmosphere. Journal of hazardous materials **311**, 237–245 (2016)
15. Wawrzynczak, A., Berendt-Marchel, M.: Application of the artificial neural network in the forecasting of the airborne contaminant. J. Phys.: Conf. Ser. **1391**, 012092 (2019)
16. Haykin, S.: Neural networks: a comprehensive foundation. Prentice Hall PTR (1994)