20th International Conference on
Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes
14-18 June 2020, Tartu, Estonia

SYNTHETIC DATA AND DEEP NEURAL NETWORKS FOR ATMOSPHERIC DISPERSION MODELLING IN URBAN AREAS

Mouhcine Mendil[1], Sylvain Leirens[1], Patrick Armand[2] and Christophe Duchenne[2]

[1] CEA, LETI, Minatec Campus, 17 rue des Martyrs, 38054 Grenoble, France
[2]CEA, DAM, DIF, F-91297 Arpajon, France

Abstract: Currently, atmospheric dispersion 3-D modelling has reached a maturity stage. Some Computational Fluid Dynamics (CFD) approaches have a high level of spatial and temporal accuracy for complex environments where site effects due to topography and/or buildings are significant, such as in urban areas. Various models require however heavy computational resources and prolonged runtimes ranging to several hours. This time requirement can be restrictive in crisis situations. One relevant scenario is the prediction with high urgency of the dispersion from a source in case of an accidental or a malicious release. In this paper, we propose to use synthetic data generated by highly accurate but time-consuming CFD models to train offline a Deep Neural Network (DNN). Specifically, we use Parallel Micro-SWIFT-SPRAY (PMSS), a multi-scale 3-D dispersion simulator parallelized over high performance computing resources, to generate two sets of pollutant concentration maps in two French cities: Grenoble and Paris. The DNN is trained on the Grenoble dataset to capture the underlying physics of transport and dispersion. The results show that the learned model generalizes successfully, as it is able to estimate accurately the pollution map in Paris unseen conditions.

Key words: Deep learning, Synthetic Data, Hazardous Pollution, Dispersion Modelling, Simulation.

INTRODUCTION

Toxic air pollutants are substances of different natures that raise many issues for the population. Being exposed to such pollutants can provoke several effects from minor discomforts (e.g. bad odor) to serious health problems (cancer, lung infection, birth defects...). Particularly, unexpected emissions of contaminants in urban areas, be it accidental or hostile, is a high priority concern. Such incidents require an urgent crisis intervention strategy from the authorities to protect the population and limit the material and environmental casualties. Dispersion modeling serves to estimate the concentration of a pollutant at different distances and directions from the sources. The models can compare exposures to some selected benchmark, such as a state pollution standard or a level with a known health effect, providing recommendations to decision makers (Sorensen, 2004).

Nowadays, Computational Fluid Dynamics (CFD) can model 3-D atmospheric dispersion with highly realistic accuracy for the most complex environments such as urban areas. However, such models require heavy computational resources and prolonged simulation time ranging to several hours (Zannetti, 2013). Alternatively, many researchers have explored the idea of introducing learning models into atmospheric dispersion prediction (Cabaneros, Calautit, & Hughes, 2019). In fact, such models have proven their ability to approximate extremely complex systems from observational data with highly nonlinear transformations. Compared to CFD, a trained neural network is generally time-efficient and low-demanding in terms of computational resources. These properties make it a suitable solution for predicting the impact of a hazardous release, where time for assessment is a serious restriction. However, the main drawback in machine learning is low quality and/or quantity of data. This is a crucial issue as learning models find recurring patterns on the training data to infer governing relationship between the input and output variables. The more situations are covered, the more powerful the inference capability the model will have. In case of dispersion modeling, data are usually collected from real size experiments or small-scale experiments in wind tunnels, which are both expensive (sensors, release systems, gases to

be released, etc.), slow (long acquisition campaigns), with predefined weather conditions (Blocken, Stathopoulos, Saathoff, & Wang, 2008).

In this paper, we propose to exploit a highly realistic CFD simulator to generate datasets that will serve to train and validate a deep learning model. In the simulation environment, the weather characteristics, the properties of the chemical release and the geometry of the terrain can be configured with no constraints. The learning models are trained on such synthetic data with the intention to use them in real crisis situations.

SYNTHETIC DATA GENERATION

We use Parallel Micro-SWIFT-SPRAY (PMSS) (Tinarelli, et al., 2013) (Oldrini, et al., 2017) as numerical modeling system of atmospheric transport and dispersion. It is the parallel version of Micro-SWIFT-SPRAY composed of 1) Micro-SWIFT: an analytically modified mass consistent interpolator over complex terrain. Given topography, meteorological data and buildings, a mass consistent 3-D wind field is generated. It is also able to derive diagnostic turbulence parameters to be used by Micro-SPRAY inside the flow zones modified by obstacles and 2) Micro-SPRAY: a Lagrangian particle dispersion model able to take into account the presence of obstacles. The dispersion of a pollutant is simulated following the trajectories of a large number of fictitious particles. The trajectories are obtained by integrating in time the particle velocity, which is the sum of a transport component defined by the local averaged wind provided generally by Micro-SWIFT, and a stochastic component, standing for the dispersion due to the atmospheric turbulence.

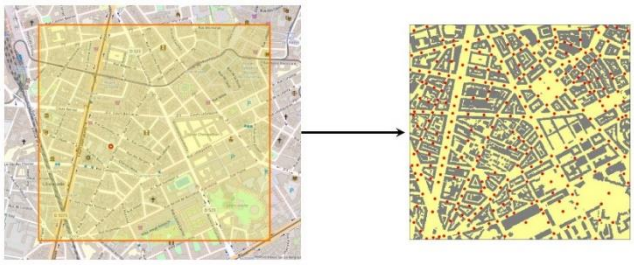

Figure 1. Left: street map of a part of the French city Grenoble and the considered bounding box for the simulation domain; Right: raster representation. The different hypothetical source locations are pinpointed with red dots.

The dataset to be used in the learning process is composed of 14,796 instances of integrated concentration over 2 hours, generated by PMSS for the following configuration. As shown in **Figure 1**, the urban area that constitutes the computation domain in PMSS is a neighborhood of the French city Grenoble located in the bounding box $x, y \in [913301, 914301] \times [6457391, 6458391]$, expressed in Lambert 93 coordinate reference system. It is a 500×500 grid with a space resolution of 2 m. The emission source is considered in 274 different hypothetical locations given 54 stationary weather conditions, built from a combination of 18 values of wind direction $\theta\,[°] \in \{0, 20, 40, \ldots, 340\}$ and 3 values of wind speed $v\,[m.s^{-1}] \in \{1.5, 3, 6\}$.

A point source produces an instantaneous fictitious emission of a unit mass of the pollutant (gas or particle matter). For a given initialization (source location and wind conditions), PMSS simulates the wind and dispersion fields for 2 hours.

DEEP LEARNING MODEL

A Deep Neural Network (DNN) is a non-linear statistical data modeling system that transforms a set of inputs into a set of outputs through multiple layers (sequence) of computations. The goal of DNNs is to infer from data the underlying phenomena that process the input values into the resulting outputs. The learning process consists in calibrating the parameters of the network to decrease the model error. The error function represents a distance metric between the model output and the observational data. In the learning phase, these parameters are updated by applying a gradient-descent algorithm using the training

dataset. After each training cycle, the DNN model is validated on a disjoined set of data —validation dataset—. Lastly, the performance of the model is evaluated on another set —test dataset—, never used during training or validation. Note that the validation step is necessary to adjust the complexity of the DNN, which depends on its hyper-parameters, such as number of layers, units (called neurons) and connections between neurons of consecutive layers. Vapnik introduces the notion of model capacity, which conceptually represents the space of functions the DNN can fit (Vapnik, 1999). For example, increasing the number of layers and/or their neural density enables to fit more complex non-linear transformations. However, the obtained function may model the random noise in the training data rather than the governing principles. This problem, called overfitting, results in building models that explain well the data at hand, but fail in out-of-sample predictions. On the other hand, DNNs with low capacity are impractical to solve complex tasks and tend to underfit. Both overfitting and under-fitting models fail to forecast the correct output for unencountered data.

In this work, the aim is to predict the integrated concentration field (2-D section) at the source height, over a time window starting from the instant of release up to 2 hours. We suppose that the atmospheric transport and dispersion depends on 1) The wind speed v and direction θ above the urban canopy and 2) a 2-D raster of the urban street map. These constitute respectively the output and inputs of a multivariate multiple regression problem.

We proposed the learning model represented in **Figure 2**. Further details about the choice of the hyper-parameters are motivated in the next section. The architecture is composed of a sequence of layers that incrementally build the pollution cartography through multiple non-linear transformations. The preprocessor performs several operations to prepare the data for the network's next stage. First, it applies a bounding box of size $\mathbb{R}^{m\times m}$ $(m = 200)$ centered on the location of the source. By doing so, the source coordinates are not needed as input parameters, thereby simplifying the learning model. Other transformations are subsequently applied such as data scaling, data augmenting and vectorization of the input data. For the next stage, we use an encoder/decoder structure. The encoder block contains four identical sub-blocks composed each of a full connected layer (512 neurons, RELU activation and HeNormal initializer) and a batch normalizer. It casts the problem in a lower dimensional space, hence reducing the complexity of learning. Next, the decoder reconstructs the data in the original space through a full-connected layer of m^2 units. Finally, a postprocessor rescales the data into the final output. Note that the loss function considered for the model training is the Mean Squared Error (MSE). The MSE of predictions on the training and validation datasets are monitored until convergence. An early stopping regularization method is adopted to avoid overfitting.

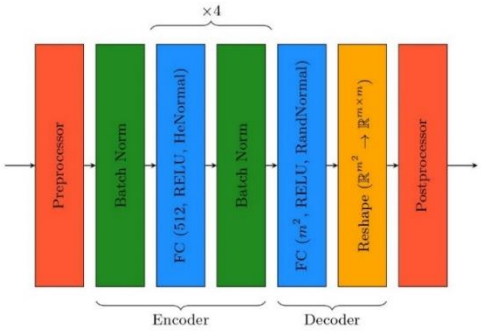

Figure 2. Deep Neural Network Architecture for learning atmospheric transport and dispersion from synthetic data.

PERFORMANCES ANALYSIS

Hyper-parameters tuning

As mentioned before, the learning capacity of a DNN is tightly linked to its hyper-parameters. A poor tuning leads to underfitting or overfitting, both reducing the inference performance of the trained model. Therefore, we propose a straightforward approach based on a grid search to select the number of layers

and units of the encoder block. The DNN is trained for different combinations of layers count and their neural density. After convergence, the MSE is evaluated on the validation dataset, normalized with respect to the highest error obtained. We compile the results in **Figure 3**. We observe that the lowest MSE is achieved by an encoder composed of four hidden layers, each containing 512 neurons. Compared to the considered configurations, this model is the best estimator of pollution concentration subsequent to a point hazardous release and will be used in the remaining of the paper.

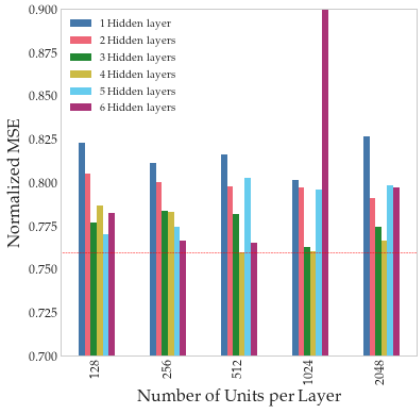

Figure 3. Grid search benchmark. The best model (lowest normalized MSE) corresponds to the encoder block composed of four hidden layers, each containing 512 neurons.

Use case study: Fictive hazardous release in the Opera district (Paris)

We consider several fictive hazardous pollutant releases in the urban district of Opera in Paris. We want to use the previously trained model to urgently estimate the exposure dose in the area. Similarly as before, we generate the concentration maps associated with these situation using the highly realistic simulator PMSS. The computation domain is the Opera district of Paris located in the bounding box $(x, y) \in [650250, 6863050] \times [651450, 6864050]$ (**Figure 4**). It is a 600×500 grid with a space resolution of 2 m. The emission source is considered in 222 different locations in 54 stationary weather conditions.

Figure 4. Left: street map of the Opera district in Paris and the considered bounding box for the simulation domain; Right: raster representation. The different hypothetical source locations are pinpointed with red dots.

To evaluate its accuracy, we use our trained model to predict the integrated concentration maps in about 4000 different situations. First, the total computation time is 3 seconds (that is, 0.75 ms per prediction). The prediction MSE averaged over the 4000 instances is 0.96. This demonstrates the highly precise and fast prediction capability of the proposed DNN, reaching the realism of PMSS nearly instantaneously. Some examples of predicted and simulated integrated concentrations are shown in **Figure 5**. Because of the complex topography of urban areas, the airflow can be very turbulent. In particular, the interactions between the buildings and the airflow as well as the effect of street intersections that change the trajectory of the pollutant are accurately modeled by the trained DNN. Additionally, pollutant mass distribution throughout the whole domain is faithfully reproduced by the prediction. The learning model can therefore successfully determine the high risk areas (risky concentration levels of the pollutant that require evacuation for example).

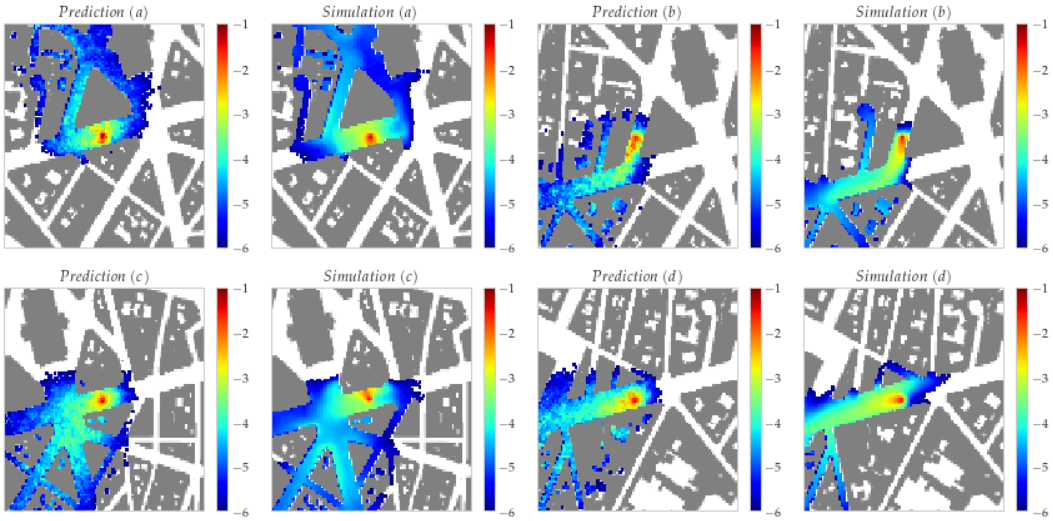

Figure 5. Integrated concentration map for different wind conditions and location of Opera district (logarithmic scale). Predictions and simulations are produced respectively by the trained DNN and PMSS.

CONCLUSION

This paper evaluates the learning potential of DNNs to model atmospheric transport and dispersion in complex urban areas using synthetic data. These data are generated using PMSS for several conditions in the French city of Grenoble. They are used to train offline a DNN to capture the physics of the dispersion phenomenon and forecast integrated concentration maps in real crisis events, given weather conditions and an urban geometry. Once the learning phase completed, the proposed DNN predicted nearly instantaneously and with high accuracy the concentration cartography in the newly encountered district of Opera (Paris).

The actual DNN forecasts a 2-D horizontal concentration field at the height of the pollution source. In future works, we plan to fully utilize the 3-D simulations of PMSS (and not just 2-D sections) to teach a DNN how to predict a 3-D concentration field, jointly estimating the horizontal and vertical distributions of the pollutant.

REFERENCES

Blocken, B., Stathopoulos, T., Saathoff, P., & Wang, X. (2008). Numerical evaluation of pollutant dispersion in the built environment: comparisons between models and experiments. *Journal of Wind Engineering and Industrial Aerodynamics, 96*, 1817--1831.

Cabaneros, S. M., Calautit, J. K., & Hughes, B. R. (2019). A review of artificial neural network models for ambient air pollution prediction. *Environmental Modelling & Software, 119*, 285-304.

Oldrini, O., Armand, P., Duchenne, C., Olry, C., Moussafir, J., & Tinarelli, G. (2017). Description and preliminary validation of the PMSS fast response parallel atmospheric flow and dispersion solver in complex built-up areas. *Environmental Fluid Mechanics, 17*, 997-1014.

Sorensen, J. (2004). Planning for protective action decision making. *Journal of Hazardous Materials, 109*, 1-11.

Tinarelli, G., Mortarini, L., Castelli, S. T., Carlino, G., Moussafir, J., Olry, C. a., & Anfossi, D. (2013). Review and validation of MicroSpray, a Lagrangian particle model of turbulent dispersion. *Lagrangian modeling of the atmosphere, 200*, 311-327.

Vapnik, V. N. (1999). An overview of statistical learning theory. *IEEE transactions on neural networks, 10*, 988-999.

Zannetti, P. (2013). *Air pollution modeling: theories, computational methods and available software.* Springer Science & Business Media.