August 27-31, 2023
FLORENCE - ITALY
ICWE 16
16th International Conference on Wind Engineering

Dimensionality reduction for pollutant forecasting on a real urban area test case

Christos Nixarlidis[1,2,3], Léo Cotteleer[1,2], Alessandro Gambale[4], Tim de Troyer[2,3], Alessandro Parente[1,2]

[1]*Université Libre de Bruxelles (ULB), Laboratory of Aero-Thermo-Mechanics, Brussels, Belgium*
[2]*Brussels Institute for Thermal-Fluid Systems and Clean Energy (BRITE), Brussels, Belgium*
[3]*Vrije Universiteit Brussel (VUB), FLOW Research Group, Brussels, Belgium*
[4]*BuildWind SPRL, Brussels, Belgium*

SUMMARY:
This study focuses on the application of dimensionality reduction modelling techniques, such as Principal Component Analysis (PCA), Auto Encoders (AE), and Convolutional Auto Encoders (CAE), to a real urban area test case for pollutant forecasting. Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) architectures of neural networks are employed to develop a digital twin of Ixelles in Belgium and predict the pollutant concentrations in real-time. An extensive sensitivity analysis was conducted to define the optimal parameters for the implemented methodologies. PCA as a reduced order modelling approach was shown to outperform both of the proposed Auto Encoder methodologies in both time and prediction accuracy. CNN captured the temporal patterns more accurately than LSTM in most methodology combinations. As future research, the potential use of Dynamic Modal Decomposition in capturing the temporal trends and patterns instead of neural networks is discussed and a data-driven CFD modelling approach for pollutant dispersion coupled with the forecasting framework is proposed.

Keywords: Pollutant prediction, digital twins, machine learning, dimensionality reduction

1. INTRODUCTION

In recent years, the developments in transportation and industrialization have constituted air pollution as one of the main widespread environmental and health issues. Air quality forecasting has become an important aspect of smart city planning and management (Dembski et al., 2020). Digital twin technologies, such as sensor networks coupled with machine learning algorithms, can be used to predict and monitor air pollution levels in real-time. This way, government agencies and organizations can take proactive measures to reduce air pollution, to identify its sources and target interventions accordingly to mitigate its negative effects on human health and the environment (Fenger, 1999).

In the context of SPICECO (Secure and open Platform for an Intelligent City ECOsystem) project, an urban air quality prediction framework is being developed that will be openly available to the public as a digital twin of the district of Ixelles in Brussels, Belgium, using state-of-the-art Machine/Deep Learning (ML/DL) methodologies to predict the concentrations of key air pollutants. A sensor network monitors the local meteorological conditions and concentrations. Dimensionality reduction techniques such as Principal Component Analysis (PCA), Auto Encoders (AE) and

Convolutional Auto Encoders (CAE) are employed to reduce the vast volume of input data and therefore drastically decrease the training time of the proposed models, in order to predict the local air quality in real-time.

2. METHODOLOGY

Throughout the district of Ixelles in Brussels, Belgium, a network of 11 QSENSE-Air sensors by Macq are installed at key locations that cover a total test case area of $\sim$1 km^2. The local meteorological conditions (Temperature, Pressure, Humidity, Precipitation, Wind Direction and Speed) and pollutant concentrations ($PM_{2.5}$, PM_{10}, NO_2 and O_3) are measured with a 10-minute sampling frequency. The date and time of the samples are then converted to numerical values (such as day of the year, day of the week, minute of the day etc.) before they are used as input for the proposed methodologies, so that the different seasonal/temporal patterns can be captured.

The aim of the present work is to enable the real-time prediction of the pollutant concentrations ($\mathbf{y}$) as a function of the meteorological conditions ($\mathbf{x}$) and time $\mathbf{t}$. Without loss of generality, the dynamic system can be expressed as:

$$\mathbf{y} = \mathbf{f}(\mathbf{x}(t), \mathbf{t}) \tag{1}$$

where $t \in [0,T]$ denotes the time with final time $T \in \mathbb{R}_+$, $\mathbf{x} : [0,T] \to \mathbb{R}^6$ and $\mathbf{f} : \mathbb{R}^6 \times [0,T] \to \mathbb{R}^4$ denotes an unknown operator linking $\mathbf{x}$ and $\mathbf{y}$.

Two different approaches have been used in this research. The first one is to teach the direct mapping $\widetilde{\mathbf{f}} : \mathbf{x} \mapsto \mathbf{y}$ with $\widetilde{\mathbf{f}} \approx \mathbf{f}$ considering the full meteorological and time input data. For this purpose, both Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) neural network architectures are employed. The second methodology relies on the assumption that a reduced and optimal basis exists to represent the input data. Therefore, the mapping is trained from a subset $\mathbf{x}^{\text{red}}$ of the data to reduce the computational cost: $\widetilde{\mathbf{f}} : \mathbf{x}^{\text{red}} \mapsto \mathbf{y}$. Hence, two steps are required: the reduction of the input data and the regression from the latent space to the output $\mathbf{y}$. As far as the dimensionality reduction is concerned, Principal Component Analysis (PCA) (Jolliffe and Cadima, 2016), Auto Encoders (AE) (Wang et al., 2016) and Convolutional Auto Encoders (CAE) (Masci et al., 2011) have been used. Subsequently, the regression is performed with the two aforementioned methodologies.

An extensive sensitivity analysis study was conducted to fine-tune the parameters of the above methodologies, such as the depth of the network, number of trainable parameters etc., an overview of which is given in Table 1 below.

Table 1. Overview of the sensitivity analysis parameters setup.

Methodology	Train. Parameters	Batch Size	Latent Dimension	Depth of Layers
CNN	32$\sim$4096	–	–	1$\sim$3
LSTM	100$\sim$800	–	–	1$\sim$3
PCA	–	–	4$\sim$8	–
AE	512$\sim$1024	1$\sim$32	4$\sim$8	2$\sim$4
CAE	32$\sim$128	1$\sim$32	4$\sim$8	2$\sim$4

3. RESULTS

The sensitivity analysis suggested a network depth of 3 layers as the optimal for all NN architectures as a compromise between training time and retained information. Since PCA is linear and does not require training, only the two AEs are compared in Fig. 1 (right). A batch size of 4 and a latent dimension size of 7 (compared to the original 9) were chosen for both AEs and CAEs dimensionality reduction methodologies. Regarding the CNN and LSTM architectures, the number of trainable parameters was kept to 1024 and 512 respectively. All the results discussed below were implemented with the aforementioned configurations and the same dataset of 30-minute observations over ∼3 months was used as input.

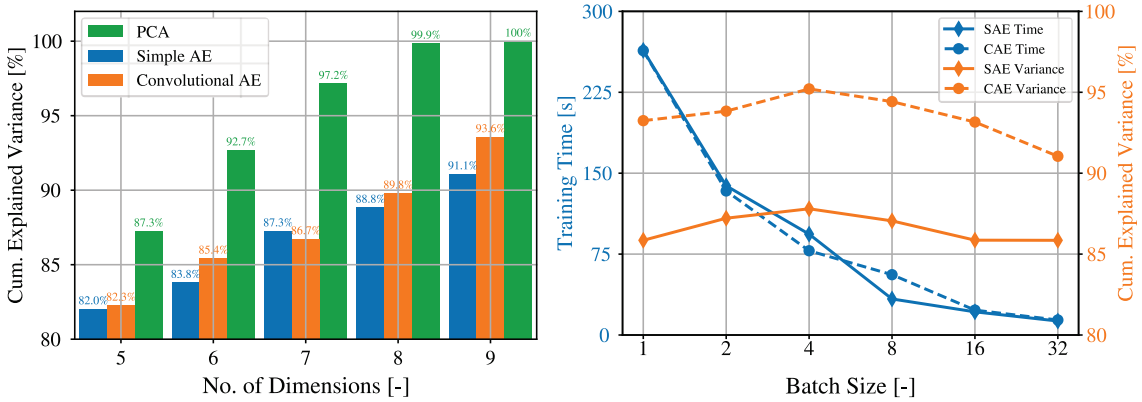

Figure 1. Cumulative explained variance and training time per methodology for different combinations of batch size and latent dimension size.

One can observe that PCA reduces the training time of both neural network architectures almost by half, while Auto Encoders may actually increase it, as in the case of AE-LSTM. As seen in Fig. 2 (left). The predicted values for $PM_{2.5}$ in Fig. 2 (right) are indicative to showcase the capability of the forecasting framework. Similar predictions are performed for PM_{10}, NO_2 and O_3.

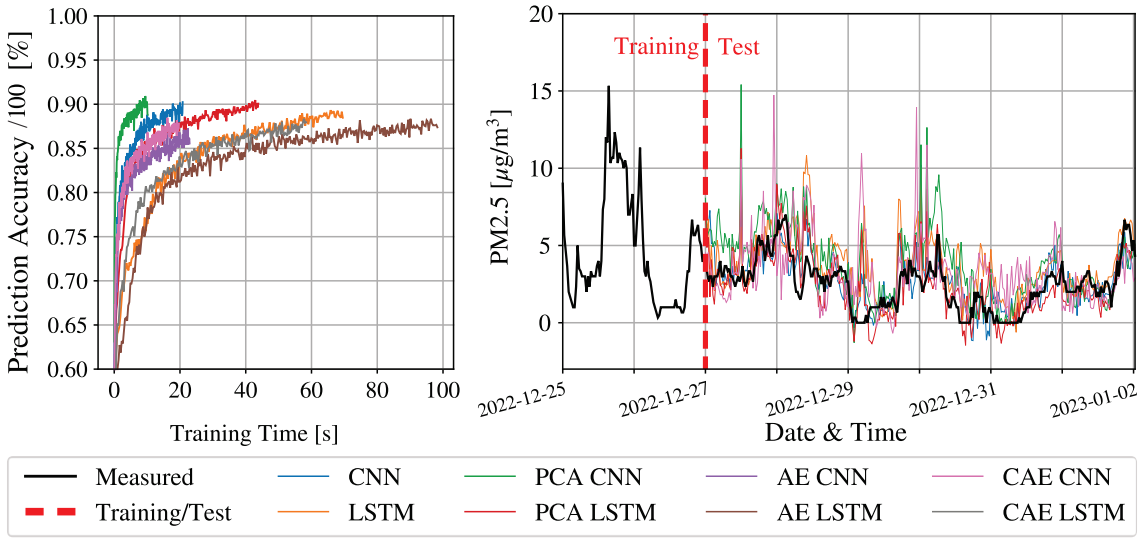

Figure 2. Prediction accuracy vs training time of the proposed methodologies (left) and $PM_{2.5}$ forecasting results using CNN/LSTM with PCA for ROM compared to measured values (right).

4. CONCLUSIONS & FUTURE WORK

The present study aims to develop a real-time pollutant concentration prediction framework. CNN and LSTM neural network architectures coupled with three different dimensionality reduction methodologies are implemented to map the local meteorological and date-time inputs to four key air pollutants. A sensitivity analysis was conducted to identify the optimal parameters of the models. Even though all NN and ROM model combinations demonstrated a prediction accuracy of over 85%, PCA with CNN was shown to outperform both of the proposed Auto-Encoder methodologies in both time and prediction accuracy.

As future research work, we could cite the implementation of modal decomposition methodologies for time-series predictions. For example, High-Order Dynamic Modal Decomposition (HODMD) (Vega and Le Clainche, 2020) has been shown to adequately capture the trends and temporal patterns in time-dependent fluid dynamics cases (Corrochano et al., 2023; Le Clainche and Vega, 2018) and could have potential in such applications. Finally, Reynolds-Averaged Navier-Stokes (RANS) simulations are able to model pollutant dispersion in complex urban flows, but the required turbulence models (Bellegoni et al., 2023; Parente et al., 2011) introduce uncertainty in the results (for example due to the wrong assumption of the turbulent Schmidt number Sc_t (Longo et al., 2020)), which may limit their application in the decision making process. A data-driven modelling approach could couple the developed pollutant forecasting framework with the dispersion fluid dynamics simulations to counter this by solving the inverse problem and therefore reducing the otherwise unavoidable uncertainty.

ACKNOWLEDGEMENTS

This research project was financially supported by Innoviris Brussels (Grant Ref. No. 2021-RDIR-17b) in the context of SPICECO Project, and this support is gratefully acknowledged.

REFERENCES

Bellegoni, M., Cotteleer, L., Srikumar, S. K. R., Mosca, G., Gambale, A., Tognotti, L., Galletti, C., and Parente, A., 2023. An extended SST k- ω framework for the RANS simulation of the neutral Atmospheric Boundary Layer. Environmental Modelling & Software 160, 105583.

Corrochano, A., D'Alessio, G., Parente, A., and Clainche, S. L., 2023. Hierarchical Higher-Order Dynamic Mode Decomposition for Clustering and Feature Selection. arXiv preprint arXiv:2301.07976.

Dembski, F., Wössner, U., Letzgus, M., Ruddat, M., and Yamu, C., 2020. Urban digital twins for smart cities and citizens: The case study of Herrenberg, Germany. Sustainability 12, 2307.

Fenger, J., 1999. Urban air quality. Atmospheric environment 33, 4877–4900.

Jolliffe, I. T. and Cadima, J., 2016. Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences 374, 20150202.

Le Clainche, S. and Vega, J. M., 2018. Analyzing nonlinear dynamics via data-driven dynamic mode decomposition-like methods. Complexity 2018, 1–21.

Longo, R., Bellemans, A., Derudi, M., and Parente, A., 2020. A multi-fidelity framework for the estimation of the turbulent Schmidt number in the simulation of atmospheric dispersion. Building and Environment 185, 107066.

Masci, J., Meier, U., Cireşan, D., and Schmidhuber, J., 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. Proceedings of Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21. Springer, 52–59.

Parente, A., Gorlé, C., Van Beeck, J., and Benocci, C., 2011. Improved k–ε model and wall function formulation for the RANS simulation of ABL flows. Journal of wind engineering and industrial aerodynamics 99, 267–278.

Vega, J. M. and Le Clainche, S., 2020. Higher order dynamic mode decomposition and its applications. Academic Press.

Wang, Y., Yao, H., and Zhao, S., 2016. Auto-encoder based dimensionality reduction. Neurocomputing 184, 232–242.