Correcting Air Quality Forecasts with Machine Learning Algorithms

Hervé Petetin[*1], Albert Soret[*2], Carlos Pérez Garcia-Pando[*3]

[*] *Barcelona Supercomputing Center (BSC)*

{[1]herve.petetin, [2]albert.soret, [3]carlos.perez}@bsc.es

***Keywords*— air quality, forecast, machine learning**

EXTENDED ABSTRACT

Air pollution is a major environmental problem affecting human health and ecosystems. Mitigating the effect of pollution episodes requires reliable air quality forecasting (AQF) systems for both warning vulnerable populations and taking short-term measures of emission control (e.g. traffic reduction, interruption of some industrial facilities). The rise of chemistry-transport models (CTMs) over the last decades have allowed major improvements in AQF, supplanting the use of purely statistical forecasting systems. However, AQF systems remain affected by numerous sources of uncertainty (e.g. emissions, meteorology, initialization). In order to improve the forecasts, so-called model output statistics (MOS) methods are commonly applied to correct CTM outputs where continuous observations are available, typically at surface AQ monitoring stations. Various methods have been proposed in the literature, including moving averages (MA), Kalman filter (KF) and, more recently, analogs (AN) (e.g. Delle Monache et al., 2006; Kang et al., 2010; Djalalova et al., 2010, 2015; Huang et al., 2017). Although these methods allow removing a large part of the bias, they still suffer from some limitations, in particular when it comes to the detection of fast changes in atmospheric conditions.

In this work, we explore the use of machine learning (ML) techniques for correcting the raw AQF produced by CTMs based on various types of ancillary information, including meteorological features, past forecast errors, or temporal features. In ML terms, this corresponds to a problem of supervised regression in which the past observations along with their associated features are used to train a model that estimates the future pollutant concentrations on the basis of new (unseen) features. Several popular ML algorithms are tested, including gradient boosting machine (GBM), random forest (RF) and support vector machine (SVM). Results are compared with other MOS methods proposed in the literature.

The analysis is applied on the AIRE-CDMX operational AQF system of Mexico City. Recently built by the Earth Science Department, this AQF system is based on the combine use of the Community Multiscale Air Quality (CMAQ) chemistry-transport model and the Weather Research and Forecasting (WRF) meteorological model. The study focuses on the correction of the coarse particulate matter (PM_{10}, particles with aerodynamical diameter below 10 µm) forecasts over the period ranging from August 2017 to December 2018. We mimic an operational AQF system in which new forecasts and observations are obtained continuously. The first ML model is trained at the 30th day and then updated every 30 days. Different sets of features are tested. Both the training and tuning of the hyper-parameters are performed only based on the past available data. This is done independently at all surface stations of the local AQ monitoring network.

Among the traditional MOS methods, the AN method is found to give the best results, closely followed by the KF method. Concerning the MOS-ML methods, results show that they are able to compete with these other MOS methods when an appropriate set of features is taken into account. Relatively poor during the first months due a very short training period, their performance shows substantial improvement with time as the size of the training dataset increases (see Fig. 1). Among the different ML algorithms tested, best results are obtained with GBM. Compared to the raw forecast, the best GBM configuration gives a reduction of the root mean square error from 71 to 48%, and an increase of the correlation from 0.48 to 0.62.

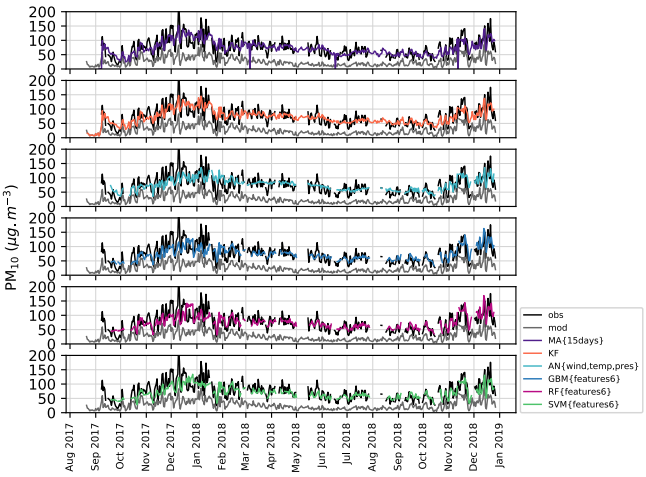

Fig. 1 Daily PM_{10} concentrations at a surface station of the AQ monitoring network. The lines show the observed concentrations ("obs"), the concentrations simulated by the AQF system without any correction ("mod"), and the concentrations corrected by the different MOS methods : moving averages ("MA"), Kalman filter ("KF"), Analogs ("AN"), gradient boosting machine ("GBM"), random forest ("RF") and support vector machine ("SVM").

One important objective of the AQF systems is to be able to predict in advance the occurrence of pollution episodes (here defined as PM_{10} concentrations exceeding a given threshold concentration) with a good reliability. The different MOS methods have been compared on a set of metrics related to episode detection skills. The raw forecasts given by the AQF system show a poor skill for detecting PM_{10} episodes due to a strong negative bias. The KF method shows a good ability to detect episodes but with numerous false detections. In other terms, many of the observed episodes are predicted by the AQF system but a forecasted episode has relatively low chance to actually happen. Conversely, most ML methods detect a lower number of pollution episodes but with a better confidence. Such a characteristic can be interesting when the short-term measures of emission control have a high financial cost and/or a low social acceptance.

A. Conclusion and Perspectives

These preliminary results highlight the potential of ML algorithms for improving the AQF provided by geophysical models. Considering the small size of the training datasets, these results are considered promising. In order to assess more precisely the ability of ML to correct AQF better than more established methods, longer datasets are required. Such an analysis is currently ongoing with a multi-year simulation from the NMMB-MONARCH model developed at BSC. This new dataset will allow to monitor how the performance of the MOS-ML methods evolves in time.

B. ACKNOWLEDGEMENTS

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement H2020-MSCA-COFUND-2016-754433.

References

[1] L. Delle Monache, T. Nipen, X. Deng, Y. Zhou, and R. stull, "Ozone ensemble forecasts:2. A Kalman filter predictor bias correction", J. Geophys. Res., Vol. 111(D5), D05308, 2006.

[2] D. Kang, R. Methur, S. T. Rao, and S. Yu, "Bias adjustment techniques for improving ozone air quality forecasts", J. Geophys. Res., Vol. 111(D23), D23308, 2008.

[3] I. Djalalova, J. Wilczak, S. McKeen, G. Grell, S. Peckman, M. Pagowski, L. Delle Monache, J. McQueen, Y. Tang, and P. Lee, "Ensemble and bias-correction techniques for air quality model forecasts of surface O3 and PM2.5 during the TEXAQS-II experiment of 2006", Atmos. Environ., Vol. 44(4), Pp. 455-467, 2010.

[4] I. Djalalova, L. Delle Monache , and J. Wilczak, "PM2.5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model", Atmos. Environ., Vol. 108, Pp. 76-87, 2015.

[5] C. Borrego, A. Monteiro, M. T. Pay, I. Ribeiro, A. I. Miranda, S. Basart, and J. M. Baldasano, "How bias-correcction can improve air quality forecasts over Portugal, Atmos. Environ., Vol. 45(37), Pp. 6629-6641, 2011

[6] R. Zeng, W. Dietzel, R. Zettler, J. Chen, and K. U. Kainer, "Microstructure evolution and tensile properties of friction-stir-welded AM50 magnesium alloy," Trans. Nonferrous Met. Soc. China, Vol. 18, Pp. s76–s80, Dec. 2008.

[7] J. Huang, J. McQueen, J. Wilczak, I. Djalalova, I. Stajner, P. Shafran, D. Allured, P. Lee, L. Pan, D. Tong, H.-C. Huang, G. DiMego, S. Upadhayay, and L. Delle Monache, "Improving NOAA NAQFC PM2.5 predictions with a bias correction approach", Weather Forecast, Vol. 32(2), Pp. 407-421, 2017.

Author biography

Hervé Petetin was born in Caen, France in 1986. He holds an engineering diploma from the Ecole Centrale de Lille, France, a M.Sc. in Mechanics and fluid dynamics from the University of Science and Technology Lille, a M.Sc. in Atmospheric physics and chemistry from the University of Paris Est Creteil, France, and a Ph.D. in Atmospheric physics and chemistry from the University of Paris Diderot, France.
His research has first focused on the study of air quality and more specifically the aerosol pollution in large megacities, using regional chemistry-transport models and in-situ observations. As a member of the IAGOS European Research Infrastructure, he then investigated during several years the variability and trends of two important intermediate-lifetime gaseous compounds, namely ozone and carbon monoxide, in the troposphere based on airborne in-situ observations provided by the IAGOS fleet.
In 2018, he obtained a postdoctoral funding at the BSC from the STARS program (Marie-Sklodowska-Curie Action COFUND program) for working on the improvement of air quality forecasts with machine learning techniques.