AN INTERACTIVE WEB-BASED GIS APPLICATION FOR THE ANALYSIS OF POLLUTANT DISPERSION MODELS

P. Beikos[1], M. Kavousanakis[1]*

[1]School of Chemical Enginnering, National Technical University, Athens, Greece

(**mihkavus@chemeng.ntua.gr*)

ABSTRACT

The escalating concerns regarding air quality necessitate the development of advanced tools for the analysis and visualization of pollutant dispersion in the environment. This paper presents an innovative Interactive Web-Based application designed for the analysis of pollutant dispersion models. The integration of Geographic Information System capabilities into the application enhances the spatial understanding of pollution patterns, contributing to more effective environmental management and decision-making. Also, by simultaneously connecting the application's user interface to multiple APIs one can obtain real time climate and environmental data that can be integrated into the modelling process.

The application leverages insights from existing literature on air pollution modeling[1], numerical simulations[2] and advancements in air quality assessment using machine learning[3, 4]. Its user-friendly interface allows for dynamic visualization of pollutant dispersion, enabling policymakers, researchers, and the public to make informed decisions related to air quality management. The accessibility of the application via the web ensures widespread usability and encourages community involvement in environmental monitoring and protection.

KEYWORDS: Pollutant dispersion modelling, Web-based gis, Machine learning

INTRODUCTION

The present paper introduces a novel methodology for performing pollutant dispersion analysis in polluted geographical areas. The methodology consists of an interactive web application, a machine learning model in the backend, qualitative and quantitative evaluation of conclusions derived from time series plots, geolocation data and other visualization tools. The proposed methodology was evaluated with real emission~~s~~ data acquired from the following sources: Hourly concentrations of NOx emissions at the Lorraine Granado Community Park were taken from the Colorado University air quality data repository[5]. Hourly NOx emissions data were taken from the power sector emissions database of the United States Environmental Agency protection[6]. All the meteorological data were acquired from the open-meteo open weather historical api[7] and all the geodata from the google maps api[8]. Relative work has been done by Rongjin Yang, Lizeyan Yin, Xuejie Hao, Lu Liu, Chen Wang, Xiuhong Li & Qiang Liu, who developed a model to predict emissions by using low-cost microstation data and machine learning[9]. The proposed methodology in this paper can be used for prioritizing possible pollution sources and for estimating pollutant emissions without using measurement equipment or sources emission data, if a dataset of accurate pollutant concentration measurements has been acquired for some months at the location of interest and if no new major pollutant emission sources are placed near the location of interest after the model has been trained. The developed machine learning model that is used for predicting the pollutant concentration per hour, receives as input only hourly meteorological data.

METHODOLOGY

At the first step of the proposed methodology the user selects a geographical location of interest for which there are available accurate hourly pollutant concentration data for a time period of at least six months and several surrounding geographical locations in which pollution sources exist and affect the pollutant concentration measurements at the site of interest. In the case study deployed here, the location of interest is the Lorraine Granado Community Park at Colorado United States, and the user selected pollution sources are energy production units that are at most 3 hours away by car from the location of interest. The chosen power stations that are located far away from the location of interest are coal fired power generating stations with very high reported emissions, like for example the Craig power station, located at 2101 S Ranney St, Colorado. The power station that is nearest to the location of interest is the Cherokee power plant, which is a natural gas-fired power plant 4.3 miles away from the community park and produces significantly lower reported NOx emissions as compared to the Craig coal fired power plant (about 15 times lower). Another power plant located very near the location of interest is the Sancor power plant, which was closed during the time the NOx concentration measurements were acquired, between January 2023 and April 2023, so it was not included in the study. Sancor, the company that owns that power plant has agreed to the largest air pollution penalty in Colorado history[10] for air quality related violations and for that specific power plant there are generally no available data in the emissions database of the United States Environmental Agency protection, while for all the rest nearby power plants data can be found in the database.

At the second step, once the user selects all the locations, he can then get all the relative meteorological data for these sites from the open-meteo api by performing a simple GET request through the interactive Angular frontend. The data are stored in a PostgreSQL database and can be used and queried after that. The user can create interactive time series visualization plots like the one shown in Figure 1 and view all time series combined one under the other at a multiplot setup with common time axis. This helps the user to reach into useful initial qualitative estimations as shown in Figure 1. By visual inspection of the plots it can be observed that when the mean of the reported factories NOx emissions starts to drop significantly (after March 12), the NOx concentration mean also drops. Also, when there are spikes observed in the factories locations wind speeds and at the same time in their respective NOx emissions, usually spikes are also observed in the NOx concentrations at the measurement site with some delay that varies between several hours to a view days. All these observations and many others need to be verified by a model that will be able to predict the NOx concentrations with acceptable accuracy, given the rest of the time series data as input. Ofcourse there are many more NOx emission sources that contribute more or less to the NOx concentration time series, like for example cars, construction activity and other industries. Part of these NOx emission sources are indirectly considered in the model, because the input data of the model is a large amount of meteorological data that correspond to the factories locations, so if for example wind speed at some factory location plays an important role in the predictablility of the model, then this may be attirbuted not only to the factory NOx emissions, but also maybe to some high car traffic in the roads in between the factory and the measurement site. This is why the usefuleness of the developed model is comparative and tries to answer the question: “to which direction should we **first** look for, when there is high NOx pollution at a measurement site?” By taking into consideration traffic counts in the nearby driveways and the existence or not of any other nearby industries the user can get useful insights.

Next the user can train and use a machine learning model that will have as input the meteorological data derived from the previous GET request and as output any available pollutant concentration measurements at the measurement site. The model that is currently available in the backend of the platform is a neural network built by using Keras in Python. The model is a sequential model, meaning layers are stacked on top of each other sequentially. The first layer is a Dense layer, which is a fully connected layer. The number of units is 256 and the activation function is the 'relu' activation function. The data pipeline procedure consists of min-max scaling the data and then applying non linear pca before inserting them into the neural network. The model was applied to the dataset acquired at the Lorraine Granado community park at 2023 and the result was a mape (mean absolute percentage error) of 48%, which is considered acceptable and a very good ability to detect spikes as it is shown in FIgure 3. Cross-validation and a train-test split 60:40 was applied on the dataset.

After performing perturbation analysis, where the input data of various locations were removed from the input data and the drops at mape values were recorded, it was found that removing all power station data and leaving only the meteorological data at the measurement site as input data, made the model unacceptable with a mape of 124% and with poor ability to detect spikes as shown in Figure 2. More importantly, removing the nearby Cherokee power station data and the JM Shaffer power station data, while keeping all the rest of the power station data, led to the highest drop of mape (7% and 5% respectively) as compared to removing the data of other power stations. This aligns with the results one would expect after using the Gaussian Plume model with the reported emission rates, the respective distances and the meteorological conditions between the selected factories and the location of interest. Finally, by approximately reversing PCA the most important variables were revealed: Temperature at the nearby Cherokee power plant and wind speed along with direct radiation at the JM Shafer power plant were the most influential variables of the model.

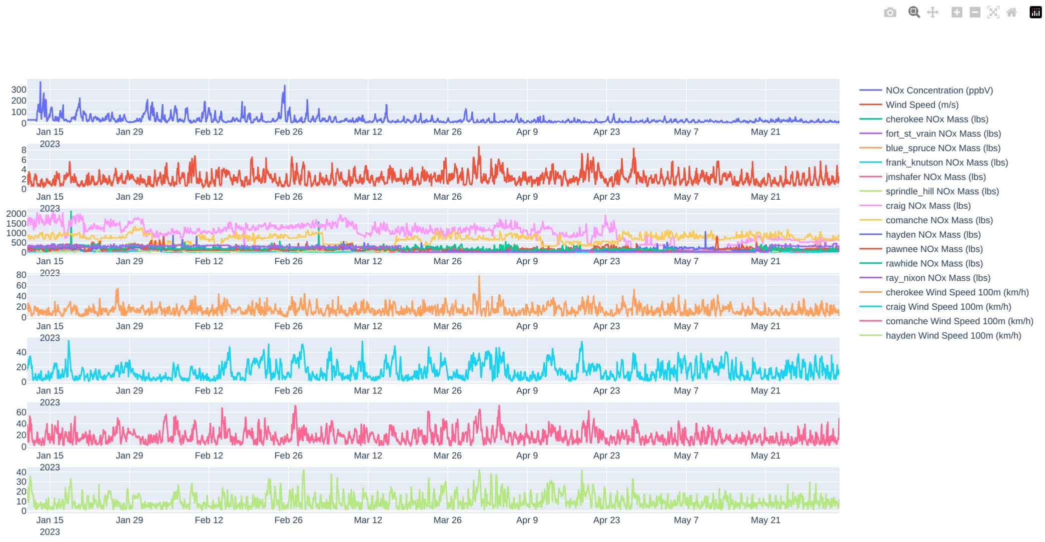

Figure 1. *All data that correspond to the user selected locations appear in a combined plot with common time axis for visual inspection of the interactions between the time series and for trend detection. The time series plot that appears first from the top shows the NOx concentration measurements at Lorraine Granado community park. The plot that is placed second starting from the top, shows the wind speed at the measurement location. The time series plot that appears third from the top shows reported NOx emissions from the user selected geographical locations, which in this case are nearby gas power factories and coal power factories at most 3 hours away by car from the measurement site. The rest of the plots are wind speed time series at the factories locations.*

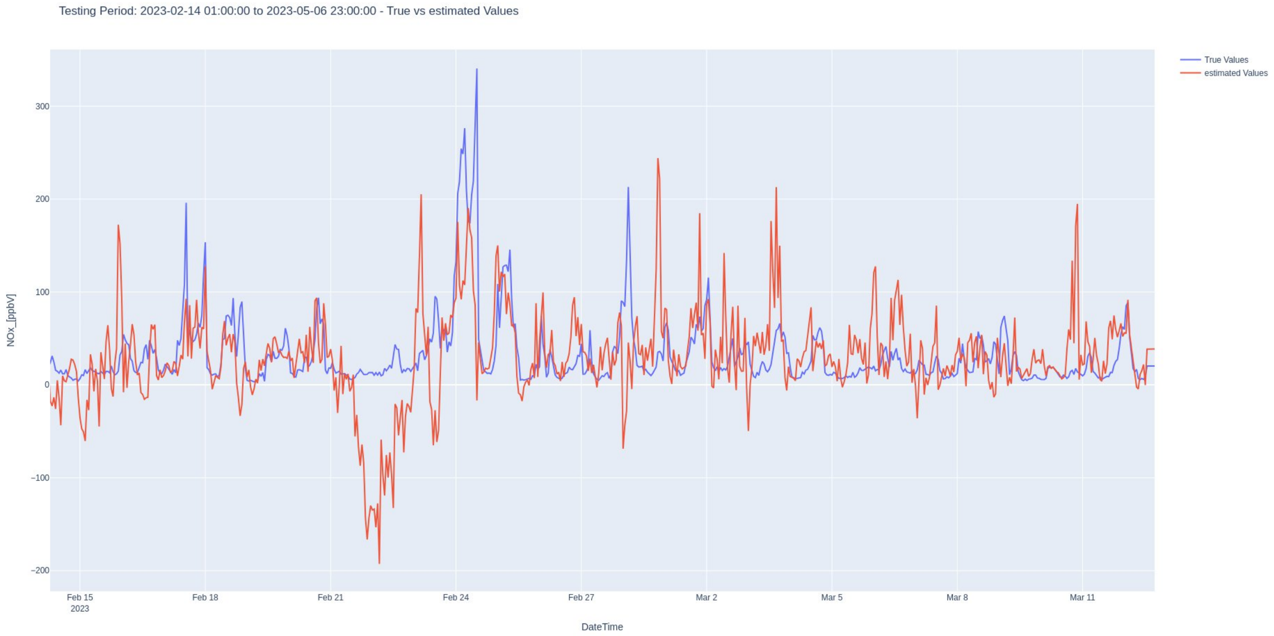

Figure 2. *The figure shows the relatively poor predictability of the developed model when the only input it receives are the meteorological data at the NOx concentrations measurement site. The mape is 124% and the model is considered overall unacceptable.*

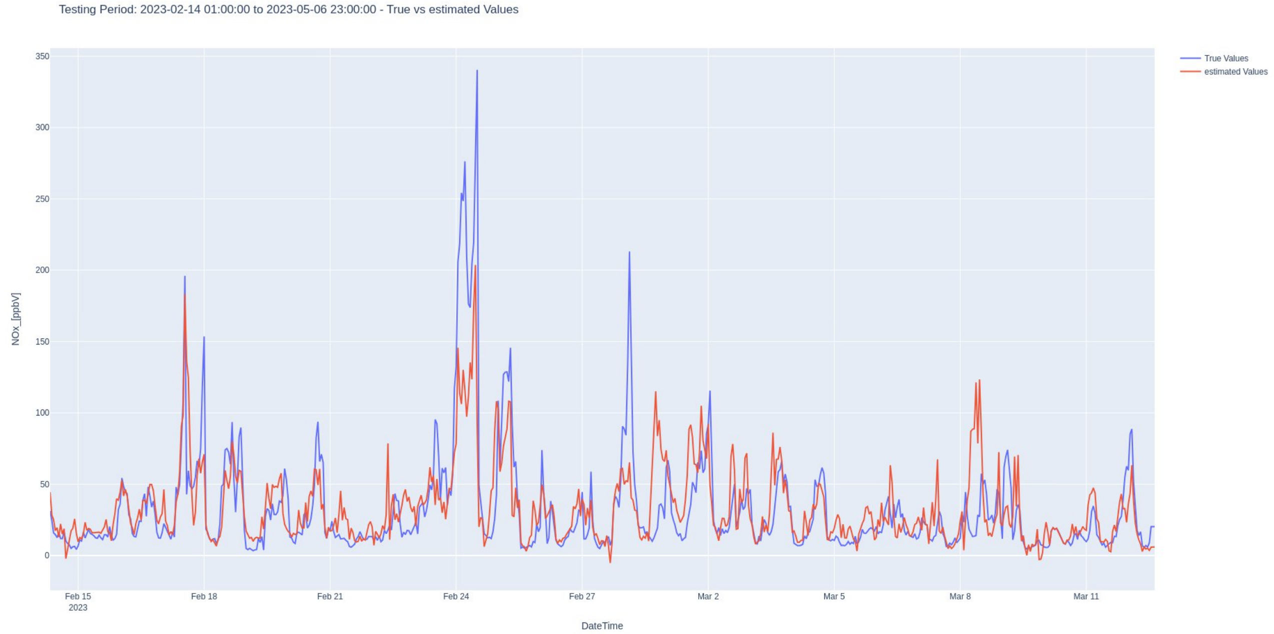

Figure 3. *The figure shows the good predictability of the developed model when the input it receives are the meteorological data at the NOx concentrations measurement site and the meteorological data at the locations of the surrounding power stations. As it is evident spikes can be detected and the mape is 48%. The model is considered acceptable to be used for predicting NOx concentrations.*

RESULTS AND DISCUSSION

A methodology that consists of a fully connected (dense) neural network that receives as input hourly meteorological data acquired from open-meteo, based on user selected geo-data acquired from the google maps api can be used to predict the hourly concentration of NOx emissions at polluted sites if no new emission sources are introduced near the measurement site after the model's training phase and if accurate measurements were acquired and used for the training phase of the model. It is significant that the model does not require pollutant emissions data from the possible pollution sources to estimate pollution concentration at the site of interest, either in the training, test or usage phase. The proposed methodology can also be used for pollutant source detection through prioritization of the datasets acquired from the user selected geographical data points, by using perturbation analysis to evaluate the effect of each dataset to the mean absolute percentage error of the model and by also visually inspecting the prediction results through interactive data visualization plots.

REFERENCES

[1] Daly, A. and P. Zannetti. 2007. Air Pollution Modeling – An Overview. Chapter 2 of AMBIENT AIR POLLUTION (P. Zannetti, D. Al-Ajmi, and S. Al-Rashied, Editors). Published by The Arab School for Science and Technology (ASST) (http://www.arabschool.org.sy) and The EnviroComp Institute (http://www.envirocomp.org/).

[2] O O Ajayi et al 2021 IOP Conf. Ser.: Mater. Sci. Eng. 1036 012017.

[3] Kawka, Marcin, Joanna Struzewska, and Jacek W. Kaminski. 2023. "Downscaling of Regional Air Quality Model Using Gaussian Plume Model and Random Forest Regression" Atmosphere 14, no. 7: 1171. https://doi.org/10.3390/atmos14071171.

[4] Li, L., Wang, J., Franklin, M. et al. Improving air quality assessment using physics-inspired deep graph learning. npj Clim Atmos Sci 6, 152 (2023). https://doi.org/10.1038/s41612-023-00475-3.

[5] https://www.colorado.gov/airquality/air_toxics_repo.aspx#og.

[6] https://campd.epa.gov/data.

[7] https://open-meteo.com/en/docs/historical-weather-api.

[8] https://developers.google.com/maps.

[9] Yang, R., Yin, L., Hao, X. et al. Identifying a suitable model for predicting hourly pollutant concentrations by using low-cost microstation data and machine learning. Sci Rep 12, 19949 (2022). https://doi.org/10.1038/s41598-022-24470-5.

[10]https://www.cpr.org/2024/02/05/suncor-energy-agrees-to-largest-air-pollution-penalty-in-colorado-history/