# NO2 Forecasting using Sentinel 5P and Weather Forecast data: a Deep Learning Approach

Huriel Reichel

## Introduction

Nitrogen dioxide ($NO_2$ ) is a greenhouse gas strongly related to respiratory diseases (Heinrich et al., 1999, Hagen et al., 2000) often used as reference for pollution indicators together with carbon dioxide ($CO_2$ ). Monitoring the presence of this gas in the atmosphere is therefore of high health importance, and the same could be said of its forecasting.

Classic weather (including air quality) forecasting makes use of a series of data-sets, including weather stations, satellite imagery, radar data, etc., and includes them for computation in physical models requiring super computers ((Rasp et al. 2020)). However, other strategies have been investigated that demand less computational cost and can reach similar results ( Rasp and Thuerey (2021) ). These strategies involve using deep learning for the predictive analysis, mainly using time-series data from weather stations ( Kurt et al. (2008) ; Russo and Soares (2013) ; Tsai, Zeng, and Chang (2018) ; Kaselimi et al. (2020) ; Masood and Ahmad (2021) ; Samal et al. (2021) ; Heydari et al. (2021) ).

By considering deep learning as a predictive model for pollution, one cannot ignore its ability to detect multiple variability and abstractness in data. Not only that, but it can take into account different sources of data. And when considering the availability of data related to $NO_2$ monitoring, one cannot ignore the imagery from Copernicus's SENTINEL 5P (S5P) satellite. This spacecraft has a daily temporal resolution by covering the entire earth, which makes of it a substantial possibility for a world-wide forecasting scenario, especially when using deep learning. Making use of this dataset and deep learning's abstractness capabilities, it is possible to forecast the whole pollution surface (pixel by pixel) in different future times, in spite of considering the aggregated time-series singularly.

As a primary aim, a deep learning model is trained and evaluated using beyond time-series data, but the series of satellite imagery from SENTINEL 5P for $NO_2$ daily forecasting. The architecture to be applied is inspired by the one used by Shi et al., 2015, Liu et al., 2018 and

Wen et al., 2019, and consists of a Convolutional Long Short-Term Neural Network (CNN-LSTM). The main rationale is to make use of both space (through the convolution - CNN) and time (through the LSTM) to forecast new “frames” in SENTINEL 5P images. This has already been tested in similar approaches, as the one from Heydari et al. (2021 Heydari et al. (2021)), although what has not been explored yet is the use of weather related covariables that could aid the forecasting goal. Many weather related phenomena affect the chemical reactions behind $NO_2$ formation in the atmosphere, which are also variables measured spatially and temporally. The goal is then to compare a CNN-LSTM model with weather related covariables and without them in therms of computational complexity and mostly accuracy in daily forecasts.

Nevertheless, before one reaches that, the next topic is here to present to give some deeper background on the topic of pollution forecasts and Deep Learning for the purpose of forecasting it.

## Background

### Weather and Pollution Forecasting

As before mentioned, $NO_2$ is strongly associated with respiratory diseases (Zhang, Wang, and Lu (2018); Jo et al. (2021)), hence it is deeply studied and often considered a target variable in pollution oriented research. This is only more clear, when looking at works focused on forecasting pollution (Shams et al. (2021) ; Pawul and Śliwka (2016) ; Ebrahimi Ghadi, Qaderi, and Babanezhad (2018) ; Mohammadi et al. (2016) ), where regardless of the model used, $NO_2$, among other possible chemicals, is central in these modelling works. In that sense, it is of high importance to first better understand this chemical, and mostly how it is formed into the atmosphere.

Nitrogen dioxide ($NO_2$) is a gaseous air pollutant that forms in the atmosphere primarily through the oxidation of nitrogen oxide ($NO$) in the presence of sunlight and other atmospheric constituents. The formation of $NO_2$ is closely linked to several chemical and physical processes, and weather conditions play a significant role in its formation, transformation, and dispersion in the atmosphere.

To what concerns the formation mechanism of this chemical, sunlight is crucial for the so called process of photochemical convertion of $NO$ into $NO_2$ when combined with $O_3$ (Ozone) (Wood et al. (2009)). This is the case even considering that anthropogenic sources like vehicle emissions, industrial activities and power generation. Hence sunlight availability is of root importance for the presence of NO2 in the atmosphere. Not only that, temperature acts as a catalyst for chemical reactions, and usually the higher, the higher the chance for $NO_2$ presence (Wood et al. (2009) ). Other factor, just as important are phenomena directly linked to the atmosphere’s stability, causing the movement of the gas among the different air layers. Those variables can be indicated by measures of wind speed, wind direction and precipitation, for example. Grundström et al (2015 - Grundström et al. (2015) ) appoints wind speed, vertical temperature gradient and weather type as variables with a strong relationship with the presence of $NO_2$ .

A similar work is presented by Samal et al (2021 Samal et al. (2021) ), which compares the use of several different models to forecast PM2.5, another pollutant indicator, making use of weather related variables as supporters, being them wind speed, temperature, wind direction, rainfall, and humidity. In their work, this relationship is tested in those models and it is left as clear how important they are and that weather variables should be used for more accurate pollution forecasting. In summary, one could add that weather is extremely relevant for explaining pollution and it should be considered when attempting to forecast it in a data-driven approach.

### Time Series Forecasting

Pollution forecasting is from many perspectives a time-series challenge (Freeman et al. (2018); Samal et al. (2021) ) and it has been approached as so by researchers. A time-series is essentially a set of observations stored in a sequence, which finally corresponds to time Chatfield (2000). Regardless of the model, if within a data-driven approach, the logic is to use past information to take assumption on the future. Any new information will be given based on past one. This is the approach present in the works of Zhu et al. (2018) and Kumar and Jain (2009) that make use of ARIMA based models, but also of works that developed neural networks for similar tasks (Freeman et al. 2018; Agirre-Basurko, Ibarra-Berastegi, and Madariaga 2006; Kurt et al. 2008; Samal et al. 2021; Tsai, Zeng, and Chang 2018).

In general, given the abstraction level of the phenomenon, deep learning approaches tend to reach more accurate results (Samal et al. (2021) ; Pawul and Śliwka (2016); Shams et al. (2021); Rasp et al. (2020) ) than standard statistical or even other machine learning approaches ("shallow" ones). Nevertheless, the limit of what be done with deep learning models is still far to be reached, and research reaching more and more accurate results are constant. Exploring this further is therefore a systemic scientific work.

Moreover, Wen et al. (2019) makes use of a sophisticated neural network architecture for the forecast of pollutants, whilst Samal et al. (2021) goes further and uses a similar model to forecast the same variable, although now using weather supportive variables. Although, one important aspect is missing in all of these works. The analysis of the spatial pattern of results. Space is completely ignored in the analysis of results, and every model, pehaps even includes the spatial component in the modelling itself, but aggregates it to a typical times series in the end. Questions related to the spatial pattern of the forecast errors or the presence of spatial dependence in forecasts are all neglected when the topic of pollution forecasting is dismantled into having only the temporal component.

Whenever decision have to be made regarding pollution, this decision has boundaries associated to political frontiers. When considering the temporal component only, space and important aspects of it are neglected in the analysis. Hence, whenever research has been made about the use of deep learning for pollution forecasting, it is either space which is forgotten in the model or in the results, or weather that is not being considered as a covariable. A gap in literature is therefore present.

### Spatial Data Cubes and Sentinel 5P

Once more, it was mentioned that past work is above all focused on “time-only” models. The input data for these models is based overall on sensors on-board weather stations spread all over Earth’s surface or even other systems constantly measuring the desired variable. For that reason, if these stations are not considered into a single model, space is ignored. Nevertheless, now considering a space oriented model, what is the kind of input data one can expect?

Fortunately, Earth Observation technology has been substantially advancing and the deployment of satellites in closer ranges are constantly imaging Earth’s and other planets surfaces for full coverage data. These includes RGB imagery, Infrared imagery, Water Coverage, Albedo, Digital Elevation data, and many others. Among the several missions to do so, an important spotlight shall be given to the Copernicus Missions for the European Space Agency (ESA). The completely free-access terabytes of data derive mainly from the Sentinel satellites aimed at giving Europe autonomous capacity in plenty of Earth Observation tasks (Jutz and Milagro-Pérez (2020)). For the task of pollution forecasting, there is Sentinel 5P (S5P), which on-boards the TROPOMI sensor, the most advanced multispectral imaging spectrometer to date (2023). With its global coverage, TROPOMI measures the ultraviolet and visible, near-infrared and shortwave infrared spectral bands, what supports a high accurate capture of pollutants such as $NO_2$, $O_3$ , $CH_2O$ , $SO_2$ , $CH_4$ and $CO$ .

To what refers to weather related data, considering the same format, *i.e.,* preserving the spatial component, an important source of that is the European Centre for Medium-Range Weather Forecasts (ECMWF). Among their many datasets, a very popular one is ERA5 (Reanalysis v5), which has also global coverage and ranges from today to 1940. It contains several weather related data which can all be freely accessed through an API or a web-portal (Copernicus Climate Change Service (2023) ).

- Spacetime cubes…

### Spacetime Deep Learning

- How do we include these cubes into a model
- The challenge of big data and deep learning (GPU…)
- CNN-LSTM and other models

Figure 1: Illustrative Figure of the S5P satellite and the variables its sensor (TROPOMI) is constantly measuring. Source: ESA

### Current Work

- Fill the gapS, use weather
- Analyse the space further

### Objectives

- clean

Therefore, we aim to investigate the feasibility of using SENTINEL 5P data to train a CNN-LSTM network and, consequently, analyse whether this architecture suits this purpose when adding weather related covariables to it. A comparison in therms of accuracy to both physical and data-driven models will be followed, taking into consideration the different purposes of the models cited.

Moreover, we wish to evaluate the accuracy decay in time and, hence, verify, how far can predictions go into the future.

### Research Questions

The study will be answering the following research questions: To what extent can one make use of a CNN-LSTM network to forecast daily $NO_2$ levels with SENTINEL 5P Imagery and weather related covariables? How does a CNN-LSTM network using SENTINEL 5P and weather data data performs compared to the same model without weather covariables, and how does the accuracy decay behaves when trying to predict further in time ?

## Methodology

- Describe it horizontally...
- Figure of Methods

### S5P Data Acquisition and Wrangling through openEO

- openeo, bbox... area figure
- Cloud Cover Interpolation

### ERA5 Data Acquisition through ECWMF

- web-platform and bbox...
- Weather Feature Engineering

### Development of the Neural Network

- Torch Tensor and Windows (composing the data into a single thing - warping, etc)
- RNN vs LSTM vs CNN vs CNN-LSTM
- Image of Architecture
- Torch Implementation
- Early Stopping
- Hyper parameter tuning

### Validation Procedure

- Performance Analysis
- RMSE
- Aggreagated RMSE
- Qualitative Spatial Pattern Analysis

### Hardware and Reproducibility

- PALMA, GPU, R version…

## Results & Discussion

- Performance comparison
- RMSE (comparison to other works)
- Figure of errors
- RMSE aggreagted.. (comparison- better because it makes sense)
- time series plot
- Spatial patterns plot

## Conclusion

Agirre-Basurko, E., G. Ibarra-Berastegi, and I. Madariaga. 2006. “Regression and Multilayer Perceptron-Based Models to Forecast Hourly O3 and NO2 Levels in the Bilbao Area.” *Environmental Modelling & Software* 21 (4): 430–46. https://doi.org/10.1016/j.envsoft.2004.07.008.

Chatfield, Chris. 2000. *Time-Series Forecasting*. CRC press.

Copernicus Climate Change Service. 2023. “Complete ERA5 Global Atmospheric Reanalyis.” ECMWF. https://doi.org/10.24381/CDS.143582CF.

Ebrahimi Ghadi, M., F. Qaderi, and E. Babanezhad. 2018. “Prediction of Mortality Resulted from NO2 Concentration in Tehran by Air Q+ software and Artificial Neural Network.” *International Journal of Environmental Science and Technology* 16 (3): 1351–68. https://doi.org/10.1007/s13762-018-1818-4.

Freeman, Brian S., Graham Taylor, Bahram Gharabaghi, and Jesse Thé. 2018. “Forecasting Air Quality Time Series Using Deep Learning.” *Journal of the Air & Waste Management Association* 68 (8): 866–86. https://doi.org/10.1080/10962247.2018.1459956.

Grundström, M., C. Hak, D. Chen, M. Hallquist, and H. Pleijel. 2015. "Variation and Co-Variation of PM10, Particle Number Concentration, NOx and NO2 in the Urban Air – Relationships with Wind Speed, Vertical Temperature Gradient and Weather Type." *Atmospheric Environment* 120 (November): 317–27. https://doi.org/10.1016/j.atmosenv.2015.08.057.

Heydari, Azim, Meysam Majidi Nezhad, Davide Astiaso Garcia, Farshid Keynia, and Livio De Santoli. 2021. "Air Pollution Forecasting Application Based on Deep Learning Model and Optimization Algorithm." *Clean Technologies and Environmental Policy* 24 (2): 607–21. https://doi.org/10.1007/s10098-021-02080-5.

Jo, Sungyang, Ye-Jee Kim, Kye Won Park, Yun Su Hwang, Seung Hyun Lee, Bum Joon Kim, and Sun Ju Chung. 2021. "Association of $NO_2$ and Other Air Pollution Exposures With the Risk of Parkinson Disease." *JAMA Neurology* 78 (7): 800. https://doi.org/10.1001/jamaneurol.2021.1335.

Jutz, S., and M. P. Milagro-Pérez. 2020. "Copernicus: The European Earth Observation Programme." *Revista de Teledetección*, no. 56 (November). https://doi.org/10.4995/raet.2020.14346.

Kaselimi, Maria, Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Demitris Delikaraoglou. 2020. "A Causal Long Short-Term Memory Sequence to Sequence Model for TEC Prediction Using GNSS Observations." *Remote Sensing* 12 (9): 1354. https://doi.org/10.3390/rs12091354.

Kumar, Ujjwal, and V. K. Jain. 2009. "ARIMA Forecasting of Ambient Air Pollutants (O3, NO, NO2 and CO)." *Stochastic Environmental Research and Risk Assessment* 24 (5): 751–60. https://doi.org/10.1007/s00477-009-0361-8.

Kurt, Atakan, Betul Gulbagci, Ferhat Karaca, and Omar Alagha. 2008. "An Online Air Pollution Forecasting System Using Neural Networks." *Environment International* 34 (5): 592–98. https://doi.org/10.1016/j.envint.2007.12.020.

Masood, Adil, and Kafeel Ahmad. 2021. "A Review on Emerging Artificial Intelligence (AI) Techniques for Air Pollution Forecasting: Fundamentals, Application and Performance." *Journal of Cleaner Production* 322 (November): 129072. https://doi.org/10.1016/j.jclepro.2021.129072.

Mohammadi, Mohammad Javad, Gholamreza Goudarzi, Sahar Geravandi, Ahmad Reza Yari, Bashir Ghalani, Saeed Shirali, Elahe Zallaghi, and Mehdi Esmaili. 2016. "Dispersion Modeling of Nitrogen Dioxide in Ambient Air of Ahvaz City." *Health Scope* 5 (2). https://doi.org/10.17795/jhealthscope-32540.

Pawul, Małgorzata, and Małgorzata Śliwka. 2016. "APPLICATION OF ARTIFICIAL NEURAL NETWORKS FOR PREDICTION OF AIR POLLUTION LEVELS IN ENVIRONMENTAL MONITORING." *Journal of Ecological Engineering* 17 (4): 190–96. https://doi.org/10.12911/22998993/64828.

Rasp, Stephan, Peter D. Dueben, Sebastian Scher, Jonathan A. Weyn, Soukayna Mouatadid, and Nils Thuerey. 2020. "WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting." *Journal of Advances in Modeling Earth Systems* 12 (11). https://doi.org/10.1029/2020ms002203.

Rasp, Stephan, and Nils Thuerey. 2021. "Data-Driven Medium-Range Weather

Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench." *Journal of Advances in Modeling Earth Systems* 13 (2). https://doi.org/10.1029/2020ms002405.

Russo, Ana, and Amílcar O. Soares. 2013. "Hybrid Model for Urban Air Pollution Forecasting: A Stochastic Spatio-Temporal Approach." *Mathematical Geosciences* 46 (1): 75–93. https://doi.org/10.1007/s11004-013-9483-0.

Samal, K. Krishna Rani, Ankit Kumar Panda, Korra Sathya Babu, and Santos Kumar Das. 2021. "An Improved Pollution Forecasting Model with Meteorological Impact Using Multiple Imputation and Fine-Tuning Approach." *Sustainable Cities and Society* 70 (July): 102923. https://doi.org/10.1016/j.scs.2021.102923.

Shams, Seyedeh Reyhaneh, Ali Jahani, Saba Kalantary, Mazaher Moeinaddini, and Nematollah Khorasani. 2021. "Artificial Intelligence Accuracy Assessment in NO2 Concentration Forecasting of Metropolises Air." *Scientific Reports* 11 (1). https://doi.org/10.1038/s41598-021-81455-6.

Tsai, Yi-Ting, Yu-Ren Zeng, and Yue-Shan Chang. 2018. "Air Pollution Forecasting Using RNN with LSTM." *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, August. https://doi.org/10.1109/dasc/picom/datacom/cyberscitec.2018.00178.

Wen, Congcong, Shufu Liu, Xiaojing Yao, Ling Peng, Xiang Li, Yuan Hu, and Tianhe Chi. 2019. "A Novel Spatiotemporal Convolutional Long Short-Term Neural Network for Air Pollution Prediction." *Science of The Total Environment* 654 (March): 1091–99. https://doi.org/10.1016/j.scitotenv.2018.11.086.

Wood, E. C., S. C. Herndon, T. B. Onasch, J. H. Kroll, M. R. Canagaratna, C. E. Kolb, D. R. Worsnop, et al. 2009. "A Case Study of Ozone Production, Nitrogen Oxides, and the Radical Budget in Mexico City." *Atmospheric Chemistry and Physics* 9 (7): 2499–2516. https://doi.org/10.5194/acp-9-2499-2009.

Zhang, Zili, Jian Wang, and Wenju Lu. 2018. "Exposure to Nitrogen Dioxide and Chronic Obstructive Pulmonary Disease (COPD) in Adults: A Systematic Review and Meta-Analysis." *Environmental Science and Pollution Research* 25 (15): 15133–45. https://doi.org/10.1007/s11356-018-1629-7.

Zhu, Min, Jing Xia, Xiaoqing Jin, Molei Yan, Guolong Cai, Jing Yan, and Gangmin Ning. 2018. "Class Weights Random Forest Algorithm for Processing Class Imbalanced Medical Data." *IEEE Access* 6: 4641–52. https://doi.org/10.1109/access.2018.2789428.