Machine Learning for Air Quality Prediction Using High-Resolution TEMPO Remote Sensing Data

Sarah Scott[1,3], Alexander Radkevich[1,2], Hazem Mahmoudh[1,2]

(1) ASDC NASA Langley Research Center, Hampton, VA, (2) ADNET-SYSTEMS (3) NASA OSTEM Intern

sarah.r.scott@nasa.gov , alexander.radkevich@nasa.gov , hazem.mahmoud@nasa.gov

North Carolina Space Grant Symposium| April 11, 2025

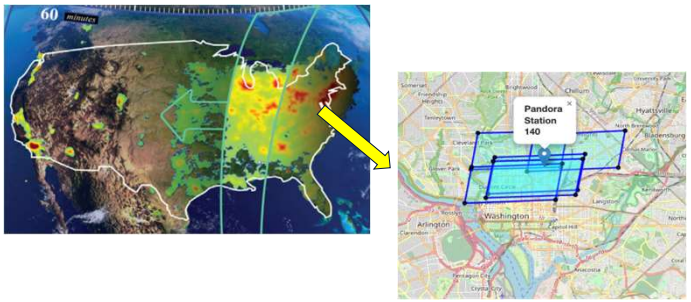
National Aeronautics and Space Administration

Introduction to TEMPO Mission

NASA's Tropospheric Emissions: Monitoring of Pollution (TEMPO) instrument provides advanced measurements of key pollutants—ozone, nitrogen dioxide, formaldehyde, sulfur dioxide, and aerosols—using an ultraviolet and visible-light grating spectrometer. From geostationary orbit, TEMPO delivers high temporal and spatial resolution, with an hourly pixel resolution of 2 km (North/South) and 4.7 km (East/West) over North America. This fine-scale data enables near real-time tracking of pollution patterns on an urban scale, offering insights into hourly air quality dynamics. We aim to utilize machine learning to predict tropospheric nitrogen dioxide (NO_2)during the validation phase of the TEMPO project. These data should be considered as provisional products per the Provisional Product Maturity level defined in the TEMPO validation plan. These data are at provisional maturity, which means that product performance has been demonstrated through a large, but still (seasonally or otherwise) limited number of independent measurements. We initialize our training with Random Forests to predict in the absence of TEMPO measurements, validating with validated ground-based Pandora NO2 measurements, ensuring consistency with real-world observations. To further enhance predictions, we propose a hybrid approach integrating a Random Forest with a Convolutional Neural Network (CNN).

Proof of Concept: Prediction during TEMPO Special Operations.

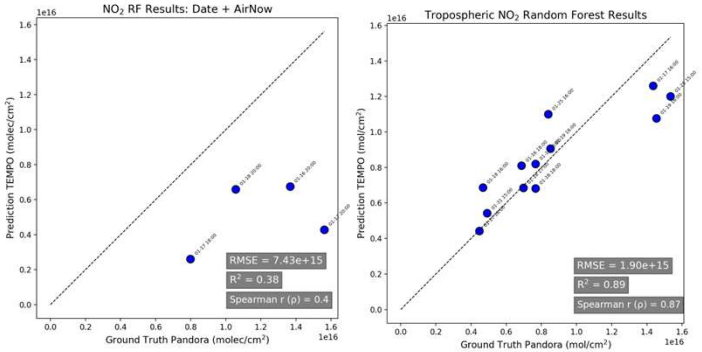
We define our period of interest for TEMPO validation during the January 2025 Los Angeles Wildfires, where the TEMPO instrument underwent special operations to focus on the fires, performing continent-wide scans every other hour, resulting in every other hour missing data from 01/16-01/19. We define our area of interest to be Washington D.C., due to ground sensor data availability, and high variability in NO2 distributions throughout the day.



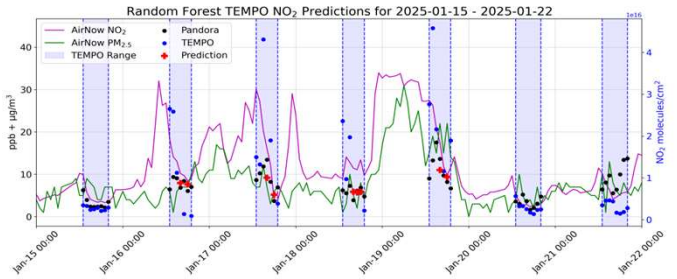
(A) Aeriel view of TEMPO coverage across North America, mapping high-resolution measurements hourly. (B) Spatial coverage of TEMPO pixels over several days at 12:00 with reference Pandora Station 40 in Washington DC (38.9218,-77.0124) in January 2025. Spatial resolution is consistent over ground sensor, ensuring continuous coverage for evaluation.

Current Results

Initial validation of the Random Forest model against ground-based Pandora sensor data demonstrates promising results, highlighting the strengths in Random Forests for feature selection and nonlinear interpretability. As we incorporate additional training data, and temporal based averages of TEMPO data, we observe a significant improvement in predictive performance, reflected in a reduction of Root Mean Square Error (**RMSE**) against Pandora ground measurements. Current features involve day, hour, AirNow NO2 , AirNow PM2.5, Temperature, Relative Humidity, Pressure, Wind Speed 10M, Wind Direction 10M, rolling daily TEMPO average, and lagged TEMPO.



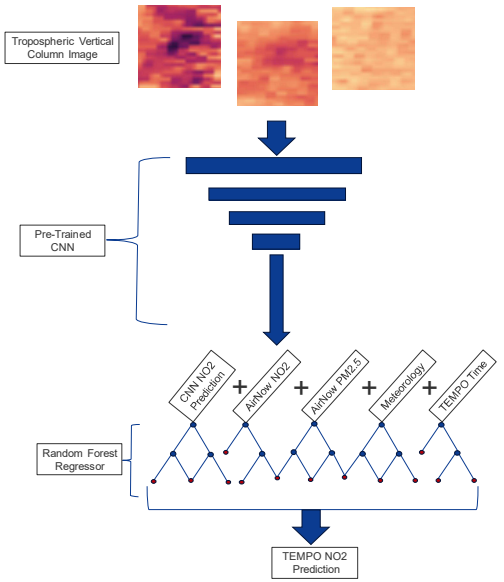
(A). Baseline RF predictions with only date, and AirNow inputs for the dates of 01/16-01/18 . (B). Current RF predictions with date, AirNow, Meteorology, and rolling temporal statistics as input features over all of January 2025.



Comparison of measurement trends over time period between Pandora NO2, TEMPO column NO2, and AirNow ground sensor NO2 and PM2.5. Random Forest predictions depicted in red.

In Progress—Predictions with CNN + RF

To leverage the spatial feature extraction capabilities of CNNs, we propose integrating CNN-based predictions alongside the existing RF feature inputs. This hybrid approach aims to enhance our model's ability to capture both spatial and temporal patterns in NO2 distributions. The CNN will be trained on TEMPO images collected over the Washington, D.C. area in 2024, ensuring exposure to a wide range of seasonal and meteorological variations of NO2. to capture as much variation in NO2 images as possible. Once trained, we will infer over our current period of interest to assess any improvement in predictive performance.



Future Work

With the evaluation of our hybrid model, we aim to extend our spatial prediction capabilities by incorporating Gaussian Processes, enabling more robust spatial estimations in regions with missing data. This will further validate the model's reliability in predicting concentrations across different locations and time periods with incomplete observations. Additionally, since TEMPO provides measurements exclusively during daylight hours, developing a complementary model for nighttime and lunar-based observations is crucial for achieving a continuous representation of NO2 distributions. To support this effort, we are actively working on obtaining access to lunar measurements via Pandora.

Hazem Mahmoud, Ph.D.
ASDC DAAC Scientist
hazem.mahmoud@nasa.gov

Scan the QR code to access science enabling tool for TEMPO data validation:
https://github.com/nasa/ASDC_Data_and_User_Services/blob/main/TEMPO/additional_drafts/ASDC_Data_Processing_ML_v0.ipynb

POWER Prediction Of Worldwide Energy Resources

Atmospheric Science Data Center

EARTHDATA