



1 **Estimating Hourly Surface NO₂ Across the Contiguous United**
2 **States Using TEMPO Observations and Machine Learning**

3 Steven Aarons¹, Wenge Ni-Meister¹, and Prathap Ramamurthy²

4 ¹Department of Geography and Environmental Science, Hunter College of CUNY, New York, NY 10065, United
5 States of America

6 ²Department of Mechanical Engineering, City College of New York, NY 10031, United States of America

7 *Correspondence to:* Steven Aarons (steven.aarons48@hunter.cuny.edu)

8



9 **Abstract.** Ground-level nitrogen dioxide (NO_2) poses significant health risks, yet hourly concentrations have not
10 been estimated due to the limited temporal sampling of polar-orbiting satellites. NASA's Tropospheric Emissions:
11 Monitoring of Pollution (TEMPO) geostationary instrument provides the first hourly tropospheric NO_2 observations
12 over North America, creating new opportunities for estimating surface NO_2 throughout the daytime. We develop
13 XGBoost models using TEMPO tropospheric NO_2 , ERA5 meteorology, road density, impervious surface percentage
14 (ISP), and temporal features to predict hourly surface NO_2 at 474 EPA stations across the contiguous United States
15 from August 2023 through June 2025, and map concentrations at TEMPO's L3 resolution. Separate TEMPO
16 (daylight) and no-TEMPO (nighttime/imputation) models achieve test $R^2 = 0.78$ (RMSE = 3.33 ppb) and $R^2 = 0.76$
17 (RMSE = 4.36 ppb), respectively. TEMPO column NO_2 is the strongest predictor by gain, while ISP dominates
18 spatial prediction and boundary layer height is the most influential meteorological predictor. Importantly, TEMPO
19 provides complementary spatial information beyond meteorological and temporal predictors, increasing R^2 across all
20 spatial cross-validation folds by an average of 17% and improving performance during data-scarce periods. Spatial
21 cross-validation ($R^2 = 0.48$) further highlights opportunities to improve predictions at unseen locations through
22 higher-resolution spatial inputs, while extending satellite coverage into the evening could improve representation of
23 the evening NO_2 peak. To our knowledge, this is the first study to predict hourly surface NO_2 across CONUS using
24 geostationary satellite data, demonstrating the potential of TEMPO to support acute exposure assessments at
25 previously infeasible temporal resolutions.

26



27 1. Introduction

28 Nitrogen dioxide (NO_2) is a trace gas byproduct of anthropogenic combustion and natural processes such as soil
 29 nitrification and lightning. Near-surface NO_2 is transferred aloft by vertical mixing and synoptic processes or cycled
 30 via photolysis, producing NO and O to drive tropospheric ozone formation, and ultimately removed by reacting with
 31 OH to form nitric acid (HNO_3), the dominant daytime terminal sink (Mucha & Mainka, 2026; Seinfeld & Pandis,
 32 2006). Thermal inversions, stagnant winds, and topographical effects retain NO_2 near the surface and contribute to
 33 steep vertical gradients (Chang et al., 2025; Harkey et al., 2015). These processes, and the presence of VOCs or
 34 hydroxyl radicals, link NO_2 to the formation of secondary pollutants such as ground-level ozone and $\text{PM}_{2.5}$ with
 35 important environmental and health implications (Jeon et al., 2024; U.S. Environmental Protection Agency, 2025b).

36 Ground-level NO_2 poses a serious health risk and is one of six criteria air pollutants under the U.S. Environmental
 37 Protection Agency's (EPA) National Ambient Air Quality Standards (NAAQS) (U.S. Environmental Protection
 38 Agency, 2025c). Long-term exposure to NO_2 has been linked to increased risk of lung cancer, cardiovascular mortality,
 39 and respiratory mortality, with an estimated 170,850 premature deaths in the United States each year (Camilleri et al.,
 40 2023; Hamra et al., 2015; Huang et al., 2021). Acute or short-term exposure is also associated with multiple adverse
 41 health effects including: respiratory hospital admissions as well as all-cause, respiratory, and cardiovascular mortality,
 42 with an estimated 24,000 premature deaths in the EU per year (European Environment Agency, 2022a, 2022b; Mainka
 43 & Žak, 2022; Mucha & Mainka, 2026; Wang et al., 2021). Additionally, NO_2 exposure disproportionately affects
 44 marginalized communities with large racial disparities, particularly among Black communities who face ~47% higher
 45 mortality rates than the CONUS-wide average (Camilleri et al., 2023). Recent epidemiological studies reveal acute
 46 health effects from hourly NO_2 variability that coarser temporal aggregations fail to capture, highlighting the critical
 47 need for ground-level models that operate at fine temporal scales (Shukla et al., 2026).

48 Ground-based NO_2 monitoring provides critical data for understanding health effects. However, the EPA's NO_2
 49 monitoring network is designed primarily for NAAQS compliance in populated areas and not for spatial
 50 representation. As a result, AQS monitors are heavily clustered in urban and metropolitan areas, with large gaps across
 51 the rural interior of the United States (U.S. Environmental Protection Agency, 2024).

52 Satellite remote sensing extends coverage to areas without ground monitors; however, it measures total column NO_2 ,
 53 rather than surface concentrations, which must be transformed or used as predictors to infer ground-level NO_2 (Tang
 54 et al., 2024). Space-based tropospheric NO_2 observation began with the Global Ozone Monitoring Experiment
 55 (GOME) aboard ERS-2 (1995) and SCIAMACHY aboard ENVISAT (2002), but their coarse resolution and multi-
 56 day global revisit time limited their use for surface estimation (Bovensmann et al., 1999; Burrows et al., 1999). The
 57 Ozone Monitoring Instrument (OMI), aboard NASA's Aura satellite, launched in 2004, was the first to pair daily
 58 global coverage with a footprint fine enough (13 x 24 km at nadir) to resolve urban NO_2 gradients, enabling numerous
 59 studies to estimate ground-level concentrations (Kim et al., 2024; Levelt et al., 2018; Tang et al., 2024). GOME-2
 60 (MetOp-A, 2006) followed with near daily global coverage but a coarser 80 x 40 km footprint. The launch of



61 TROPOMI aboard Sentinel-5P in 2017 marked a substantial improvement in spatial resolution and signal-to-noise
62 ratio (Veefkind et al., 2012). These improvements have led to increased predictive power for models that attempt to
63 estimate surface NO₂ (Nawaz, Goldberg, et al., 2025). Despite these advances, these instruments are polar-orbiting
64 and provide once-daily temporal resolution, from mid-morning for GOME, SCIAMACHY, and GOME-2 to early
65 afternoon for OMI and TROPOMI, leaving the diurnal variation of surface NO₂ unresolved.

66 Surface NO₂ can be estimated using a range of approaches, including chemical transport models (CTMs), machine
67 learning (ML) (including deep learning), spatial interpolation, and land-use regression. CTMs are physics-based
68 models that simulate emissions, chemical reactions, transport, and deposition of atmospheric pollutants through
69 interactions between pollutants and meteorology (Gao & Zhou, 2024). CTM-derived vertical profiles have been
70 combined with satellite observations to estimate surface NO₂ across China and to evaluate trends in NO₂ across the
71 U.S with good correlations against ground observations (Gu et al., 2017; Lamsal et al., 2015). However, CTMs are
72 computationally expensive and subject to bias (Montgomery et al., 2023).

73 Deep learning such as convolutional-neural-networks can capture complex spatial relationships between predictors
74 and often requires less computational resources than CTMs with comparable or improved accuracy (Salman et al.,
75 2024; Xing et al., 2020). However, deep learning remains computationally intensive, less interpretable, and often
76 requires extensive preprocessing (e.g., re-gridding) (Ghahremanloo et al., 2021; Peinado et al., 2025). In contrast,
77 simple approaches such as linear regressions are computationally efficient, but cannot capture complex nonlinear
78 relationships between predictors and surface NO₂ (Kim et al., 2024; Wei et al., 2025).

79 Machine learning (ML) offers a middle ground between these approaches and has been widely applied to estimate
80 surface NO₂ using satellite vertical column NO₂ measurements, meteorology, and land use factors (Ahmad et al.,
81 2024; Larkin et al., 2017; Li et al., 2022; Nawaz, Huber, et al., 2025; Peinado et al., 2025; Wei et al., 2025). ML
82 approaches typically outperform traditional statistical models in complex and highly dynamic urban environments,
83 and when data are affected by structural noise (Peinado et al., 2025). However, studies over large domains in the
84 United States are limited to daily, monthly, or annual scales due to the temporal resolution constraints of polar-orbiting
85 satellites such as TROPOMI (Cao, 2023; Ghahremanloo et al., 2021; Nawaz, Goldberg, et al., 2025). Surface
86 NO₂ exhibits strong diurnal variability, with peaks during morning and evening hours, that are not captured by
87 TROPOMI's single early afternoon overpass (~13:30 local time) (Acker et al., 2025; Mijling & van der A, 2012).

88 As a result, a major gap remains in estimating hourly surface NO₂ across the contiguous United States (CONUS).
89 Salman et al. (2024) developed a DL model to emulate the Community Multiscale Air Quality (CMAQ) CTM using
90 meteorology, emissions, and land use data to produce hourly estimates of ground-level NO₂ for summer months in
91 2017. However, this study relied on outputs from CMAQ as training targets, rather than direct ground observations
92 and was limited to summer months when NO₂ concentrations are typically lower due to a shorter photolytic lifetime.
93 To our knowledge, no prior study has produced observation-based, hourly surface NO₂ estimates across CONUS.



94 NASA's Tropospheric Emissions: Monitoring of Pollution (TEMPO), launched in April 2023, provides improved
 95 temporal resolution with hourly and sub-hourly scans of vertical column pollutant concentrations for North America
 96 during daylight. A comparative analysis between TEMPO and AQS stations found, for annual hourly averages,
 97 TEMPO shows good agreement with AQS stations ($R^2 = 0.51$), with increased correlations in winter ($R^2 = 0.57$) and
 98 for stations that are not part of the AQS near-road network ($R^2 = 0.65$) (Nawaz, Huber, et al., 2025). When compared
 99 to TROPOMI measurements, TEMPO better captures the full range of surface NO_2 variability, with strongest
 100 agreement during midday (~10:00-14:00 LT) (Acker et al., 2025). In contrast to prior hourly approaches that use CTM
 101 output as ground truth, this study provides the first evaluation of hourly geostationary satellite NO_2 -based predictions
 102 against ground measurements across CONUS. By providing high-frequency observations, TEMPO addresses the
 103 temporal limitations of polar-orbiting satellites and enables new opportunities for hourly surface NO_2 estimation.

104 In this study, we examine the feasibility of developing a machine learning model to predict hourly surface NO_2 using
 105 meteorological variables, land use features, and TEMPO vertical column NO_2 , trained based on ground-based NO_2
 106 measurements. We first characterize the diurnal patterns of NO_2 across different levels of urbanization and examine
 107 how TEMPO's hourly tropospheric NO_2 columns capture these surface-level diurnal cycles under different
 108 meteorological conditions. We then develop and evaluate the machine learning model and assess the added values of
 109 TEMPO observations by comparing the model performance with and without TEMPO data. By capturing the short
 110 lifetime and pronounced diurnal variability of NO_2 , this study provides a framework for estimating hourly surface
 111 NO_2 across the United States using TEMPO observations, with potential applications for improved air pollution
 112 exposure assessment and air quality management.

113 **2 Data and Methods**

114 **2.1 Study Domain and Monitoring Sites**

115 The study domain encompasses the contiguous United States (CONUS), with 474 ground-based NO_2 monitoring
 116 stations in the U.S. Environmental Protection Agency's (EPA) Air Quality System (AQS) across the domain (Fig. 1).
 117 The AQS network is concentrated in urban and metropolitan regions, with comparatively sparse coverage in rural
 118 areas. We use impervious surface percentage (ISP) to quantify the degree of urbanization at each station. Stations are
 119 classified into three categories based on ISP: rural (<20%), suburban (20-50%), and urban (>50%). Of the 474
 120 stations, 133 are classified as rural, 160 as suburban, and 181 as urban.

121 Mean NO_2 concentrations exhibit distinct spatial patterns across CONUS. Rural stations in the Rocky Mountain region
 122 exhibit particularly low mean concentrations, while Northeast stations, most of which are urban, display elevated
 123 mean NO_2 ; however, stations with the highest mean NO_2 concentrations (>20 ppb) were found in Denver, Phoenix,
 124 and Los Angeles, from lowest to highest mean NO_2 .

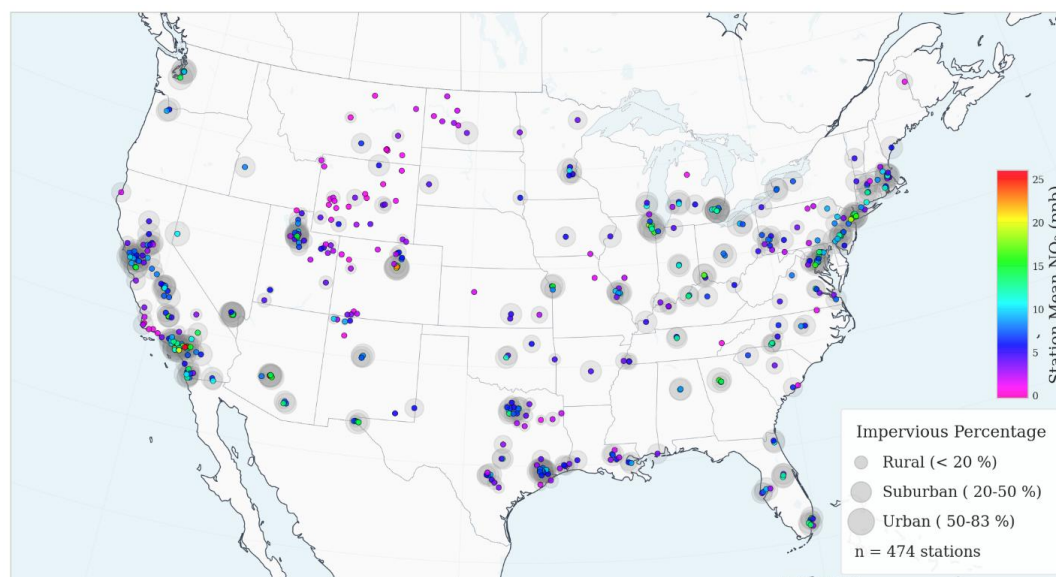


Figure 1. Locations of EPA AQS ground-level NO₂ monitoring stations across the contiguous United States. Stations are colored by mean ground-level NO₂ over the study period. The grey halo size represents impervious surface percentage (ISP) (0-83%) derived from NLCD data. Thresholds of 20% and 50% ISP are used to classify stations as rural, suburban, or urban.

2.2 Datasets

2.2.1 EPA AQS Ground-Level NO₂

The EPA AQS collects hourly data about the ambient concentrations of pollutants including NO₂, from over 5,000 monitors across the U.S. In general, different sets of pollutants are monitored at different stations, limiting the number of stations that collect ground-level NO₂ concentrations to 474 across CONUS for the study period. The EPA generates yearly-collated, validated, hourly NO₂ concentrations (U.S. Environmental Protection Agency, 2025a). We acquired the data for years 2023-2025. As of this time, the 2025 AQS National NO₂ network has sample measurements up to October 1, 2025; however, after July 1, 2025, the minimum number of stations reporting measurements daily drops from 368 to 174. Therefore, we limit the analysis from August 2, 2023 to June 30, 2025.

Fifty-three stations have multiple NO₂ monitors, consisting of chemiluminescent and cavity-attenuated phase-shift (CAPS) instruments, resulting in duplicate values per hour and station. Duplicate values cause the supervised model to struggle because it is tasked with predicting two different NO₂ concentrations given the same features. Chemiluminescent monitors have a known positive bias in NO₂ measurements (Lamsal et al., 2015; U.S. Environmental Protection Agency, 2024). In addition, CAPS monitors also have a lower minimum detection level (MDL) on average (0.08 ppb, std: 0.15) compared to chemiluminescent monitors (1.94 ppb, std: 1.12). Given the bias in chemiluminescent monitors and higher sensitivity of CAPS instruments (lower MDL) we retain CAPS instrument records to remove duplicates.



147 Due to instrumental drift, under very clean conditions, NO₂ measurements fall below zero. Approximately one percent
 148 (0.987%) of surface NO₂ observations in the study period are negative, AQS discourages the replacement of these
 149 negative values with zero as this introduces statistical bias to computations therefore we retain them (U.S.
 150 Environmental Protection Agency, 2011).

151 Given the pronounced right skewed distribution of surface NO₂ measurements, we perform a log-transformation of
 152 the hourly data (offset to ensure positive values) when training the model to deal with data imbalance and reduce
 153 heteroscedasticity (Tang et al., 2024). We retain local time (LT) throughout the data processing.

154 **2.2.2 NASA TEMPO Column NO₂**

155 The NASA Tropospheric Emissions: Monitoring of Pollution (TEMPO) instrument is a UV-Visible (293–494, 538–
 156 741 nm) spectrometer with a spectral resolution of 0.6 nm, which measures or derives vertical column concentrations
 157 of pollutants, such as, ozone, nitrogen dioxide, formaldehyde, aerosols, and others. TEMPO is the first geostationary
 158 mission to provide continuous air pollutant monitoring over North America during daylight. TEMPO has an hourly
 159 temporal resolution with shorter 40-minute (sub-hourly) scans in either the eastern (morning) or western (evening)
 160 half of the continent, depending on which side of the continent is in darkness (Nowlan et al., 2025). This hourly
 161 temporal resolution is a major improvement from daily sensors, including TROPOMI and OMI (Levelt et al., 2006;
 162 Veefkind et al., 2012). The spatial resolution is also finer than current satellite instruments at 2km/pixel North-South
 163 by 4.75 km/pixel East-West at the center of field of view (33.5° N, 89.2° W). One limitation of TEMPO is that, due
 164 to its reliance on solar backscatter, observations are limited to daylight hours whose hourly range is seasonally
 165 dependent, restricting the number of comparable ground-level NO₂ readings to that range.

166 The latest TEMPO data (Level 3, Version 3) was acquired from the NASA Earth Access API (Barrett et al., 2024).
 167 We sampled TEMPO variables at AQS station locations for all available times between August 2, 2023, the beginning
 168 of the TEMPO data release, and June 30, 2025 (Zoogman et al., 2017). Because TEMPO scans do not occur exactly
 169 on the hour, we floor rounded scan times to align the data with AQS temporally. For duplicate timestamps, resulting
 170 from sub-hourly scans, we use the mean. Negative values are also found in TEMPO observations. Nonphysical
 171 retrievals are flagged as “bad” in the main data quality flag. TEMPO NO₂ is derived from the ratio of the spectrum of
 172 Earth radiance to a clean reference radiance. For scans with almost no NO₂, instrument noise can result in
 173 measurements slightly below zero. We remove observations flagged as “bad” by the main data quality flag, but keep
 174 negative measurements, not flagged as “bad”, as their removal would lead to statistical bias and is therefore
 175 discouraged (Nowlan et al., 2025), consistent with the treatment of AQS data.

176 Cloud contamination is a recognized source of uncertainty in satellite NO₂ retrievals. Under cloudy conditions,
 177 scattering and reflection reduce instrument sensitivity to the near-surface layer and lower the tropospheric air mass
 178 factor, inflating the retrieved vertical column relative to surface concentrations (Goldberg et al., 2025). Goldberg et
 179 al. (2025) report in situ measurements averaging +17% higher across all days than on clear-sky days alone, and +36%
 180 higher on cloudy versus clear days (at 13:30 local), with the normalized mean difference increasing per hour from



181 ~+8% at 08:30 to +52% at 15:30 under TEMPO-based cloud classification. Screening clouds would therefore bias the
 182 sampled record toward the lower-concentration clear-sky regime, discarding the observations that coincide with the
 183 highest surface NO₂. This compounds a volume cost: clear-sky filtering removes 20–70% of CONUS satellite NO₂
 184 retrievals depending on region (Goldberg et al., 2025).

185 We therefore retain cloud-affected retrievals. To control the liability they carry, we drop TEMPO column values at
 186 the 98th percentile of the training distribution, clipping only the extreme, high-uncertainty tail: 0.27% of the ~15
 187 billion valid retrievals across 2024, retaining 99.73%. This tail is overwhelmingly cloud-driven: the removed cells
 188 have a median effective cloud fraction of 1.0 and a mean of 0.89, whereas retained cells have a mean effective cloud
 189 fraction of 0.35 (interquartile range across weeks: 0.30–0.40). By comparison, a conventional cloud screen would
 190 discard a large share of the domain: an effective-cloud-fraction threshold of ≥ 0.2 would remove 47.5% of cells, and
 191 even a less restrictive ≥ 0.5 threshold would remove 30.0%. Retaining cloud-affected observations and clipping only
 192 the extreme tail therefore preserves nearly half the domain that a standard 0.2 screen would eliminate, at the cost of
 193 removing just 0.27% of cells.

194 To characterize retrieval reliability, we also derive a signal to noise (SNR) feature defined as the ratio of the
 195 troposphere vertical column NO₂ to troposphere NO₂ uncertainty.

196 Due to a technical issue, no radiance scans were observed from 09–22 August 2024 (Smithsonian Astrophysical
 197 Observatory, 2026).

198 **2.2.3 ERA5 Meteorology Reanalysis**

199 The ERA5 reanalysis dataset, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) is
 200 widely used for historical estimates of atmospheric characteristics. Different ERA5 datasets are available at 137
 201 vertical layers, exclusively over land, and at single pressure levels. We use single-pressure-level products, which are
 202 most relevant for near-surface processes influencing ground-level NO₂ (Hersbach et al., 2023).

203 Following Nawaz, Huber, et al. (2025), we select boundary layer height (BLH), 10m *u*- (East-West) and 10m *v*-
 204 (North-South) wind components (*u*₁₀, *v*₁₀). Additionally, we introduce surface pressure, 2m temperature (*T*_{2m}), 2m
 205 dewpoint temperature (*D*_{2m}), surface shortwave radiation downwards, and friction velocity. We include surface
 206 shortwave radiation downwards as it drives diurnal and seasonal BLH and other meteorological variability. Friction
 207 velocity is an ERA5-derived parameter that represents the magnitude of mechanical stress exerted over a surface by
 208 air flowing above (Hersbach et al., 2023). Friction velocity quantifies the near-surface turbulent intensity, as
 209 turbulence is the primary mechanism for vertical mixing in the atmospheric boundary layer. We derive wind speed
 210 and wind direction from the *u*₁₀ and *v*₁₀ wind components and calculate relative humidity using the 2m temperature
 211 and dewpoint temperature.



ERA5 has a spatial resolution at $0.25^\circ \times 0.25^\circ$, which is relatively coarse compared to the spatial density of monitoring stations. As a result, nearby AQS stations may sample the same ERA5 grid cell. We use bilinear interpolation to resample ERA5 data to match the TEMPO grid.

2.2.4 Land Cover Features

We acquire impervious surface percentage (ISP) data from the National Land Cover Dataset (NLCD) 2023 and 2024 products provided by the Multi-Resolution Land Characteristics Consortium (MRLC), as an indicator of urban or near-road conditions. ISP is available at 30 m spatial resolution and ranges from 0 to 100%. To ensure consistency with satellite observations, we spatially aggregate ISP to the TEMPO L3 grid ($0.02^\circ \times 0.02^\circ$) at each AQS station by averaging all 30 m pixels within the corresponding grid cell. Because 2025 data are not yet available and land cover changes gradually over time, we use 2024 ISP values to represent 2025 conditions (U.S. Geological Survey, 2025). We also acquire the most recent (2025) U.S. census TIGER\Line shapefiles for primary (divided, limited-access highways) and secondary (main arteries within the U.S., State, or County Highway system) roads for CONUS and calculate road density in km/km^2 for each TEMPO grid cell (TIGER/Line Shapefiles, 2025). This produces a TEMPO-scale representation of urbanization at each monitoring location.

2.2.5 Temporal features

Surface NO_2 has distinct diurnal and seasonal patterns driven by emission cycles and atmospheric processes (Mucha & Mainka, 2026). Using the local datetime, we create features for day of week, hour of day, and month to capture these systematic patterns. In addition, we create a binary holiday indicator (“Is Holiday”) using the pandas *USFederalHolidayCalendar* function.

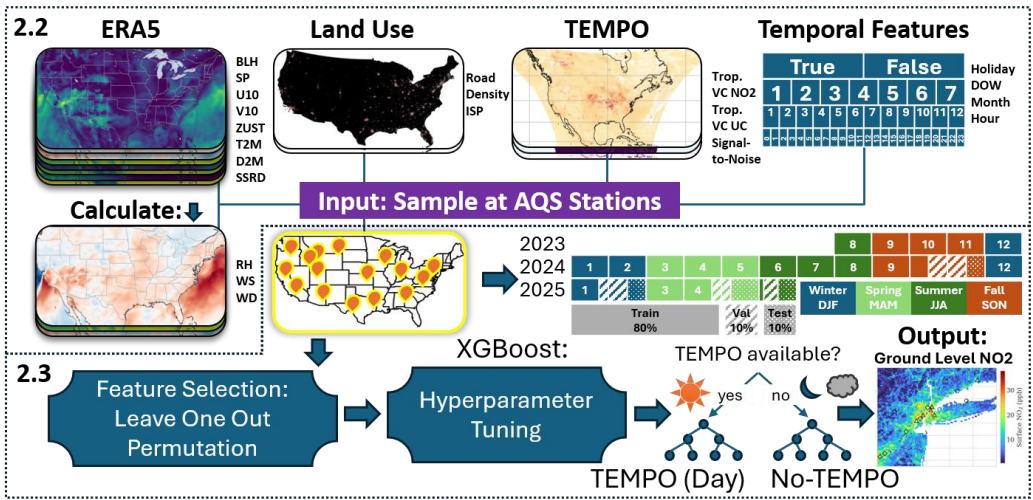
2.3 Analysis and Model Configuration

2.3.1 Diurnal Analysis

To characterize diurnal relationships between predictor variables and ground-level NO_2 concentrations, we compute the mean and standard deviation (SD) for each local hour of the day using the gridded data sampled at monitoring station locations. To make sure that each hour is equally represented, we limit the analysis to stations that have records for all 24 hours. For each station, we first calculate the hourly average across the study period, then average across stations to analyze the intra-site variance, using the standard deviation. For TEMPO features, we limit values below the 98th percentile for each station to remove outliers. We also calculate the Pearson correlation coefficients (r) using all hourly observations to account for decreased variance from diurnal averaging. This combined approach allows us to separately assess diurnal structure and overall statistical relationships.

2.3.2 Model Development

Overview of the modeling framework for estimating ground-level NO_2 is described in Fig. 2. Below provides the details of each component.



244

245 **Figure 2.** Overview of the modeling framework for estimating ground-level NO₂. Gridded predictor variables (satellite, land
246 cover, and meteorology) are sampled at AQS monitoring stations and combined with temporal features. Each season within
247 the dataset is split chronologically into training (80%), validation (10%), and test (10%) sets. Feature selection is performed
248 using a leave-one-out ablation with retraining approach, followed by hyperparameter tuning of an XGBoost model on
249 TEMPO-inclusive samples. A parallel model is trained without any TEMPO-derived features to predict hours where
250 TEMPO is unavailable (night, quality-flagged, or above the 98th training percentile). The trained models are used to predict
251 surface NO₂ concentrations.

252 ML model: For our ML model, we utilize XGBoost (eXtreme Gradient Boosting), a scalable end-to-end tree boosting
253 model developed by Chen and Guestrin (2016) that builds a strong predictive model by sequentially combining weak
254 decision trees through gradient boosting. XGBoost utilizes a regularized objective function that systematically
255 prevents overfitting by including a penalty term to control model complexity in the number of tree leaves and the
256 magnitude of leaf weights (Ahmad et al., 2024; Chen & Guestrin, 2016). We choose XGBoost due to its low
257 computational complexity, ability to handle missing values and multicollinearity, avoidance of overfitting, and ease
258 of set up. XGBoost showed best performance in estimating NO₂ concentrations from column measurements when
259 compared to other machine learning models (Ahmad et al., 2024; Wei et al., 2025). We train two separate models, one
260 for samples when TEMPO is available, hereon referred to as the TEMPO model, and one without TEMPO features,
261 used to estimate surface NO₂ at night and when TEMPO is unavailable due to data quality or when values exceed the
262 98th percentile of the training dataset, usually as a result of high cloud coverage. The latter model is hereon referred to
263 as the no-TEMPO model.

264 Model Training and Testing: The dataset is split sequentially while maintaining seasonal representation. Each season
265 is defined as 3 months starting in December for winter. Within each season observations are ordered chronologically
266 then partitioned sequentially into training (80%), validation (10%), and test (10%) subsets. These seasonal subsets are
267 then recombined to form the final training, validation, and test datasets. This approach preserves temporal structure
268 and prevents data leakage while ensuring each season is represented across all splits.



269 Feature Selection: To select features we first create a baseline model with all acquired and derived features and then
 270 remove one feature at a time and retrain the model to determine the feature's impact on model performance metrics.
 271 We remove features that decrease the coefficient of determination and increase the RMSE. We find that dewpoint
 272 temperature, day of year, year, season, and categorical national land cover class decrease model performance. TEMPO
 273 features such as stratosphere column NO_2 and radiance features like solar zenith angle, relative zenith angle, and
 274 viewing zenith angle also reduce model performance. The features included in the final model are: road density, ISP,
 275 BLH, u_{10} and v_{10} , wind speed, wind direction, friction velocity, surface pressure, surface shortwave radiation
 276 downwards, T2m, relative humidity, TEMPO troposphere NO_2 , TEMPO troposphere NO_2 uncertainty, TEMPO signal-
 277 to-noise ratio, and temporal features including: day of week, hour, month, and the binary holiday indicator. The no-
 278 TEMPO model contains all of the features above, excluding those from TEMPO.

279 Hyperparameter Tuning: To tune the hyperparameters of the model, we first use a 10-fold random CV search for the
 280 following parameter grid using the *scikit-learn* package (Pedregosa et al., 2012). We fit 10 folds for 100 candidates,
 281 totaling 1000 fits using the training dataset (Table 1). We further investigate important parameters such as max depth,
 282 minimum child weight, learning rate, number of estimators, subsample, and column sample by tree using a 10-fold
 283 grid search. We keep gamma, L1 and L2 regularization and column sample by level constant as the rounded results
 284 from the random CV search. The no-TEMPO model uses the same hyperparameters as the TEMPO model. This
 285 ensures performance differences reflect TEMPO's contribution rather than differential hyperparameter optimization,
 286 preserving the interpretability of model metrics. The results with the best R^2 are listed below.

287 **Table 1. Hyperparameter search range and optimized values for the XGBoost model obtained via 10-fold random cross-**
 288 **validation (CV) and grid search**

<i>Parameter</i>	<i>Range</i>	<i>Optimized Values</i>
<i>n estimators</i>	100 – 1100	1050
<i>max depth</i>	3 - 15	12
<i>learning rate</i>	0.01 – 0.3	0.04
<i>subsample</i>	0.6 – 1.0	0.8
<i>min child weight</i>	1 – 10	5
<i>Col sample by tree</i>	0.6 – 1.0	0.85
<i>Col sample by level</i>	0.6 – 1.0	0.8
<i>gamma</i>	0 – 0.5	0.04
<i>L1 regularization</i>	0 – 1	0.2
<i>L2 regularization</i>	0 – 3	3

289



290 Model Evaluation: Model performance is assessed using RMSE, mean absolute error (MAE), mean bias error (MBE),
 291 and coefficient of determination (R^2) (equations listed in Appendix A). We compare predicted and observed values
 292 across training, validation, and test datasets to understand how well the combined model performs. To evaluate
 293 temporal behavior, we average all predicted and actual values across stations per hourly timestep and season and
 294 compare the time series of the average actual versus predicted values and the between-site variance of predicted and
 295 observed values. This allows us to view how well the model captures the diurnal cycles of NO_2 across seasons and
 296 whether there is any systematic drift in the model. To examine how model performance and observed NO_2 variability
 297 differs across the urban gradient, we compute the mean observed and predicted NO_2 for each site and hour. We
 298 aggregate per-site hourly means by urbanization class, with reported statistics representing across-site mean ± 1 SD
 299 at each hour of the day.

300 To assess performance at individual monitoring sites, we plot station R^2 and RMSE and the relationship between the
 301 MBE and ISP, as well as the relationship between the station mean NO_2 and RMSE, shading pairs by ISP to see how
 302 metrics differ across environments. We also showcase a representative case study by selecting a day demonstrating
 303 the median RMSE of the full test dataset, November 27, 2024. We selected the New York City metropolitan region
 304 for detailed analysis due to its dense monitoring network spanning urban, suburban, and rural stations, enabling
 305 thorough validation of model performance across the full urbanization gradient. For this day, we plotted gridded model
 306 predictions, station observed values, and TEMPO observations at three key hours: 07:00 (AM rush), 12:00 (noon),
 307 and 17:00 (PM rush). Additionally, we extracted time series of observed versus predicted NO_2 across 24 hours at three
 308 representative stations, one from each urbanization class to assess typical hourly model performance over time.

309 Model Generalization across Space and Time: In addition to the 80/10/10 train-test-split for model development and
 310 evaluation, split to reduce data leakage, tune hyperparameters, and report metrics on a true hold out set, we implement
 311 three cross-validation strategies to assess model generalizability:

312 Random 10-fold CV: To evaluate the model's overall robustness, using sklearn's Kfold, we first perform a 10-fold
 313 cross-validation, randomly dividing the entire dataset into ten non-overlapping sets of samples. We train the model
 314 using nine out of ten sets and then test the model on the remaining set, repeating this process ten times to ensure every
 315 set of samples is evaluated.

316 Temporal CV: We then perform a 10-fold temporal cross-validation (temporal CV) using sklearn's *TimeSeriesSplit*,
 317 which preserves the chronological order of observations to assess sensitivity to training period length. In each
 318 successive fold, the training set expands by ~ 2 months (10%) of the data while the model is evaluated on the
 319 subsequent ~ 2 months (following 10%). This expanding window approach allows us to assess how much historical
 320 data the model requires to achieve stable predictive performance.

321 Spatial CV: To assess model's spatial transferability, we use the spatial-kfold package and split stations into 100 by
 322 100 km hexagon blocks. Then we randomly sample blocks to create ten regions of an equal number of blocks and
 323 roughly equal number of stations. Using nine regions we train the model and test on the remaining region or unseen



324 set of stations. We repeat this process ten times to test the model on each set of unseen stations and evaluate the
 325 model's transferability across stations.

326 Model performance metrics are expected to be lower than those reported in previous studies that relied on daily
 327 average NO₂. This is attributable to the increased variability of hourly observations that are smoothed out in daily
 328 aggregations. While this presents a more demanding prediction task, the hourly resolution enables characterization of
 329 within-day NO₂ patterns that are relevant for acute exposure assessment. We calculate the percent relative standard
 330 deviation (%RSD) of surface NO₂ values across all observations (hourly) and across site-day combinations (daily) to
 331 determine the extent of variability in relation to the mean of the study period.

332 Model Interpretation: ML models are often considered a “black box” and model interpretation is often ignored in air
 333 pollutant ML modeling studies; however it is of tremendous practical importance: guiding feature engineering,
 334 diagnosing models, and improving prediction outcomes (Tang et al., 2024). To improve interpretability, we analyze
 335 feature importance using XGBoost built-in feature importance metrics, including gain and cover. *Gain* is the average
 336 reduction in loss (RMSE) by a feature when used in a split, while *cover* reflects the number of samples affected by
 337 splits on that feature, i.e. how many training samples flow through the split point based on that feature. We also apply
 338 SHapley Additive exPlanations (SHAP) to quantify how much each feature contributes to a model's prediction by
 339 comparing it to the average prediction across the training dataset. They are calculated by considering every
 340 combination of features and measuring the marginal contribution of adding a given feature, then averaging across all
 341 combinations. We use SHAP to understand how feature values influence predicted values (Lundberg & Lee, 2017)

342 Impact of TEMPO on Diurnal Performance: Finally, we compute model performance metrics (RMSE, MAE, and R²)
 343 as a function of hour of the day, stratified by the availability of TEMPO NO₂ observations to assess the impact of
 344 TEMPO data coverage on model prediction accuracy. We also map daytime and nighttime surface NO₂ across CONUS
 345 and selected metropolitan regions to examine spatial patterns in predicted surface NO₂.

346 **3. Results**

347 We first examine the diurnal relationship among meteorological variables, TEMPO features, and surface NO₂ (Section
 348 3.1). We then evaluate hourly model performance across all stations, metropolitan regions, land use types, and levels
 349 of urbanization for both the TEMPO and no-TEMPO models (Section 3.2). Model robustness and generalizability are
 350 assessed using random, temporal, and spatial cross-validation together with a TEMPO feature ablation model (Section
 351 3.3). Feature importance and contribution of TEMPO data to model performance are presented in Sections 3.4 and
 352 3.5, respectively. Finally, we map daytime and nighttime surface NO₂ annual means across CONUS and selected
 353 metropolitan regions and compare the resulting spatial patterns with ground-based observations (Section 3.6).

354 **3.1 Diurnal Analysis**

355 Figure 3 shows the local time diurnal patterns of the mean surface NO₂ and mean selected predictor variables. Results
 356 are restricted to stations with complete 24-hour coverage ($n = 436$ stations), and TEMPO features are filtered below



the 98th percentile range per station to reduce the influence of outliers and visualize the data under our cloud coverage approach.

Surface NO₂ concentrations exhibit a pronounced diurnal cycle, highest at early morning, 5:00-7:00, coinciding with peak traffic-related NO_x emissions when boundary layer height and wind speed are lowest, relative humidity is highest, and solar-driven mixing and photolysis are minimal.

During daylight, increasing solar radiation enhances atmospheric mixing through boundary layer height growth, increased friction velocity, wind speeds, and 2m temperature (with a noticeable lag). This, coupled with photolysis, leads to a reduction in surface NO₂ concentrations during 8:00-20:00, with a minimum between 13:00-14:00. At 15:00, concentrations begin to rise as solar radiation and therefore BLH and winds decrease. Surface NO₂ experiences a second, more gradual peak between 19:00-21:00.

Meteorological variables associated with atmospheric mixing, including BLH, surface shortwave radiation downwards, friction velocity, wind features, and T2m, show strong inverse relationships with surface NO₂. Vertical column troposphere NO₂ and relative humidity loosely follow the same trend as NO₂. Among these, wind speed shows the strongest relationship with ground-level NO₂ ($r = -0.939$). When there is no wind, ground-level NO₂ concentrations build up, and as wind speed increases, more atmospheric mixing occurs and pollutants disperse farther from their sources (Harkey et al., 2015; Román-Cascón et al., 2023).

Boundary layer height has the second strongest relationship per hour of the day ($r = -0.934$). This relationship was first quantified by Harkey et al. (2015), confirming earlier suggested dilution theories. They find that a shallow boundary layer traps pollution near the surface; and as the boundary layer height increases, there is more volume for pollutants to disperse into, increasing atmospheric mixing and diluting NO₂ concentrations. Surface shortwave radiation is the primary driver for boundary layer height growth and has a similar diurnal relationship as BLH to surface NO₂.

Friction velocity has the fourth highest diurnal correlation ($r = -0.904$). In local studies, increased friction velocity generally resulted in increased vertical mixing and therefore decreased ground-level NO₂ concentrations (Román-Cascón et al., 2023).

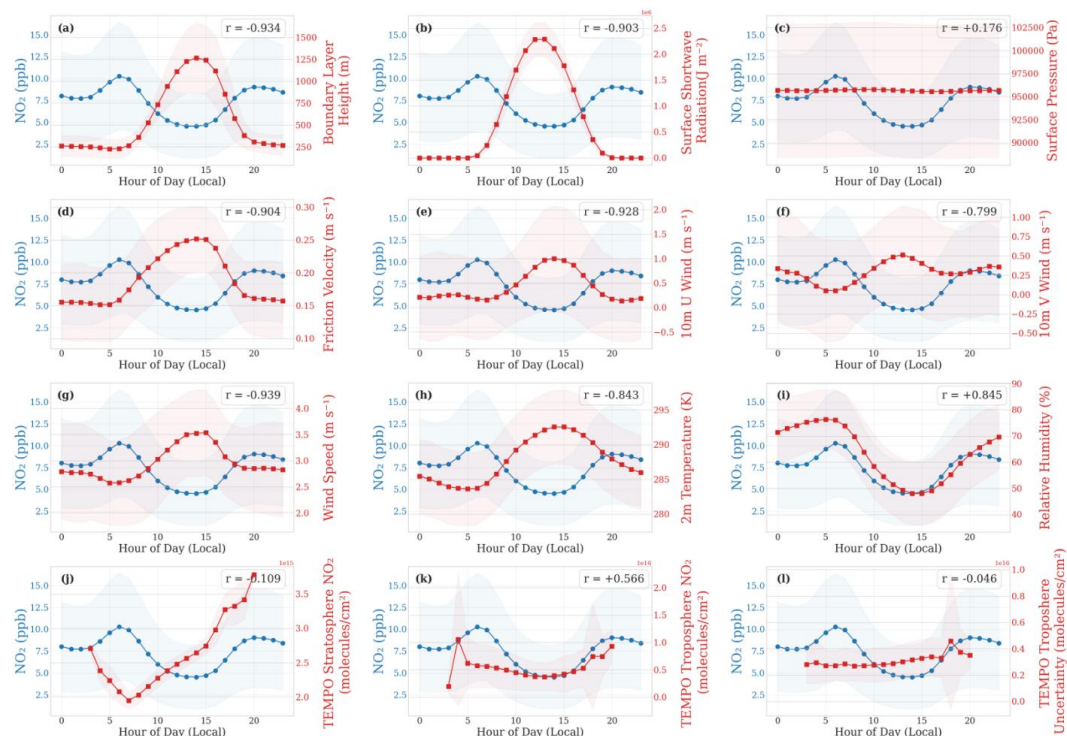
Across all observations (i.e., without diurnal averaging), correlations are substantially weaker due to increased variability, with BLH ($r = -0.311$) and friction velocity ($r = -0.302$) remaining among the strongest predictors (Fig. A2).

Mean troposphere vertical column NO₂, measured by TEMPO, shows good agreement with mean hourly surface NO₂ ($r = 0.566$) and the strongest agreement across all observations ($r = 0.425$), representing the strongest satellite-derived predictor. Retrieval quality is highest between 09:00 and 13:00, when over 97% of observations are flagged as good and uncertainty is lowest. Outside this window, quality degrades: at 04:00, 59% of retrievals are marked bad, the second highest following 20:00, with only 19% flagged good, producing the anomalous dip in panel (k). Uncertainty



increases throughout the day from 7:00-16:00, then spikes sharply at 18:00 before declining as more samples are
 flagged as “bad” and filtered out. High solar zenith angles (SZA) degrade retrieval quality by inflating air mass
 fractions used to derive vertical column densities from slant column densities (Nowlan et al., 2025). Ghahremanloo et
 al. (2025) found worst agreement between Pandora column pollutant ground sensors and TEMPO NO₂ measurements
 at SZA above 70°. However, filtering observations where SZA exceeds 70° limits temporal coverage from ~17 to ~13
 hours of the day, which prevents a full diurnal cycle analysis in critical hours of peak surface NO₂.

In contrast, average stratosphere vertical column NO₂ from TEMPO has a lower correlation at $r = -0.109$. Its
 correlation across all observations ($r = -0.285$) is lower than that of certain meteorological variables. This feature was
 ultimately removed from the model as it decreased performance. TEMPO stratosphere column NO₂ also suffers from
 bias when past troposphere measurements are high relative to GEOS-CF predictions (Nowlan et al., 2025).



400

Figure 3. Diurnal cycles of ground-level NO₂ (blue: right y-axis) versus selected predictor variables (red: left y-axis). Values represent mean concentrations per local hour across stations; shaded areas denote ±1 standard deviation. Panels (a–l) show relationships with meteorology variables and TEMPO NO₂ products with Pearson correlation coefficients (r) shown in each panel.

Finally, variability analysis indicates that the hourly NO₂ concentrations exhibit substantially greater dispersion than daily averages. Across all hours the percent Relative Standard Deviation (%RSD) has a mean of 101.2, with a median %RSD of 100.6 (the 25th-75th percentile: 92.0-110.6). In contrast, daily averages show lower variability, with a mean %RSD of 82.7, a median %RSD of 83.1, and an interquartile range of 78.6-87.0. This supports the idea that hourly



values have high and variable heterogeneity, while daily values are lower and more stable, decreasing variability by 19% and highlighting the importance of understanding surface NO₂ at high spatial and temporal resolutions.

3.2 Model Performance

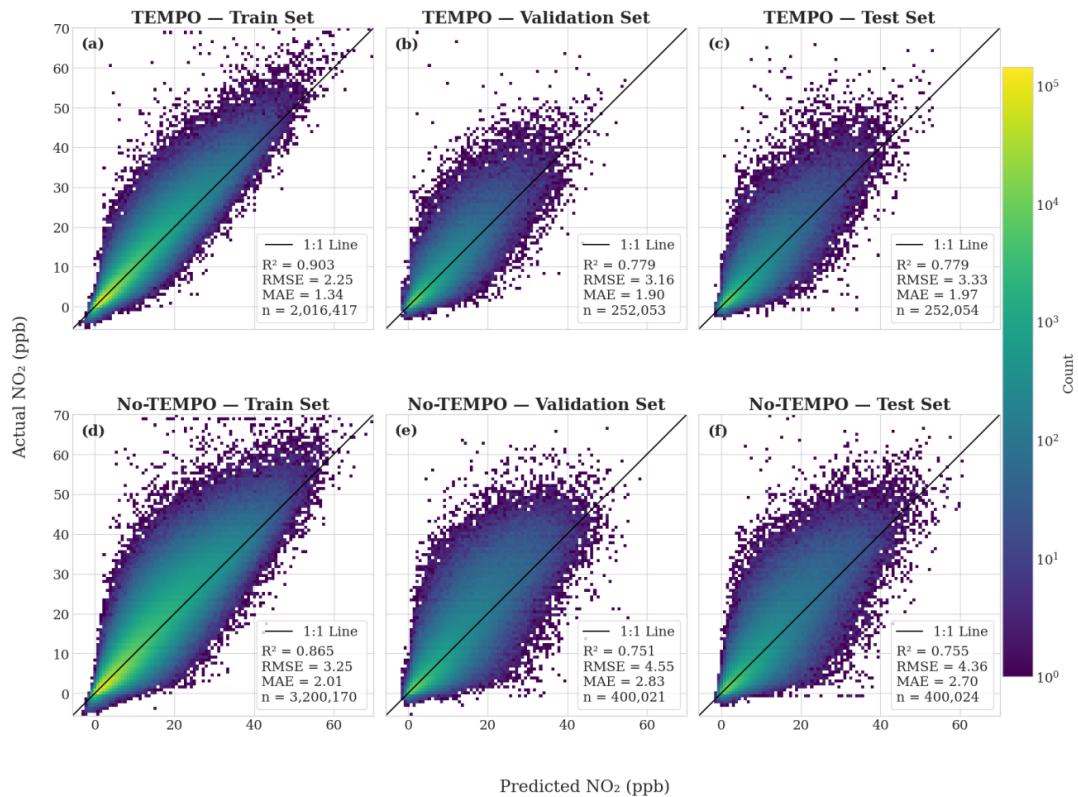
3.2.1 Overall Model Performance

Figure 4 shows the relationship between observed and predicted ground-level NO₂ concentrations for the training, validation, and test datasets for both models. In the test datasets, four observations have a ground-level NO₂ value greater than 70 ppb; for purpose of visualization, we limit the axes to 70.

The TEMPO model, used during most daylight hours, predicts hourly ground-level NO₂ well, with a test dataset R² of 0.78 and an RMSE of 3.33 ppb. The training dataset achieved an R² of 0.903 and an RMSE of 2.24 ppb. Across all three sets, predictions cluster tightly along the 1:1 line, particularly at lower concentrations.

The no-TEMPO model is used when TEMPO is not available, generally at night, when NO₂ concentrations are generally higher but less variable than during the daytime ($\bar{x} = 8.32$, %RSD = 105.64%, compared to day $\bar{x} = 6.26$, %RSD = 113.03%). In contrast, the no-TEMPO model shows weaker performance with a test dataset R² of 0.755 (vs 0.78 for the TEMPO model) and an RMSE of 4.36 ppb. Without a direct satellite-based NO₂ concentration signal as an input, the predicted values show a wider spread relative to the observations (Fig 4.d-f).

At higher NO₂ values (above ~30 ppb), the models, particularly the no-TEMPO model, tend to underpredict, as indicated by points falling above the 1:1 line. Extreme values are often underrepresented in data and models tend to regress towards the mean, leading to a persistent challenge of underestimation in air quality regression models (Gu et al., 2017; Tang et al., 2024). The issue of underestimation may be further affected by logarithmic transformation on the target feature, which lowers the weights of high values in the loss function, despite improved model predictive performance (Tang et al., 2024). Performance degradation at higher concentrations is consistent with prior air quality prediction studies (Gu et al., 2017; Larkin et al., 2017; Nawaz, Goldberg, et al., 2025; Salman et al., 2024; Wei et al., 2025). This underestimation of high-concentration events has important implications, since high pollutant concentration events are often the most relevant for a public health and regulatory perspective (Mainka & Žak, 2022).



433

434 **Figure 4.** Observed versus predicted hourly ground-level NO₂ concentrations for the training (left), validation (middle), and
435 test (right) datasets, split sequentially by time. The top row presents TEMPO model results, while the bottom row shows
436 no-TEMPO model results. Colors indicate point density (log-scaled counts). The black line represents the 1:1 relationship.
437 Model performance metrics (R^2 , RMSE, and sample size) are shown in each panel. Axes are limited to 70 ppb for
438 visualization.

439 3.2.2 Diurnal Patterns at Metropolitan Regions

440 To evaluate how well the model captures regional-scale diurnal patterns of ground-level NO₂ across seasons, we select
441 four metropolitan-regions (New York City (NYC), Chicago, Denver, Los Angeles (LA)) and average the observed
442 and predicted values across stations at each timestep within each region (Fig. 5). For each region, stations within a
443 1.6° x 2.55° geographic extent centered on the urban core are included (Fig. 12). This extent was chosen to provide a
444 stable and meaningful regional mean, although the number and spatial distribution of monitoring stations vary across
445 regions.

446 Model skill (R^2), error (RMSE), and temporal tracking (r) are generally strong but varies among regions and seasons.
447 Across regions, R^2 is highest for LA (mean = 0.79), followed by NYC (0.76), and lowest for Chicago (0.63) and
448 Denver (0.59). These regional differences generally correspond to differences in station density and spatial coverage.
449 LA has the densest monitoring network, with 27 stations, followed by NYC with ~18 stations. However, Denver and
450 Chicago have considerably fewer stations, with 9-10 and 5-8 stations per season respectively. In addition, Denver



451 stations are heavily clustered, while stations in other regions are more broadly distributed across the study domains
452 (Fig. 12).

453 Per-region subsampling analysis (Fig. A3) was conducted to distinguish differences in model performance from
454 station averaging effect. Drawing n stations at random within each region and recomputing average R^2 , all four regions
455 show steep gains from $n = 1$ to $n \approx 5$, confirming that a substantial part of raw cross-city R^2 spread reflects the number
456 of stations averaged rather than model quality. At the single station level, regional differences are far less ($R^2 = 0.53$ -
457 0.68) than average values suggest. However, at matched n , R^2 remains consistently highest in LA, intermediate in
458 Chicago and NYC, and lowest in Denver across the shared range of station counts. The R^2 values approach
459 approximately 0.91 for LA, 0.80 for Chicago, 0.79 for NYC, and 0.66 for Denver, indicating that regional differences
460 persist after accounting for the station averaging effect.

461 RMSE shows a consistent seasonal gradient, independent of region. Errors are largest in winter (3.9-5.2 across regions)
462 and lowest in summer (1.1-2.4), with spring and fall intermediate. This gradient mirrors the seasonal NO_2
463 concentration cycle: winter peaks reach 40-50 ppb in NYC, Chicago, and Denver, while summer values rarely
464 exceeded 20 ppb, leading to absolute error scaling with concentration even where relative skill was comparable.

465 Model performance varied across regions, as indicated by the divergence between r and R^2 . Denver was characterized
466 by lower correlation in every season ($r = 0.73$ - 0.88) with R^2 closely tracking r^2 , indicating that the model captured the
467 general NO_2 concentration but reproduced hour-to-hour and day-to-day variability poorly. Chicago in fall, and LA in
468 winter showed the opposite signature: high correlation alongside a substantially lower R^2 (Chicago-fall $r = 0.83$, $R^2 =$
469 0.40 ; LA-winter $r = 0.90$ but $R^2 = 0.63$). The predicted series reproduces the diurnal and weekly pattern but
470 underpredicts peaks (LA and NYC winter) or is systematically offset from the observations (overprediction in
471 Chicago). The model performed best in LA spring, summer and fall.

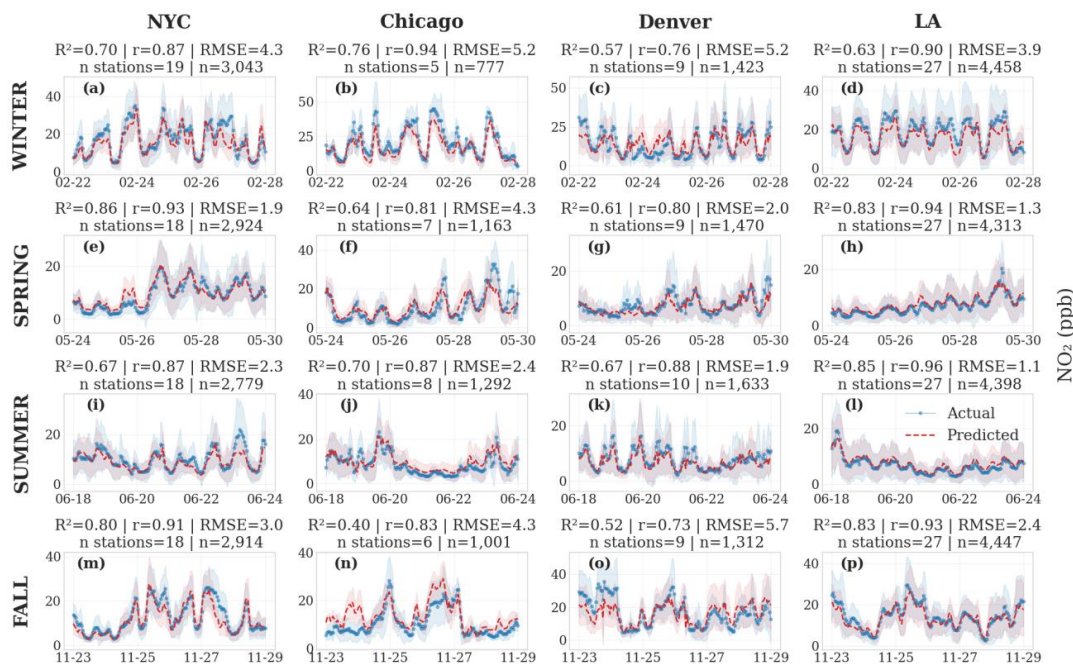
472 The between-site variation in observed versus predicted values at each timestep, presented as standard deviation bands,
473 shows substantial overlap for most seasons. This suggests the model not only accurately predicts the mean value but
474 also approximates the degree of site-to-site variation at each timestep reasonably well. In winter, peak concentrations
475 are systemically underpredicted.

476 For days that exhibit bimodal peaks, the lower (evening) surface NO_2 peak is not captured well by the model with a
477 pattern of underestimation. The morning peak is associated with strong concurrent gradients of ERA5-derived
478 variables, including BLH, surface radiation, wind speed, and friction velocity, all rapidly changing during morning
479 hours, while the evening peak period (~19:00-23:00 LT) lacks this equivalent meteorological variability. Most
480 meteorological features exhibit low hour-to-hour variability, with relatively flat profiles representing minimal changes
481 between mean values. This means features cannot effectively signal the distinction of the evening emission spike from
482 surrounding nighttime hours.



483 This limitation is further compounded by the temporal coverage of TEMPO satellite retrievals. While observations
484 fully capture the entire morning peak, they have a final observation at approximately 20:00 LT, truncating coverage
485 during the evening peak. As a result, the model loses the only direct NO₂ observational inputs during this period,
486 leaving the model dependent on the remaining features, which, as mentioned, exhibit little variability. In addition,
487 TEMPO measurements are also subject to increasing uncertainty throughout the day and unfavorable viewing
488 conditions including a high solar zenith angle at this time, with a higher proportion of retrievals flagged bad (Acker
489 et al., 2025; Nowlan et al., 2025). This may lead to larger disagreement between column and ground measurements
490 or missing samples and therefore less predictive power in times where TEMPO is available during the second peak.

491 The inclusion of a binary “Is Holiday” temporal feature reduces the largest errors in the time series which occurred on
492 holidays (for example: Thanksgiving Day: 28 November). Without this feature, the model would dramatically
493 overestimate surface NO₂ values. On average across stations, holidays are marked by lower NO₂ concentrations
494 throughout the day and more subdued diurnal patterns, with the evening peak values closer to morning peak than
495 during non-holidays.



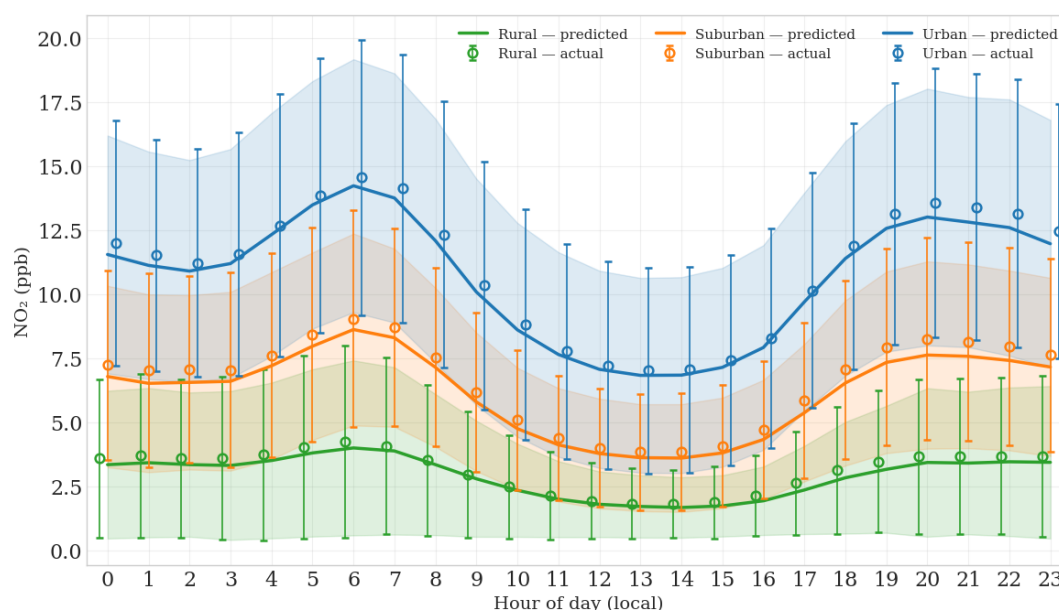
496

497 **Figure 5. Time series comparison of observed and predicted hourly ground-level NO₂ concentrations per season at four**
498 **metropolitan regions: New York City, Chicago, Denver, Los Angeles (columns). Lines represent mean values averaged**
499 **across stations at each region, and shading indicates ± 1 standard deviation. Blue: observed; red: predicted. Rows show**
500 **winter, spring, summer, and fall. Within each season, one week with the most sensors active across regions was selected**
501 **from each season to visualize typical performance. Metrics are computed on that selected week with R^2 , r , RMSE, n stations**
502 **and n samples labeled above each panel. X-axis ticks are formatted: Month-Date. Fall: 2024 data. Winter, spring, summer:**
503 **2025 data.**



3.2.3 Diurnal Patterns by Urbanization Class

To understand how model performance and NO₂ daily variability differs across urbanization we plot the mean observed (circle) and predicted (line) surface NO₂ per hour of the day across urbanization classes. All three classes exhibit the bimodal diurnal NO₂ cycle, found in Figure 3. The morning peak modestly exceeds the evening peak in every class. Concentrations are ordered urban, suburban, rural at every hour without exception, and the amplitude of the diurnal range scales with the class mean, from roughly 7.5 ppb minimum-to-maximum in urban stations (~14.5 to ~7 ppb) down to about 2.5 ppb in rural stations (~4 to ~1.6 ppb). Across all 24 hours and all three classes, the predicted lines pass through or are immediately adjacent to the observed points, indicating that the model captures both the phase and the amplitude of the cycle rather than the mere daily mean, with stronger agreement in daylight hours. Notably, the within-class ± 1 STD bands overlap substantially between adjacent classes, showing that site-to-site heterogeneity within a single land-use designation exceeds the mean differences between designations.



515

516 **Figure 6.** Composite mean diurnal cycle of surface NO₂ by hour of local day, stratified by station land-use class (rural:
 517 green, suburban: orange, urban: blue). Solid lines show model predictions and open circles show observations, both
 518 computed as the mean across all stations within each class at each hour. Shaded bands and error bars denote ± 1 standard
 519 deviation of the station-to-station distribution within each class (predicted and observed, respectively).

3.2.4 Spatial Variation in Station-Level Performance

To evaluate if station performance differs spatially or with increasing urbanization we plot R^2 and RMSE at each station (Fig. 7). Results combined from both models using the 80/10/10 split, show a median R^2 of 0.596 and median RMSE of 3.57 (mean $R^2 = 0.517$, mean RMSE = 3.61) across 441 stations included in the test dataset. Panels a and b show that these two metrics are organized spatially in opposite ways. Skill is highest along the densely monitored



525 Northeast corridor and in coastal and southern California, where most stations exceed R^2 of 0.7, and remains moderate
 526 to high across Texas, the Gulf Coast, and the industrial Midwest. While most stations have an R^2 above 0.5, stations
 527 in the Rocky Mountain region exhibit particularly diminished R^2 s. Their apparent limited skill is a result of the low
 528 observed variance relative to their modest mean concentrations, making even small absolute errors produce a low R^2 .

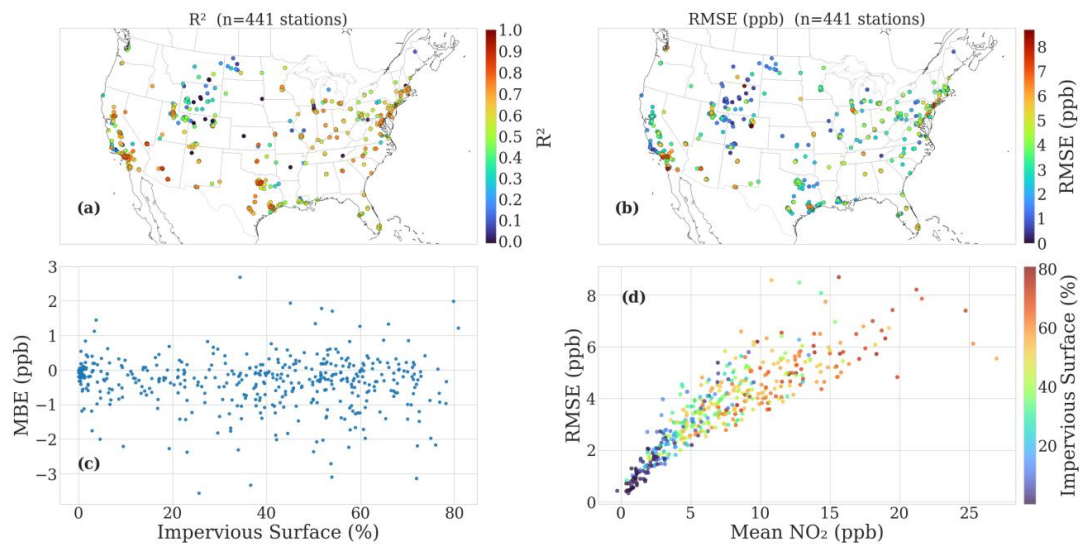
529 Panel d makes this scaling explicit: stations with higher mean NO_2 have higher mean RMSE, indicating
 530 heteroscedasticity. Stations in cleaner environments, where NO_2 is below 5 ppb, generally achieve RMSE values under
 531 3 ppb, while stations with mean NO_2 above 10 ppb frequently exceed 4-5 ppb RMSE.

532 This pattern is closely linked to urbanization. In general, stations with low mean NO_2 have low ISP, and vice versa.
 533 Rural stations, (including those from the Rocky Mountain region) with ISP below 20%, are clustered in the lower left
 534 corner where concentrations and errors are both low. Stations with higher ISP generally exhibit higher mean
 535 NO_2 values and higher model errors, however the relationship between errors and mean values is more scattered.

536 Panel c shows that error direction, unlike error magnitude, is largely decoupled from urbanization. Station MBE is
 537 centered near zero (median -0.33 ppb, mean -0.44 ppb) with no significant trend against impervious surface (-0.0025
 538 ppb per %ISP, $r = -0.077$, $p = 0.108$). Across equal-count ISP deciles, median bias steps from -0.18 ppb at stations
 539 with essentially no impervious cover to between -0.32 and -0.52, then varies without order across the remaining
 540 gradient, with the largest decile median falling in a mid-ISP bin rather than at the urban end. Because the least
 541 developed stations are also the cleanest, even that step from -0.18 to -0.32 is more plausibly a concentration effect
 542 than an urbanization effect.

543 Bias does scale weakly with concentration, declining at -0.050 ppb per ppb against station mean observed NO_2 ($r = -$
 544 0.312), with decile medians deepening from -0.10 to -0.74 ppb. Because MBE is computed from the same observations
 545 used as the predictor, this slope is partly inflated by regression toward the mean; regressing instead against mean
 546 predicted NO_2 halves the slope and correlation (-0.026 ppb per ppb; $r = -0.160$), though the relationship persists,
 547 indicating genuine underprediction at the most polluted sites. This pattern points to variance compression rather than
 548 unresolved urban processes: the model underpredicts as a function of concentration, not land cover, as expected when
 549 a regression model shrinks extremes toward the conditional mean. Unresolved street-scale emission dynamics would
 550 instead produce bias deepening systematically with ISP.

551 The bias is small relative to random error. In the most polluted decile, a median bias of -0.74 ppb against a median
 552 RMSE of 5.79 ppb means squared bias accounts for 1.6% of mean squared error. RMSE is therefore driven by variance
 553 rather than offset, and predictions can be treated as approximately unbiased for applications aggregating over space
 554 or time, with the caveat that long-term means at the most polluted urban sites may run low by a few tenths of a ppb.



555

556 **Figure 7. Individual station performance across 441 stations included in the test dataset, colored by station R^2 (a) and RMSE**
557 **(b). (c) Relationship between Mean Bias Error and impervious surface percentage (ISP). (d) Relationship between mean**
558 **NO_2 concentration and model error (RMSE). Each point represents a station. Panel (d) is colored by ISP.**

559 **3.2.5 Representative Performance Hourly Case Study**

560 To examine typical hourly performance, we select a day exhibiting median performance with respect to RMSE (27
561 November 2024). We show a representative case study over the New York metropolitan region at three key hours
562 7:00 (AM rush hour), 12:00, and 17:00 LT (PM rush hour) (columns; Fig. 8). Station points are shown on row 2 and
563 share the color range as model predictions. Row 3 shows the nearest TEMPO scan used in the TEMPO model for the
564 first two hours. We also plot a time series of station observed and predicted values for each urbanization class for all
565 hours of the selected date (bottom-right panel).

566 The time series of the day's observed and predicted NO_2 shows best agreement for all classes with the inclusion of
567 TEMPO. The selected AM rush hour is a morning peak hour for multiple urban and suburban stations. Two of three
568 rural stations show only weak morning emission spikes, while the remaining station has elevated concentrations
569 throughout the night, peaking at 2:00 LT. The station-mean TEMPO trend shows good agreement with the suburban
570 and rural stations. In urban stations, TEMPO NO_2 increases dramatically in morning hours as traffic emissions are
571 carried aloft and model predictions overestimate more severely than other hours. Values are slightly overpredicted
572 during morning hours at suburban stations as well. Past noon, TEMPO values continue to fall and underprediction
573 occurs at 2 stations instead. In hours without TEMPO, the predicted values follow the trend observed in the urban and
574 suburban stations, although performance is diminished, predominantly in the morning. The rural stations in particular
575 show best agreement when TEMPO is available and fail to track variability in other hours with predicted values
576 remaining mostly flat. Most notably, the trend and magnitude of two rural station's evening peaks at 20:00 are severely
577 underpredicted and mean TEMPO values are highly variable at rural stations.



Predictions are elevated over urban cores and roadways at rush hour and suppressed at midday, revealing that predicted hourly departures are spatially organized by land use as opposed to a spatially uniform offset. Row 1 shows that elevated NO_2 concentrations are confined to the city's center at 12:00, with rural stations exhibiting markedly lower concentrations. During rush hours, traffic-related emissions spread elevated concentrations much further as validated by the station values.

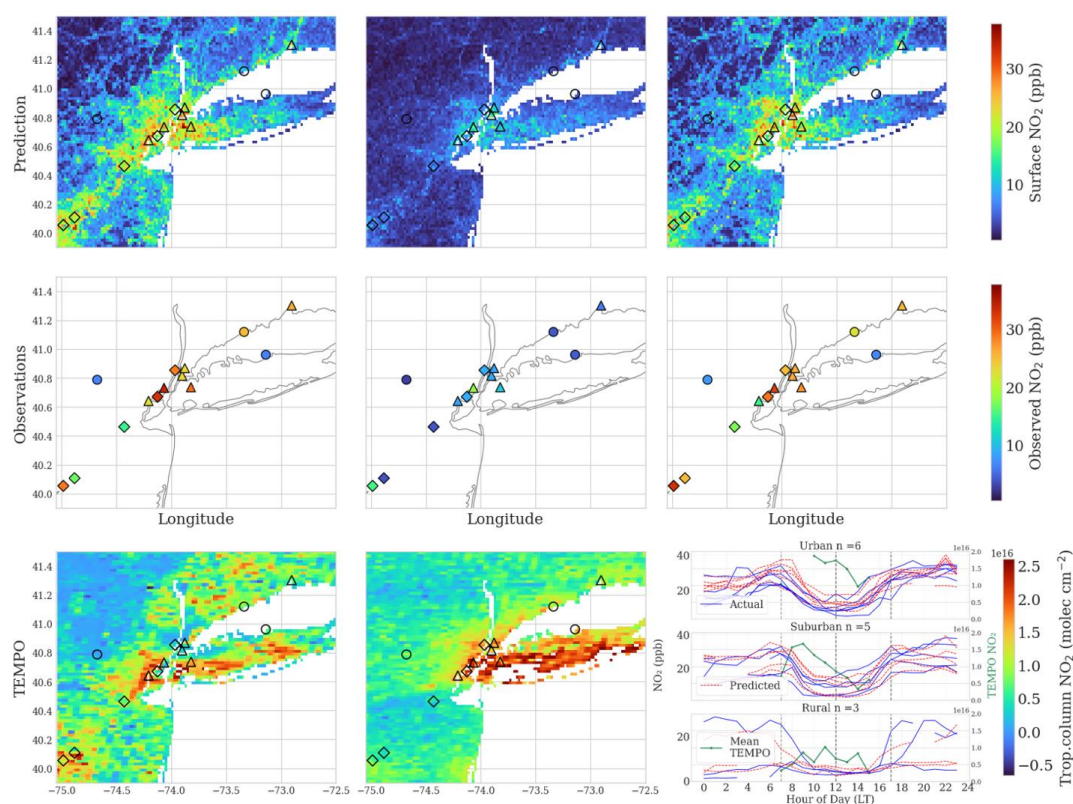


Figure 8. Case study for November 27, 2024 (median RMSE day of the test set) over the New York metropolitan region, at 07:00 LT (AM rush), 12:00 LT (noon), and 17:00 LT (PM rush). Top row: predicted surface NO_2 (ppb). Middle row: AQS station observed surface NO_2 (ppb), sharing the same color scale as above (stations are not colored in rows 1 & 3). Bottom row: TEMPO tropospheric NO_2 column at 07:00 and 12:00, and observed (solid, blue) versus predicted (dashed, red) hourly NO_2 over 24 h for stations at each class with station-mean TEMPO NO_2 (green) shown on right y-axis. Circles denote urban stations, diamonds: suburban, triangles: urban.

3.3 Model Transferability

To evaluate generalizability, we apply multiple cross-validation (CV) strategies. Our expanding window approach in the temporal CV allows us to estimate how much temporal data the model needs to achieve stable performance and our spatial CV assesses the model's ability to generalize to unseen stations. We describe the TEMPO model results (Table 2) and TEMPO feature ablation (Table 3) for daylight samples here, while no-TEMPO results can be found in Table A1.



596 3.3.1 Random Cross-Validation

597 As a baseline, we first conduct a 10-fold random CV. Model performance remains stable in each fold. We achieve a
 598 mean R^2 of 0.84 (ranging from 0.838 to 0.842) and a mean RMSE of 2.93 ppb (ranging from 2.91 to 2.96 ppb).
 599 Removing TEMPO, model performance is minimally diminished ($\sim 1\%$ for R^2) due to the strong autocorrelative nature
 600 of hourly values and data leakage in random CV.

601 3.3.2 Temporal Cross-Validation

602 For the temporal evaluation, model performance generally improves as more training data is available, with R^2 growing
 603 from 0.61 using 10% of samples to 0.79 with $\sim 60\%$ of samples, when evaluated on summer months. Performance
 604 decreases slightly when the testing period includes winter months, R^2 falls to 0.76 and RMSE increases to 3.95 in the
 605 following fold. Metrics decrease further in fold 8 ($R^2 = 0.75$; RMSE = 4.13), reflecting increased variability and higher
 606 NO_2 concentrations during these conditions. In fold 9 (April to June 2025), with $\sim 80\%$ of samples used for training,
 607 R^2 increases back up to 0.79. In the last fold, with $\sim 90\%$ of samples for training, R^2 decreases slightly (0.77), which
 608 might indicate mild overfitting and reduced generalization to unseen conditions when trained on a disproportionately
 609 large fraction of data relative to the test set. Interestingly, the inclusion of TEMPO features provides the largest
 610 contribution in the first two most data-scarce folds (R^2 % change = 9–7; RMSE = -6%–5%) indicating that a direct,
 611 high-resolution NO_2 signal matters most when little historical data is available.

612 Notably, R^2 jumps from 0.60 to 0.75 from fold 2 (4 months) to fold 3 (6 months) with RMSE dropping from 5.09 to
 613 3.43 ppb. The results suggest the model benefits from more seasonal coverage, while RMSE tends to increase in folds
 614 where the test dataset is composed of winter months. In fold 5, removing TEMPO leads to nearly identical performance
 615 when tested on summer months where more active photochemistry lowers NO_2 and compresses TEMPO's dynamic
 616 range toward instrument-noise levels (Nawaz, Huber, et al., 2025). Overall, the model generalizes reasonably well
 617 forward in time using at least one year of training data and TEMPO provides the most benefit in winter and fall.

618 3.3.3 Spatial Cross-Validation and TEMPO Ablation

619 In spatial CV evaluation, model performance dropped significantly, with a mean spatial R^2 of 0.48 (ranging from 0.35
 620 to 0.6) and a mean RMSE of 4.95 ppb (ranging from 4.05 to 6 ppb). This is consistent with many surface
 621 NO_2 prediction studies and has practical implications indicating reduced generalizability at unseen stations (Cao,
 622 2023; Ghahremanloo et al., 2021; Wei et al., 2025). Despite reduced performance, spatial cross-validation makes
 623 TEMPO's added value clear. TEMPO features raise R^2 in every one of the ten folds, by 10% to 37% (mean +17%,
 624 $\Delta R^2 = +0.07$), and lowers RMSE by a mean of 6%. The gain is largest in fold 3 (R^2 +37%, $\Delta R^2 = +0.14$; RMSE -12%).
 625 In the 80/10/10 split, ablation leads to a minor reduction in R^2 of 1.03% (0.772 versus 0.779) and RMSE increase of
 626 1.77% (3.33 versus 3.38) in overall metrics for daylight samples. This spatial improvement stands in contrast to the
 627 temporal fold and 80/10/10 split, where TEMPO's contribution faded once sufficient training history accumulated and
 628 indicates that TEMPO is most valuable for generalizing to unseen locations, where the model cannot rely on learned
 629 local patterns and a direct column observation supplies information coarse meteorology cannot.



630 **Table 2. Random, Temporal (Expanding), and Spatial Cross-Validation Results for the TEMPO Model**

<i>Fold</i>	<i>Random</i>			<i>Temporal</i>		<i>Spatial</i>	
	R ²	RMSE	End Date	R ²	RMSE	R ²	RMSE (ppb)
<i>1</i>	0.84	2.95	07 October 2023	0.61	5.02	0.60	4.15
<i>2</i>	0.84	2.91	11 December 2023	0.60	5.09	0.60	4.23
<i>3</i>	0.84	2.94	22 February 2024	0.75	3.43	0.52	6.00
<i>4</i>	0.84	2.95	23 April 2024	0.76	2.78	0.44	5.65
<i>5</i>	0.84	2.93	19 June 2024	0.76	2.81	0.49	4.05
<i>6</i>	0.84	2.96	06 August 2024	0.79	3.21	0.38	5.56
<i>7</i>	0.84	2.92	12 October 2024	0.76	3.95	0.44	5.16
<i>8</i>	0.84	2.92	25 December 2024	0.75	4.13	0.35	4.84
<i>9</i>	0.84	2.94	05 March 2025	0.79	3.19	0.53	5.18
<i>10</i>	0.84	2.92	01 May 2025	0.77	2.70	0.44	4.66
<i>Mean</i>	0.84	2.93		0.74	3.63	0.48	4.95

631

632 **Table 3. Random, Temporal (Expanding), and Spatial Cross-Validation Ablation Percent Change from the TEMPO Model**
633 **Samples Trained without TEMPO Features**

<i>Fold</i>	<i>Random</i>			<i>Temporal</i>		<i>Spatial</i>	
	R ²	RMSE	End Date	R ²	RMSE	R ² (%)	RMSE (%)
<i>1</i>	1.14	-0.30	07 October 2023	9.04	-6.00	11.59	-6.92
<i>2</i>	1.20	-0.72	11 December 2023	7.32	-4.76	10.28	-6.38
<i>3</i>	1.14	-0.45	22 February 2024	0.93	-1.39	37.30	-12.07
<i>4</i>	0.89	0.26	23 April 2024	0.50	-0.79	22.44	-6.53
<i>5</i>	1.04	-0.21	19 June 2024	-0.06	0.10	18.73	-6.76
<i>6</i>	1.07	0.55	06 August 2024	1.74	-3.15	11.28	-2.88
<i>7</i>	1.21	-0.89	12 October 2024	2.65	-3.93	16.94	-5.18
<i>8</i>	1.04	-0.74	25 December 2024	2.51	-3.42	11.04	-2.59
<i>9</i>	0.93	0.37	05 March 2025	0.72	-1.31	13.14	-6.07
<i>10</i>	1.07	-0.16	01 May 2025	0.79	-1.27	18.23	-5.63
<i>Mean</i>	1.07	-0.23		2.61	-2.59	17.10	-6.10



634 3.4 Feature Importance

635 Figure 9 shows the importance of model features based on the metrics: gain, cover, and SHAP mean values.

636 ISP, BLH, road density, closely followed by TEMPO troposphere column NO₂, and day of week have the 1st-5th
 637 highest SHAP mean values, respectively. However, in terms of gain, TEMPO NO₂ dominates as the feature
 638 contributing the most to the average reduction in loss, followed by ISP and BLH.

639 The dominance of static spatial features such as ISP and road density reinforces the finding from the spatial error
 640 analysis that land use is the primary driver of spatial NO₂ variability. It is important to note that heavy reliance on
 641 these static features may limit generalizability for stations not included in the training or validation datasets.

642 Among meteorological predictors, boundary layer height (BLH) is the most influential. This variable has the strongest
 643 correlation of the ERA5 variables across all observations and is available for all hours in the model.

644 The TEMPO signal-to-noise feature is ranked 5th in terms of gain and substantially lower than TEMPO NO₂ in terms
 645 of cover and SHAP value, possibly due to this feature being derived from features already included in the model.

646 The divergence in gain/SHAP versus cover metrics of temporal features does not indicate that these features have little
 647 predictive power, but rather that their role is structural in the model's decision making.

648 The SHAP beeswarm plot, shown right in Figure 9, represents the range and magnitude of SHAP values across
 649 observations. For example, impervious surface percentage spreads across the entire SHAP value range, with low
 650 feature values (blue) decreasing the predicted value, and high feature values (red) increasing feature predictions. By
 651 max SHAP values, shown in red on the right side of Figure 9, ISP influences predictions the most. Impactful features,
 652 such as BLH, ISP, TEMPO NO₂, and road density have a wide spread, indicating these features have a large effect on
 653 model predicted values.

654 High BLH, hour of day, solar radiation, 2m temperature, relative humidity, and wind speed values decrease surface
 655 NO₂ predictions, while high ISP, and TEMPO NO₂ values increase model predicted values. The ISP SHAP range
 656 spreads far beyond those of other features, making them look relatively less influential despite high feature importance
 657 according to XGBoost feature importance metrics.

658 No-TEMPO model and TEMPO-ablated model feature importance metrics can be found in Figures A1 and A2,
 659 respectively. For these two models, features are ranked in the same general order as in Figure 9 with ISP first, followed
 660 by BLH and road density. In both, temporal features rise in rank. The daytime, TEMPO ablated model focuses on
 661 daily/hourly indicators (holiday, day of week, and hour of day rank above most meteorological variables); while the
 662 no-TEMPO model shifts to seasonal signals, with month ranked 4th by gain. Surface pressure importance rises in both
 663 models. In the no-TEMPO night model, features related to dispersion (friction velocity, wind speed) rank higher, while



lowest in the ablated day model. Surface shortwave radiation downwards importance decreases in both models, most likely due to its relationship with solar-backscattered TEMPO observations.

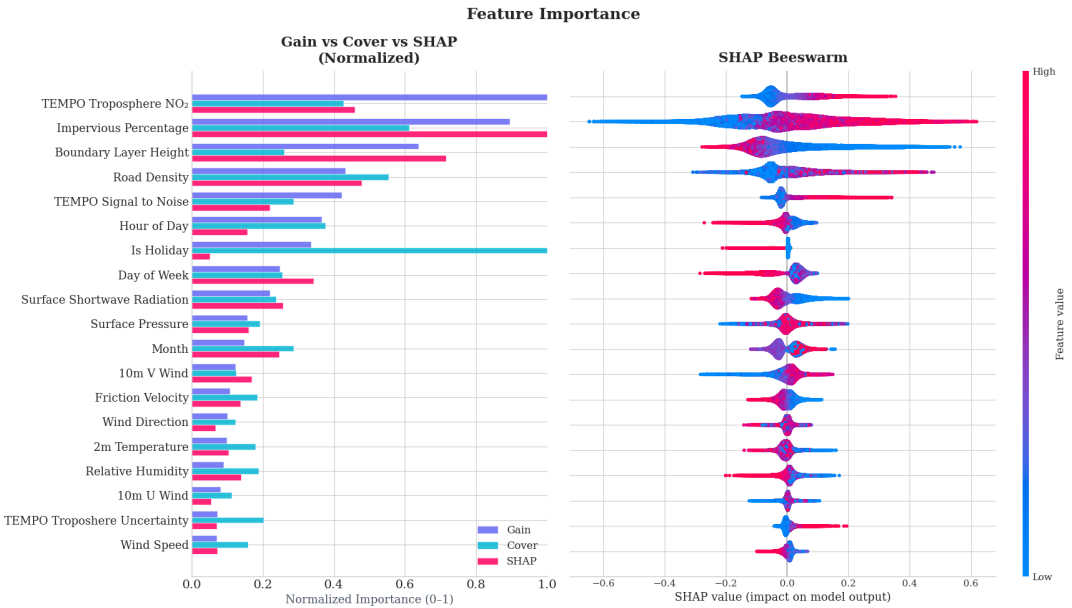


Figure 9. (A) Feature importance for the XGBoost TEMPO model based on model gain (purple), cover (blue), and SHAP values (red), normalized to 0-1 scale. (B) SHAP Beeswarm plot showing the distribution of feature contributions to model predictions. Each point represents a single prediction with color indicating the feature value (blue: low, red: high). The width or spread of the distribution signifies the magnitude of each feature’s impact on model predictions, while the sign of the SHAP value indicates the direction of that impact, with positive values increasing the predicted value and negative values decreasing it.

3.5 Impact of TEMPO on Diurnal Performance

To assess the impact of limited TEMPO data availability on model performance, we evaluate hourly variations of the mean errors (RMSE and MAE) and R^2 to gain a better understanding of the importance of TEMPO features. Fig. 10 shows that model performance tracks TEMPO availability at all hours. R^2 peaks (0.79) in hours where TEMPO coverage is highest (86%) and drops to ~0.72 at night when TEMPO is unavailable. Hours 8:00-10:00 show the largest improvement in both R^2 and RMSE when compared to without TEMPO. R^2 is lowest in hours, 1:00-3:00, with a local minimum in hour 16:00 where the inclusion of TEMPO provides little benefit. The gap in errors shown between no-TEMPO and TEMPO model hours stems from evaluation on different samples. Predicting TEMPO model samples with the no-TEMPO model removes disjoint error curves and mirrors the ablation results below when compared to the TEMPO model (Fig. A4).

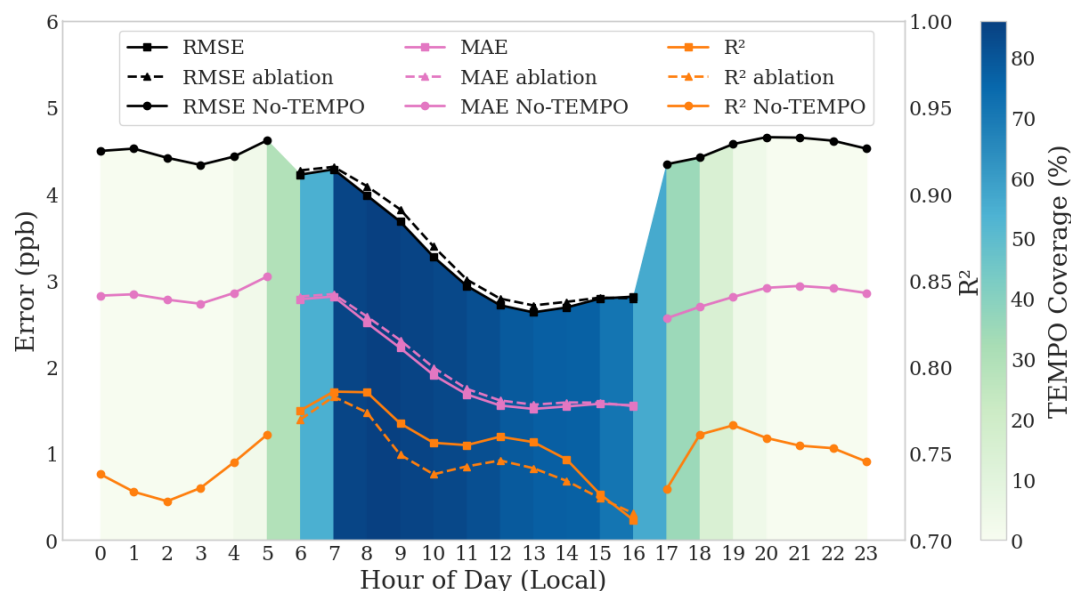
Interestingly, R^2 exhibits a second, less extreme, peak in hours 17:00 and 19:00. This time marks the increase towards the second daily NO_2 peak (20:00). As shown by the time series, the model captures the temporal trend during this period, but tends to underpredict the magnitude, leading to a lower R^2 than in hours with full TEMPO coverage. This



686 result suggests that while the temporal structure of the evening rise can be predicted from land use and meteorological
 687 features alone, accurate magnitude estimation relies on quality TEMPO observations to signal the concentration range.

688 The model error shows consistent diurnal patterns. RMSE and MAE peak during early morning (5:00-6:00), and late
 689 evening (19:00-23:00) and decrease when TEMPO variables are available, with the lowest errors in midday hours
 690 (12:00-15:00). RMSE increases in hours 16:00-22:00 as TEMPO availability falls from 72 to 56% and
 691 NO₂ concentrations build up as daylight ends. Model errors per hour of the day, particularly MAE, largely reflect the
 692 magnitude and variability of NO₂ concentrations during those hours rather than a fundamental degradation in model
 693 skill, mirroring the diurnal pattern of NO₂.

694 Of 474 AQS stations, 436 stations provide complete 24-hour NO₂ observations. In stations that lack 24-hour coverage,
 695 early morning hours (2:00-4:00) have the fewest observations, with 459, 465, and 467 stations reporting data,
 696 respectively. The reduced number of stations during these hours likely decreases variability and may account for the
 697 decreased error during these hours.



698

699 **Figure 10. RMSE Model Error by local Hour of Day (black line). Hours 17-5 show no-TEMPO model test results (circles),**
 700 **while hours 6-16 show the results from both the TEMPO model (squares) and from a model trained on the same samples**
 701 **without TEMPO features (dashed line, triangles). MAE is shown as a pink line. R² is shown as an orange line and labeled**
 702 **on the left y-axis. Hours colored by TEMPO Troposphere Vertical Column Availability (%).**



703 3.6 Daytime and Nighttime Surface NO₂ Mapping

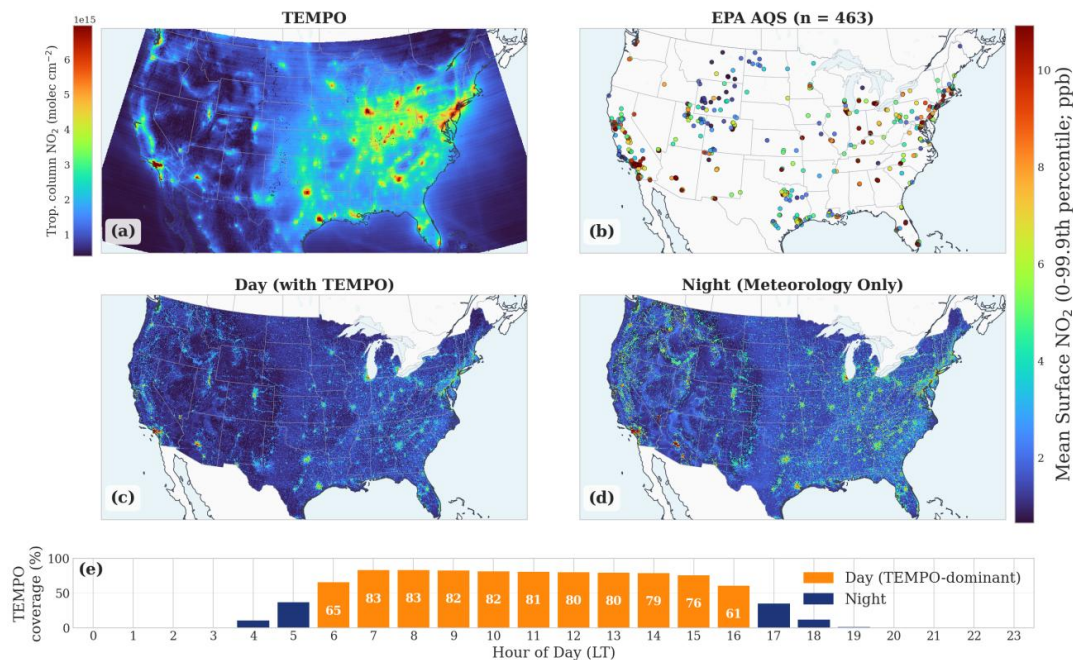
704 3.6.1 Annual Mean Differences Across CONUS

705 To examine spatial differences in day versus night predictions, we show the annual mean predicted surface NO₂ across
 706 CONUS at 0.02° resolution, separated into daytime (c) and nighttime (d) composites, together with the hourly fraction
 707 of grid cells for which TEMPO observations were available (e) (Fig. 11). Panel a shows TEMPO mean troposphere
 708 column NO₂ and panel b shows AQS station means for reference. Hours colored orange in panel e were selected for
 709 the day average while all other hours represent the night average.

710 Panel e shows that TEMPO observations are available for 78-84% of cells between 7:00 and 15:00 LT, declining to
 711 65% at 6:00 and 60% at 16:00 as retrieval quality diminishes. Residual coverage at 4:00-5:00 and 17:00-18:00 reflects
 712 seasonal variation in daylight hours rather than routine sampling. Between 19:00 and 3:00 LT no TEMPO observations
 713 exist and all predictions in this window derive from the no-TEMPO model. The nighttime composite in panel d is
 714 therefore produced entirely via meteorology and should be interpreted with more caution, since predictions are not
 715 constrained by TEMPO NO₂ observations, while the daytime composite is produced predominantly by the TEMPO-
 716 informed model; the sharper, more spatially compact urban enhancements captured during the day, relative to the
 717 broader, less resolved nighttime plumes, underscore the added spatial resolution TEMPO provides beyond what
 718 meteorology and static land-use features alone can recover.

719 Both composites reproduce the expected urban-rural gradient, with rural background concentrations below 2 ppb and
 720 discrete enhancements over major metropolitan areas. The Los Angeles basin is the highest-concentration region in
 721 both panels, exceeding 10 ppb, followed by the New York-New Jersey corridor, Chicago, Houston, the San Francisco
 722 Bay Area, Seattle, Denver, and Atlanta. Smaller enhancements mark secondary cities and, in the Gulf Coast and Ohio
 723 Valley, industrial point sources, spatially tracking the TEMPO column field. Predicted magnitudes in (c-d) are
 724 consistent with the co-located AQS station means in (b), which range from under 2 ppb at rural sites to above 10 ppb
 725 in the Los Angeles basin.

726 Nighttime concentrations exceed daytime concentrations across the domain, and the spatial character of the two
 727 composites differs. Daytime enhancements are compact and centered on urban cores, whereas the nighttime composite
 728 shows an extended road network along the interior West and Great Plains that is largely absent in the TEMPO column,
 729 along with broader plumes surrounding each urban center. Both features are consistent with nocturnal boundary layer
 730 collapse: the same emissions are confined to a shallower mixing layer, raising surface concentrations and preserving
 731 near-source gradients that daytime mixing would otherwise dilute. The cessation of NO₂ photolysis after sunset further
 732 extends its daytime-limited lifetime.



733

734 **Figure 11.** Annual mean predicted surface NO₂ over CONUS for 2024, composited by time of day. (a) TEMPO troposphere
735 column NO₂ mean. (b) AQS station mean. (c) Daytime mean (6:00–16:00 LT) and (d) nighttime mean (17:00–05:00 LT),
736 both at 0.02° resolution. Panels b–d share the color scale at right. (e) Percentage of grid cells with valid TEMPO tropospheric
737 column retrievals by hour of local time; hours with coverage above 50% are labeled and selected for the daytime mean.
738 Predictions during hours without TEMPO availability are generated by the no-TEMPO model, so the nighttime composite
739 in (d) derives entirely from that model, while (c) is produced predominantly by the TEMPO-informed model. Values are
740 limited up to the 99.9th percentile of the combined predictions and station values to visualize the spatial distribution.
741 TEMPO column NO₂ is limited between 0 and the 98th percentile.

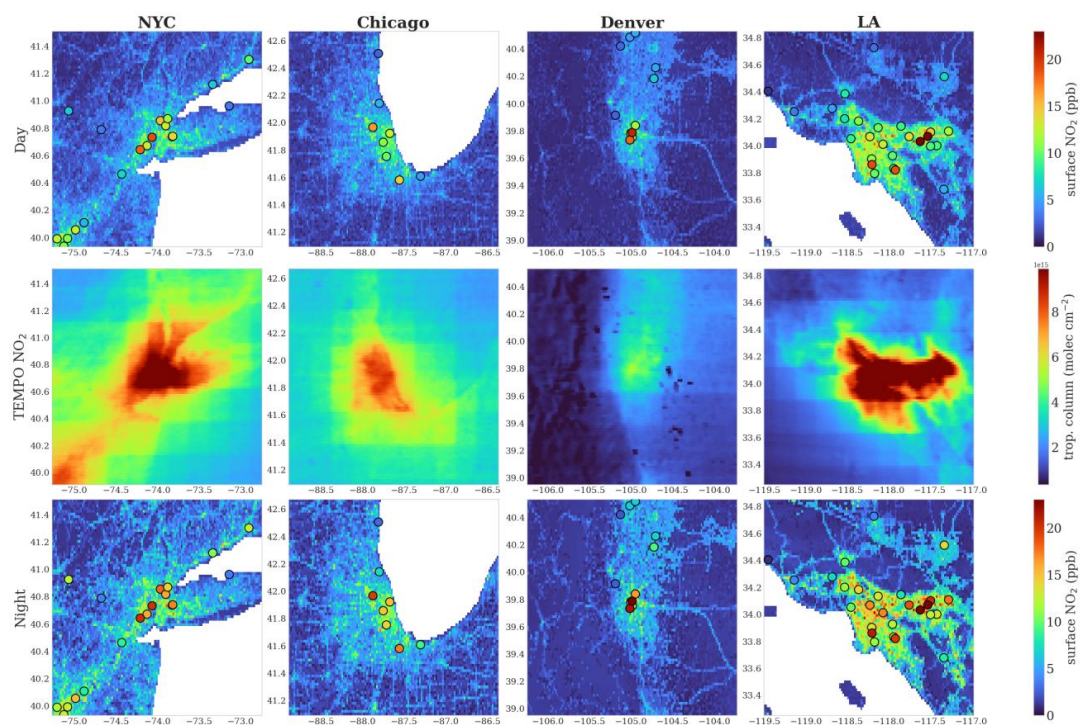
742 3.6.2 Metropolitan-Scale Agreement with Station Observations

743 Figure 12 offers a closer look at four metropolitan areas, New York, Chicago, Denver, and Los Angeles, to examine
744 the agreement of individual station annual means. For each city the annual daytime (top row) and nighttime (bottom
745 row) predicted surface NO₂ fields are shown with EPA AQS station means overlaid as points on the same color scale.
746 The middle row shows the annual mean TEMPO tropospheric column NO₂ that is used to inform the TEMPO model.

747 At this scale we observe how predictions are informed by TEMPO. Broadly, the column maxima and surface maxima
748 are co-located over each urban center, but the surface field is sharper, resolving near-source gradients and the arterial
749 road network that the vertically integrated column instead renders as a diffuse plume over most metropolitan areas.
750 The road network is fully resolved at night and higher NO₂ concentrations spread further from the city center,
751 signifying a shift from the finer scale TEMPO input to natively-coarser ERA5 meteorology and static urbanization
752 features.



753 The transition from high-concentration urban cores to background NO_2 is accurately captured and validated by the
 754 station means, which grade from predicted maxima at the city center to the rural background in step with the underlying
 755 field. This transition is more extended at night, consistent with boundary-layer collapse concentrating emissions within
 756 a shallower mixing layer and preserving near-source gradients that daytime convective mixing would otherwise dilute.



757
 758 **Figure 12.** Detailed view of the annual mean predicted surface NO_2 field (Fig. 11) over four representative metropolitan
 759 areas: New York, Chicago, Denver, and Los Angeles (columns). Top row: daytime surface NO_2 (ppb). Middle row: annual
 760 mean TEMPO tropospheric NO_2 column (molecules cm^{-2}), used to inform the daytime model. Bottom row: nighttime
 761 surface NO_2 (ppb). Filled points denote EPA AQS station annual means on the same color scale as the underlying surface
 762 field. Note that the nighttime surface panel's maximum value is slightly higher than that of daytime.

763 4 Discussion

764 Hourly surface NO_2 estimation across CONUS has not been possible due to the temporal limitations of polar-orbiting
 765 satellites. This study demonstrates that hourly surface NO_2 can be predicted across CONUS using XGBoost models
 766 trained on TEMPO satellite observations, ERA5 meteorology, and temporal and land use features. The daytime
 767 (TEMPO) model achieves a test R^2 of 0.779 and RMSE of 3.33 ppb; while the no-TEMPO night model achieves a
 768 lower R^2 of 0.755 and higher RMSE of 4.36.

769 When spatially averaging predictions per timestep, results vary by region and season. Across seasons, skill is highest
 770 when concentration and variability are lowest. In winter, a shallow boundary layer, longer NO_2 photochemical



lifetime, and elevated heating emissions produce the highest concentrations and the sharpest peaks, which the model struggles to capture. This leads to the largest RMSEs and widest gaps between r and R^2 . The summer-minimum RMSE reflects deeper daytime mixing and rapid photolytic loss, which flatten concentrations and reduce the dynamic range, the strong LA-summer performance indicates that low concentrations do not inherently degrade performance ($R^2 = 0.85$). We separate regional differences by those attributable to station sample size and environmental characteristics. Denver's low correlations point to unresolved local variability, most likely in the form of complex topography, and a tightly-clustered monitor network, Chicago's high- r , low- R^2 fall panel instead reflects a systematic offset error on a smaller station sample ($n = 6$), where a single poorly represented site can dominate the aggregated mean. Denver's performance necessitates additional features or finer spatial representation (however inclusion of GMRT topography led to persistent artifacts) whereas the Chicago/LA-winter results show the temporal structure remains intact, calling for bias-correction.

Temporal cross-validation reveals that with at least a year of data, the model generalizes relatively well forward in time, with improved coefficients of determination and TEMPO proving particularly valuable when data are scarce or during fall and winter months.

Stations exhibit different behavior when stratified by land cover class and hour of day, showing lower concentrations and less variability from urban to suburban to rural. Agreement is strongest in local hours 8-17 from the benefit of TEMPO and minimal concentrations.

Feature importance reveals impervious surface percentage dominates spatial prediction and boundary layer height is the most important dynamic predictor. TEMPO column NO_2 ranks first by gain and fourth by max SHAP values as its influence is concentrated in a one-sided tail of high-column samples rather than spread across the record (Fig. 9B), unlike impervious surface percentage which influences nearly all predictions. With TEMPO removed, temporal features rise in importance for both the day and night models, however the day model emphasizes daily/hourly patterns while the night model relies on monthly/seasonal patterns, consistent with daytime NO_2 exhibiting greater variability on short timescales (%RSD), requiring finer temporal resolution to capture. This suggests that TEMPO contributes both a direct signal and implicit temporal information, now recovered through calendar proxies. In the no-TEMPO model, representing mostly night samples, mixing features rise in importance, indicating that dispersion dynamics play a key role in nocturnal NO_2 prediction. While originally meant to serve as a proxy for photolysis, shortwave solar radiation downwards ranks higher with TEMPO included, suggesting it may inform the model of TEMPO retrieval quality. Hourly analysis reveals model performance is strongly modulated by TEMPO data availability. R^2 peaks at 0.79 during hours with greater than 86% TEMPO coverage and drops to approximately 0.72 at night when no satellite NO_2 concentrations are available. This pattern is consistent across RMSE and indicates that TEMPO provides information about surface NO_2 concentrations and variability that meteorological and temporal features alone cannot replicate. The small performance difference when removing TEMPO suggests partial compensation when averaged across hours of day under 80/10/10 seasonal split. In contrast, for spatial cross-validation, the inclusion of TEMPO proves vital to generalizing at unseen stations, leading to a 17% increase in R^2 and 6% reduction in RMSE.



806 Spatial generalizability is limited: spatial CV yields R^2 of 0.48 versus 0.84 for random CV, indicating reliance on
807 station-specific patterns learned through static spatial features. Given the gridded format of most features and
808 limitations exposed during spatial cross-validation, future work could include developing an image-based ML or deep
809 learning model, such as a CNN, which was used with good results in daily studies (Cao, 2023; Ghahremanloo et al.,
810 2021). In addition, higher spatial resolution meteorological variables may be necessary to capture the extreme local
811 gradients of NO_2 dispersion from emission sources (Mucha & Mainka, 2026).

812 The evening NO_2 peak, which coincides with declining TEMPO coverage, increasing retrieval uncertainty, and
813 meteorological features that exhibit minimal hour-to-hour variability, is systematically underpredicted, which may be
814 resolved with high-quality traffic data. Originally, we evaluated Federal Highway Administration's (FHWA) Traffic
815 Monitoring and Surveys (TMAS) hourly traffic volumes, which showed moderate correlation with surface NO_2 ($r =$
816 0.287, stronger during daylight and downwind conditions). However, inclusion degraded model performance, likely
817 due to inconsistent station availability causing high day-to-day variability in nearest-monitor averages. As of this time,
818 TMAS data is only available through 2024 which would have reduced the study period.

819 This work is fundamentally constrained by sparse monitoring across a large area for a pollutant with steep distance
820 decay and a short diurnal lifecycle. Other key limitations include excessive cloud coverage where recommended
821 thresholds would remove 47.5% of samples on average per week and the use of L3 vertical column densities rather
822 than slant column densities, which correlate more strongly with surface NO_2 (Nawaz, Huber, et al., 2025); and lack of
823 reliable hourly traffic data to capture evening emission dynamics. In theory, observed NO_2 concentrations from ground
824 sensors or satellite instruments should not fall below zero, but often do due to instrumental noise or drift. Future work
825 could involve quantifying the trend of this drift and correcting for negative values per sensor. This could lead to more
826 realistic distributions of ground-level NO_2 , while mitigating statistical bias. High concentrations are also generally
827 underpredicted in the model which may be addressed through strategies such as oversampling high-concentration
828 periods, asymmetric loss functions, or ensemble approaches that weigh tail predictions more heavily.

829 Direct comparison of our model's performance with existing literature is constrained by the absence of prior studies
830 that achieve hourly temporal resolution and continental-scale spatial coverage. The closest CONUS-wide hourly study
831 (Salman et al., 2024) uses CMAQ output as opposed to observed surface concentrations as ground truth. Daily studies
832 report stronger agreement, but this is expected given that temporal aggregation to daily resolution substantially
833 smooths out the diurnal variability that our model attempts to capture. This study establishes the first empirical
834 benchmark for using geostationary satellite NO_2 to estimate hourly surface concentrations with continental coverage
835 across the US, quantifies TEMPO's added value relative to meteorological features alone, and identifies where the
836 model fails (evening hours, unseen stations, high-concentration events) to guide future work. This study aims to
837 support epidemiological studies of acute NO_2 health effects and exposure assessment at temporal resolutions that have
838 not previously been feasible at this scale.



1. Equations

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

$$r = \frac{\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2} \sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2}}$$

Where n is the total sample size, i is the sample index, y_i is the i^{th} observed value, and $\hat{y}_i$ is the i^{th} predicted value, $\bar{y}$ is the mean of observed values, and $\bar{\hat{y}}$ is the predicted mean.

Appendix A

Table A1. Random, Temporal (Expanding), and Spatial Cross Validation Results for the No-TEMPO Model

Fold	Random		Temporal		Spatial		
	R ²	RMSE	End Date	R ²	RMSE	R ²	RMSE (ppb)
0	0.82	3.83	9/29/2023	0.53	6.78	0.51	6.34
1	0.82	3.82	11/29/2023	0.57	6.52	0.43	6.63
2	0.81	3.81	1/25/2024	0.66	5.31	0.36	6.33
3	0.82	3.80	3/21/2024	0.70	4.05	0.42	6.90
4	0.81	3.82	5/26/2024	0.68	3.71	0.45	6.47
5	0.82	3.81	8/10/2024	0.74	4.08	0.49	6.35
6	0.81	3.83	10/12/2024	0.75	4.95	0.53	6.20
7	0.82	3.81	12/7/2024	0.73	5.45	0.44	6.39
8	0.82	3.81	2/5/2025	0.74	4.74	0.54	6.68
9	0.82	3.82	4/15/2025	0.72	3.58	0.53	5.93
Mean	0.82	3.82		0.68	4.92	0.47	6.42

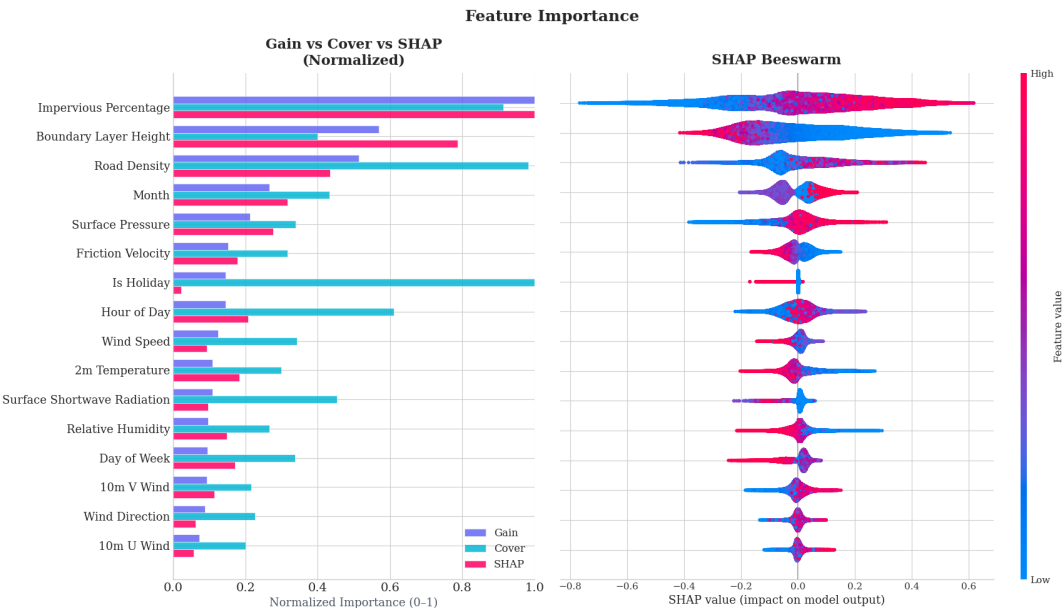
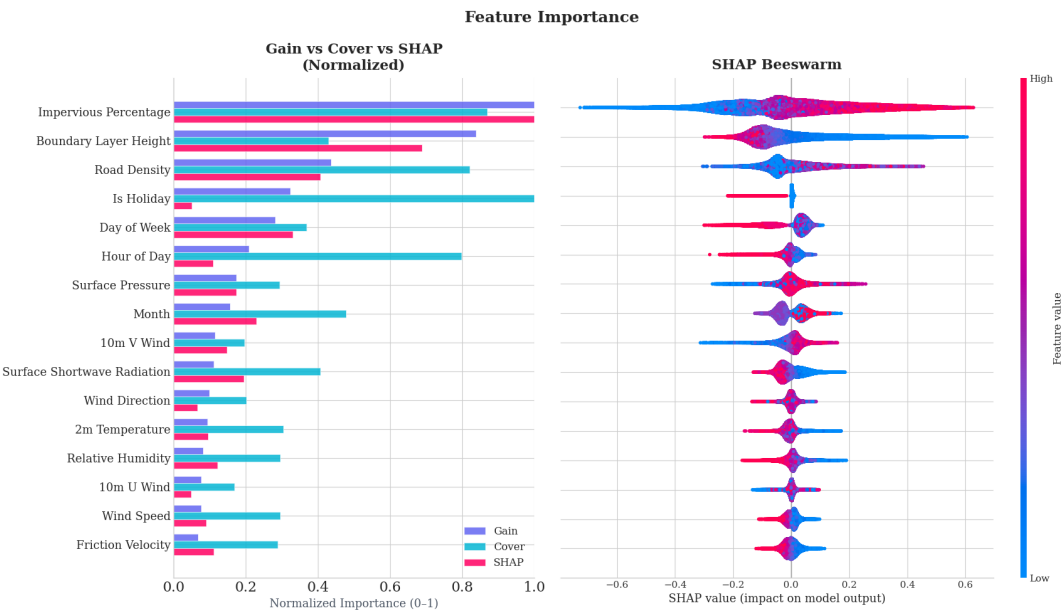
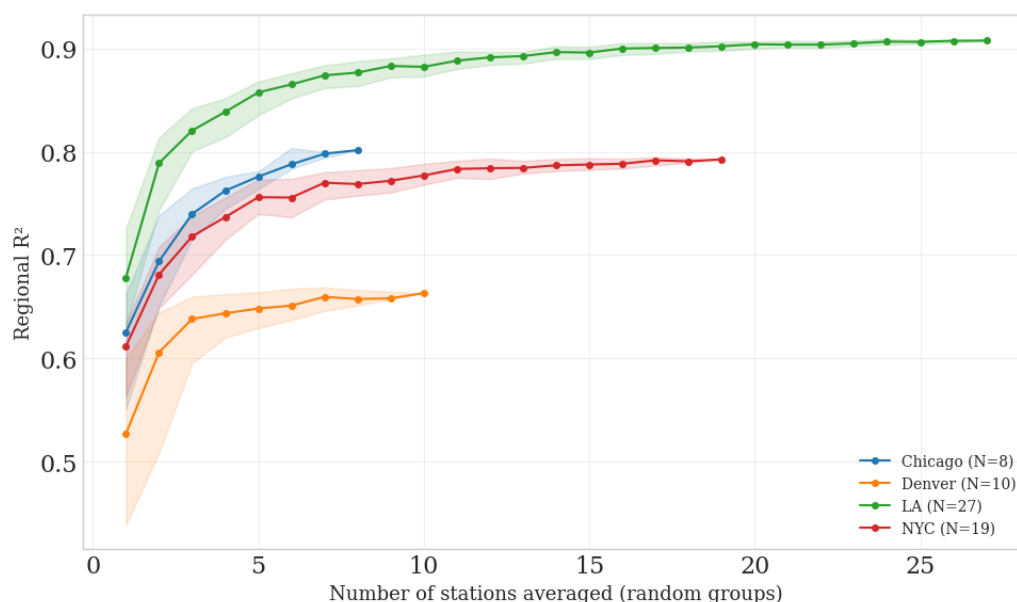


Figure A1. (A) Feature importance for the XGBoost no-TEMPO model based on model gain (purple), cover (blue), and SHAP values (red), normalized to 0-1 scale. (B) SHAP Beeswarm plot showing the distribution of feature contributions to model predictions. Each point represents a single prediction with color indicating the feature value (blue: low, red: high). The width or spread of the distribution signifies the magnitude of each feature's impact on model predictions, while the sign of the SHAP value indicates the direction of that impact, with positive values increasing the predicted value and negative values decreasing it.





858 **Figure A2. (A) Feature importance for the XGBoost model trained on TEMPO-available samples with TEMPO features**
 859 **removed based on model gain (purple), cover (blue), and SHAP values (red), normalized to 0-1 scale. (B) SHAP Beeswarm**
 860 **plot showing the distribution of feature contributions to model predictions.**



861
 862 **Figure A3. Regional NO₂ prediction skill as a function of the number of stations averaged by metropolitan region (Chicago:**
 863 **blue, Denver: orange, LA: green, NYC: red). For each region, n monitoring stations were drawn at random (100 draws per**
 864 **n, without replacement) and the regional R² was computed between the station-averaged and predicted NO₂. Curves show**
 865 **the median across draws and shaded bands: the interquartile range. Each curve extends to that region's total station count**
 866 **(LA = 27, NYC = 19, Denver = 10, Chicago = 8). All four regions gain steeply from n = 1 to n ≈ 5, indicating that a substantial**
 867 **part of the difference in raw regional R² reflects the number of stations averaged rather than model quality. At matched n,**
 868 **the ordering LA > Chicago > NYC > Denver holds without curves crossing, indicating a genuine skill hierarchy independent**
 869 **of station count. Denver's early plateau near R² ≈ 0.66, largely unchanged by additional stations is consistent with**
 870 **unresolved local variability rather than sampling noise.**

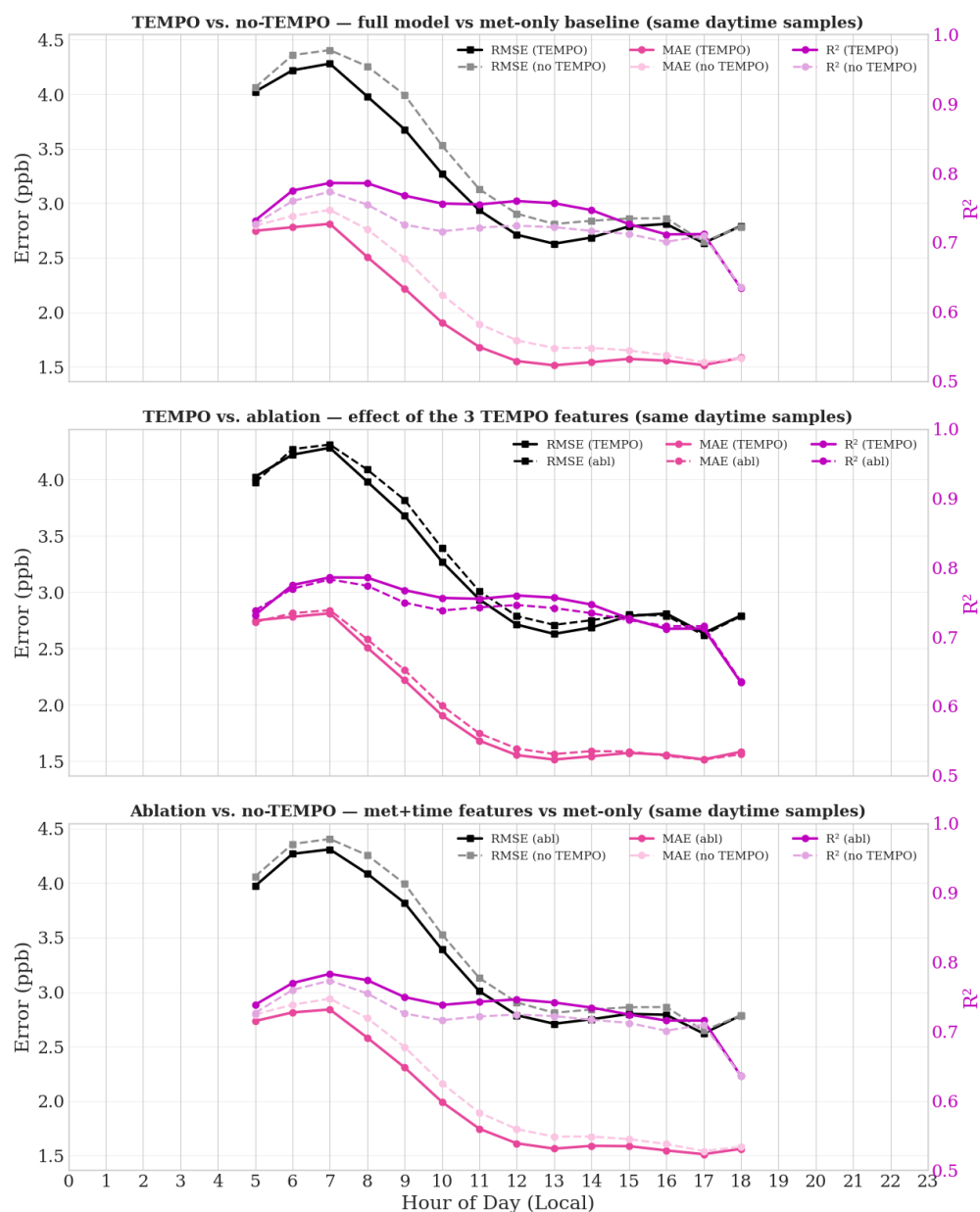


Figure A4. Hourly daytime error and model skill for three model configurations evaluated on identical TEMPO-available (daytime) test samples. Metrics are computer per local hour over the daytime TEMPO retrieval window (~5:00-18:00). Left axis: RMSE (black) and MAE (magenta), ppb; right axis: R² (purple).

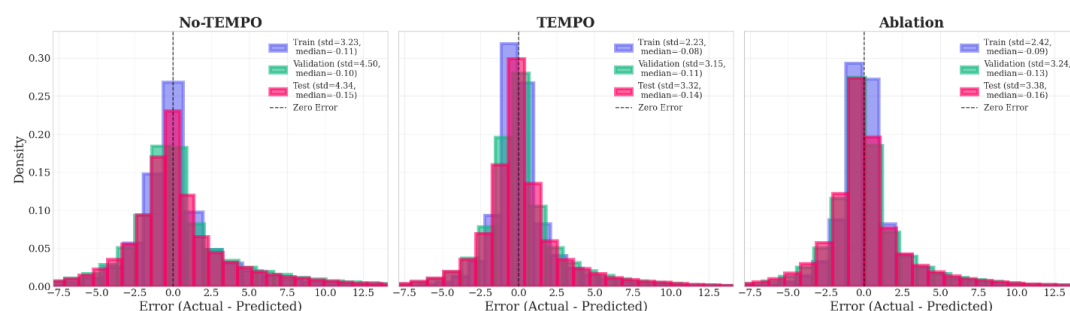


Figure A5. Distribution of residual errors (actual – predicted) for the train, validation, and test sets, clipped at the 99th percentile, for the No-TEMPO, TEMPO and TEMPO-ablated model (columns).

Across all three models the residual distributions are sharply peaked and centered near zero, with median errors between -0.08 and -0.16 ppb, indicating that predictions are essentially unbiased at the median with only a slight tendency toward overprediction. The TEMPO model produces the tightest distributions (test $\sigma = 3.32$ ppb), narrower than both the no-TEMPO model (test $\sigma = 4.34$ ppb) and the TEMPO-ablated version (test $\sigma = 3.38$ ppb), with more errors centered at 0, confirming TEMPO observations help the model predict the concentration range more accurately. For every model the training residuals are modestly narrower than the validation and test residuals, a small generalization gap consistent with good out-of-sample behavior rather than substantial overfitting. All distributions share a mild right skew, with a longer positive tail (actual > predicted) reflecting residual underprediction of high-concentration episodes, consistent with the evening and boundary-layer effects discussed in the main text.

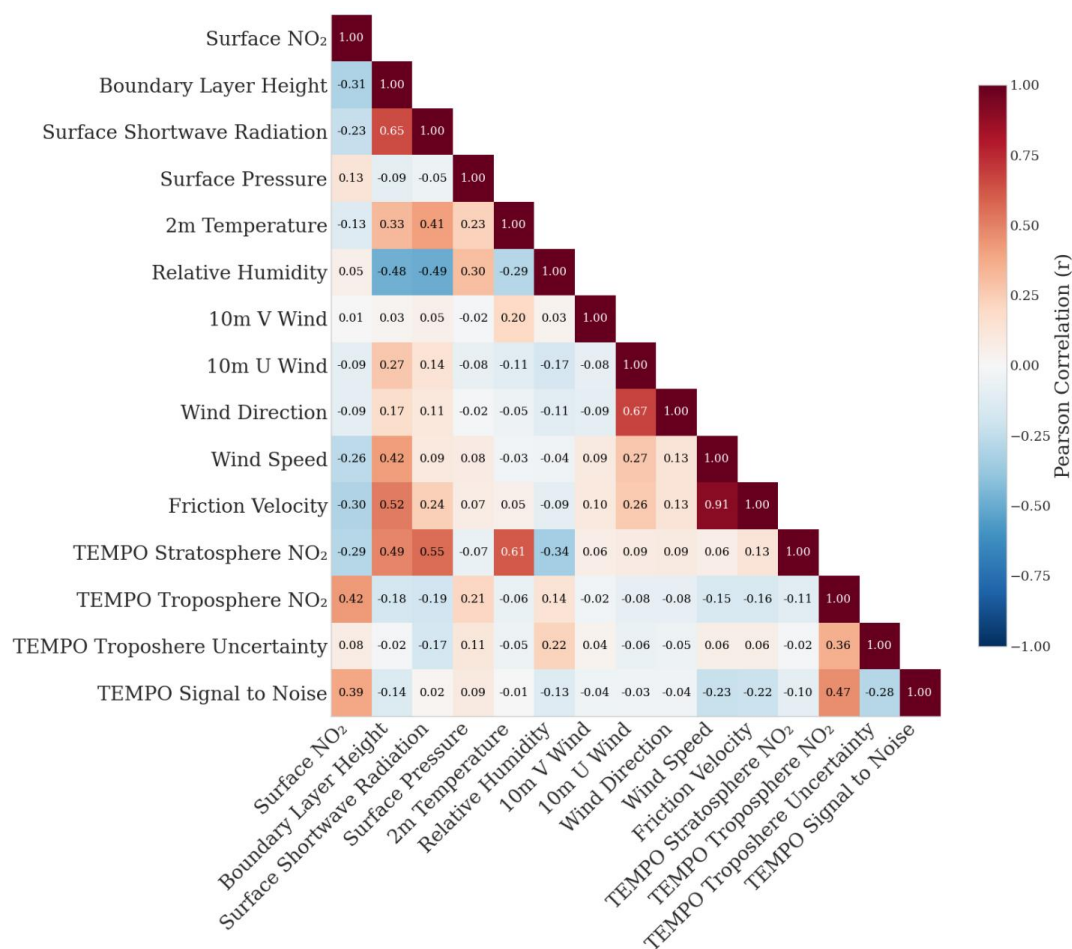


Figure A6. Correlation matrix of meteorology and TEMPO features across all observations. TEMPO features are limited to data between the 1st and 99th percentile to remove outliers.

Code and data availability

The code and data will be made available by the authors upon request.

Author contributions

SA, WN, and PR conceptualized the study and developed the methodology. WN provided computing resources. SA acquired and processed the data, wrote the software, performed the formal analysis and validation, produced the visualizations. SA and WN wrote the original draft. All authors reviewed and edited the manuscript.



896 **Competing interests**

897 The contact author has declared that none of the authors has any competing interests.

898 **Financial support**

899 We gratefully acknowledge funding from NASA grants: 80NSSC25M0070 and 80NSSC21K0194.



900 References

- 901 Acker, S., Holloway, T., & Harkey, M. (2025). Satellite detection of NO₂ distributions using
902 TROPOMI and TEMPO and comparison with ground-based concentration measurements.
903 *Atmospheric Chemistry and Physics*, 25(14), 8271-8288. <https://doi.org/10.5194/acp-25-8271-2025>
- 904 Ahmad, N., Lin, C., Lau, A. K. H., Kim, J., Zhang, T., Yu, F., Li, C., Li, Y., Fung, J. C. H., & Lao, X. Q. (2024).
905 Estimation of ground-level NO₂ and its spatiotemporal variations in China using GEMS
906 measurements and a nested machine learning model. *Atmospheric Chemistry and Physics*, 24(16),
907 9645-9665. <https://doi.org/10.5194/acp-24-9645-2024>
- 908 Barrett, A., Battisto, C., Bourbeau, J., Fisher, M., Kaufman, D., Kennedy, J., Lopez, L., Lowndes, J., Scheick, J.,
909 & Steiker, A. (2024). *earthaccess* In (Version v0.10.0) Zenodo.
- 910 Bovensmann, H., Burrows, J. P., Buchwitz, M., Frerick, J., Noël, S., Rozanov, V. V., Chance, K. V., & Goede, A.
911 P. H. (1999). SCIAMACHY: Mission Objectives and Measurement Modes. *Journal of the Atmospheric*
912 *Sciences*, 56(2), 127-150. [https://doi.org/10.1175/1520-](https://doi.org/10.1175/1520-0469(1999)056<0127:SMOAMM>2.0.CO;2)
913 [0469\(1999\)056<0127:SMOAMM>2.0.CO;2](https://doi.org/10.1175/1520-0469(1999)056<0127:SMOAMM>2.0.CO;2)
- 914 Burrows, J. P., Weber, M., Buchwitz, M., Rozanov, V., Ladstätter-Weissenmayer, A., Richter, A., DeBeek, R.,
915 Hoogen, R., Bramstedt, K., Eichmann, K.-U., Eisinger, M., & Perner, D. (1999). The Global Ozone
916 Monitoring Experiment (GOME): Mission Concept and First Scientific Results. *Journal of the*
917 *Atmospheric Sciences*, 56(2), 151-175. [https://doi.org/10.1175/1520-](https://doi.org/10.1175/1520-0469(1999)056<0151:TGOMEG>2.0.CO;2)
918 [0469\(1999\)056<0151:TGOMEG>2.0.CO;2](https://doi.org/10.1175/1520-0469(1999)056<0151:TGOMEG>2.0.CO;2)
- 919 Camilleri, S. F., Kerr, G. H., Anenberg, S. C., & Horton, D. E. (2023). All-Cause NO₂-Attributable Mortality
920 Burden and Associated Racial and Ethnic Disparities in the United States. *Environ Sci Technol Lett*,
921 10(12), 1159-1164. <https://doi.org/10.1021/acs.estlett.3c00500>
- 922 Cao, E. L. (2023). National ground-level NO₂ predictions via satellite imagery driven convolutional neural
923 networks [Original Research]. *Frontiers in Environmental Science*, Volume 11 - 2023.
924 <https://doi.org/10.3389/fenvs.2023.1285471>
- 925 Chang, B., Liu, H., Zhang, C., Xing, C., Tan, W., & Liu, C. (2025). Relating satellite NO₂ tropospheric columns
926 to near-surface concentrations: implications from ground-based MAX-DOAS NO₂ vertical profile
927 observations. *npj Climate and Atmospheric Science*, 8(1), 1. [https://doi.org/10.1038/s41612-024-](https://doi.org/10.1038/s41612-024-00891-z)
928 [00891-z](https://doi.org/10.1038/s41612-024-00891-z)
- 929 Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System* Proceedings of the 22nd ACM
930 SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco,
931 California, USA. <https://doi.org/10.1145/2939672.2939785>
- 932 European Environment Agency. (2022a). *Air pollution and health*.
933 <https://www.eea.europa.eu/en/analysis/publications/zero-pollution/health/air-pollution-and-health>
- 934 European Environment Agency. (2022b). *Air quality in Europe 2022*.
935 <https://www.eea.europa.eu/en/analysis/publications/air-quality-in-europe-2022>
- 936 Gao, Z., & Zhou, X. (2024). A review of the CAMx, CMAQ, WRF-Chem and NAQPMS models: Application,
937 evaluation and uncertainty factors. *Environmental Pollution*, 343, 123183.
938 <https://doi.org/10.1016/j.envpol.2023.123183>
- 939 Ghahremanloo, M., Lops, Y., Choi, Y., & Yeganeh, B. (2021). Deep Learning Estimation of Daily Ground-Level
940 NO₂ Concentrations From Remote Sensing Data. *Journal of Geophysical Research: Atmospheres*,
941 126(21), e2021JD034925. <https://doi.org/10.1029/2021JD034925>
- 942 Ghahremanloo, M., Nowlan, C., González Abad, G., Garraffo, C., Liu, X., Wang, H., Henderson, B., Baumann,
943 E., Valin, L., Geddes, J. A., & Zhao, X. (2025). Comprehensive Analysis of Bias in TEMPO NO₂ Column
944 Densities Through Pandora Observations. *Journal of Geophysical Research: Atmospheres*, 130(15),
945 e2025JD044150. <https://doi.org/10.1029/2025JD044150>
- 946 Goldberg, D. L., Nawaz, M. O., Lyu, C., He, J., Carlton, A. G., Kondragunta, S., & Anenberg, S. C. (2025). Clear-
947 sky and cloudy-sky differences in NO₂ concentrations over the United States: implications for satellite
948 measurement applications. *Atmos. Chem. Phys.*, 25(22), 16287-16302. [https://doi.org/10.5194/acp-](https://doi.org/10.5194/acp-25-16287-2025)
949 [25-16287-2025](https://doi.org/10.5194/acp-25-16287-2025)



- 950 Gu, J., Chen, L., Yu, C., Li, S., Tao, J., Fan, M., Xiong, X., Wang, Z., Shang, H., & Su, L. (2017). Ground-Level NO₂
951 Concentrations over China Inferred from the Satellite OMI and CMAQ Model Simulations. *Remote*
952 *Sensing*, 9(6), 519. <https://www.mdpi.com/2072-4292/9/6/519>
- 953 Hamra, G. B., Laden, F., Cohen, A. J., Raaschou-Nielsen, O., Brauer, M., & Loomis, D. (2015). Lung Cancer and
954 Exposure to Nitrogen Dioxide and Traffic: A Systematic Review and Meta-Analysis. *Environ Health*
955 *Perspect*, 123(11), 1107-1112. <https://doi.org/10.1289/ehp.1408882>
- 956 Harkey, M., Holloway, T., Oberman, J., & Scotty, E. (2015). An evaluation of CMAQ NO₂ using observed
957 chemistry-meteorology correlations. *Journal of Geophysical Research: Atmospheres*, 120(22),
958 11,775-711,797. <https://doi.org/10.1002/2015JD023316>
- 959 Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu,
960 R., Rozum, I., Schepers, D., Simmons, A., Soci, C., Dee, D., & Thépaut, J.-N. (2023). *ERA5 hourly data*
961 *on single levels from 1940 to present* (<https://doi.org/10.24381/cds.adbb2d47>)
- 962 Huang, S., Li, H., Wang, M., Qian, Y., Steenland, K., Caudle, W. M., Liu, Y., Sarnat, J., Papatheodorou, S., & Shi,
963 L. (2021). Long-term exposure to nitrogen dioxide and mortality: A systematic review and meta-
964 analysis. *Sci Total Environ*, 776, 145968. <https://doi.org/10.1016/j.scitotenv.2021.145968>
- 965 Jeon, J.-W., Park, S.-W., Han, Y.-J., Lee, T., Lee, S.-H., Park, J.-M., Yoo, M.-S., Shin, H.-J., & Hopke, P. K. (2024).
966 Nitrate formation mechanisms causing high concentration of PM_{2.5} in a residential city with low
967 anthropogenic emissions during cold season. *Environmental Pollution*, 352, 124141.
968 <https://doi.org/10.1016/j.envpol.2024.124141>
- 969 Kim, E. J., Holloway, T., Kokandakar, A., Harkey, M., Elkins, S., Goldberg, D. L., & Heck, C. (2024). A Comparison
970 of Regression Methods for Inferring Near-Surface NO₂ With Satellite Data. *Journal of Geophysical*
971 *Research: Atmospheres*, 129(16), e2024JD040906.
972 <https://doi.org/10.1029/2024JD040906>
- 973 Lamsal, L. N., Duncan, B. N., Yoshida, Y., Krotkov, N. A., Pickering, K. E., Streets, D. G., & Lu, Z. (2015). U.S.
974 NO₂ trends (2005–2013): EPA Air Quality System (AQS) data versus improved observations from the
975 Ozone Monitoring Instrument (OMI). *Atmospheric Environment*, 110, 130-143.
976 <https://doi.org/10.1016/j.atmosenv.2015.03.055>
- 977 Larkin, A., Geddes, J. A., Martin, R. V., Xiao, Q., Liu, Y., Marshall, J. D., Brauer, M., & Hystad, P. (2017). Global
978 Land Use Regression Model for Nitrogen Dioxide Air Pollution. *Environ Sci Technol*, 51(12), 6957-6964.
979 <https://doi.org/10.1021/acs.est.7b01148>
- 980 Levelt, P. F., Joiner, J., Tamminen, J., Veefkind, J. P., Bhartia, P. K., Stein Zweers, D. C., Duncan, B. N., Streets,
981 D. G., Eskes, H., van der A, R., McLinden, C., Fioletov, V., Carn, S., de Laat, J., DeLand, M., Marchenko,
982 S., McPeters, R., Ziemke, J., Fu, D., ...Wargan, K. (2018). The Ozone Monitoring Instrument: overview of
983 14 years in space. *Atmos. Chem. Phys.*, 18(8), 5699-5745. <https://doi.org/10.5194/acp-18-5699-2018>
- 984 Levelt, P. F., Oord, G. H. J. v. d., Dobber, M. R., Malkki, A., Huib, V., Johan de, V., Stammes, P., Lundell, J. O. V.,
985 & Saari, H. (2006). The ozone monitoring instrument. *IEEE Transactions on Geoscience and Remote*
986 *Sensing*, 44(5), 1093-1101. <https://doi.org/10.1109/TGRS.2006.872333>
- 987 Li, M., Wu, Y., Bao, Y., Liu, B., & Petropoulos, G. P. (2022). Near-Surface NO₂ Concentration Estimation by
988 Random Forest Modeling and Sentinel-5P and Ancillary Data. *Remote Sensing*, 14(15), 3612.
989 <https://www.mdpi.com/2072-4292/14/15/3612>
- 990 Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions* Proceedings of the
991 31st International Conference on Neural Information Processing Systems, Long Beach, California,
992 USA.
- 993 Mainka, A., & Žak, M. (2022). Synergistic or Antagonistic Health Effects of Long- and Short-Term Exposure to
994 Ambient NO₂ and PM_{2.5}: A Review. *International Journal of Environmental Research and Public*
995 *Health*, 19(21), 14079. <https://www.mdpi.com/1660-4601/19/21/14079>
- 996 Mijling, B., & van der A, R. J. (2012). Using daily satellite observations to estimate emissions of short-lived air
997 pollutants on a mesoscopic scale. *Journal of Geophysical Research: Atmospheres*, 117(D17).
998 <https://doi.org/10.1029/2012JD017817>
- 999 Montgomery, A., Schnell, J. L., Adelman, Z., Janssen, M., & Horton, D. E. (2023). Simulation of Neighborhood-
1000 Scale Air Quality With Two-Way Coupled WRF-CMAQ Over Southern Lake Michigan-Chicago Region.
1001 *Journal of Geophysical Research: Atmospheres*, 128(6), e2022JD037942.
1002 <https://doi.org/10.1029/2022JD037942>



- 1003 Mucha, W., & Mainka, A. (2026). Exposure to Nitrogen Dioxide (NO₂) Emitted from Traffic-Related Sources:
1004 Review. *Applied Sciences*, 16, 859. <https://doi.org/10.3390/app16020859>
- 1005 Nawaz, M. O., Goldberg, D. L., Kerr, G. H., & Anenberg, S. C. (2025). TROPOMI Satellite Data Reshape NO₂ Air
1006 Pollution Land-Use Regression Modeling Capabilities in the United States. *ACS ES&T Air*, 2(2), 187-
1007 200. <https://doi.org/10.1021/acsestair.4c00153>
- 1008 Nawaz, M. O., Huber, D. E., Kerr, G. H., Judd, L. M., Acker, S. J., & Goldberg, D. L. (2025). A Comparative Analysis
1009 of TEMPO NO₂ Remote Sensing With Surface-Level Monitoring Through Diurnal and Seasonal Trends,
1010 Meteorology, and Monitor Characteristics. *Journal of Geophysical Research: Atmospheres*, 130(20),
1011 e2025JD043923. <https://doi.org/https://doi.org/10.1029/2025JD043923>
- 1012 Nowlan, C. R., González Abad, G., Liu, X., Wang, H., & Chance, K. (2025). *TEMPO Nitrogen Dioxide Retrieval*
1013 *Algorithm Theoretical Basis Document*. NASA Algorithm Publication Tool.
1014 <https://doi.org/10.5067/WX026254FI2U>
- 1015 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,
1016 Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M.,
1017 Duchesnay, E., & Louppe, G. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine*
1018 *Learning Research*, 12.
- 1019 Peinado, L. B., Guarda, T., Herrera-Vidal, G., Minnaard, C., & Coronado-Hernández, J. R. (2025). Statistical and
1020 Machine Learning Models for Air Quality: A Systematic Review of Methods and Challenges. *Algorithms*,
1021 18(12), 783. <https://www.mdpi.com/1999-4893/18/12/783>
- 1022 Román-Cascón, C., Yagüe, C., Ortiz-Corral, P., Serrano, E., Sánchez, B., Sastre, M., Maqueda, G., Alonso-
1023 Blanco, E., Artiñano, B., Gómez-Moreno, F. J., Díaz-Ramiro, E., Fernández, J., Martilli, A., García, A. M.,
1024 Núñez, A., Cordero, J. M., Narros, A., & Borge, R. (2023). Wind and turbulence relationship with NO₂
1025 in an urban environment: a fine-scale observational analysis. *Urban Climate*, 51, 101663.
1026 <https://doi.org/https://doi.org/10.1016/j.uclim.2023.101663>
- 1027 Salman, A. K., Choi, Y., Park, J., Mousavinezhad, S., Payami, M., Momeni, M., & Ghahremanloo, M. (2024). Deep
1028 learning based emulator for simulating CMAQ surface NO₂ levels over the CONUS. *Atmospheric*
1029 *Environment*, 316, 120192. <https://doi.org/https://doi.org/10.1016/j.atmosenv.2023.120192>
- 1030 Seinfeld, J. H., & Pandis, S. N. (2006). *Atmospheric Chemistry and Physics: From Air Pollution to Climate*
1031 *Change* (2nd ed.). John Wiley & Sons.
- 1032 Shukla, N., Kwok, A., Topham, A., Gallier, S., Sapey, E., Percy, A., Bayliss, M., Nagakumar, P., Hansell, A. L.,
1033 Bartington, S. E., & Lucas, T. C. (2026). Does hourly variation in ambient nitrogen dioxide increase the
1034 risk of acute asthma hospitalisations? A time-series study of Birmingham and Solihull, UK. *BMJ Open*,
1035 16(8), e110612. <https://doi.org/10.1136/bmjopen-2025-110612>
- 1036 Smithsonian Astrophysical Observatory. (2026). TEMPO Daily Log during Nominal Operation (10/17/2023–
1037 present). https://github.com/Smithsonian/TEMPO-Observations-log/blob/main/daily_log.md
- 1038 Tang, D., Zhan, Y., & Yang, F. (2024). A review of machine learning for modeling air quality: Overlooked but
1039 important issues. *Atmospheric Research*, 300, 107261.
1040 <https://doi.org/https://doi.org/10.1016/j.atmosres.2024.107261>
- 1041 TIGER/Line Shapefiles ((2025). [https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-](https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html)
1042 [line-file.html](https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html)
- 1043 U.S. Environmental Protection Agency. (2011, February 28). AQS Addendum to Negative values.
1044 https://19january2017snapshot.epa.gov/aqs/aqs-addendum-negative-values_.html
- 1045 U.S. Environmental Protection Agency. (2024). Overview of Nitrogen Dioxide (NO₂) Air Quality in the United
1046 States. https://www.epa.gov/system/files/documents/2024-08/no2_2023.pdf
- 1047 U.S. Environmental Protection Agency. (2025a). Air Quality System (AQS) hourly NO₂ data, 2023–2025,
1048 (https://aqs.epa.gov/aqswweb/airdata/download_files.html
- 1049 U.S. Environmental Protection Agency. (2025b, June 13). Health Effects of Ozone and Particulate Matter.
1050 <https://www.epa.gov/advance/health-effects-ozone-and-particulate-matter>
- 1051 U.S. Environmental Protection Agency. (2025c). NAAQS Table. Retrieved 2/15 from
1052 <https://www.epa.gov/criteria-air-pollutants/naaqs-table>
- 1053 U.S. Geological Survey. (2025). Annual National Land Cover Database (NLCD) Collection 1 Fractional
1054 Impervious Surface Conterminous United States (<https://doi.org/10.5066/P94UXNTS>



- 1055 Veeffkind, J. P., Aben, I., McMullan, K., Förster, H., de Vries, J., Otter, G., Claas, J., Eskes, H. J., de Haan, J. F.,
1056 Kleipool, Q., van Weele, M., Hasekamp, O., Hoogeveen, R., Landgraf, J., Snel, R., Tol, P., Ingmann, P.,
1057 Voors, R., Kruizinga, B.,...Levelt, P. F. (2012). TROPOMI on the ESA Sentinel-5 Precursor: A GMES
1058 mission for global observations of the atmospheric composition for climate, air quality and ozone
1059 layer applications. *Remote Sensing of Environment*, 120, 70-83.
1060 <https://doi.org/https://doi.org/10.1016/j.rse.2011.09.027>
- 1061 Wang, M., Li, H., Huang, S., Qian, Y., Steenland, K., Xie, Y., Papatheodorou, S., & Shi, L. (2021). Short-term
1062 exposure to nitrogen dioxide and mortality: A systematic review and meta-analysis. *Environ Res*, 202,
1063 111766. <https://doi.org/10.1016/j.envres.2021.111766>
- 1064 Wei, Q., Song, W., Dai, B., Wu, H., Zuo, X., Wang, J., Chen, J., Li, J., Li, S., & Chen, Z. (2025). Spatiotemporal
1065 estimation of surface NO₂ concentrations in the Pearl River Delta region based on TROPOMI data and
1066 machine learning. *Atmospheric Pollution Research*, 16(3), 102353.
1067 <https://doi.org/https://doi.org/10.1016/j.apr.2024.102353>
- 1068 Xing, J., Zheng, S., Ding, D., Kelly, J. T., Wang, S., Li, S., Qin, T., Ma, M., Dong, Z., Jang, C., Zhu, Y., Zheng, H.,
1069 Ren, L., Liu, T.-Y., & Hao, J. (2020). Deep Learning for Prediction of the Air Quality Response to
1070 Emission Changes. *Environmental Science & Technology*, 54(14), 8589-8600.
1071 <https://doi.org/10.1021/acs.est.0c02923>
- 1072 Zoogman, P., Liu, X., Suleiman, R. M., Pennington, W. F., Flittner, D. E., Al-Saadi, J. A., Hilton, B. B., Nicks, D.
1073 K., Newchurch, M. J., Carr, J. L., Janz, S. J., Andraschko, M. R., Arola, A., Baker, B. D., Canova, B. P.,
1074 Chan Miller, C., Cohen, R. C., Davis, J. E., Dussault, M. E.,...Chance, K. (2017). Tropospheric
1075 Emissions: Monitoring of Pollution (TEMPO). *J Quant Spectrosc Radiat Transf*, 186, 17-39.
1076 <https://doi.org/10.1016/j.jqsrt.2016.05.008>
1077