





RESEARCH ARTICLE

10.1029/2023JD040676

Inferring Surface NO₂ Over Western Europe: A Machine Learning Approach With Uncertainty Quantification

Wenfu Sun^{1,2} , Frederik Tack¹, Lieven Clarisse², Rochelle Schneider³, Trissevgeni Stavrakou¹, and Michel Van Roozendael¹ 

¹Royal Belgian Institute for Space Aeronomy (BIRA-IASB), Brussels, Belgium, ²Spectroscopy, Quantum Chemistry and Atmospheric Remote Sensing (SQUARES), Université Libre de Bruxelles (ULB), Brussels, Belgium, ³Φ-Lab, European Space Agency (ESA), Frascati, Italy

Key Points:

- A novel and reliable machine learning model with uncertainty quantification is applied to infer surface NO₂ levels over Western Europe
- Our work uncovers how predictors impact model inference of various surface NO₂ levels differently
- Our approach identifies areas of high uncertainty in surface NO₂ mapping and potential environmental risks from overlooking uncertainty

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

W. Sun,
wenfu.sun@aeronomie.be

Citation:

Sun, W., Tack, F., Clarisse, L., Schneider, R., Stavrakou, T., & Van Roozendael, M. (2024). Inferring surface NO₂ over Western Europe: A machine learning approach with uncertainty quantification. *Journal of Geophysical Research: Atmospheres*, 129, e2023JD040676. <https://doi.org/10.1029/2023JD040676>

Received 24 DEC 2023

Accepted 30 SEP 2024

Author Contributions:

Conceptualization: Wenfu Sun, Frederik Tack, Lieven Clarisse, Rochelle Schneider, Michel Van Roozendael

Data curation: Wenfu Sun

Formal analysis: Wenfu Sun, Frederik Tack, Lieven Clarisse, Rochelle Schneider, Trissevgeni Stavrakou, Michel Van Roozendael

Funding acquisition: Frederik Tack, Michel Van Roozendael

Investigation: Wenfu Sun, Frederik Tack, Lieven Clarisse, Rochelle Schneider, Trissevgeni Stavrakou, Michel Van Roozendael

© 2024 The Author(s).

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial License](https://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

Abstract Nitrogen oxides (NO_x = NO + NO₂) are of great concern due to their impact on human health and the environment. In recent years, machine learning (ML) techniques have been widely used for surface NO₂ estimation with rapid developments in computational power and big data. However, the uncertainties inherent to such retrievals are rarely studied. In this study, a novel ML framework has been developed, enhanced with uncertainty quantification techniques, to estimate surface NO₂ and provide corresponding data-induced uncertainty. We apply the Boosting Ensemble Conformal Quantile Estimator (BEnCQE) model to infer surface NO₂ concentrations over Western Europe at the daily scale and 1 km spatial resolution from May 2018 to December 2021. High NO₂ mainly appears in urban areas, industrial areas, and roads. The space-based cross-validation shows that our model achieves accurate point estimates ($r = 0.8$, $R^2 = 0.64$, root mean square error = 8.08 μg/m³) and reliable prediction intervals (coverage probability, PI-50%: 51.0%, PI-90%: 90.5%). Also, the model result agrees with the Copernicus Atmosphere Monitoring Service (CAMS) model. The quantile regression in our model enables us to understand the importance of predictors for different NO₂ level estimations. Additionally, the uncertainty information reveals the extra potential exceedance of the World Health Organization (WHO) 2021 limit in some locations, which is undetectable by only point estimates. Meanwhile, the uncertainty quantification allows assessment of the model's robustness outside existing in-situ station measurements. It reveals challenges of NO₂ estimation over urban and mountainous areas where NO₂ is highly variable and heterogeneously distributed.

Plain Language Summary Inferring surface NO₂ concentrations is an effective way to monitor and mitigate NO_x pollution which is of great concern due to its impact on human health and the environment. Machine learning (ML) techniques have been widely used for surface NO₂ estimation with rapid developments in computational power and big data. However, such estimations can be uncertain due to inherent errors in the data, and this uncertainty is rarely studied. We develop a novel ML framework to estimate surface NO₂ concentrations and provide corresponding uncertainty information. We infer surface NO₂ levels over Western Europe at the daily scale and 1 km spatial resolution from May 2018 to December 2021. Our model's performance is reliable as verified by in-situ station measurements and an independent physics-based model. We observe NO₂ hotspots over urban areas, industrial areas, and major roads. The uncertainty quantification (UQ) techniques allow us to analyze the influence of different input data on estimating different NO₂ levels. The UQ also helps to identify potential NO₂ exceedances of the WHO 2021 limit, which have not been observed in previous research. Additionally, we assess the model's robustness outside of in-situ stations and witness the challenge of NO₂ estimation over urban and mountainous areas.

1. Introduction

Nitrogen oxides (NO_x = NO + NO₂) have received considerable attention due to their adverse effects on human health and air quality. Epidemiological studies have shown that NO_x can lead to allergic reactions, respiratory symptoms, and asthma (Achakulwisut et al., 2019; F. Zhu et al., 2019). In addition, NO_x plays a key role in atmospheric chemistry and affects the environment. For example, high concentrations of NO_x can contribute to acid rain, surface ozone, and PM_{2.5} (Grennfelt et al., 2020; Hodan & Barnard, 2004; Ren et al., 2022). Also, it then can contribute to the deposition of nitrogen which leads to other environmental issues such as eutrophication and biodiversity loss (Stevens et al., 2018). Surface NO_x pollution is primarily caused by anthropogenic activities at the ground level (e.g., transport, energy production, industrial, residential, biomass burning, and agricultural

Methodology: Wenfu Sun, Frederik Tack, Rochelle Schneider

Project administration: Michel Van Roozendael

Resources: Wenfu Sun, Frederik Tack, Rochelle Schneider, Michel Van Roozendael

Supervision: Frederik Tack, Lieven Clarisse, Michel Van Roozendael

Validation: Wenfu Sun, Frederik Tack, Lieven Clarisse, Rochelle Schneider, Trissevgeni Stavrakou, Michel Van Roozendael

Visualization: Wenfu Sun

Writing – original draft: Wenfu Sun

Writing – review & editing: Frederik Tack, Lieven Clarisse, Rochelle Schneider, Trissevgeni Stavrakou, Michel Van Roozendael

activities), with relatively small contributions from natural sources (e.g., soil and wildfires) (Crippa et al., 2018; Oikawa et al., 2015; Song et al., 2021). NO₂ is recognized as a robust indicator of NO_x, and therefore assessing surface NO₂ concentrations and understanding its spatiotemporal pattern is an effective way of monitoring NO_x pollution and supporting the development of mitigation strategies.

NO₂ is routinely monitored from air quality measurement stations operated by environmental agencies (Guerreiro et al., 2014; Kong et al., 2021), providing accurate and long-term records of (near-)surface concentrations. However, the stations' spatial distribution is sparse and uneven, resulting in limited coverage. Spaceborne observations (Douros et al., 2023; Lamsal et al., 2021; van Geffen et al., 2020; Veefkind et al., 2012) provide daily global coverage of NO₂ observations at relatively coarse resolution (up to 5 × 5 km²). As satellite sounders measure integrated columns, rather than surface concentrations, extra methods and assumptions are required to infer surface NO₂ concentrations from such measurements (Cooper et al., 2022; Lamsal et al., 2008). Chemical transport models (CTMs) can incorporate knowledge of physical and chemical processes to estimate surface NO₂ concentrations, and the estimation can be constrained by assimilating real observations (Inness et al., 2019; Kumar et al., 2012; Poraicu et al., 2023; Vira & Sofiev, 2015). However, CTM models are typically affected by uncertainties from mechanisms (e.g., simplification of some processes) and data (e.g., uncertainty in the emission data). Also, they require considerable computational resources, leading to a compromise between resolution and efficiency. As an alternative, statistical and machine learning (ML) models offer a data-driven approach that directly connects influential factors with surface NO₂ levels, enabling high-resolution mapping of the latter. Nonetheless, the traditional statistical models typically struggle to adequately address complex nonlinear or higher-order interactive relationships (Fan et al., 2019).

ML techniques offer a promising option to address the challenges mentioned above, especially in light of the rapid development of computational power and the increased availability of Earth system data (Reichstein et al., 2019). At present, different studies have demonstrated the predictive power of ML on surface NO₂ concentration at high spatiotemporal resolution (e.g., 100 m and hourly (Kim et al., 2021), 1 km and daily (Ghahremanloo et al., 2023; Wei et al., 2022)). Most studies employ supervised learning with in-situ data to train ML models of the relationship between surface NO₂ and multiple predictors, before upscaling the site-level surface NO₂ concentrations to larger spatial areas where measurements are not available. The predictors encompass different variables from observations, modeled data (e.g., meteorology), and inventories. While the ML model does not explicitly encode the underlying chemical and physical processes, it is capable of establishing a complex mapping relationship from these predictor variables to the surface NO₂ concentrations. Although that relationship does not strictly adhere to the underlying physics, the construction of a robust model still necessitates a comprehensive understanding of the chemical and physical properties of NO₂ by the model builder. This knowledge enables the identification of optimal predictors and the enhancement of model generalization performance. For instance, given that NO₂ is highly reactive, solar radiation is typically employed as an indicator for photochemistry (Balamurugan et al., 2023; Di et al., 2020), which is also considered in this study. Plus, other studies use surface ozone as a predictor for surface NO₂ prediction (L. Li & Wu, 2021) or link surface ozone prediction to surface NO₂ prediction in the multi-task ML model (Yang et al., 2023). It is essential to note that the success of ML models does not imply the replacement of more established methodologies, which often generate predictors for ML models and provide full coverage data.

Currently, the popular ML algorithms for surface NO₂ estimation can be divided into three classes. These are the tree-based model, the neural-network-based model, and the ensemble models of the former two. The tree-based models mainly use random forest (de Hoogh et al., 2019; M. Li et al., 2022; Pan et al., 2021; Qin et al., 2020), extremely randomized trees and deep forest (Wei et al., 2022), and gradient boosting decision trees (Balamurugan et al., 2023; Chi et al., 2022; Kang et al., 2021; Kim et al., 2021; Liu & Chen, 2022; Wang et al., 2021). The neural-network-based models either use the normal neural networks (Chan et al., 2021) or the deep learning models (Ghahremanloo et al., 2021, 2023; L. Li & Wu, 2021; Scheibenreif et al., 2022; Zhang et al., 2022). Ensemble models (Di et al., 2020; He et al., 2022) assemble different types of ML models mentioned above to make the predictions. No matter what approach is chosen to infer surface NO₂ concentrations, errors are inevitable. In ML models, these are related to inherent randomness and errors in the data applied, out-of-regime errors due to data limitations, and errors in the model itself (Haynes et al., 2023; Kiureghian & Ditlevsen, 2009). To properly exploit NO₂ estimates, it is essential from a user's perspective to indicate the expected uncertainty associated with each prediction. As part of the model training, ML models typically go through cross-validation (CV) to assess their predictive skills. While this informs on the statistical performance of the model, it is not a

substitute for a single prediction uncertainty estimate. Some studies used the ensemble model or Monte Carlo dropout method (Di et al., 2020; L. Li & Wu, 2021; Scheibenreif et al., 2022), but the derived uncertainty mainly relates to the out-of-regime and model errors (Haynes et al., 2023) where the uncertainty information is derived from the variance between predictions from sub-models. Currently, research on Uncertainty Quantification (UQ) is still lacking, particularly regarding errors related to the input data (e.g., variability and representativeness). Note that these data-induced uncertainties are generally independent of the specific model that is used, but are often expected to dominate the error budget, especially for the purely data-driven ML models.

This study develops an ML framework enhanced with UQ techniques to estimate surface NO₂, using a quantile regression strategy to address uncertainties arising from the data. We select the eXtreme Gradient Boosting (XGBoost, v2.0.0) model (Chen & Guestrin, 2016), which is a powerful and efficient gradient boosting decision tree model, as the core ML model of this framework, though other models are also applicable. Instead of only making point estimates (i.e., giving an expected surface NO₂ concentration) by minimizing the mean squared error (MSE) loss function (or cost function), we quantify the uncertainty by minimizing the quantile loss function (i.e., constructing quantile regression models (Koenker & Bassett, 1978)) during ML model training. This method can directly provide boundary values of prediction intervals (PIs) of the target variable, and it does not need assumptions on the distribution of the target variable, thereby flexibly adapting to a wide range of data distributions (e.g., non-normal, skewed distributions) (Haynes et al., 2023; Koenker, 2005; Takeuchi et al., 2006). To ensure the effectiveness of PI, we follow Romano et al. (2019) to incorporate conformal prediction to make the PI have a reliable coverage probability for true values. Meanwhile, to mitigate the randomness inherent in model predictions, we leverage the ensemble prediction strategy to improve result stability. Overall, this proposed UQ-enabled ML framework is composed of XGBoost, quantile regression, conformal prediction, and ensemble prediction, which we name BEnCQE (Boosting Ensemble Conformal Quantile Estimator).

The BEnCQE model is employed to infer surface NO₂ concentrations over Western Europe since this region is characterized by a dense population, extensive urbanization, and a thriving industry, which requires reliable air quality management. In this work, we infer the daily mean of surface NO₂ concentrations and the associated uncertainty from May 2018 to December 2021 at 1 km resolution. To ensure the effectiveness of the data applied, we use SHapley Additive exPlanations (SHAP) (Lundberg & Lee, 2017) to interpret how our model utilizes various predictors. Additionally, we compare our NO₂ estimates with independent data from the Copernicus Atmosphere Monitoring Service (CAMS) European air quality reanalysis, using it as an external validation for our model's accuracy and reliability.

We intend to address the following questions: (a) How can we estimate surface NO₂ levels at high spatiotemporal resolution over Western Europe, and specifically quantify and integrate the uncertainties arising from the data into these estimations? (b) How does incorporating UQ enhance our understanding of the dynamics of surface NO₂ in Western Europe, which is likely overlooked in previous modeling approaches? (c) How can insights gained from UQ guide future developments in the model, and what implications does this have for practical applications in air quality monitoring and regulation?

The remainder of this paper is organized as follows. The study region and different data sets are introduced in Section 2. The BEnCQE model setup and associated methodology are discussed in Section 3. The main results are presented in Section 4 and discussed in Section 5. The conclusions and outlook are stated in Section 6.

2. Study Area and Data Set

This study focuses on part of Western Europe from 5°W to 9°E and 42°N to 54°N, including the Netherlands, Belgium, Luxembourg, France, and Western Germany (Figure 1). Western Europe is a critical region for surface NO₂ research as it features high levels of urbanization and industrialization and notable air pollution challenges. The domain includes important megacities such as Paris, Brussels, Amsterdam, and Cologne. It also encompasses coastal cities like Rotterdam and Antwerp whose ports are among the largest petrochemical clusters in the world (Van den Berghe et al., 2023), and strongly industrialized areas such as the Ruhr and the Alsace-Lorraine. In addition to NO₂ hotspots, the domain includes large rural areas and diverse topography (e.g., low-lying flat areas, plains, plateaus, river valleys, and mountains), leading to a highly heterogeneous NO₂ distribution.

To effectively infer surface NO₂ concentrations over Western Europe, we use measured surface NO₂ concentrations from European Environmental Agency (EEA) air quality stations (orange dots in Figure 1) as the target

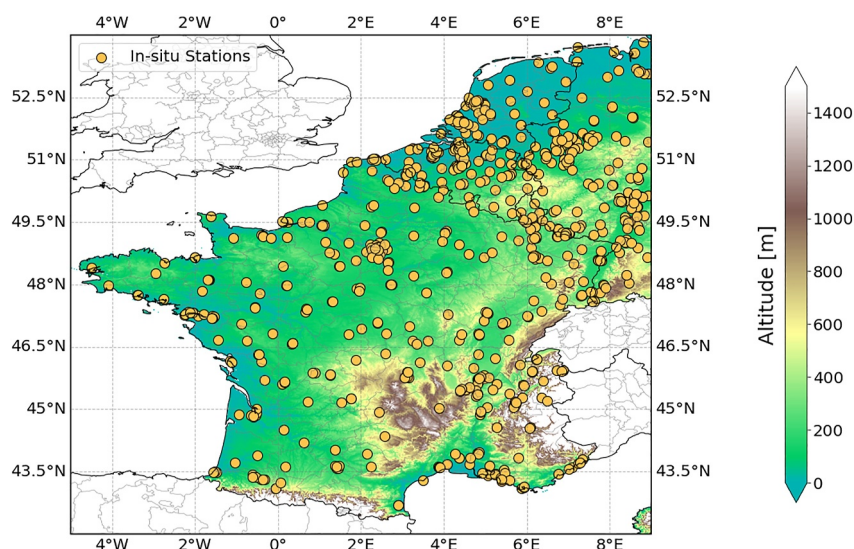


Figure 1. Topography and EEA station distribution over the study area. The study area is the non-masked region of Western Europe, covering the Netherlands, Belgium, Luxembourg, France, and Western Germany. Geographically, the area extends from 5°W to 9°E and from 42°N to 54°N. The orange dots represent the locations of the EEA air quality stations within the domain.

and multiple source data sets as predictors to construct the ML model. All data sets are reprojected on a 1 km resolution grid with bilinear interpolation or averaging approach and converted to the daily scale. The target variable and the predictors are paired within the same grid. Detailed information on the used predictors is provided in Table 1. In this study, selected predictors are comprised of satellite observations, meteorological factors, emissions, land information, spatially lagged (SL) data, and time information. We analyze the predictors' importance (see Section 3.3) and keep only those predictors in the model with importance greater than 1% (Text S1 in Supporting Information S1). The diversity of the data sets ensures that our model captures the variance in surface NO_2 as much as possible. After data preparation, we have 871,007 samples in total, collected from different stations over the study period (i.e., from May 2018 to December 2021), with each sample (daily and 1 km) consisting of one target value and 29 predictor values. More information about the data sets is provided in the following sections.

2.1. Surface NO_2 Measurements

Surface NO_2 measurements from the official air quality networks are taken as the target for the model. The data set is acquired between May 2018 and December 2021 from the EEA air quality database (European Environment Agency, 2022). A total of 737 stations within the study area (Figure 1) are reprojected on 1 km resolution grids, and data are averaged for the grid cells containing multiple stations. Consequently, 727 grids contain NO_2 measurements categorized into industrial (95), background (452), and traffic (180, 3 of the grids contain both traffic and background stations but are taken as traffic grids) groups based on the EEA classification (European Environment Agency, 2023a). Data set values are 24-hr averages calculated from hourly measurements from local time 00:00 to 23:59.

2.2. NO_2 TVCDs and Ancillary Data

Tropospheric vertical column densities (TVCDs) of NO_2 are retrieved from satellite observations and are of great relevance for ML-based surface NO_2 prediction (Kang et al., 2021; Kim et al., 2021; Wei et al., 2022). In this work, we use three types of NO_2 TVCDs:

1. Daily TROPospheric Monitoring Instrument (TROPOMI) NO_2 TVCDs based on CAMS a-priori profiles ($7 \times 3.5 \text{ km}^2$, $5.5 \times 3.5 \text{ km}^2$ since August 2019, PAL + OFFL v2.3.1). The daily TROPOMI NO_2 TVCDs provide information on the dynamics of NO_2 columns. This data is obtained from Douros et al. (2023) which differs from the operational TROPOMI NO_2 product (van Geffen et al., 2020), where the original TM5-MP

Table 1

Summary of Predictors Used for Model Estimation

Group	Predictors	Unit	Spatial resolution	Temporal resolution	Data source
NO ₂ TVCD	TROPOMI TVCD long-term	molecules/cm ²	1 km	–	Operational TROPOMI NO ₂ product with TMS- MP a-priori profile (van Geffen et al., 2020) (PAL + OFFL v2.3.1)
	CAMS TVCD	–	10 km	Daily	Modified TROPOMI NO ₂ product for Europe with regional CAMS a-priori profiles (Douras et al., 2023) (PAL + OFFL v2.3.1)
	TROPOMI TVCD daily	–	7 × 3.5 km ² , 5.5 × 3.5 km ² after August 2019	–	–
	TROPOMI TVCD precision ^a	–	–	–	–
	TROPOMI QA ^a	–	–	–	–
	Cloud fraction	–	–	–	–
Meteorology	Boundary layer height	m	25 km	Hourly	ERA5 (Hersbach et al., 2020)
	Rain	m of water equivalent	9 km	Hourly	ERA5-land (Muñoz-Sabater et al., 2021)
	Evaporation	m of water equivalent	–	–	–
	Surface pressure	pa	–	–	–
	Solar radiation	J/m ²	–	–	–
	Dewpoint Temperature	K	–	–	–
	Temperature	–	–	–	–
	V wind	m/s	–	–	–
	U wind	m/s	–	–	–
	Road-highways	m/km ²	1 km	–	Global Roads Inventory Project (GRIP) global roads database (Meijer et al., 2018)
Emissions	Road-primary	–	–	–	–
	Road-secondary	–	–	–	–
	Road-tertiary	–	–	–	–
	Road-local	–	–	–	–
	Night light	nW/cm ² /sr	500 m	Yearly	VIIRS (Elvidge et al., 2021)
	Population	Number of people/km ²	1 km	–	JRC-GEOSTAT 2018 (Silva et al., 2021)
	Emission inventory	Kt/year/(100 km ²)	10 km	–	EEA National Gridded Data of Emissions by Source Category (GNFR) (European Environment Agency, 2023b)
	Wildfires ^a	Kg/m ² s	10 km	Daily	CAMS Global Fire Assimilation System (GFAS)
	Leaf area index-low ^a	m ² /m ²	9 km	Hourly	ERA5-land
	Land cover	–	100 m	–	CORINE Land Cover (CLC) 2018 inventory (European Environment Agency, 2020)
Land	Elevation	m	90 m	–	Multi-Error-Removed Improved-Terrain digital elevation models (MERIT DEM) (Yamazaki et al., 2017)

Table 1
Continued

Group	Predictors	Unit	Spatial resolution	Temporal resolution	Data source
Spatially lagged (SL) data	SL-TRA-local	ug/m ³	1 km	-	Processed from European Environmental Agency (EEA) air quality database (European Environment Agency, 2022)
	SL-TRA-regional				
	SL-BKGINd-local				
	SL-BKGINd-regional				
Time	Day of year	-	-	-	
	Day of week				

^aPredictors with the importance of less than 1% for the BEnCQE model are not considered in the final model (Text S1 in Supporting Information S1).

(100 km) a-priori profile is replaced by the regional CAMS (10 km) profile for the European domain. This product significantly improves the satellite sensitivity to NO₂ hotspots (Tack et al., 2021). In this study, we did not filter the daily TROPOMI NO₂ TVCDs by quality assurance (QA) to avoid severe data loss (around 50% loss when QA > 0.75 is applied, as shown in Figure S4c in Supporting Information S1), and avoid subsequent gap-filling work. Meanwhile, we use the cloud fraction as ancillary data to indicate the data quality. The impact of this data processing on model estimation performance and uncertainty is discussed in Section 5.1. Initially, the QA flag and TROPOMI precision data were included as predictors, but analysis showed that these did not meaningfully impact the results (Figure S2 in Supporting Information S1) and so these have been excluded in the final model.

2. Long-term oversampled TROPOMI NO₂ TVCDs (1 km). This is obtained by oversampling the operational daily TROPOMI NO₂ TVCDs (TM5-MP profile, PAL + OFFL v2.3.1) Level 2 pixels with QA greater than 0.75 over 1 km scale grids from May 2018 to December 2021. This aims at providing an observed average NO₂ pattern and supplement information on the NO₂ emissions. It is more straightforward to implement oversampling of the official operational product than oversampling the CAMS profile-equipped TROMOPI data, given that the latter requires substantial pre-processing even though its profile has a higher resolution.
3. Simulated daily NO₂ TVCDs from regional CAMS (10 km). This data is used to provide the model with noise-free NO₂ columns.

2.3. Meteorology

Meteorological predictors are derived from the European Centre for Medium-Range Weather Forecast (ECMWF) ERA5 (Hersbach et al., 2020, 25 km) and ERA5-land ((Muñoz-Sabater et al., 2021), 9 km) reanalysis data sets. These include wind (*u*-component, *v*-component at 10 m elevation), surface temperature, dewpoint temperature, total precipitation, surface net solar radiation, evaporation amount, and boundary layer height (BLH). All meteorological variable data sets are 24-hr averages representing the daily average meteorology. This is done in order to assist the model in capturing the daily average NO₂ level. Total precipitation, surface net solar radiation, and evaporation are accumulated variables, while the others are instantaneous variables.

2.4. Emission Data Set

The anthropogenic NO₂ emission inventory is obtained from the EEA National Gridded Data of Emissions by Source Category (European Environment Agency, 2023b, 10 km), which represents the period of 2019 and aggregates various emission sectors such as power plants, industry, road transport, and livestock. In addition, we use the 1km-scale road density data set (meters of road per grid unit) through rasterizing road vectors from the Global Roads Inventory Project (GRIP) global road database of different road types (i.e., highways, primary roads, secondary roads, tertiary roads, and local roads) (Meijer et al., 2018). Also, we use the population density from the JRC-GEOSTAT 2018 gridded population (Silva et al., 2021, 1 km) and the night lights (500 m) from the annual global Visible Infrared Imaging Radiometer Suite (VIIRS) nighttime lights data set (Elvidge et al., 2021), as other indicators of anthropogenic activity. Although NO₂ can also be emitted from wildfires (Wan et al., 2023), the wildfire data from the CAMS Global Fire Assimilation System offers almost no contribution to the model estimates (Figure S2 in Supporting Information S1) and is therefore excluded.

2.5. Land Information

To introduce land information in our model, we use the elevation data from Multi-Error-Removed Improved-Terrain digital elevation models (MERIT DEM) (Yamazaki et al., 2017, 90 m) and land cover data from CORINE (Coordination of Information on the Environment) Land Cover (CLC) 2018 inventory (European Environment Agency, 2020, 100 m). The land cover is reaggregated from 44 classes into five main classes: anthropogenic surfaces, agricultural areas, forest and semi-natural areas, wetlands, and water bodies. While we initially used the leaf area index of low and high vegetation types from ERA5-land to represent the land vegetation cover, the data showed little importance to the model (Figure S2 in Supporting Information S1), and so this data was not retained in the end.

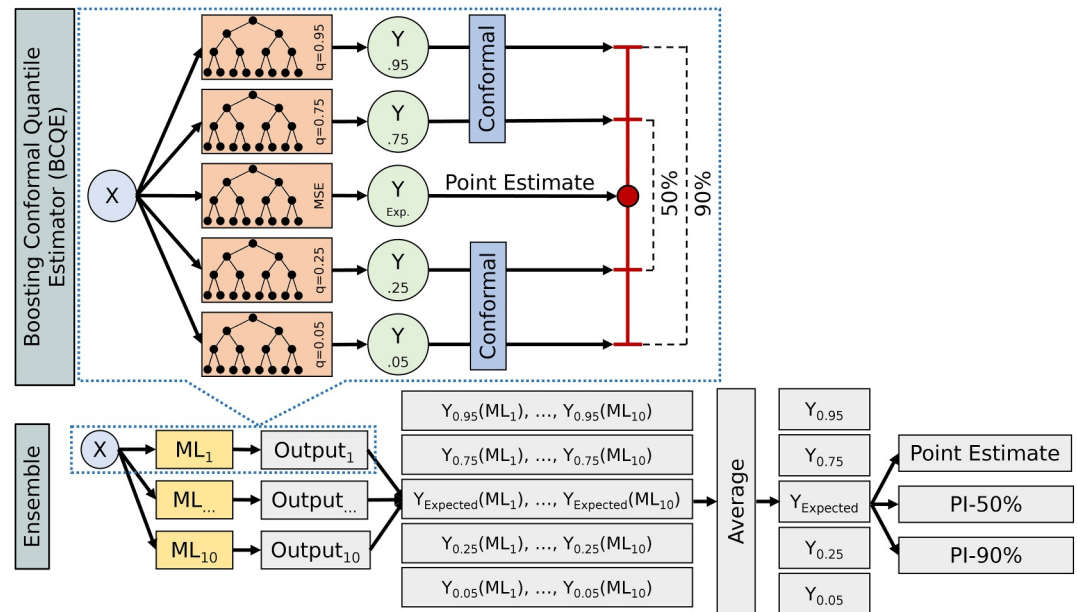


Figure 2. Schematic structure of BEnCQE for surface NO_2 estimation. Each member model (i.e., BCQE) of the BEnCQE model consists of five XGBoost models with different objectives (i.e., point estimate and quantiles). The XGBoost model with the MSE loss function gives an expected NO_2 value, and other XGBoost models with different quantile objectives (i.e., $q = 0.05, 0.25, 0.75,$ and 0.95) provide quantile estimates. These quantile outputs are then conformalized to provide PI-50% (by quantiles of 0.25 and 0.75) and PI-90% (by quantiles of 0.05 and 0.95). The BEnCQE model assembles 10 BCQE models, and the final output for each objective is the average of the outputs from all the sub-models.

2.6. Spatially Lagged NO_2 Concentrations

We follow Schneider et al. (2020) to generate four variables from in-situ surface NO_2 measurements to introduce SL NO_2 averaged concentrations. The stations are categorized by EEA classification (European Environment Agency, 2023a) into BKGIND (background-industrial) to represent general NO_2 levels and TRA (traffic) to represent high NO_2 levels. For each grid with measurement, we average the measurements over the study period. We then use the inverse-distance weighted leave-one-out (IDW-LOO) approach with two different weights to calculate the regional NO_2 level (weight = 1, inverse distance, more weight to distant stations) and local NO_2 level (weight = 2, inverse squared distance, more weight to near stations) for every grid in the study region. Thus, four new variables are spatially-lagged-BKGIND-regional, spatially-lagged-BKGIND-local, spatially-lagged-TRA-regional, and spatially-lagged-TRA-local. The respective distributions are shown in Figure S3 in Supporting Information S1. These variables can assist the model in capturing the heterogeneity across in-situ stations and leverage the spatial autocorrelation to capture the NO_2 distribution from real measurements (Di et al., 2020; Schneider et al., 2020). We have applied a data mask strategy to generate the different SL data for CV training and testing, which avoids the target leakage during model evaluation.

2.7. Time Information

Day of week and day of year are used to indicate the weekly and seasonal cycle of surface NO_2 .

3. Methodology

In this study, we develop a novel BEnCQE (Boosting Ensemble Conformal Quantile Estimator) model (Figure 2) to infer surface NO_2 concentrations and associated uncertainties. The BEnCQE model is an ensemble of 10 Boosting Conformal Quantile Estimators (BCQE) (blue dotted frame in Figure 2), where each BCQE consists of five XGBoost models (orange rectangles) with different objectives (green circles). The number selection of the ensemble is explained in Section 3.1. The XGBoost model with the MSE loss function provides the point estimate (i.e., expected NO_2 value) and other XGBoost models with different quantile objectives (i.e., 0.05, 0.25, 0.75, and 0.95) provide the corresponding PI-50% ($q = 0.25, 0.75$), and PI-90% ($q = 0.05, 0.95$). In addition, the four

quantile estimates are refined with conformal prediction (blue rectangles) to enhance the PI's coverage guarantee of the true values. The output of this ensemble model for each objective is the average of the outputs of all members for that objective (gray rectangles). Further details are given in the following sections.

3.1. Boosting Ensemble Conformal Quantile Estimator

The BCQE model consists of five ML models with different objectives. In this study, we used XGBoost, a gradient boosting decision tree model that incrementally combines weak decision tree learners to minimize the loss function and improve the overall model performance (Chen & Guestrin, 2016), as the core model in the estimator. XGBoost has a powerful predictive ability and has been widely used in air components estimation, including surface NO₂ estimation (Balamurugan et al., 2023; Chi et al., 2022; Kim et al., 2021). In the BCQE model, the XGBoost model trained with the MSE loss function provides the point estimate (i.e., expected value) of the surface NO₂ concentration. Likewise, the XGBoost models trained with different quantile loss functions behave as quantile regression models and estimate the quantiles of surface NO₂. The idea of the quantile regression, proposed by Koenker and Bassett (1978), is to provide a quantile value for the target variable, below which the actual value is expected to lie within a certain probability. For example, given input data and a quantile objective of 0.95, the quantile regression model outputs a value below which there is a 95% probability that the actual value will fall. The basic idea of quantile loss functions in ML model training is that the ML model is penalized more for overestimation and less for underestimation when the quantile objective is less than 0.5, and vice versa. To capture the uncertainty of surface NO₂ estimation, we trained four types of XGBoost models with different quantile objectives (i.e., 0.05, 0.25, 0.75, and 0.95) to generate PI-50% ($q = 0.25$ and $q = 0.75$) and PI-90% ($q = 0.05$ and $q = 0.95$), delineating the intervals within which the true value might reasonably be expected to fall (Haynes et al., 2023; Koenker, 2005; Koenker & Bassett, 1978; Takeuchi et al., 2006). This method can quantify the spread probability of surface NO₂. Quantile points of 0.25 and 0.75 are widely used to form the interquartile range which measures the spread of the data with a skewed distribution, while the quantile points of 0.05 and 0.95 provide a conservative estimate of the possible values range. This method does not need assumptions on the distribution of the target variable, thereby flexibly adapting to a wide range of data distributions (e.g., non-normal and skewed distributions).

To guarantee the coverage probability of the generated PIs, we followed Romano et al. (2019) to incorporate conformal prediction to refine the boundaries of PIs. The underlying principle of conformal prediction is that new inputs, which are less similar to the training data, should lead to less certain estimates. The conformal prediction yields conformal scores during the model training process to represent the dissimilarity between the training and new data set and uses this score to adjust the PI boundary (Figure 2, blue rectangles). The details and equations for constructing the PIs are described in Appendix A. In this way, we can obtain both the point estimate and the effective PIs.

To mitigate the disturbance of model randomness on model estimates, we applied an ensemble prediction strategy where multiple BCQE model outputs are averaged for each model objective (gray rectangles in Figure 2), as in the work of Jensen et al. (2022). We decided to assemble 10 BCQE models (blue dotted frame in Figure 2) based on model output robustness and computational efficiency (Text S2 and Table S1 in Supporting Information S1). For this purpose, we randomly divided the training data set into training and validation data sets, based on station sites, (i.e., 70% stations for training and 30% stations for validation) for each BCQE model training. Therefore, each model learns from different aspects of the total training data set and generates different structures. In this way, the BEnCQE model provides a reliable point estimate and associated uncertainty, with a reduced impact of the model randomness.

Theoretically, quantile regression can be trained for any choice of quantile points, and a cumulative density function can then be obtained through the interpolation of a set of consecutive estimated quantile points. Due to the considerable resource demand of the ensemble approach, and the limitations of computational power, we decided not to expand the scale of the BEnCQE model to encompass many quantile points. Consequently, only four quantile points and one point estimate were included.

3.2. Model Training, Optimization, and Evaluation

For this study, we first trained, optimized, and evaluated the model on the space-based 10-fold CV (cross-validation) data set before training the final model on the total data set. The CV data set was generated by dividing the

total data set randomly into 10 groups based on stations (detail is described in Text S3 in Supporting Information S1). Nine groups were used for training and optimization, leaving out one group for testing, which was used to examine the spatial generalization of the ML model. After preparing the CV data set, we optimized the XGBoost hyperparameters using the Optuna framework (Akiba et al., 2019) for the point estimate and each quantile objective, respectively. The optimization process was conducted based on the training part of the CV data set which was further divided into a training and validation data set. Every set of randomly generated hyperparameters was first used for model training and then evaluated on the validation data set. The optimization process was replicated over 10 different CV training data sets and the score (root mean square error (RMSE) for point estimates and quantile loss for quantile objectives) for the hyperparameter set was the average of 10 validation results. The optimization was iterated 100 times, and we chose the hyperparameter sets with the best scores for point estimate and different quantile objectives, respectively. Once the hyperparameters were determined, we used them to train and test 10 BEnCQE models over the CV data sets and calculated the final CV evaluation score from the CV test results. Hereafter, we trained the final BEnCQE model with the same hyperparameters. The optimal hyperparameter searching process was accelerated by the NVIDIA A30 GPU, with the search process for one objective taking approximately 7 hr. Such acceleration also helps model training for CV evaluation and final model generation, where one BEnCQE model training takes around 20 min.

The performance (i.e., spatial generalization) of the BEnCQE model was evaluated in terms of the accuracy of point estimates and the coverage probability of PIs. The point estimates performance was evaluated by aggregating all CV test results and calculating the Pearson correlation coefficient (r), the coefficient of determination (R^2), and RMSE. The coverage probabilities of PI-50% and PI-90% are compared with their design confidence level, respectively, to verify their effectiveness. The final coverage is the average of observed coverages over 10 CV results. The observed coverage probability of PI should closely match the designed confidence level. If the observed coverage probability is lower than the designed level, it indicates that the model underestimates uncertainty, and if it is higher, the model overestimates uncertainty. This assessment provides an understanding of the reliability of the model performance and its capability to encapsulate the true value within its PIs. The evaluation result is shown in Section 4.1.

3.3. Predictor's Importance

The predictor's importance is calculated by the SHAP (SHapley Additive exPlanations) which is an advanced and popular ML model explanatory technology (Lundberg & Lee, 2017). We applied the SHAP to analyze the final BEnCQE model based on the entire training data set and derived the absolute SHAP value for each predictor in each member model of the BEnCQE. For each objective (i.e., point estimate and quantiles), we determined the importance of the predictor as the relative proportion of the predictor's mean absolute SHAP value in each member model, and then took the average across the corresponding member models as the final importance value. The result of the SHAP analysis is provided in Section 5.1, and it reveals the global contribution of each predictor to the model estimate.

3.4. Uncertainty Quantification

The PIs in the BEnCQE model can serve as a tool to quantify the uncertainty associated with individual instance estimates. As a conservative range, the PI-90% represents a "very likely" possibility of having the true NO₂ value, and thus we calculated the absolute uncertainty and the relative uncertainty based on the PI-90%. The absolute uncertainty is the full-length of the PI-90% (Equation 1) and the relative uncertainty is the ratio of the absolute uncertainty to the point estimate (Equation 2). Given the lack of knowledge regarding the distribution of surface NO₂ concentrations, it is recommended to utilize the full-length of the prediction interval rather than the half-length. This is because the latter is typically employed for data exhibiting a symmetric distribution (e.g., normal distribution and t-distribution).

Note that high NO₂ predictions usually come with high absolute uncertainty, whereas low NO₂ values almost invariably have a high relative uncertainty (see Section 4.4). Here we introduce an adjusted uncertainty to remove these dependencies. The adjusted uncertainty is calculated by removing the dependence of the absolute uncertainty on the NO₂ magnitude where the dependence variable (slope " a " in Equation 3) is obtained by linearly fitting the absolute uncertainty and point estimate over the whole study area and period (not only the training data set). As such, the adjusted uncertainty is independent of the NO₂ magnitude and can be used as a metric for

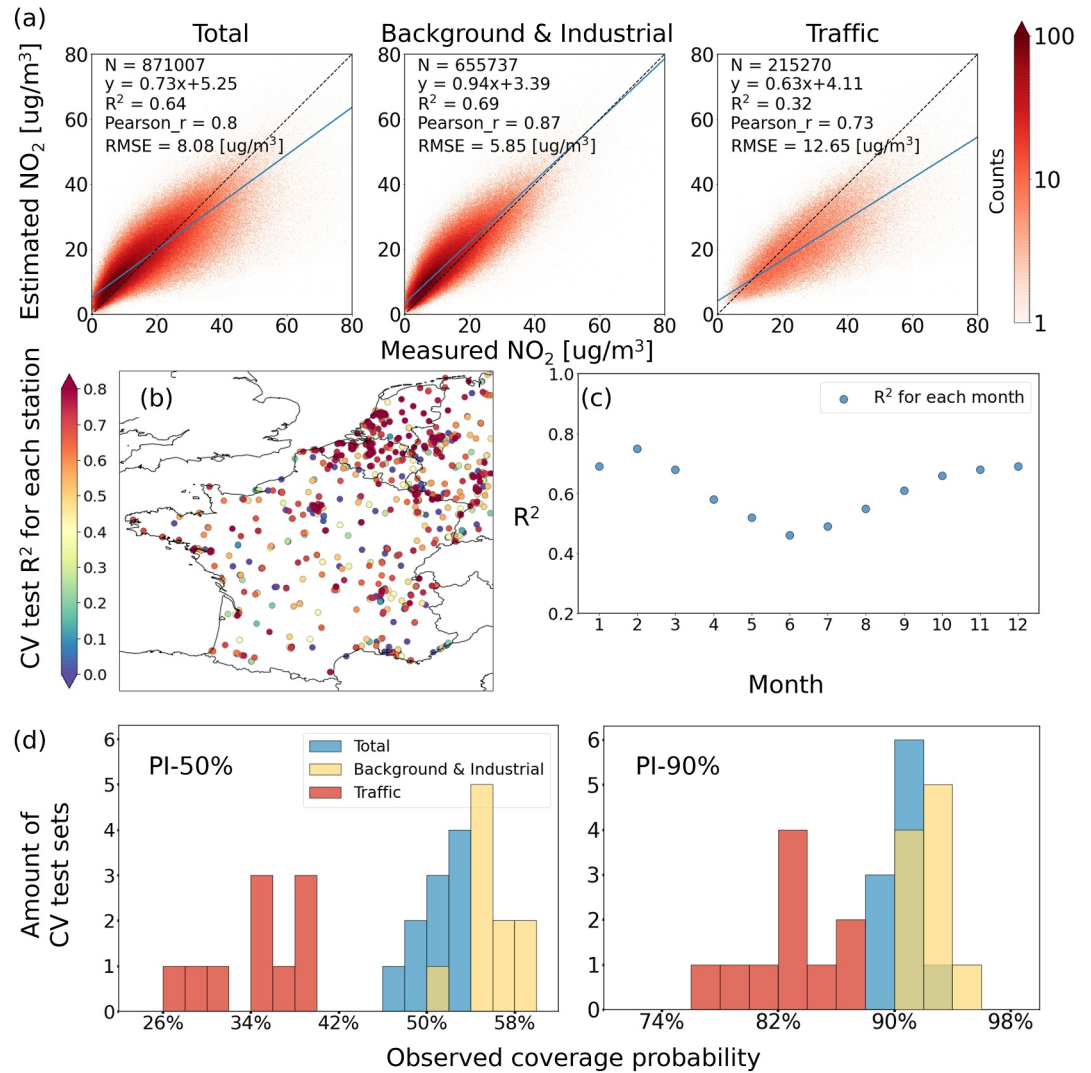


Figure 3. Ten-fold space-based CV test results. Panel (a) shows the CV results of the BEncQE model point estimates for total, background and industrial, and traffic stations. The point estimates performance is evaluated by aggregating all CV test results and calculating r , R^2 , and root mean square error. The fitting line of the scatter points is calculated by orthogonal linear regression. Panel (b) shows the distribution of CV test R^2 for each station, and panel (c) shows the monthly CV test R^2 for total stations. Panel (d) illustrates the CV results of PIs' coverage probability compared with the designed coverage for total, background and industrial, and traffic stations. The statistics for each fold of the CV are shown in Table S3 in Supporting Information S1.

gauging the uncertainty level. Recognizing that the selection of a prediction interval to quantify uncertainty is dependent on the task requirement and PI-90% offers a conservative uncertainty result, we also provide the uncertainty calculated based on PI-50% in the supplement (Figure S11 in Supporting Information S1) as a reference.

$$\text{absolute uncertainty}(i) = \text{Length of (PI - 90\%)} \quad (1)$$

$$\text{relative uncertainty}(i) = \frac{\text{absolute uncertainty}(i)}{\text{point estimate}(i)} \times 100\% \quad (2)$$

$$\text{adjusted uncertainty}(i) = \text{absolute uncertainty}(i) - a \times \text{point estimate}(i) \quad (3)$$

4. Results

4.1. Station-Based Model Evaluation

We used the 10-fold space-based CV to examine the performance of the point estimates and PIs of the BEnCQE model for a set of in-situ stations that were not included in the model training. Figure 3 shows that an overall good model performance is achieved for the point estimates when all stations are considered ($r = 0.80 \pm 0.02$, the error term denotes the standard deviation of 10-fold CV results, $R^2 = 0.64 \pm 0.03$, RMSE = $8.08 \pm 0.42 \mu\text{g}/\text{m}^3$). These statistics are comparable with those reported in other studies on Germany ($R^2 = 0.68$, Balamurugan et al., 2023) and the Alpine domain ($R^2 = 0.59$, Kim et al., 2021). The model accurately captures the NO_2 concentration over industrial and background stations ($r = 0.87 \pm 0.02$, $R^2 = 0.69 \pm 0.03$, RMSE = $5.85 \pm 0.33 \mu\text{g}/\text{m}^3$). While the model has a limited capacity to capture the NO_2 magnitude of traffic stations ($R^2 = 0.32 \pm 0.15$, RMSE = $12.65 \pm 1.70 \mu\text{g}/\text{m}^3$), it still learns a good correlation ($r = 0.73 \pm 0.05$). The statistics for each fold of the CV are shown in Table S3 in Supporting Information S1. The discrepancy can be attributed to the mismatched spatial representativeness of in-situ traffic station measurements to target grids. Traffic stations are typically positioned near major roads (European Environment Agency, 2023a) and predominantly capture local NO_2 levels (Maiheu & Janssen, 2019), which are highly influenced by local factors such as proximate vehicular emissions, road layout, and traffic behavior (Y. Zhu et al., 2020), making it difficult to capture the measured NO_2 from 1 km scale predictors. The R^2 distribution over each station and the monthly R^2 variation (Figures 3b and 3c) demonstrate that our model point estimate shows the best accuracy for the northern part of the study area, urban areas, and winter months, indicating a good predictive performance for the relatively high NO_2 levels (NO_2 distribution is shown in Section 4.2), except for the traffic situation. This implies an improved estimation when the NO_2 concentration dynamic is high.

In terms of PI's coverage probability, Figure 3d illustrates that the BEnCQE model successfully constructs valid PIs for all test samples that approximate the design confidence level (PI-50%: 51.0%, PI-90%: 90.5%). It is noteworthy that the conformal prediction successfully calibrates the boundaries of the PIs, as the original PI without conformal prediction does not fulfill the confidence level (PI-50%: 40.9%, PI-90%: 85.9%, not shown). However, we find that the model struggles to satisfy the desired coverage for traffic stations (PI-50%: 34.7%, PI-90%: 83.1%) and consequently provides a relatively conservative coverage for industrial and background stations (PI-50%: 56.2%, PI-90%: 92.8%) to compensate. Although estimating traffic NO_2 is challenging, quantile regression enhances the model's perception of such scenarios by focusing on the data distribution tails. Table S2 in Supporting Information S1 reveals that traffic NO_2 concentrations predominantly fall within the range between 75th and 95th quantiles (41.5% of measurements). Given that this part of the measurements is typically underestimated by the point estimates (Figure 3a), the 75th–95th quantile range can be recognized as an important reference for estimating traffic NO_2 . Additionally, Figure S6 in Supporting Information S1 displays the time series of CV test results from some stations, providing a glimpse of how quantile ranges encapsulate the real measurements for different types of stations. Employing the quantile regression method to estimate the extreme cases of NO_2 deserves further investigation, given that similar applications have been already conducted in other disciplines, such as studies of floods and droughts (Abbas et al., 2019), and extreme temperatures (Gao & Franzke, 2017).

As an alternative training approach, we also tuned and trained the model with the period-based CV to examine the temporal generalization of the ML model, where the data-splitting strategy was similar to the space-based CV but based on dates. The period-based CV results are shown in Figure S7 in Supporting Information S1, where the point estimates accuracy for total stations ($r = 0.91 \pm 0.02$, $R^2 = 0.83 \pm 0.06$, RMSE = $5.65 \pm 1.02 \mu\text{g}/\text{m}^3$), for background and industrial stations ($r = 0.9 \pm 0.02$, $R^2 = 0.8 \pm 0.05$, RMSE = $4.66 \pm 0.90 \mu\text{g}/\text{m}^3$), and for traffic stations ($r = 0.86 \pm 0.03$, $R^2 = 0.73 \pm 0.11$, RMSE = $7.94 \pm 1.60 \mu\text{g}/\text{m}^3$) are all higher than that of space-based CV. Meanwhile, the PIs have effective coverage for all stations (PI-50%: 47.5%, PI-90%: 87.0%), better coverage for background and industrial stations (PI-50%: 49.6%, PI-90%: 88.0%), and still insufficient coverage for traffic stations (PI-50%: 41.1%, PI-90%: 83.9%).

The results demonstrate that the BEnCQE model exhibits robust performance in estimating the NO_2 concentration for the stations over unknown periods. One of the model's applications is to detect the missing NO_2 variation at a given location. It is, however, essential to distinguish between the model developed using space-based CV and the model developed using period-based CV, as the choice of data-splitting strategy significantly impacts the model's nature and generalizability throughout the training, optimization, and evaluation

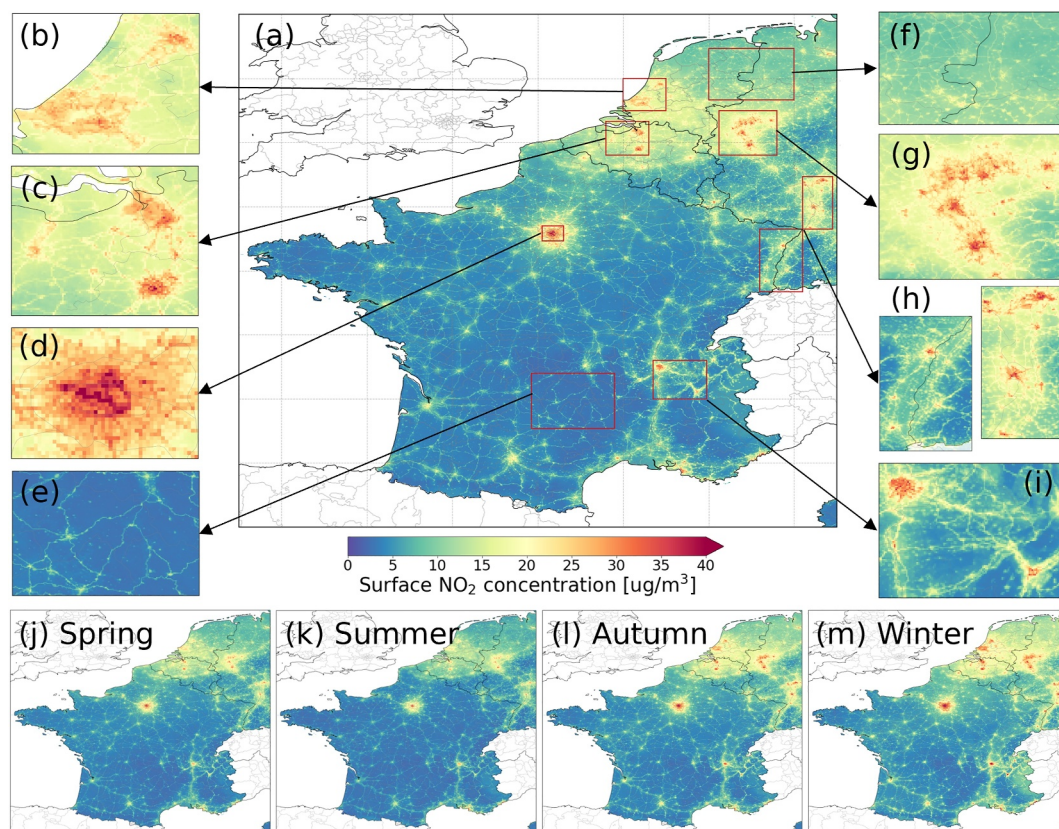


Figure 4. Spatial distribution of estimated surface NO_2 concentrations over Western Europe averaged from May 2018 to December 2021. Panel (a) shows the NO_2 distribution over the study area covering the Netherlands, Belgium, Luxembourg, France, and Western Germany (from 5°W to 9°E and from 42°N to 54°N). Panels (b)–(i) show NO_2 distributions for specific regions of interest and panels (j–m) show the NO_2 distributions for different seasons. The surface NO_2 level is estimated by the BENCQE model point estimate. Masked areas are out of the research scope.

processes. Both the training data and the data in the application area should be independent and identically distributed (IID). In the case of unevenly distributed stations, space-based CV is designed to make the training data satisfy the requirements of IID for unknown application areas. It is challenging for the period-based CV or the randomly sample-based CV to meet such IID requirements. The autocorrelation within the training samples may lead to an overly optimistic assessment of the model's ability to map the surface NO_2 distribution over the unknown area (Meyer & Pebesma, 2022). Consequently, we continue to utilize the BENCQE model constructed with the space-based CV strategy, as this CV strategy attempts to approximate the distributions of both the training and test sets to that of the unknown area.

4.2. Estimated Surface NO_2 in Western Europe

Figure 4a displays the mean distribution of surface NO_2 ($7.14 \pm 4.41 \mu\text{g}/\text{m}^3$, from May 2018 to December 2021, where the error term denotes the standard deviation of the spatial mean) in Western Europe, as inferred by the BENCQE model (daily distribution samples are provided in Figure S5 in Supporting Information S1). It shows that the largest NO_2 concentrations are mainly located in urban and industrial areas. The NO_2 signals on major roads are also visible. Notable high NO_2 regions include the western Netherlands (Figure 4b, $19.05 \pm 3.77 \mu\text{g}/\text{m}^3$), northern Belgium (Figure 4c, $17.20 \pm 4.10 \mu\text{g}/\text{m}^3$), and the Rhine-Ruhr region in western Germany (Figure 4g, $17.64 \pm 4.70 \mu\text{g}/\text{m}^3$). The region shown in panel (f) still has overall higher NO_2 levels (Figure 4f, $10.73 \pm 1.80 \mu\text{g}/\text{m}^3$) than the south, despite being distant from major NO_2 hotspots. A similar difference is also shown in the average TROPOMI NO_2 column distributions. The relatively higher NO_2 in this region can be attributed to the transport of NO_2 from nearby high NO_2 regions by the frequent southwest wind (Figure S10 in Supporting Information S1). For the entire domain, Paris is the largest NO_2

hotspot (Figure 4d, $25.46 \pm 4.83 \mu\text{g}/\text{m}^3$). In the South, high NO_2 levels are found also in the southern Rhine (Figure 4h, $11.77 \pm 5.38 \mu\text{g}/\text{m}^3$), Lyon (Figure 4i, $24.88 \pm 5.42 \mu\text{g}/\text{m}^3$), and Grenoble (Figure 4i, $12.15 \pm 6.75 \mu\text{g}/\text{m}^3$). These areas have strong NO_2 emissions which are trapped by natural barriers (valleys). More remote areas have low surface NO_2 levels (e.g., Figure 4e, the Massif Central, $3.46 \pm 1.35 \mu\text{g}/\text{m}^3$) with the highest NO_2 concentrations found near major roads.

Figures 4j–4m show the seasonal variation of surface NO_2 . The NO_2 levels are moderate in spring (Figure 4j, $6.52 \pm 3.98 \mu\text{g}/\text{m}^3$, March, April, and May) and low in summer (Figure 4k, $5.63 \pm 3.57 \mu\text{g}/\text{m}^3$, June, July, and August). The NO_2 levels are higher in autumn (Figure 4l, $7.72 \pm 4.87 \mu\text{g}/\text{m}^3$, September, October, and November), reaching a peak in winter (Figure 4m, $8.92 \pm 5.45 \mu\text{g}/\text{m}^3$, December, January, and February). The increase in NO_2 during the cooler seasons can be attributed to a longer NO_2 lifetime (less photolysis), weaker dispersion under unfavorable weather conditions, and more anthropogenic emissions related to energy production (Y. Shen et al., 2021).

4.3. Intercomparison With CAMS Model Data

To assess the physical plausibility of the estimated surface NO_2 patterns, we also compared the patterns from the BEnCQE point estimates with those from the physics-based CAMS (Copernicus Atmosphere Monitoring Service) European air quality reanalyses data set, which is based on an ensemble of eight to ten air quality data assimilation systems across Europe (Meleux et al., 2023). The CAMS NO_2 data is the ensemble median to ensure, on average, better performance than individual model products (Peuch et al., 2022). The comparison was operated at the spatial resolution of the CAMS EU product (10 km), in a grid-to-grid manner, from May 2018 to December 2020 (data for the year 2021 are not available at the last access time: 15 June 2023). We averaged the BEnCQE outputs to match the same spatial resolution, and the CAMS hourly data were aggregated to daily means.

Figures 5a and 5b show that the BEnCQE model and CAMS generate very similar NO_2 distributions, but the BEnCQE results reveal more spatial structures and are less smooth than the CAMS results. This difference can largely be attributed to the different resolutions of the input data sets and the impact of transport mechanisms in the CAMS model. Figure 5c depicts the difference between the two. The mean difference is mostly small over Western Europe (differences between -2 and $2 \mu\text{g}/\text{m}^3$ for 89.5% of the grids), while our model gives higher estimates over NO_2 hotspot areas and mountainous regions. This might be due to the difference in observation data used for model optimization, as the CAMS data assimilation is mainly conducted over background fields (Inness et al., 2019). Meanwhile, the 10 km resolution increases the CAMS uncertainty for local hotspots and makes CAMS underestimate NO_2 levels for urban monitor stations (Meleux et al., 2023). The concentrations estimated by the BEnCQE model are lower than CAMS over much of northern and eastern France. Figure 5d presents the Pearson correlation between the two data sets, which is above 0.8 in the northern half of the domain. It is lower (0.5–0.75) in the southern rural area and drops to about 0.3 in the southern mountainous regions. In the area with lower correlation, the surface NO_2 might be impacted by soil emissions (Oikawa et al., 2015) and affected by the complex terrain in the mountains. Although we used land type and elevation data sets, these influences might not be well accounted for in the model due to a lack of measurements. Figure S15 in Supporting Information S1 shows that the Pearson correlation increases when the TROPOMI data is filtered using the QA flag (QA > 0.75), especially for the southern rural area (0.6–0.8) and mountainous regions (about 0.5), although the mean difference remains almost the same. This indicates that the BEnCQE model, when using higher quality TROPOMI observations, behaves more consistently with the physical model.

Figures 5e–5l show the temporal variations of the BEnCQE results and the CAMS results for the regions of interest in Figure 4, demonstrating the strong agreement between both data sets in magnitude. Despite the lower correlation over the mountainous region, Figure 5h (the Massif Central in France) shows that the two model results are close in absolute values for the whole time series. Furthermore, although the BEnCQE's point estimates sometimes deviate from CAMS, the corresponding PI-50% still encapsulates the CAMS results. Additionally, this robust agreement persists even during anomalous conditions. As illustrated in Figure S14 in Supporting Information S1 for early 2020, the BEnCQE estimates are consistent with the CAMS results. This period includes exceptionally high temperatures (February) and the COVID-19 lockdown (March and April), during which surface NO_2 levels were found to be significantly reduced compared to the same months in 2019 (Barré et al., 2021; Guevara et al., 2021). Both the BEnCQE model and the CAMS model successfully capture this abrupt change in NO_2 . While we have not found direct evidence of how CAMS

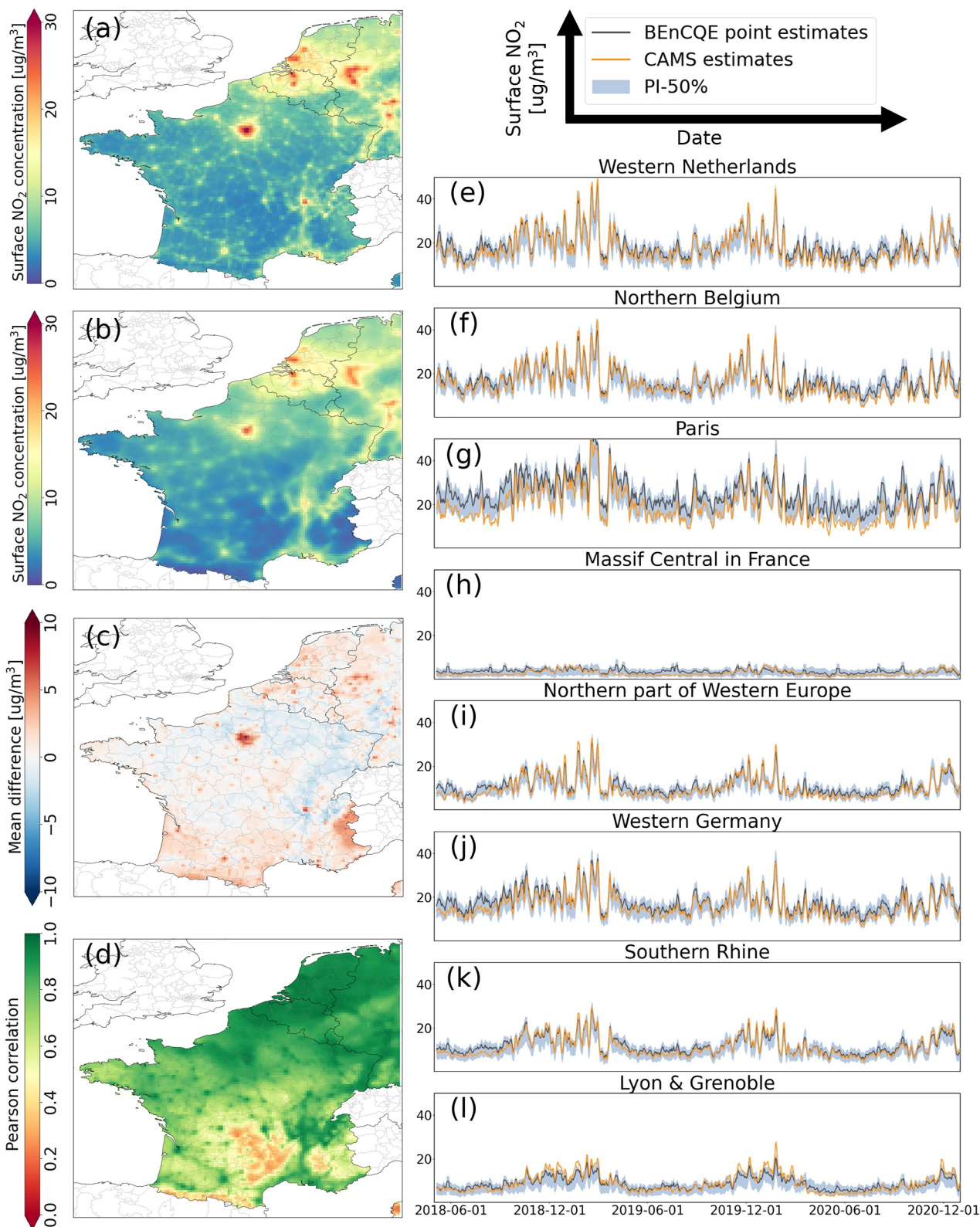


Figure 5.

adapts to such unprecedented scenarios, the data assimilation of EEA ground measurements (Meleux et al., 2023) is speculated to enable CAMS regional reanalysis data to reflect actual situations. Also, the performance of CAMS for the year 2020 has been validated to be as stable as in previous years with routine evaluation processes (Meleux et al., 2023).

4.4. Uncertainty of Surface NO₂ Estimates

Here, we analyze the BEnCQE NO₂ estimates uncertainty which mainly arises from the data. In this work, we calculate the absolute, relative, and adjusted uncertainty for each estimate (Section 3.4). The corresponding distributions are shown in Figure 6. Generally, the absolute uncertainty (Figure 6a) follows the distribution of estimated NO₂ concentrations, with the largest uncertainties over regions with high NO₂. Conversely, the relative uncertainty (Figure 6b) is largest for background regions (rural and mountainous areas) and lower for urban areas, industrial areas, and roads. To eliminate the dependence of these uncertainties on the NO₂ values, we calculate the adjusted uncertainty using Equation 3 (slope $a = 0.75$), which represents the residual absolute uncertainty after removing the mean NO₂ magnitude-dependent part of the absolute uncertainty. This type of uncertainty provides a consistent and balanced uncertainty measure across different NO₂ levels and is more reflective of the model's estimation capability rather than the inherent variability of the NO₂ values. The corresponding distribution is shown in Figure 6c.

The largest adjusted uncertainties (ranging from 15 to 20 $\mu\text{g}/\text{m}^3$) are found over NO₂ hotspot areas and the Alpine mountains. This indicates the challenge of robust estimation in such areas, where surface NO₂ is highly variable and spatially heterogeneous. Figures 6d–6g depict the seasonal variation of the adjusted uncertainty. High uncertainty is mainly present in NO₂ hotspot areas in spring and summer, and over mountainous areas in autumn and winter. We speculate that this is due to the conflict between the spatial and the temporal patterns of surface NO₂ that the model learned from the static and dynamic data respectively, as the uncertainty of the model is largest over high NO₂ areas in the low NO₂ season, and vice versa (low NO₂ areas in high NO₂ seasons). Given that the spatial patterns are strongly affected by the static emission data, making the emission data dynamic may help to reduce the model uncertainty.

Such an uncertainty analysis provides a nuanced perspective on the reliability of ML model estimates of surface NO₂. Note that the PI selection for quantifying uncertainty should be task-dependent because it will influence the judgments of model reliability. This study chooses a wide PI (i.e., PI-90%), resulting in a larger uncertainty value, which may be too conservative for NO₂ product usage. A comparison of the uncertainty calculated based on PI-50% (Figure S11 in Supporting Information S1) with that calculated based on PI-90% reveals that both types of uncertainty present similar distribution patterns of absolute and relative uncertainty. However, the adjusted uncertainty of PI-50% in the north does not exhibit a similar high relative value as seen with PI-90%.

5. Discussion

5.1. Importance of Predictors in Point and Quantiles Estimates

We used SHAP to calculate the importance of predictors (Section 3.3) to our final model, which has both point estimate and quantile estimates objectives. The SHAP values were calculated across the entire training data set, encompassing data from all stations over the study period. Quantile estimates such as Q-0.05, Q-0.25, Q-0.75, and Q-0.95, correspond to the 5th, 25th, 75th, and 95th quantiles of the estimated NO₂ concentration distribution, respectively, ranging from low to high concentrations. Therefore, the model objectives correspond to different NO₂ levels, and such an analysis not only elucidates the contribution of the predictors to the NO₂ point estimate but also provides crucial insights into how predictors influence the model estimates for different NO₂ levels. Figure S1 in Supporting Information S1 shows the importance percentage of each predictor, and we find that the

Figure 5. Comparison of surface NO₂ concentrations estimated by the BEnCQE and CAMS European air quality reanalysis data set. The comparison is conducted at the spatial resolution of CAMS (10 km), with the BEnCQE results reprojected from a 1 km scale to this resolution. Panels (a) and (b) show the surface NO₂ concentration distributions estimated by the BEnCQE (10 km) and CAMS (10 km), respectively, which are averaged from May 2018 to December 2020 (not the entire study period). Their mean difference (BEnCQE results minus CAMS results) is shown in panel (c). Panel (d) displays the distributions of Pearson correlation between two estimates, which are calculated for each grid during the period. Panels (e–l) present the comparison of the time series of regional NO₂ estimates, with a moving average window of 5 days, between the BEnCQE model and CAMS for regions shown in Figure 4 in Supporting Information S1, and the shaded area represents the regional average of the BEnCQE PI-50%.

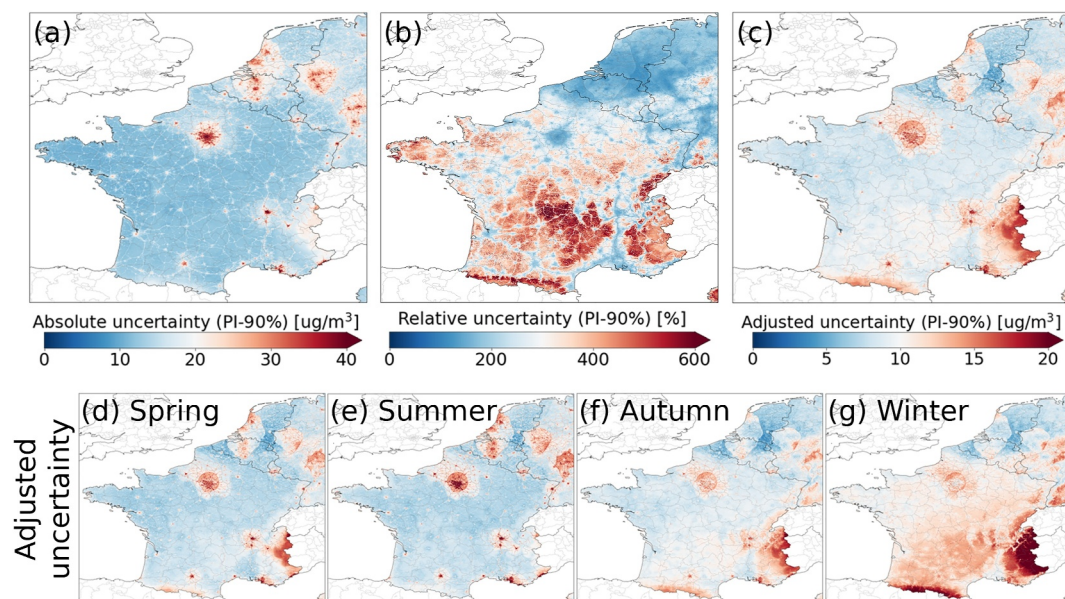


Figure 6. Spatial distribution of the uncertainty of the BEnCQE point estimates. Panel (a) shows the mean absolute uncertainty distribution for the entire study period, where the uncertainty is defined as the length of PI-90% (Equation 1). Panel (b) shows the average distribution of relative uncertainty which is the ratio of absolute uncertainty to point estimate (Equation 2). Panel (c) depicts the mean distribution of adjusted uncertainty which is calculated by removing the influence of point estimate magnitude from absolute uncertainty (Equation 3, slope $a = 0.75$). Panels (d)–(g) present the mean distribution of adjusted uncertainty for different seasons.

distributions are similar across different objectives. The daily TVCD data from TROPOMI and CAMS are very important, though their importance diminishes at the higher NO_2 quantile estimate. The BLH acts as the most important predictor for almost all objectives, in agreement with previous studies (Balamurugan et al., 2023; Kim et al., 2021). Additionally, the night light, population, and road networks contribute significantly, emphasizing the influence of anthropogenic activities on surface NO_2 variations, which aligns with findings by Balamurugan et al. (2023), Di et al. (2020), and Kim et al. (2021). The day of week is another important predictor, reflecting the strong weekly cycle of NO_2 (Stavrakou et al., 2020).

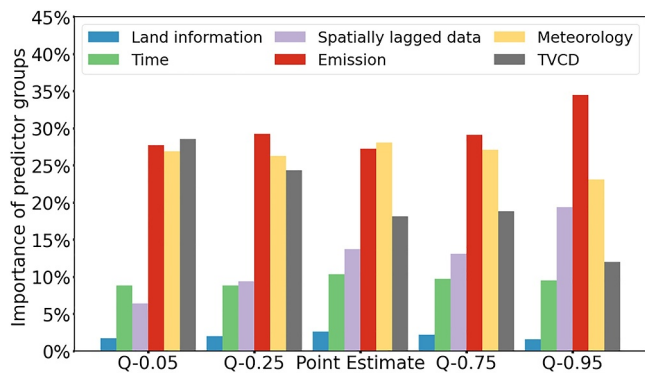


Figure 7. Importance of predictor groups for the BEnCQE estimates. The importance of the predictor is calculated using SHAP, an ML model explanation technique. SHAP is used to explain the BEnCQE model based on the entire training data set, and the absolute SHAP value is derived for each predictor in each member model of the BEnCQE. For each objective (i.e., point estimate and quantiles), we calculate the importance of the predictor as the relative proportion of the predictor's mean absolute SHAP value in each member model, and then take the average across the corresponding member models as the final importance value. The importance of predictor groups aggregates the importance of associate predictors. Details of the grouping and the importance of each predictor are shown in Table 1 and Figure S1 in Supporting Information S1.

Figure 7 provides an integrated overview of these results, grouping the predictors into different classes (see Table 1 and Figure S1 in Supporting Information S1). In general, the model estimation is dominated by emissions (29.5%), meteorology (26.3%), and TVCD (20.4%), while the SL data (12.4%) and time (9.4%) also play influential roles. Land information has the smallest contribution (2.0%), which differs from previous research (Balamurugan et al., 2023; Kim et al., 2021; Zhang et al., 2022), possibly due to different study areas. As quantiles increase, the importance of emissions and SL data increases noticeably. The contribution of meteorology declines moderately, while the influence of TVCD diminishes the most. As shown in Table S2 in Supporting Information S1, the measured NO_2 concentration for background and industrial stations is primarily within the range between the 5th quantile and the point estimate, while the traffic station measured NO_2 values mainly fall in the range between the point estimate and the 95th quantile. Considering that different types of stations have different representative areas and traffic stations usually measure the locally high concentration of NO_2 , this suggests that TVCD and meteorology assist the model in capturing general and broadly distributed NO_2 levels, while emissions and SL

data improve the model's perception of high and heterogeneously distributed NO₂ levels. This could be attributed to the fact that high NO₂ levels typically result from local emissions (Di et al., 2020). Moreover, the resolution is also a key factor in estimating high NO₂ levels, since coarse resolution reduces data sensitivity to NO₂ hotspots, as demonstrated by the relatively lower importance of the emission inventory (Figure S1 in Supporting Information S1) which has a resolution of 10 km.

One should note that the daily TROPOMI NO₂ TVCD (Figure S1 in Supporting Information S1, 7.9%) does not appear to have a dominant contribution to the point estimates, which differs from some previous works (Kim et al., 2021; Wei et al., 2022). This discrepancy can be attributed to three reasons. First, we did not filter the daily TROPOMI pixels by QA values in order to avoid substantial missing data. Nevertheless, this operation preserves the noise in the TROPOMI data and reduces the correlation between TROPOMI data and the surface NO₂ concentration, as illustrated in Figures S4a–S4b in Supporting Information S1. Nonetheless, our additional experiment indicates that using only QA75 data, defined as the data set of which the daily TROPOMI NO₂ TVCD has the QA value greater than 0.75, to train the model does not improve the model point estimates accuracy or reduce absolute uncertainty (Figures S8 and S9 in Supporting Information S1). Although the importance of both the daily TROPOMI NO₂ TVCD and the CAMS-simulated TVCD rises (Figure S12 in Supporting Information S1), the model performance stays unchanged. This might be due to the reduced number of training samples and the saturated explanatory power of NO₂ TVCD data to the surface NO₂ dynamic. Additionally, the relatively coarse spatial resolution reduces the sensitivity of TROPOMI TVCD to the surface NO₂, particularly in traffic-related NO₂ hotspots.

Second, the importance of the daily TROPOMI TVCD to the ML model has been partially displaced by the less noisy CAMS-simulated TVCD. To examine the impact of TROPOMI noise on the model, we conducted another model training experiment using QA75 data with noise. During the training process, random normal Gaussian noise was added to the daily TROPOMI NO₂ TVCD. As illustrated in Figures S8 and S9 in Supporting Information S1, the BEnCQE models trained on respective QA75 data and QA75 data with noise achieve similar accuracy and uncertainty. This is attributed to the fact that the model transfers its dependence from the daily TROPOMI TVCD to the CAMS-simulated TVCD, as evidenced in the comparison between Figures S12 and S13 in Supporting Information S1. This also demonstrates the robustness of the BEnCQE model to the noise in the daily TROPOMI data.

Third, the NO₂ transport behaves differently in columns and near the surface, as the transport in the latter is more damped. Despite this, the TROPOMI NO₂ TVCD is indispensable because it can directly observe the NO₂ variation, especially for unexpected changes. In our experiment analyzing predictors' changes during the COVID period (Text S4 in Supporting Information S1), most of the other predictors remain normal or static while the TROPOMI NO₂ TVCD and night light change significantly, in line with previous research (Levelt et al., 2022; Xu et al., 2021). Therefore, we speculate that these two predictors can assist the model in recognizing the sudden change in NO₂ levels.

5.2. Observed Potential NO₂ Exceedance

The 2021 World Health Organization Air Quality Guidelines (WHO AQGs) set a daily NO₂ limit of 25 µg/m³. The guidelines aim to assist in making a clean air policy and mitigating the health effects of air pollution (Hoffmann et al., 2021). Most ML-based NO₂ studies assess NO₂ exceedance by determining only whether point estimates exceed the WHO limit. However, this approach may underestimate NO₂ exceedances by ignoring the inherent uncertainty of the data. Here, we evaluate the frequency of NO₂ exceedances of the WHO limit during the study period using both the BEnCQE point estimates and the PI-50% range to identify potential NO₂ exceedances that have not been detected in previous ML-based surface NO₂ studies. Figure 8a illustrates that the BEnCQE point estimate indicates an overall low frequency of NO₂ exceedances in Western Europe, except for the hotspot areas (e.g., major cities and industrial areas) where the population density is high.

Coupled with the uncertainty, we define the potential NO₂ exceedance as when the point estimate is below the WHO limit, but the expected range of NO₂ concentration (i.e., PI-50%) crosses over this limit. As shown in Figure 8b, the potential exceedance is frequent (ranging from 10% to 30%) and expands beyond the known NO₂ hotspots to other regions, such as smaller cities, suburban areas, and roads within mountainous regions. In certain regions, potential exceedance is more frequent than the determined exceedance by approximately 2% (red areas in Figure S16 in Supporting Information S1). These regions would be inadequately considered if the policy-making

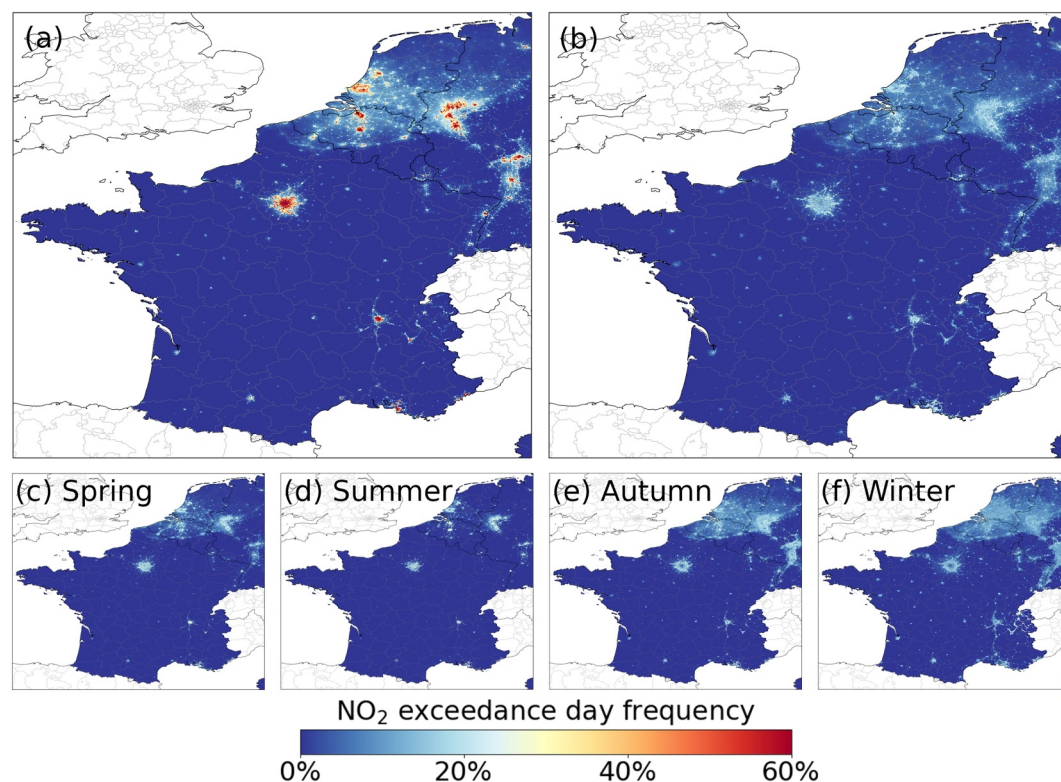


Figure 8. Spatial distribution of determined surface NO_2 exceedance frequency and potential exceedance frequency. The NO_2 exceedance is defined as the daily surface NO_2 concentration greater than $25 \mu\text{g}/\text{m}^3$ (WHO guidelines). Panel (a) shows the distribution of the determined NO_2 exceedance frequency over the study period calculated based on the BENCQE point estimates. Panel (b) shows the distribution of the potential NO_2 exceedance where the point estimate is below $25 \mu\text{g}/\text{m}^3$ but the expected range (PI-50%) crosses the threshold. Panels (c)–(f) show the distribution of the frequency of potential NO_2 exceedances over four seasons.

process is solely based on point estimates. Figures 8c–8f show the seasonal variation of potential exceedances. The potential exceedance concentrates in the NO_2 hotspot areas during the spring and summer, while it distributes more widely during autumn and winter. This suggests that we should be cautious of a possible underestimation of NO_2 exceedance in the hotspot area during the warm seasons, even though such periods usually have lower NO_2 levels. In addition, we also need to be mindful of the NO_2 exceedance in other non-hotspot areas during the cool seasons. Although ML is becoming popular for surface NO_2 estimation, the pure data-driven ML model usually tends to be conservative for high NO_2 estimates, as shown by the scatter of the BENCQE point estimate in Figure 3. This may be due to factors such as imbalanced training samples and the nature of the loss functions, which lead the model to concentrate on the overall accuracy for all predictions and sacrifice the accuracy for a small group of high NO_2 samples. Therefore, incorporating uncertainty information is crucial to avoid overlooking potential NO_2 pollution, which benefits air quality risk management.

5.3. Implication of Uncertainty Quantification for Model Improvement

One prevailing thought in the field of surface NO_2 estimation using ML models has been that station distribution and density are primary determinants of the robustness of prediction performance (Kim et al., 2021; L. Li & Wu, 2021; Wei et al., 2022). This belief presumes that a higher density of monitoring stations will capture more of the NO_2 variability, thereby strengthening the ability of the ML model to make accurate predictions. In this study, we aim to examine this assumption by exploring whether station density is a sufficient and reliable indicator of the predictive robustness of the ML model.

To conduct the analysis, we coupled the spatial patterns of both the adjusted uncertainty and the EEA in-situ station density. Figure 9a shows the normalized station density distribution computed using a 2D Gaussian convolution with a kernel size of 100 km. We adopted an 80% quantile threshold to distinguish between low and

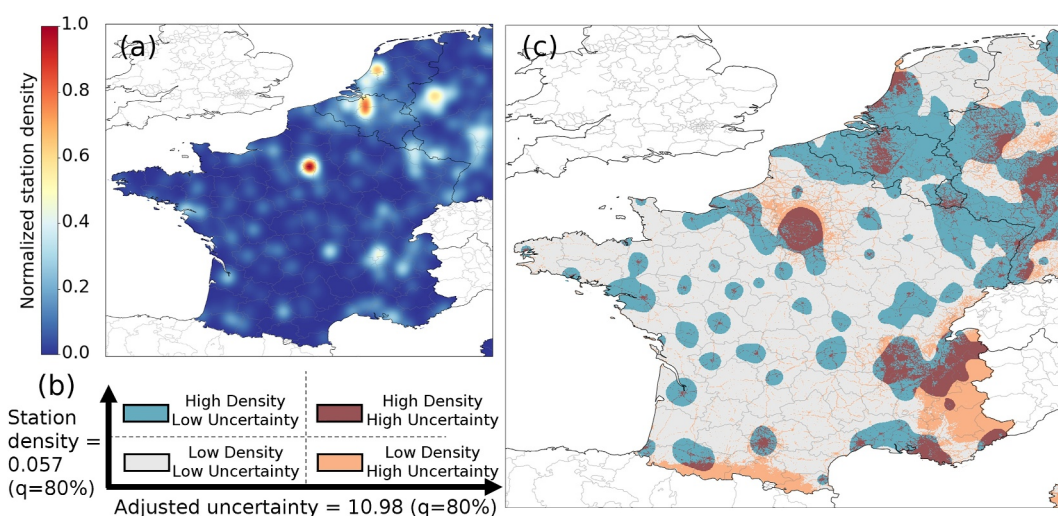


Figure 9. Intersections of adjusted uncertainty and the EEA in-situ station density. Panel (a) shows the normalized station density distribution which is computed using a 2D Gaussian convolution with a kernel size of 100 km. Panel (b) shows four interaction scenarios where low and high station densities intersect with low and high adjusted uncertainties. An 80% quantile threshold is used to differentiate between low and high scenarios. Panel (c) shows the distribution of the four intersection scenarios.

high station density and adjusted uncertainty. Consequently, we categorized four distinct scenarios as depicted in Figure 9b: (a) low-density-low-uncertainty (LDLU), (b) high-density-low-uncertainty (HDLU), (c) low-density-high-uncertainty (LDHU), and (d) high-density-high-uncertainty (HDHU). The distributions of the four scenarios are shown in Figure 9c. LDLU shows up in most rural areas, and HDLU mainly surrounds urban and industrial areas. LDHU tends to be concentrated in mountainous areas, whereas HDHU predominantly appears in NO_2 hotspots and parts of mountainous areas. In general, LDLU and HDLU, which cover most of the study region, can be recognized as the “safe” zones. However, caution is still needed for LDLU areas due to the lack of measurements, where part of the NO_2 distribution is possibly unknown to the model. The distribution of LDHU suggests a need to increase the sampling density in the mountainous region to promote the model to learn more reliable relationships between predictors and surface NO_2 in this area. Crucially, the appearance of HDHU means that the uncertainty persists in such areas even with relatively dense sampling, especially in the urban area. This directs us to review the predictors used in our model, as key predictors might be missing to accurately explain the high variability of surface NO_2 in the city. For instance, the traffic volume is essential for analyzing local NO_2 hotspots in urban areas (Y. Zhu et al., 2020), but this variable was not included in our predictors due to the difficulty of acquiring such data. The absence of such a key variable can confuse the model when different types of stations, such as background and traffic stations, are located close to one another and share similar predictors but record different NO_2 levels. The cause underlying the HDHU scenario is complex, emphasizing the need for a more nuanced investigation.

6. Conclusions

This work proposes a UQ-enabled ML framework and presents the BENCQE model for inferring surface NO_2 concentrations (daily and at a 1 km resolution) while also providing information on the uncertainty of each estimate. We analyze the uncertainty primarily arising from the inherent randomness and errors in the data. Our BENCQE model provides a reliable and plausible estimate of surface NO_2 over Western Europe, including accurate point estimates ($r = 0.80 \pm 0.02$, $R^2 = 0.64 \pm 0.03$, $\text{RMSE} = 8.08 \pm 0.42 \mu\text{g}/\text{m}^3$), reliable PIs coverage probabilities (PI-50%: 51.0%, PI-90%: 90.5%), and good agreement with CAMS results ($r = 0.8$ in the northern half of the domain, $0.5 < r < 0.75$ in the southern rural area, and $r = 0.3$ in the southern mountainous regions but absolute values are close).

However, the model estimation over the traffic scenario needs to be further investigated, as the current low-performance statistics are not only related to the model prediction capacity but also to the representativeness of the traffic stations for 1 km grids. In addition, the coarse resolution and the afternoon overpass (13:30 local

time) of the TROPOMI satellite observation restrict the model's capacity to capture the traffic NO₂ dynamic. The upcoming Sentinel-4 (hourly temporal resolution (European Space Agency, 2017)) and Sentinel-5 (morning overpass at 9:30 local time (European Space Agency, 2020)) missions might enrich the observational data and strengthen the model's ability to capture the high-dynamic NO₂.

Apart from regression using the conventional loss function (i.e., MSE), the application of quantile regression not only allows for quantifying uncertainty but also helps in understanding the model behavior at different NO₂ levels. The importance distribution of the predictors suggests that emission data (29.5%), meteorology (26.3%), and satellite observations (20.4%) dominate the estimation of surface NO₂, while the influence of SL data (12.4%) and time (9.4%) comes second, and land information (2.0%) is marginal. Among them, the importance of emission data, especially at high resolution, and SL data increases for the estimation of high NO₂ levels. The uncertainty information can contribute to air quality policymaking by revealing potential NO₂ exceedances that are undetectable when relying solely on point estimates. The BEnCQE model demonstrates its resilience to noise in the input data, as evidenced by the comparison between different models trained on full data, QA75 data, and QA75 data with noise, respectively. Nonetheless, it hinders us from quantifying the uncertainty introduced by simple perturbations to the predictors, which requires the development of specific methods.

Furthermore, the uncertainty information allows us to examine the robustness of the model outside of in-situ stations. Since the absolute and relative uncertainties are largely related to the magnitude of NO₂, we have proposed the adjusted uncertainty, which removes the influence of NO₂ magnitude, for an objective uncertainty analysis. It is observed that the largest adjusted uncertainty (ranging from 15 to 20 μg/m³) occurs in mountainous areas and NO₂ hotspots. In these regions, in situ stations are sparsely distributed in the mountains, whereas they are densely clustered in urban areas. This phenomenon highlights the challenge of achieving robust estimation in areas where NO₂ is highly variable and spatially heterogeneous. For such areas, achieving a robust estimate is not only a matter of sampling frequency but also a matter of the representativeness of the predictors and the station-measured data. For instance, estimating the dynamics of local high NO₂ concentrations in the city can be uncertain if some key predictors (e.g., traffic volume) are missing or if measured NO₂ concentrations cannot represent the NO₂ levels of target grids. Also, the seasonal variation of the adjusted uncertainty suggests that the uncertainty is related to the conflict between the spatial and temporal patterns of surface NO₂ that the model has learned. Given that spatial patterns are strongly affected by static emission data, making the emission data dynamic may help to reduce the model uncertainty.

Such uncertainty analysis presents a significant perspective for improving the robustness of the purely data-driven ML model, as it emphasizes the essentiality of the explanatory power and representativeness of the data (i.e., predictors and measurements) for real surface NO₂ variations. Apart from introducing more informative data to improve the model's perception of different NO₂ levels, another possible solution is to introduce physical and chemical constraints to reduce the model's reliance on data (C. Shen et al., 2023). Although we have taken the XGBoost model as the core ML model of the BEnCQE for its proven efficiency and robustness in many previous works (Balamurugan et al., 2023; Chi et al., 2022; Kang et al., 2021; Kim et al., 2021; Liu & Chen, 2022), we acknowledge that other models such as tree-based (e.g., random forest and light gradient-boosting machine) or neural-network-based (e.g., convolutional neural network) models could also be suitable. Future work will involve a comparative analysis of these models within the BEnCQE framework and explore potential enhancements. Finally, we hope that our study will encourage the systematic incorporation of UQ in future ML approaches to atmospheric components.

Appendix A

The quantile regression is proposed by Koenker and Bassett (1978) and achieved by estimating the conditional quantile function, as shown in Equation A1:

$$q_{\alpha}(x) = \inf \{y \in \mathbb{R} : F_y(y|X = x) \geq \alpha\} \quad (\text{A1})$$

This represents the value at which at least α proportion of the conditional distribution of the target variable Y given predictors ($X = x$) lies below, defining the α -th conditional quantile of Y . Usually a pair of CQFs is used for a PI construction, and the α is seen as a mis-coverage rate and so the PI coverage probability is $(1 - \alpha)$. The equation for the PI is defined as $C_{\alpha}(x)$, shown in Equation A2:

$$C_{\alpha}(x) = [q_{\alpha_{\text{low}}}(x), q_{\alpha_{\text{high}}}(x)] \quad (\text{A2})$$

where the α_{low} is $\alpha/2$ and α_{high} is $1 - \alpha/2$.

Establishing a ML model for quantile regression objective can be achieved by training the model through optimizing quantile loss which is shown in Equation A3:

$$\text{Loss}_{\alpha,i} = \begin{cases} (1 - \alpha)(q_{\alpha}(x_i) - y_i), & q_{\alpha}(x_i) \geq y_i \\ \alpha(y_i - q_{\alpha}(x_i)), & q_{\alpha}(x_i) < y_i \end{cases} \quad (\text{A3})$$

where $q_{\alpha}(x_i)$ is the α -th quantile estimate in the i th sample given the true target (y_i) and predictors (x_i). For low quantile regressions ($\alpha < 0.5$), the ML model would be penalized more for overestimation and less for underestimation, and vice versa for high quantile regressions. Such an approach has been demonstrated to be effective by previous research (Jensen et al., 2022; Romano et al., 2019; Vasseur & Aznarte, 2021). In this study, we want to obtain PI of 50% ($\alpha = 0.5$, $\alpha_{\text{low}} = 0.25$, $\alpha_{\text{high}} = 0.75$) and 90% ($\alpha = 0.1$, $\alpha_{\text{low}} = 0.05$, $\alpha_{\text{high}} = 0.95$). Thus, we build different XGBoost models with different quantile objectives (i.e., $q = 0.05, 0.25, 0.75$, and 0.95) for the corresponding PI.

To guarantee the coverage probability of the generated PIs, we follow Romano et al. (2019) to incorporate conformal prediction to refine the boundaries of PIs. The conformal prediction strategy calculates the conformity scores, which are residuals computed on the calibration data set I_2 (i.e., validation data set), when the model is trained on training data set I_1 (Angelopoulos & Bates, 2021). For the lower and higher bounds of a PI, the conformity scores are represented as $E_{\alpha_{\text{low}}}$ and $E_{\alpha_{\text{high}}}$, as shown in Equation A4:

$$E_{\alpha_{\text{low}}} = \{q_{\alpha_{\text{low}}}(x_i) - y_i : i \in I_2\}; E_{\alpha_{\text{high}}} = \{y_i - q_{\alpha_{\text{high}}}(x_i) : i \in I_2\} \quad (\text{A4})$$

We conduct asymmetrical conformalization (Romano et al., 2019) to refine the PI (Equation 2) by controlling its left and right tails independently to enhance the overall coverage guarantee. The equation is shown in Equation A5:

$$C_{\alpha}(x_n) = [q_{\alpha_{\text{low}}}(x_n) - Q_{\alpha_{\text{cp}}}(E_{\alpha_{\text{low}}}, I_2), q_{\alpha_{\text{high}}}(x_n) + Q_{\alpha_{\text{cp}}}(E_{\alpha_{\text{high}}}, I_2)] \quad (\text{A5})$$

where the lower bound subtracts the α_{cp} -th empirical quantile of $E_{\alpha_{\text{low}}}$ and the higher bound adds the α_{cp} -th empirical quantile of $E_{\alpha_{\text{high}}}$, and α_{cp} equals α_{high} . In this way, we obtain a reliable PI for each estimate.

Data Availability Statement

The operational TROPOMI NO₂ product is accessible via the Copernicus Data Space Ecosystem (<https://data-space.copernicus.eu/>). The modified TROPOMI NO₂ product for Europe with regional CAMS a-priori profiles is available at https://www.temis.nl/airpollution/no2_cams.php. The meteorological data is provided by the fifth-generation ECMWF atmospheric reanalysis of the global climate product (ERA5 and ERA5-land), which can be accessed via <https://cds.climate.copernicus.eu/>. The GRIP global roads database can be downloaded from <https://www.globio.info/download-grip-dataset>. The VIIRS night light data can be accessed from <https://eogdata.mines.edu/products/vnl/>. The population data set is provided by <https://ec.europa.eu/eurostat/web/gisco/geodata/population-distribution/geostat>. The EEA emission inventory is retrieved from <https://cdr.eionet.europa.eu/>. The CORINE land cover data set is downloaded from <https://land.copernicus.eu/en/products/corine-land-cover/clc2018>. The MERIT DEM data is accessible via https://hydro.iis.u-tokyo.ac.jp/~yamada/MERIT_DEM/. The EEA air quality database can be downloaded from <https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm>. The CAMS European air quality reanalyses data set is accessible via <https://ads.atmosphere.copernicus.eu/>. The study data included can be accessed from Zenodo data archive (Sun et al., 2024, <https://doi.org/10.5281/zenodo.10425430>). The BEnCQE core model (i.e., XGBoost v2.0.0) can be accessed from <https://xgboost.readthedocs.io/en/stable/>.

Acknowledgments

The Belgian Federal Science Policy Office is gratefully appreciated for funding part of this work in the framework of the Terrascope-S5P PRODEX project (PEA 4000136290); L.C. is a research associate supported by the Belgian F.R.S.-FNRS; We appreciate the Belgian Interregional Environment Agency (IRCEL—CELINE) for its support on this work.

References

Abbas, S. A., Xuan, Y., & Song, X. (2019). Quantile regression based methods for investigating rainfall trends associated with flooding and drought conditions. *Water Resources Management*, 33(12), 4249–4264. <https://doi.org/10.1007/s11269-019-02362-0>

Achakulwisut, P., Brauer, M., Hystad, P., & Anenberg, S. C. (2019). Global, national, and urban burdens of paediatric asthma incidence attributable to ambient NO₂ pollution: Estimates from global datasets. *The Lancet Planetary Health*, 3(4), e166–e178. [https://doi.org/10.1016/S2542-5196\(19\)30046-4](https://doi.org/10.1016/S2542-5196(19)30046-4)

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. *arXiv e-prints*. arXiv:1907.10902. <https://doi.org/10.48550/arXiv.1907.10902>

Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv e-prints*. arXiv:2107.07511. <https://doi.org/10.48550/arXiv.2107.07511>

Balamurugan, V., Chen, J., Wenzel, A., & Keutsch, F. N. (2023). Spatiotemporal modeling of air pollutant concentrations in Germany using machine learning. *Atmospheric Chemistry and Physics*, 23(17), 10267–10285. <https://doi.org/10.5194/acp-23-10267-2023>

Barré, J., Petetin, H., Colette, A., Guevara, M., Peuch, V. H., Rouil, L., et al. (2021). Estimating lockdown-induced European NO₂ changes using satellite and surface observations and air quality models. *Atmospheric Chemistry and Physics*, 21(9), 7373–7394. <https://doi.org/10.5194/acp-21-7373-2021>

Chan, K. L., Khorsandi, E., Liu, S., Baier, F., & Valks, P. (2021). Estimation of surface NO₂ concentrations over Germany from TROPOMI satellite observations using a machine learning method. *Remote Sensing*, 13(5), 969. <https://doi.org/10.3390/rs13050969>

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Paper presented at the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA*. <https://doi.org/10.1145/2939762.2939785>

Chi, Y., Fan, M., Zhao, C., Yang, Y., Fan, H., Yang, X., et al. (2022). Machine learning-based estimation of ground-level NO₂ concentrations over China. *Science of the Total Environment*, 807, 150721. <https://doi.org/10.1016/j.scitotenv.2021.150721>

Cooper, M. J., Martin, R. V., Hammer, M. S., Levelt, P. F., Veeffkind, P., Lamsal, L. N., et al. (2022). Global fine-scale changes in ambient NO₂ during COVID-19 lockdowns. *Nature*, 601(7893), 380–387. <https://doi.org/10.1038/s41586-021-04229-0>

Crippa, M., Guizzardi, D., Muntean, M., Schaaf, E., Dentener, F., van Aardenne, J. A., et al. (2018). Gridded emissions of air pollutants for the period 1970–2012 within EDGAR v4.3.2. *Earth System Science Data*, 10(4), 1987–2013. <https://doi.org/10.5194/essd-10-1987-2018>

de Hoogh, K., Saucy, A., Shtein, A., Schwartz, J., West, E. A., Strassmann, A., et al. (2019). Predicting fine-scale daily NO₂ for 2005–2016 incorporating OMI satellite data across Switzerland. *Environmental Science & Technology*, 53(17), 10279–10287. <https://doi.org/10.1021/acs.est.9b03107>

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., et al. (2020). Assessing NO₂ concentration and model uncertainty with high spatio-temporal resolution across the contiguous United States using ensemble model averaging. *Environmental Science & Technology*, 54(3), 1372–1384. <https://doi.org/10.1021/acs.est.9b03358>

Douros, J., Eskes, H., van Geffen, J., Boersma, K. F., Compennolle, S., Pinardi, G., et al. (2023). Comparing Sentinel-5P TROPOMI NO₂ column observations with the CAMS regional air quality ensemble. *Geoscientific Model Development*, 16(2), 509–534. <https://doi.org/10.5194/gmd-16-509-2023>

Elvidge, C. D., Zhizhin, M., Ghosh, T., Hsu, F.-C., & Taneja, J. (2021). Annual Time Series of Global VIIRS Nighttime Lights Derived from Monthly Averages: 2012 to 2019. *Remote Sensing*, 13(5), 922. <https://doi.org/10.3390/rs13050922>

European Environment Agency. (2020). CORINE Land Cover 2018 (vector), Europe, 6-yearly (Version 2020_20u1) [Dataset]. European Environment Agency. <https://doi.org/10.2909/71c95a07-e296-44fc-b22b-415f42acfd0>

European Environment Agency. (2022). AirBase - The European air quality database [Dataset]. European Environment Agency. Retrieved from <https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm>

European Environment Agency. (2023a). Classification of monitoring stations and criteria to include them in EEA's assessments products. Retrieved from <https://www.eea.europa.eu/themes/air/air-quality-concentrations/classification-of-monitoring-stations-and>

European Environment Agency. (2023b). Gridded emissions 2019 [Dataset]. *Eionet Central Data Repository*. Retrieved from <https://cdr.eionet.europa.eu/>

European Space Agency. (2017). *Sentinel-4: ESA's Geostationary Atmospheric Mission for Copernicus Operational Services*. ESA Communications. Retrieved from <https://esamultimedia.esa.int/multimedia/publications/SP-1334/SP-1334.pdf>

European Space Agency. (2020). *Sentinel-5*. ESA Communications. Retrieved from https://esamultimedia.esa.int/docs/EarthObservation/SP_1336_Sentinel-5_web.pdf

Fan, J., Wu, L., Zhang, F., Cai, H., Zeng, W., Wang, X., & Zou, H. (2019). Empirical and machine learning models for predicting daily global solar radiation from sunshine duration: A review and case study in China. *Renewable and Sustainable Energy Reviews*, 100, 186–212. <https://doi.org/10.1016/j.rser.2018.10.018>

Gao, M., & Franzke, C. L. E. (2017). Quantile regression-based spatiotemporal analysis of extreme temperature change in China. *Journal of Climate*, 30(24), 9897–9914. <https://doi.org/10.1175/JCLI-D-17-0356.1>

Ghahremanloo, M., Lops, Y., Choi, Y., Mousavinezhad, S., & Jung, J. (2023). A coupled deep learning model for estimating surface NO₂ levels from remote sensing data: 15-year study over the contiguous United States. *Journal of Geophysical Research: Atmospheres*, 128(2), e2022JD037010. <https://doi.org/10.1029/2022JD037010>

Ghahremanloo, M., Lops, Y., Choi, Y., & Yeganeh, B. (2021). Deep learning estimation of daily ground-level NO₂ concentrations from remote sensing data. *Journal of Geophysical Research: Atmospheres*, 126(21), e2021JD034925. <https://doi.org/10.1029/2021JD034925>

Grennfelt, P., Englerød, A., Forsius, M., Hov, Ø., Rodhe, H., & Cowling, E. (2020). Acid rain and air pollution: 50 years of progress in environmental science and policy. *Ambio*, 49(4), 849–864. <https://doi.org/10.1007/s13280-019-01244-4>

Guerreiro, C. B. B., Foltescu, V., & de Leeuw, F. (2014). Air quality status and trends in Europe. *Atmospheric Environment*, 98, 376–384. <https://doi.org/10.1016/j.atmosenv.2014.09.017>

Guevara, M., Jorba, O., Soret, A., Petetin, H., Bowdalo, D., Serradell, K., et al. (2021). Time-resolved emission reductions for atmospheric chemistry modelling in Europe during the COVID-19 lockdowns. *Atmospheric Chemistry and Physics*, 21(2), 773–797. <https://doi.org/10.5194/acp-21-773-2021>

Haynes, K., Lagerquist, R., McGraw, M., Musgrave, K., & Ebert-Uphoff, I. (2023). Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artificial Intelligence for the Earth Systems*, 2(2), 220061. <https://doi.org/10.1175/AIES-D-22-0061.1>

He, S., Dong, H., Zhang, Z., & Yuan, Y. (2022). An ensemble model-based estimation of nitrogen dioxide in a southeastern coastal region of China. *Remote Sensing*, 14(12), 2807. <https://doi.org/10.3390/rs14122807>

- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049. <https://doi.org/10.1002/qj.3803>
- Hodan, W. M., & Barnard, W. R. (2004). Evaluating the contribution of PM_{2.5} precursor gases and re-entrained road emissions to mobile source PM_{2.5} particulate matter emissions. Retrieved from <https://www3.epa.gov/ttnchie1/conference/ei13/mobile/hodan.pdf>
- Hoffmann, B., Boogaard, H., de Nazelle, A., Andersen, Z. J., Abramson, M., Brauer, M., et al. (2021). WHO air quality guidelines 2021—Aiming for healthier air for all: A joint statement by medical, public health, scientific societies and patient representative organisations. *International Journal of Public Health*, 66, 1604465. <https://doi.org/10.3389/ijph.2021.1604465>
- Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A. M., et al. (2019). The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*, 19(6), 3515–3556. <https://doi.org/10.5194/acp-19-3515-2019>
- Jensen, V., Bianchi, F. M., & Anfinsen, S. N. (2022). Ensemble conformalized quantile regression for probabilistic time series forecasting. In *IEEE Transactions on Neural Networks and Learning Systems* (pp. 1–12). <https://doi.org/10.1109/TNNLS.2022.3217694>
- Kang, Y., Choi, H., Im, J., Park, S., Shin, M., Song, C.-K., & Kim, S. (2021). Estimation of surface-level NO₂ and O₃ concentrations using TROPOMI data and machine learning over East Asia. *Environmental Pollution*, 288, 117711. <https://doi.org/10.1016/j.envpol.2021.117711>
- Kim, M., Brunner, D., & Kuhlmann, G. (2021). Importance of satellite observations for high-resolution mapping of near-surface NO₂ by machine learning. *Remote Sensing of Environment*, 264, 112573. <https://doi.org/10.1016/j.rse.2021.112573>
- Kiureghian, A. D., & Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2), 105–112. <https://doi.org/10.1016/j.strusafe.2008.06.020>
- Koenker, R. (2005). *Quantile regression*. Cambridge University Press. Retrieved from <https://www.cambridge.org/core/books/quantile-regression/C18AE7BCF3EC43C16937390D44A328B1>
- Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, 46(1), 33–50. <https://doi.org/10.2307/1913643>
- Kong, L., Tang, X., Zhu, J., Wang, Z., Li, J., Wu, H., et al. (2021). A 6-year-long (2013–2018) high-resolution air quality reanalysis dataset in China based on the assimilation of surface observations from CNEMC. *Earth System Science Data*, 13(2), 529–570. <https://doi.org/10.5194/essd-13-529-2021>
- Kumar, U., De Ridder, K., Lefebvre, W., & Janssen, S. (2012). Data assimilation of surface air pollutants (O₃ and NO₂) in the regional-scale air quality model AURORA. *Atmospheric Environment*, 60, 99–108. <https://doi.org/10.1016/j.atmosenv.2012.06.005>
- Lamsal, L. N., Krotkov, N. A., Vasilkov, A., Marchenko, S., Qin, W., Yang, E. S., et al. (2021). Ozone Monitoring Instrument (OMI) Aura nitrogen dioxide standard product version 4.0 with improved surface and cloud treatments. *Atmospheric Measurement Techniques*, 14(1), 455–479. <https://doi.org/10.5194/amt-14-455-2021>
- Lamsal, L. N., Martin, R. V., van Donkelaar, A., Steinbacher, M., Celarier, E. A., Bucsela, E., et al. (2008). Ground-level nitrogen dioxide concentrations inferred from the satellite-borne Ozone Monitoring Instrument. *Journal of Geophysical Research*, 113(D16), D16308. <https://doi.org/10.1029/2007JD009235>
- Levelt, P. F., Stein Zweers, D. C., Aben, I., Bauwens, M., Borsdorff, T., De Smedt, I., et al. (2022). Air quality impacts of COVID-19 lockdown measures detected from space using high spatial resolution observations of multiple trace gases from Sentinel-5P/TROPOMI. *Atmospheric Chemistry and Physics*, 22(15), 10319–10351. <https://doi.org/10.5194/acp-22-10319-2022>
- Li, L., & Wu, J. (2021). Spatiotemporal estimation of satellite-borne and ground-level NO₂ using full residual deep networks. *Remote Sensing of Environment*, 254, 112257. <https://doi.org/10.1016/j.rse.2020.112257>
- Li, M., Wu, Y., Bao, Y., Liu, B., & Petropoulos, G. P. (2022). Near-surface NO₂ concentration estimation by Random Forest modeling and Sentinel-5P and ancillary data. *Remote Sensing*, 14(15), 3612. <https://doi.org/10.3390/rs14153612>
- Liu, J., & Chen, W. (2022). First satellite-based regional hourly NO₂ estimations using a space-time ensemble learning model: A case study for Beijing-Tianjin-Hebei Region, China. *Science of the Total Environment*, 820, 153289. <https://doi.org/10.1016/j.scitotenv.2022.153289>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Paper presented at the Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA*. <https://doi.org/10.5555/3295222.3295230>
- Maiheu, B., & Janssen, S. (2019). Assessing the spatial representativeness of air quality sampling points – Literature Review. Retrieved from <https://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/20191218DGenVAQWkSR5Task1LitRevFINAL.pdf>
- Meijer, J. R., Huijbregts, M. A. J., Schotten, K. C. G. J., & Schipper, A. M. (2018). Global patterns of current and future road infrastructure. *Environmental Research Letters*, 13(6), 064006. <https://doi.org/10.1088/1748-9326/aab442>
- Meleux, F., Raux, B., Ung, A., Colette, A., Gauss, M., Douros, J., et al. (2023). Annual report on the evaluation of validated reanalyses VRA2020. Retrieved from https://atmosphere.copernicus.eu/sites/default/files/custom-uploads/EQC-regional/VRA/CAMS283_2021ISC1_D83.2.2.1-2020_202303_VRA2020_evaluation_v2.pdf
- Meyer, H., & Pebesma, E. (2022). Machine learning-based global maps of ecological variables and the challenge of assessing them. *Nature Communications*, 13(1), 2208. <https://doi.org/10.1038/s41467-022-29838-9>
- Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., et al. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13(9), 4349–4383. <https://doi.org/10.5194/essd-13-4349-2021>
- Oikawa, P. Y., Ge, C., Wang, J., Eberwein, J. R., Liang, L. L., Allsman, L. A., et al. (2015). Unusually high soil nitrogen oxide emissions influence air quality in a high-temperature agricultural region. *Nature Communications*, 6(1), 8753. <https://doi.org/10.1038/ncomms9753>
- Pan, Y., Zhao, C., & Liu, Z. (2021). Estimating the daily NO₂ concentration with high spatial resolution in the Beijing–Tianjin–Hebei Region using an ensemble learning model. *Remote Sensing*, 13(4), 758. <https://doi.org/10.3390/rs13040758>
- Peuch, V.-H., Engelen, R., Rixen, M., Dee, D., Flemming, J., Suttie, M., et al. (2022). The Copernicus Atmosphere Monitoring Service: From Research to Operations. *Bulletin of the American Meteorological Society*, 103(12), E2650–E2668. <https://doi.org/10.1175/BAMS-D-21-0314.1>
- Poraicu, C., Müller, J. F., Stavrou, T., Fonteyn, D., Tack, F., Deutsch, F., et al. (2023). Cross-evaluating WRF-Chem v4.1.2, TROPOMI, APEX, and in situ NO₂ measurements over Antwerp, Belgium. *Geoscientific Model Development*, 16(2), 479–508. <https://doi.org/10.5194/gmd-16-479-2023>
- Qin, K., Han, X., Li, D., Xu, J., Loyola, D., Xue, Y., et al. (2020). Satellite-based estimation of surface NO₂ concentrations over east-central China: A comparison of POMINO and OMNO2d data. *Atmospheric Environment*, 224, 117322. <https://doi.org/10.1016/j.atmosenv.2020.117322>
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, 566(7743), 195–204. <https://doi.org/10.1038/s41586-019-0912-1>
- Ren, J., Guo, F., & Xie, S. (2022). Diagnosing ozone–NO_x–VOC sensitivity and revealing causes of ozone increases in China based on 2013–2021 satellite retrievals. *Atmospheric Chemistry and Physics*, 22(22), 15035–15047. <https://doi.org/10.5194/acp-22-15035-2022>
- Romano, Y., Patterson, E., & Candès, E. J. (2019). Conformalized quantile regression. *arXiv e-prints*. arXiv:1905.03222. <https://doi.org/10.48550/arXiv.1905.03222>

- Scheibenreif, L., Mommert, M., & Borth, D. (2022). Toward global estimation of ground-level NO₂ pollution with deep learning and remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14. <https://doi.org/10.1109/TGRS.2022.3160827>
- Schneider, R., Vicedo-Cabrera, A. M., Sera, F., Masselot, P., Stafoggia, M., de Hoogh, K., et al. (2020). A satellite-based spatio-temporal machine learning model to reconstruct daily PM_{2.5} concentrations across Great Britain. *Remote Sensing*, 12(22), 3803. <https://doi.org/10.3390/rs12223803>
- Shen, C., Appling, A. P., Gentine, P., Bandai, T., Gupta, H., Tartakovsky, A., et al. (2023). Differentiable modelling to unify machine learning and physical models for geosciences. *Nature Reviews Earth & Environment*, 4(8), 552–567. <https://doi.org/10.1038/s43017-023-00450-9>
- Shen, Y., Jiang, F., Feng, S., Zheng, Y., Cai, Z., & Lyu, X. (2021). Impact of weather and emission changes on NO₂ concentrations in China during 2014–2019. *Environmental Pollution*, 269, 116163. <https://doi.org/10.1016/j.envpol.2020.116163>
- Silva, F. B. E., Poelman, H., & Dijkstra, L. (2021). JRC-GEOSTAT 2018 [Dataset]. *Eurostat*. <https://ec.europa.eu/eurostat/web/gisco/geodata/population-distribution/geostat>
- Song, W., Liu, X.-Y., Hu, C.-C., Chen, G.-Y., Liu, X.-J., Walters, W. W., et al. (2021). Important contributions of non-fossil fuel nitrogen oxides emissions. *Nature Communications*, 12(1), 243. <https://doi.org/10.1038/s41467-020-20356-0>
- Stavrakou, T., Müller, J. F., Bauwens, M., Boersma, K. F., & van Geffen, J. (2020). Satellite evidence for changes in the NO₂ weekly cycle over large cities. *Scientific Reports*, 10(1), 10066. <https://doi.org/10.1038/s41598-020-66891-0>
- Stevens, C. J., David, T. I., & Storkey, J. (2018). Atmospheric nitrogen deposition in terrestrial ecosystems: Its impact on plant communities and consequences across trophic levels. *Functional Ecology*, 32(7), 1757–1769. <https://doi.org/10.1111/1365-2435.13063>
- Sun, W., Tack, F., Clarisse, L., Schneider, R., Stavrakou, T., & Roozendaal, M. V. (2024). Inferring surface NO₂ over Western Europe: A machine learning approach with uncertainty quantification (version 2) [Dataset]. *Zenodo*. <https://doi.org/10.5281/zenodo.10425430>
- Tack, F., Merlaud, A., Iordache, M. D., Pinaridi, G., Dimitropoulou, E., Eskes, H., et al. (2021). Assessment of the TROPOMI tropospheric NO₂ product based on airborne APEX observations. *Atmospheric Measurement Techniques*, 14(1), 615–646. <https://doi.org/10.5194/amt-14-615-2021>
- Takeuchi, I., Le, Q. V., Sears, T. D., & Smola, A. J. (2006). Nonparametric quantile estimation. *Journal of Machine Learning Research*, 7, 1231–1264. Retrieved from <http://jmlr.org/papers/v7/takeuchi06a.html>
- Van den Berghe, K., Peris, A., Meijers, E., & Jacobs, W. (2023). Friends with benefits: The emergence of the Amsterdam–Rotterdam–Antwerp (ARA) polycentric port region. *Territory, Politics, Governance*, 11(2), 301–320. <https://doi.org/10.1080/21622671.2021.2014353>
- van Geffen, J., Boersma, K. F., Eskes, H., Sneep, M., ter Linden, M., Zara, M., & Veefkind, J. P. (2020). S5P TROPOMI NO₂ slant column retrieval: Method, stability, uncertainties and comparisons with OMI. *Atmospheric Measurement Techniques*, 13(3), 1315–1335. <https://doi.org/10.5194/amt-13-1315-2020>
- Vasseur, S. P., & Aznarte, J. L. (2021). Comparing quantile regression methods for probabilistic forecasting of NO₂ pollution levels. *Scientific Reports*, 11(1), 11592. <https://doi.org/10.1038/s41598-021-90063-3>
- Veefkind, J. P., Aben, I., McMullan, K., Forster, H., de Vries, J., Otter, G., et al. (2012). TROPOMI on the ESA Sentinel-5 Precursor: A GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. *Remote Sensing of Environment*, 120, 70–83. <https://doi.org/10.1016/j.rse.2011.09.027>
- Vira, J., & Sofiev, M. (2015). Assimilation of surface NO₂ and O₃ observations into the SILAM chemistry transport model. *Geoscientific Model Development*, 8(2), 191–203. <https://doi.org/10.5194/gmd-8-191-2015>
- Wan, N., Xiong, X., Kluitenberg, G. J., Hutchinson, J. M. S., Aiken, R., Zhao, H., & Lin, X. (2023). Estimation of biomass burning emission of NO₂ and CO from 2019–2020 Australia fires based on satellite observations. *Atmospheric Chemistry and Physics*, 23(1), 711–724. <https://doi.org/10.5194/acp-23-711-2023>
- Wang, Y., Yuan, Q., Li, T., Zhu, L., & Zhang, L. (2021). Estimating daily full-coverage near surface O₃, CO, and NO₂ concentrations at a high spatial resolution over China based on SSP-TROPOMI and GEOS-FP. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175, 311–325. <https://doi.org/10.1016/j.isprsjprs.2021.03.018>
- Wei, J., Liu, S., Li, Z., Liu, C., Qin, K., Liu, X., et al. (2022). Ground-level NO₂ surveillance from space across China for high resolution using interpretable spatiotemporally weighted artificial intelligence. *Environmental Science & Technology*, 56(14), 9988–9998. <https://doi.org/10.1021/acs.est.2c03834>
- Xu, G., Xiu, T., Li, X., Liang, X., & Jiao, L. (2021). Lockdown induced night-time light dynamics during the COVID-19 epidemic in global megacities. *International Journal of Applied Earth Observation and Geoinformation*, 102, 102421. <https://doi.org/10.1016/j.jag.2021.102421>
- Yamazaki, D., Ikeshima, D., Tawatari, R., Yamaguchi, T., O'Loughlin, F., Neal, J. C., et al. (2017). A high-accuracy map of global terrain elevations. *Geophysical Research Letters*, 44(11), 5844–5853. <https://doi.org/10.1002/2017GL072874>
- Yang, Q., Yuan, Q., Gao, M., & Li, T. (2023). A new perspective to satellite-based retrieval of ground-level air pollution: Simultaneous estimation of multiple pollutants based on physics-informed multi-task learning. *Science of the Total Environment*, 857, 159542. <https://doi.org/10.1016/j.scitotenv.2022.159542>
- Zhang, C., Liu, C., Li, B., Zhao, F., & Zhao, C. (2022). Spatiotemporal neural network for estimating surface NO₂ concentrations over north China and their human health impact. *Environmental Pollution*, 307, 119510. <https://doi.org/10.1016/j.envpol.2022.119510>
- Zhu, F., Ding, R., Lei, R., Cheng, H., Liu, J., Shen, C., et al. (2019). The short-term effects of air pollution on respiratory diseases and lung cancer mortality in Hefei: A time-series analysis. *Respiratory Medicine*, 146, 57–65. <https://doi.org/10.1016/j.rmed.2018.11.019>
- Zhu, Y., Chen, J., Bi, X., Kuhlmann, G., Chan, K. L., Dietrich, F., et al. (2020). Spatial and temporal representativeness of point measurements for nitrogen dioxide pollution levels in cities. *Atmospheric Chemistry and Physics*, 20(21), 13241–13251. <https://doi.org/10.5194/acp-20-13241-2020>