https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

Open Access Earth System Science Data Discussions

High-resolution global ultrafine particle concentrations through a machine learning model and Earth observations

Pantelis Georgiades[1,2], Matthias Kohl[3], Mihalis A. Nicolaou[1], Theodoros Christoudias[2], Andrea Pozzer[2,3], Constantine Dovrolis[1], and Jos Lelieveld[2,3]

[1]Computation-based Science and Technology Research Center (CaSToRC), The Cyprus Institute, Nicosia, Cyprus
[2]Climate and Atmosphere Research Centre (CARE-C), The Cyprus Institute, Nicosia, Cyprus
[3]Department of Atmospheric Chemistry, Max Planck Institute for Chemistry, Mainz, Germany

Correspondence: Pantelis Georgiades (p.georgiades@cyi.ac.cy) and Jos Lelieveld (jos.lelieveld@mpic.de)

Abstract. Atmospheric pollution is a major concern due to its well-documented and detrimental impacts on human health, with millions of excess deaths attributed to it annually. Particulate matter (PM), comprising airborne pollutants in the form of solid and liquid particles suspended in the air, has been particularly concerning. Historically, research has focused on PM with an aerodynamic diameter less than 10 μm (PM_{10}) and 2.5 μm ($PM_{2.5}$), referred to as coarse and fine particulate matter, respectively. The long term exposure to both classes of PM have been shown to impact human health, being linked to a range of respiratory and cardiovascular complications. Recently, attention has been drawn to the lower end of the size distribution, specifically *ultrafine particles* (UFPs), with an aerodynamic diameter less than 100 nm ($PM_{0.1}$). UFPs can deeply penetrate the respiratory system, reach the bloodstream, and have been increasingly associated with chronic health conditions, including cardiovascular disease. Accurate mapping of UFP concentrations at high spatial resolution is crucial considering strong gradients near the sources. However, due to the relatively recent focus on this class of PM, there is a scarcity of long-term measurements, particularly on the global scale. In this study, we employed a machine learning methodology to produce the first global maps of UFP concentrations at high spatial resolution (1 km) by leveraging limited ground station measurements worldwide. We trained an XGBoost model to predict annual UFP concentrations for a decade (2010-2019) and utilized the conformal prediction framework to provide reliable prediction intervals. This approach makes local-to-global UFP data available to support assessments of the health implications associated with long-term exposure.

1 Introduction

The growing concern surrounding atmospheric pollution stems from its well- established, detrimental impacts on human health. Current estimates suggest that air pollution is responsible for many millions of excess deaths annually and a leading contributor to the loss of healthy years of life (Mukherjee and Agrawal, 2017; World Health Organization, 2021). Particulate matter (PM), a diverse category of airborne pollutants, consists of minute particles of solids and liquids suspended in the air, classified based on their aerodynamic diameter. Although historical evidence has long underscored the risks associated with PM exposure, recent global trends have amplified these concerns (Lelieveld et al., 2015; GBD 2021 Risk Factors Collaborators, 2024). The

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

growing population with intensifying industrialization, urbanization, as well as agricultural emissions, have collectively led to a substantial increase in atmospheric PM levels (Alemayehu et al., 2020).

Until recently, the predominant emphasis was on particulate matter (PM) with diameters less than 10 μm (PM_{10}) and 2.5 μm ($PM_{2.5}$), often referred to as *coarse* and *fine* particulate matter, respectively (Kim et al., 2015; Pozzer et al., 2023). Prolonged exposure to these particles has been demonstrated to exert adverse effects on the respiratory and cardiovascular systems. Both PM_{10} and $PM_{2.5}$ possess the capability to reach the lungs, with the smaller particles generally penetrating more deeply, and long-term exposure causes inflammation and oxidative stress, associated with disease risk, including chronic obstructive pulmonary disease (COPD), asthma, lung cancer, strokes, and heart attacks (Crouse et al., 2012; Münzel et al., 2021).

There is growing concern about the health implications of PM smaller than $PM_{2.5}$. At the lower end of the size distribution, ultrafine particles (UFPs) are those with an aerodynamic diameter less than 0.1 μm or 100 nm ($PM_{0.1}$), a subset of $PM_{2.5}$ (Donaldson et al., 2001). Despite constituting a minor proportion of PM2.5 by mass, UFPs dominate in terms of number concentrations. In fact, the total particle number concentration (PNC) is often employed as a proxy measure for the UFP prevalence (Presto et al., 2021). Natural sources of UFPs include new particle formation from inorganic and organic gases emitted by marine and forest ecosystems. The main sources of UFPs relevant to health, though, are anthropogenic and related to combustion processes, such as oil and coal combustion, notably from vehicular, marine and air traffic, and various industrial emissions (Moreno-Ríos et al., 2022).

The small size of UFPs facilitates the deep infiltration into the respiratory system, allowing them to reach the alveoli, transmigrate into the bloodstream and thereby cause adverse health effects in the vasculature and distant organs (Ohlwein et al., 2019). The large number and large surface-to-mass ratio of UFPs may expedite interactions with biological tissue, potentially instigating inflammatory responses and oxidative stress. These molecular interactions have been implicated with several health conditions, including respiratory and cardiovascular diseases, as well as carcinogenesis (Schraufnagel, 2020).

Fine-grained maps of UFP concentrations are necessary for epidemiological assessments aiming at unraveling relationships between air pollution and public health outcomes (Marval and Tronville, 2022). High-resolution mapping enables researchers to conduct detailed spatial analyses, identify vulnerable populations, and understand the complex interplay between environmental factors and health. Such maps are fundamental for policymakers to formulate targeted interventions and regulatory policies to reduce UFP exposure and mitigate associated health risks effectively (Kwon et al., 2020).

The investigation of UFPs and their impact on human health is hindered by the scarcity of measurements, especially at the global scale. Existing monitoring systems lack the spatial coverage necessary for a comprehensive understanding of UFP distributions and determining long-term exposure. Furthermore, the intricate nature of UFPs, characterized by their small size and dynamic behavior, poses challenges for traditional measurement techniques (Trechera et al., 2023). The recent literature on estimating long-term mean, spatially distributed UFP concentrations largely depends on two main methodologies: land use regression models and chemical transport models. Each of these approaches, however, comes with limitations that impact their effectiveness in various contexts.

Land use regression models are known for their ability to provide high spatial resolution, making them particularly useful for detailed local analyses. However, their utility is confined to specific geographic regions with good coverage of UFP mea-

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

surements. The reliance on local data and the necessity for model training procedures to be tailored to the particularities of each area was highlighted in studies by Saha (2021) and Jones (2020) (Saha et al., 2021; Jones et al., 2020). Such dependence on localized data sources and custom training means that extending these models beyond their original scope can be challenging. Chemical transport models offer an option to the issue of geographical coverage, as they are designed to achieve broader spatial up to global applicability. However, this extensive coverage comes at the cost of spatial resolution due to computational constraints. Typically, these models operate at coarse resolution, typically in the range of 10 to 100 kilometers (Kohl et al., 2023). This can obscure the details of UFP distributions near sources, particularly in densely populated urban areas where strong local UFP emissions are associated with steep concentration gradients.

To overcome these limitations, we employed machine learning (ML) methodologies to map UFP concentrations at high spatial resolutions on a global scale. By leveraging the limited ground station measurements available worldwide and incorporating diverse auxiliary information, such as land use, degree of urbanization, built-up volume, and anthropogenic emissions, we trained an Extreme Gradient Boosting (XGBoost) regression model. XGBoost is a well-established and widely used algorithm, chosen for its robustness and efficiency, to predict annual average UFP concentrations at a 1 km spatial resolution over land. Furthermore, the XGBoost model was integrated in a conformal prediction framework, a statistical method that provides robust and distribution-agnostic confidence intervals for ML models (Johansson et al., 2014; Barber et al., 2021; Kim et al., 2020). This integration enabled the calculation of 95% confidence intervals for the model predictions. Additionally, we implemented the SHAP (SHapley Additive exPlanations) method to provide explainability for the ML model and investigate how the model reaches its predictions in different locations with varying characteristics. This innovative approach not only addresses current data gaps but also lays the groundwork for comprehensive, data-informed assessments of the health implications associated with UFP exposure.

In the following sections, we pursue a detailed examination of long-term mean UFP concentrations, achieved through the development of a ML model designed to estimate UFP levels at high spatial resolution. Section 2, *Methods and Data*, provides an in-depth description of the datasets utilized in this study, detailing the methods employed to construct the datasets for training the ML model and conducting inference. The ML methodologies applied are outlined, including the evaluation criteria employed to assess model performance. In Section 3, *Results and Discussion*, we present and analyze the outcomes from the ML model. Finally, in Section 4, *Conclusion*, we synthesize the key insights from our study, emphasizing the implications for future research and policy considerations in addressing the UFP exposure and human health nexus.

2 Data and Methods

This section comprises two parts. In the first part, we discuss the data sources and the methods employed to standardize and homogenize them, detailing the steps taken to create the training and inference datasets. In the second part, we provide the specifics of our model, describing the training procedures and model performance evaluation using relevant metrics.

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024


Open Access Earth System Science Data Discussions

2.1 Data

2.1.1 Ultrafine particle concentrations

In acquiring the target variable for our ML model, we employed an approach fusing data from distinct sources. Initially, we accessed the EBAS and NOAA databases to retrieve particle size distribution and PNC data from ground stations in Europe and North America. For particle size distribution data, measurements of PM with a diameter less than 100 nm were aggregated to represent UFPs. PNC data served as a proxy for the UFP concentration. To ensure data representability, stations with fewer than 200 unique days of data per year were excluded from our analysis. The database was queried for data from 2000 to 2020.

Additionally, we conducted an extensive literature review to supplement the ground station data with information derived from published scientific articles presenting yearly averages worldwide (Aalto et al., 2005; Kohl et al., 2023; Saha et al., 2021). This literature review aimed to enhance the comprehensiveness of our dataset by integrating findings from diverse geographical locations and monitoring networks. By synthesizing data from the EBAS and NOAA databases and the scientific literature, we constructed a comprehensive and representative dataset for training and evaluating our machine-learning model.

2.1.2 Land Cover

The land cover maps published by the Copernicus Global Land Use (CGLS) were used to characterise the global grid in terms of land cover (Buchhorn et al., 2020). A total of 22 distinct land cover classes are available, which were mapped to 7 semantically similar classes, notably to summarize forest ecosystem types, being less relevant for pollution exposure, following the procedures in (Ashiotis et al., 2023). Land use is considered a critical determinant of UFP concentrations, i.e., the emissions strengths of UFPs and their precursors. Given the strong association between UFPs and combustion emissions, urban and industrial areas are anticipated to exhibit a pronounced correlation with elevated UFP levels (Garcia Marlès et al., 2023).

Table A1 shows the grouping for the land cover classes. The datasets are provided with an annual temporal resolution and 100 m spatial resolution. To convert the different land use cover classes into a tabular format for training and inference purposes, we identified the 100 land use grid cells corresponding to the 1 km grid cells. Subsequently, we computed the percentage coverage for each of the 7 mapped classes.

2.1.3 Global Human Settlement Layer

The Global Human Settlement Layer (GHSL) by the European Commission offers open and freely accessible data and tools for evaluating human presence and activities. The global built-up volume (GHS-BUILT-V) dataset was employed at a 5-year temporal steps and 1 km spatial resolution (Pesaresi and Politis, 2023). This dataset includes both residential and non-residential buildings to encompass industrial and commercial complexes. Datasets such as the degree of urbanization (GHS-SMOD) (Schiavina et al., 2023b) and human settlement (GHS-POP) (Schiavina et al., 2023a) were used at the same temporal and spatial resolution. The GHSL datasets were employed to provide insight into anthropogenic activities and industrialization

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

indicators, such as instances where a high built-up volume coincides with a low population density, potentially signaling the presence of industrial zones or other high-emission activities.

No temporal or spatial interpolations were conducted, and the closest year available for each of the dataset was utilized, as these variables do not change much over time. Given the strong linkage between emissions and human activity, these datasets can serve as proxies for pollution emissions.

2.1.4 Global NO_2 and $PM_{2.5}$

Two global datasets of NO_2 and $PM_{2.5}$ were incorporated into the feature set to determine the yearly average concentration of these air pollutants (Anenberg et al., 2022; van Donkelaar et al., 2021). Both datasets offer a spatial resolution of 0.01°, corresponding roughly to 1 km at the equator and a temporal resolution of one year.

These datasets provide the yearly average concentrations of NO_2 and $PM_{2.5}$ for each grid cell. We note that these datasets were initially generated for epidemiological and health burden studies, similar to the scope of this study.

The base spatial grid utilized throughout this study was constructed on the orthogonal lat-lon grid of the NO_2 dataset. Furthermore, following the latitude range of the PM2.5 dataset, this study spans latitudes ranging from 55°S to 68°N degrees.

2.1.5 Emissions

The gridded distributions of global anthropogenic emissions from the Copernicus Atmosphere Monitoring Service (CAMS) were utilized to obtain combustion-related emissions data (Denier van der Gon et al., 2023). The dataset comprises modified Copernicus Atmosphere Monitoring Service Information for the year 2023, retrieved from the Copernicus Atmosphere Data Store. The global emission inventory from the Copernicus Atmosphere Monitoring Service (CAMS) was utilized to derive proxies to estimate UFP concentrations and consider anthropogenic contributions, by including yearly average emissions of black carbon (BC), carbon monoxide (CO), carbon dioxide (CO_2) and nitrogen oxides (NO_x).

The datasets contained both individual sector emissions and the cumulative sum, with the total variable selected for each species. Emphasis was placed on emissions over land, thus grid cells classified as 100% ”open sea” when combined with the land cover data were excluded. Notably, only emissions resulting from combustion processes were considered for this study.

Data retrieval involved obtaining datasets at a spatial resolution of 0.1° and monthly temporal resolution. Subsequently, the yearly averages per grid cell were calculated utilizing the *resample* method of the Python *xarray* library. Spatial interpolations were performed to redistribute the emissions in each grid cell with respect to population density and built-up density, as described in section 2.1.8.

2.1.6 Temperature

The fifth-generation ECMWF reanalysis for global climate and weather, ERA5, served as the source for the temperature feature in our analysis, which may be viewed as a proxy for meteorological conditions. Specifically, the 2m temperature (*t2m*) variable was obtained format from the Copernicus Climate Change Service (C3S) Climate Data Store (CDS), at a spatial resolution of

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

0.25° and hourly temporal resolution (Hersbach et al., 2023). Temperature was included as a parameter due to its potential to influence and reflect atmospheric processes. Temperature can also affect UFP formation and growth through photochemical oxidation of volatile organic compounds (VOCs) and nitrogen oxides (NO_x), as well as condensation and evaporation of semi-volatile reaction products (Kaur et al., 2022; Zhao et al., 2018). The dataset contains modified Copernicus Climate Change Service information (2023), retrieved from the Copernicus Atmosphere Data Store. Yearly averages for each grid cell were computed using the *resample* method of the *xarray* library in Python 3.11, and no spatial interpolations were applied during this process.

2.1.7 Population

The global population dataset from WorldPop (www.worldpop.org) was incorporated into our analysis (WorldPop, 2018). This dataset provides population counts at a yearly temporal resolution and a spatial resolution of 1km at the a global scale. The data were obtained directly from the organization's website, without any temporal or spatial manipulation.

Data spanning the years 2000 to 2020 were retrieved to ensure a comprehensive temporal coverage for our analysis. The WorldPop population counts dataset serves as a fundamental resource in our study, offering insights into the spatial distribution of human populations across diverse regions worldwide.

2.1.8 Data Homogenization

The NO_2 dataset at 0.01° grid resolution, roughly 1 km at the equator, served as the baseline for establishing a uniform gridded dataset. This dataset functioned as the reference point for aligning the spatial resolution of other datasets, ensuring consistency throughout the training and inference processes. To integrate land use data into the uniform dataset, the 100 grid points within each 1km grid cell were identified. For each land use class, the percentage coverage was extracted, resulting in seven features.

Datasets sharing the same spatial resolution as the NO_2 dataset, such as the $PM_{2.5}$ and the GHSL datasets, were seamlessly integrated into the uniform gridded dataset, ensuring the coherence of the datasets without introducing discrepancies.

To address the spatial resolution disparity between the emissions dataset (10km grid) and other datasets (1km grid), a redistribution process was executed. This process maintained the total emissions within each 10 km grid cell while redistributing them to a 1km resolution. Redistribution was achieved by linearly weighting emissions based on population and built-up volume, ensuring harmonization with other datasets, following Kohl et al. (Kohl et al., 2023). Finally, Table 1 provides a list of the feature set employed in this study. By implementing the aforementioned procedures, we ended up with a dataset comprised of 565 examples of UFP concentration characterised by the set of 18 features, which we used for the training and evaluation procedures.

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

Open Access Earth System Science Data Discussions

Table 1. The input feature set used to train the ML models and during the inference procedures.

No.	Feature Name	Reference	No.	Feature Name	Reference
1	Forest	Buchhorn et al. (2020)	10	Degree of Urbanisation	Schiavina et al. (2023b)
2	Low Vegetation	Buchhorn et al. (2020)	11	Human Settlement	Schiavina et al. (2023a)
3	Inland Water	Buchhorn et al. (2020)	12	NO_2 concentration	Anenberg et al. (2022)
4	Cropland	Buchhorn et al. (2020)	13	$PM_{2.5}$ concentration	van Donkelaar et al. (2021)
5	Urban	Buchhorn et al. (2020)	14	Black Carbon Emissions	Denier van der Gon et al. (2023)
6	Snow/Ice	Buchhorn et al. (2020)	15	Carbon Dioxide Emissions	Denier van der Gon et al. (2023)
7	Open Sea	Buchhorn et al. (2020)	16	Carbon Monoxide Emissions	Denier van der Gon et al. (2023)
8	Population	WorldPop (2018)	17	Nitrogen Oxides Emissions	Denier van der Gon et al. (2023)
9	Build Up Volume	Pesaresi and Politis (2023)	18	Temperature	Hersbach et al. (2023)

2.2 Methodology

2.2.1 XGBoost

In this study, we apply the Extreme Gradient Boosting (XGBoost) algorithm to estimate the UFP concentrations. The XGBoost algorithm was chosen for its computational efficiency, scalability, and recognized track record in performance and flexibility. It utilizes an ensemble tree-based learning scheme, which can effectively handle mixed data types, resist outliers, and model complex, non-linear relationships without overfitting (Chen and Guestrin, 2016; Budholiya et al., 2022; Moore and Bell, 2022).

XGBoost is an implementation of gradient-boosted decision trees designed for speed and performance. The core idea of boosting involves sequentially adding predictors to an ensemble, each one correcting its predecessor's errors. Mathematically, the model is built in a stage-wise manner:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathcal{F}. \tag{1}$$

where $\hat{y}_i$ is the predicted value, K is the number of trees, f_k represents an individual tree, and $\mathcal{F}$ is the space of regression trees. Each tree f_k maps an input x_i to an output, which in our case, is the estimated UFP concentration.

The objective function of XGBoost is composed of a loss function to measure the model fit and a regularization term to penalize the complexity of the model:

$$\mathcal{L}(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2}$$

where l is the loss function (e.g., mean squared error), and Ω is the regularization term, which helps control the model complexity and prevent overfitting. The regularization term is defined as:

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{3}$$

where T is the number of leaves in the tree, w_j is the score on leaf j, and γ and λ are regularization parameters (Chen and Guestrin, 2016).

XGBoost employs an additive training strategy, where new trees are added to the model sequentially, and each new tree is trained to predict the residual errors of the sum of the previous trees. This approach allows XGBoost to efficiently handle large datasets and high-dimensional feature spaces, making it well-suited for our task of estimating UFP concentrations globally.

2.2.2 Conformal Prediction with XGBoost

To assess the prediction performance of the model, we used the conformal prediction statisctical framework, to estimate the uncertainties of the model results. Conformal prediction provides a mechanism to generate statistically valid prediction intervals associated with the results of traditional ML models. Prediction intervals in this framework are distribution-agnostic, unlike similar methods like Natural Gradient Boosting and Gaussian processes, which assume data is normally distributed, an assumption that often fails in real world datasets (Johansson et al., 2014). We used the Model Agnostic Prediction Interval Estimator (MAPIE) library in Python 3.11 to implement conformal predictions with the XGBoost Regressor implementation of the *xgboost* library.

In general, conformal predictions operates by training the base model and calculating the prediction intervals using a holdout set of data. In this study, due to the limited amount of data available, we used the Jackknife+ after Bootstrap method to enhance the robustness of our prediction intervals. This method involves the following steps:

- **Bootstrap Resampling.** In the first step of the process, the training dataset is resampled multiple times (in this case 20), to create several bootstrap samples. The XGBoost regression model is trained separately on each of these samples.
- **Leave-One-Out predictions.** For each bootstrap sample, leave-one-out (LOO) predictions are made, where each instance in the sample is left out once during the prediction process.
- **Nonconformity scores.** The nonconformity of each prediction is assessed by comparing the LOO predictions to the actual values in terms of the mean-squared error. These scores measure how well the predictions conform to the observed data.
- **Interval Calculation.** The distribution of the nonconformity scores across all boostrap samples is used to determine the bounds of the prediction invervals for new data points, based on the desired confidence intervals (in this case α=0.05, or 95% confidence interval).

The jackknife+ after bootstrap approach guarantees a coverage level (the amount of observed data that lie within the predicted confidence intervals of the model) higher that 1-2α for a target coverage level of 1-α, without any a priori assumption on the distribution of the data, where α is the confidence interval (Barber et al., 2021; Kim et al., 2020).

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

2.2.3 Training and Evaluation

To determine the optimal parameters for the model, we first randomly set aside 10% of the dataset as a validation set, ensuring that this portion of the data was not utilized during the hyperparameter tuning process. Subsequently, the remaining data was split into a 90/10 ratio for training and testing, respectively. We performed an exhaustive search in parameter space using the GridSearchCV method of *scikit-learn* and assessed the performance of the model using the test set. The set of parameters for which the grid search was performed were the number of estimators (ranging from 30 to 200), the number of parallel trees (1-15), maximum depth (3-15), learning rate (0.02 - 0.5), subsample ratio of the training instance (0.3-1) and the subsample ratio of columns when constructing each tree (0.3-1).

Once the optimal set of parameters was determined, the performance metrics were evaluated on the validation set to evaluate the model effectiveness. The model was subsequently trained on the entire dataset using the established parameters. This step ensured that the model could effectively capture the underlying patterns and relationships present in the data. Finally, to validate the generalizability of the trained model, a K-fold cross-validation analysis was performed. The dataset was randomly partitioned into K folds, with 90% of the data used for training and 10% for testing in each fold. This iterative process allowed for the evaluation of the model's performance across multiple validation sets, providing a robust measure of its effectiveness in predicting ultrafine particle concentrations at a global scale.

2.2.4 Explainability

To gain insights into the underlying fundamental operation of the ML model, we utilised the SHAP (SHapley Additive exPlanations) method. Shapley values, a commonly used approach from cooperative game theory, assesses the individual contribution of each input feature to a specific prediction, which allow us to identify and quantify the features that contribute the most to the model's output (Lundberg and Lee, 2017). The core concept behind SHAP involves comparing the model prediction for a single data point to what it would have predicted under various hypothetical scenarios, where certain features are "masked out". By aggregating these individual feature contributions, SHAP assigns an attribution value to each feature, indicating its impact on the final prediction (Lundberg et al., 2020).

Mathematically, the model is retrained on all feature subsets $S \subseteq F$, where F is the feature set. The importance value is assigned to each feature that represents the effect on the model output including that feature. To compute this effect, two models are trained, one with the feature present ($f_{S\cup\{i\}}$) and one with the feature witheld (f_S). The predictions from the two models are then compared to for each input $f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)$, where x_S represents the values of the input features in the set S. As the effect of removing a feature is dependent on other features in the model, the preceding differences are computed for all permutations of the feature subset $S \subseteq F \setminus \{i\}$. The Shapley values are subsequently computed as feature attributions and are a weighted average of all possible differences (Lundberg et al., 2020; Pezoa et al., 2023):

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)\right]. \tag{4}$$

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

A positive SHAP value suggests the feature improves the model prediction, while a negative value indicates the feature operates in the opposite direction. The magnitude of the value reflects the strength of the influence (Nohara et al., 2021).

We utilized the SHAP library in Python and the TreeExplainer method to generate beeswarm and waterfall visualisations (Lundberg et al., 2020). These plots served to elucidate the feature attributions in the model and their influence on individual predictions, respectively. SHAP is a model-agnostic framework that computes feature attributions, explaining how each feature contributes to a specific prediction. In this case, the TreeExplainer method leverages tree-based ML models to calculate these attributions. It does so by creating a set of decision trees that mimic the behavior of the original model. By analyzing how each feature splits the data within these trees, the explainer can determine the contribution of each feature to the final prediction.

Beeswarm plots visually represent the distribution of SHAP feature attributions for all instances in the dataset. Each data point is depicted as a dot, and its position along the x-axis corresponds to the feature value, while its location on the y-axis signifies the SHAP value. This visualization facilitates understanding the overall impact of each feature on the model predictions. Features with a wider spread of SHAP values across the y-axis indicate a greater influence on the model outputs, while features that cluster around zero have a minimal impact. Waterfall plots go deeper by explaining the predictions for individual data points. They illustrate the sequential effect of each feature on the model prediction, starting from the base value and progressing towards the final prediction. Each feature's contribution is depicted as a bar, either increasing or decreasing the value based on its positive or negative SHAP value. By analyzing these plots, we gain insights into how specific features interact and collectively influence the model predictions for individual instances.

3 Results and Discussion

This section presents the outcomes of our study, including both the presentation of results and in-depth discussion related to estimating the global distribution of Ultrafine Particles (UFPs) at high spatial resolution. We begin by showcasing the efficacy of our training and inference procedures in generating fine-grained estimates of UFP levels. By highlighting the successful application of our methods despite challenges like sparse monitoring networks and limited spatial coverage, we aim to demonstrate the significance of our approach.

Next, we address explainability aspects of our study. Guided by the UFP estimates, we explore the intricate interplay of various environmental, meteorological, and anthropogenic factors influencing UFP concentrations across the globe. This analysis aims to unravel the complex mechanisms controlling UFP distribution and provide valuable insights for future research and mitigation strategies.

3.1 Training and evaluation

In the development of the ML model to estimate the global concentration of UFPs at a high spatial resolution (1 km), the inclusion of environmental, land use, and anthropogenic features like population, built-up volume, and anthropogenic emissions as input variables is grounded in their expected impact on UFP concentrations (Moreno-Ríos et al., 2022). These parameters

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

Open Access Earth System Science Data Discussions

are crucial for capturing the multifaceted nature of UFP distribution and the interplay between natural processes and human activities that influence air quality.

To develop a comprehensive dataset comprising yearly averages of UFP concentrations at various global locations, we accessed open-access databases such as the European Monitoring and Evaluation Programme (EMEP) through the EBAS database and the National Oceanic and Atmospheric Administration (NOAA). Additionally, an extensive review of the literature was undertaken. From these sources, a total of 565 examples were collated and utilized to construct the training set for our ML model. Fig 1 shows the locations of the measurements included in the constructed dataset. The locations are shown in orange circles, where the diameter of the circles corresponds to the typical mean UFP concentrations at each location, illustrating the wide range of ambient conditions.

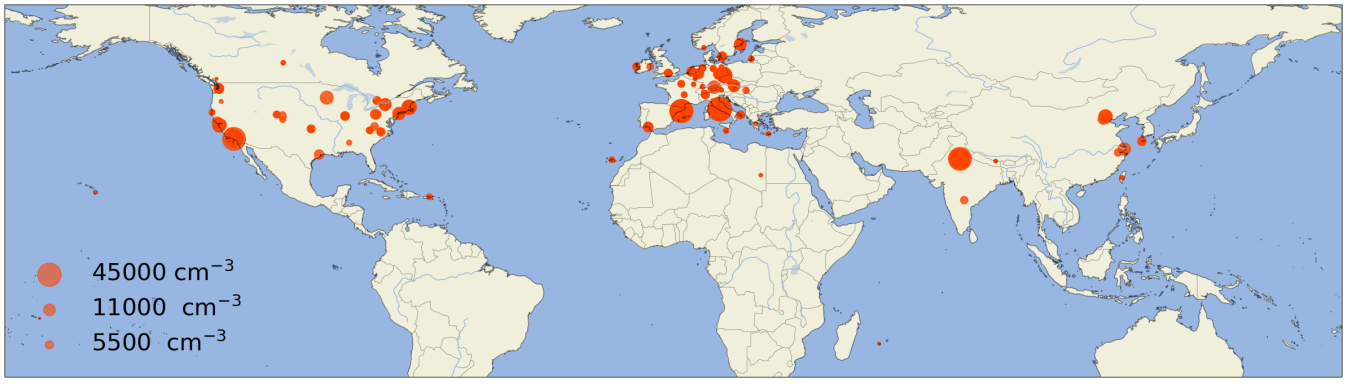

Figure 1. Map depicting the locations of examples in the constructed dataset. Each location is represented by an orange circle, with the diameter of the circle indicative of the mean UFP concentration.

The dataset was partitioned into training and validation subsets, adhering to an 90-10 split, to facilitate the development of the model. To derive the optimal model parameters, we conducted a comprehensive grid search across the parameter space, using an 90/10 training/test split on the remaining data. This optimization process resulted in the identification of the optimal parameter configuration for our XGBoost model: a total of 200 estimators, 1 parallel tree, a maximum depth of 5 a learning rate of 0.075, and 0.7 for the *subsample* and *colsample_bytree* parameters.

Following the grid search procedures, the derived optimal parameters were used to train the algorithm on the whole dataset. To further evaluate the model's ability to generalize to unseen data, a 10-fold cross-validation procedure was implemented. Fig 2 illustrates the comparison between the model predictions and the ground truth (observed) values across the full dataset. The evaluation of the ML model revealed a Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 of 956 cm^{-3}, 2102 cm^{-3}, and 0.956, respectively, for the evaluation set. Furthermore, these metrics were found to be 1270±274 cm^{-3} (MAE), 2424±836 cm^{-3} (RMSE), and 0.896±0.061 (R^2) during the 10-fold cross-validation, indicating a slight variance in the model's prediction accuracy across different folds of the validation process.

The global 95% confidence interval of the model prediction was calculated to be approximately 2000 cm^{-3} (up to a maximum of 13000 cm^{-3}). To accommodate the spatial heterogeneity of rural, suburban, urban, and highly populated urban regions,

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

Open Access Earth System Science Data Discussions

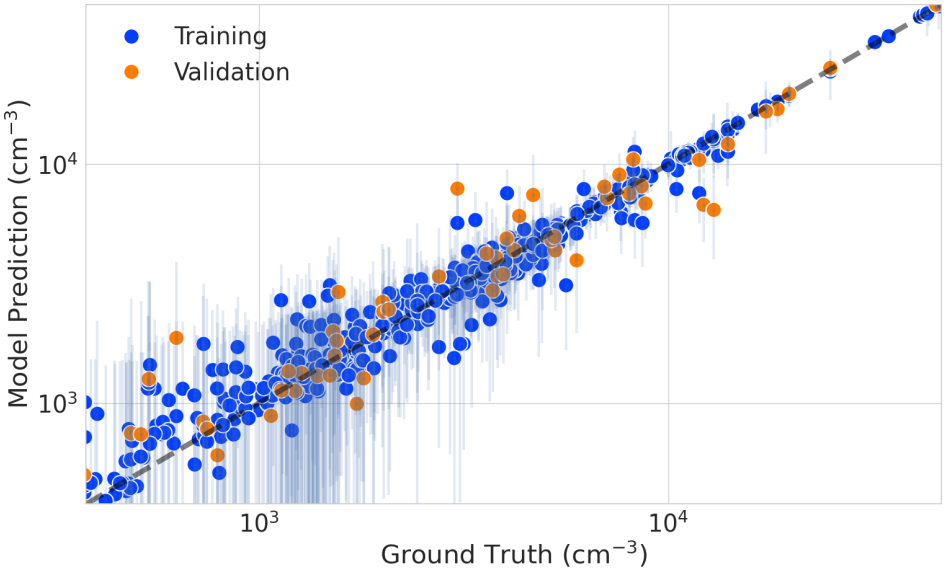

Figure 2. Comparison of observations with model predictions: Blue points represent the entire training dataset, and orange points denote the evaluation set, not seen during training. Vertical lines on each point represent the 95% confidence interval for the prediction.

thresholds of 250, 800, and 1900 inhabitants per km^2 were employed, respectively (Hanberry, 2023). Applying these thresholds, our model achieved percentage errors of approximately 50%, 43%, and 31%, respectively, for these regions. Notably, the percentage errors are considerably lower in the populated areas which represent the highest public health interest.

The selection of the conformal prediction framework for our model was based on its ability to provide statistically robust prediction intervals. Despite a potential limitation of producing wide prediction intervals (Fontana et al., 2020; Kivaranovic et al., 2020), conformal prediction ensures the integrity of outcomes under various distributional changes. We adopted the jackknife+ after bootstrap method over simpler approaches like split or naive conformal prediction. This method not only guarantees coverage of the specified level $1-2\alpha$ and produces non-constant intervals that adapt to data heterogeneity, but also offers a higher sensitivity to epistemic uncertainty (Kim et al., 2020). The computational overhead, although significant compared to the naive and split conformal methods, is justified by its superior performance in ensuring robust interval estimations and accommodating the critical public health relevance of predicted concentrations of UFPs.

Examining the percentage error provides further insights into the model performance, also in regions with minimal human habitation. Desert areas like the Sahara, the Arabian Peninsula, and parts of Australia exhibit notably high percentage errors ($\geq$100%), indicating potential limitations in model accuracy in these sparsely populated regions. However, it is crucial to contextualize these findings, considering that these areas may have limited relevance in terms of UFP exposure. Moreover, the exaggerated percentage error in certain regions is partly attributed to the alignment of elevated error rates with relatively low UFP predictions. In other words, regions with high percentage errors typically coincide with areas characterized by lower

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

predicted UFP concentrations. This underscores the challenge of accurately modeling UFP concentrations in regions with sparse empirical data.

Despite the observed increase in error metrics in sparsely populated areas, the results suggest that the selected features for training the model encapsulate critical information necessary for estimating UFP concentrations and human exposure. The model ability to generalize to unseen data is underscored by the fact that, even with increased error margins in the 10-fold cross-validation experiment, the overall performance remains within a reasonable range for practical applications in estimating UFP concentrations. This generalization capability is pivotal, especially considering the complex nature of atmospheric processes influencing UFP dynamics and the inherent variability in environmental data. The performance metrics, despite being higher for unseen data, align with expectations given the challenges associated with modelling atmospheric particulate matter.

The annual mean background concentration of UFPs at the global scale, i.e., in relatively sparsely to densely populated areas, is estimated to range from 2500 to 3000 cm^{-3} in rural settings, up to 45000–50000 cm^{-3} in highly urban areas, where vehicular emissions and other anthropogenic sources are predominant (Schraufnagel, 2020). While epidemiological research has yet to provide a basis for thresholds or guidelines at which exposure to UFPs should be considered adverse, the World Health Organization qualifies exposure to be of emerging concern in view of growing evidence of health impacts. Research increasingly suggests a pronounced association between UFPs and health outcomes (Weichenthal et al., 2017; Pieters et al., 2015; Olsen et al., 2014). Notably, our model demonstrated good performance in highly populated regions, which are of the highest epidemiological interest, emphasizing its potential as a valuable tool for assessing the impacts of UFP exposure on human health.

3.2 High resolution UFP maps

Following the evaluation of the model performance metrics, we create detailed global maps depicting ultrafine particle (UFP) concentrations between the year 2010 to 2019, with a spatial resolution of 1 km. We note that this period can be conceived as representative of long-term exposure. Fig 3 shows the global distribution of UFPs at a spatial resolution of 1 km, as predicted by our model. The zoomed views in the figure highlight several cities around the world, where the intracity distribution of UFPs is discernible, due to the high spatial resolution of our data.

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

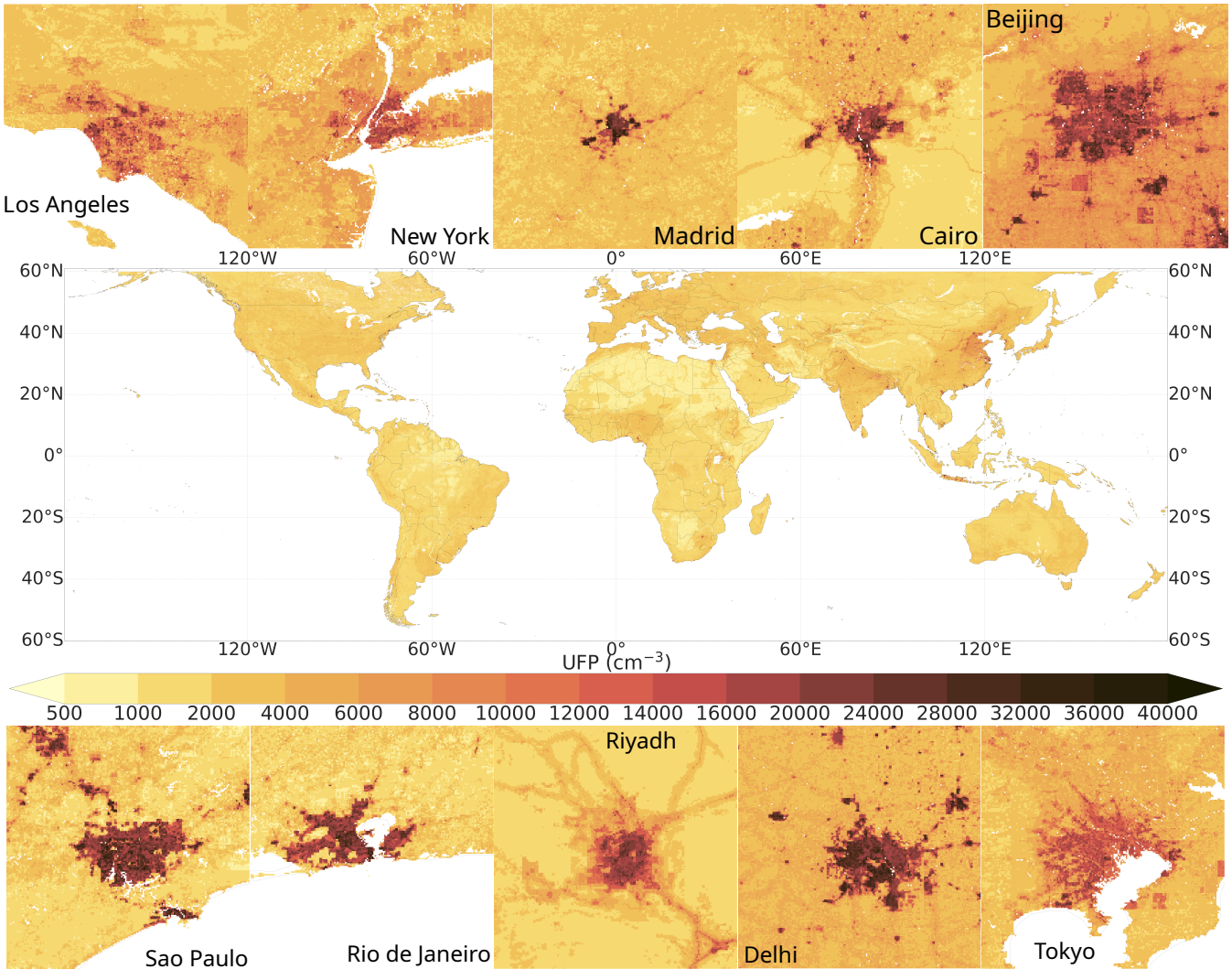

Figure 3. Global map of UFP concentrations for the year 2015. The zoomed views represent urban conglomerations around the world, illustrating the fine-scale distributions of UFPs.

The annual maps generated in this study also allow for the annual examination of UFP levels and exposure, and can be used to gain insights into the spatiotemporal patterns of urban and rural air quality and pollution trends. Such insights can be subsequently employed in guiding the development of targeted environmental and public health policies. Moreover, the annual updates of these maps provide a dynamic view of how pollution patterns change over time, highlighting areas of improvement or concern. By offering a solid foundation for informed decision-making and effective pollution mitigation strategies, these maps are available for policymakers, researchers, and urban planners. Ultimately, this temporal data offers potential for targeted interventions, promoting healthier environments and mitigating adverse health effects of air pollution.

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

While land use regression models and chemical transport models contribute valuable insights into the study of UFP concentrations, their limitations in local-to-global applicability and limited spatial resolution, respectively, highlight the need for innovative approaches that can combine the high spatial resolution of land use regression models with the broad geographical applicability of global chemical transport models.

Our ML approach incorporates diverse auxiliary information alongside land use data, enabling high spatial resolution (1 km) estimates while maintaining broader applicability compared to solely land use-based models. This approach bridges the gap between the high resolution of land use models and the broader coverage of global models, offering a valuable tool for studying UFP distributions at the scales needed to characterize air quality and exposure.

The comprehensive datasets of global UFP concentrations that we generated are accessible in the NetCDF format. The NetCDF files include yearly data from 2010 to 2019, offering the mean UFP concentrations and the associated 95% confidence intervals (estimated within the conformal prediction framework) for each 1x1 km grid cell.

3.3 Explainability

To distinguish how individual input variables influence predictions, we employed SHAP (SHapley Additive exPlanations) analysis. Figure 4 presents a summary plot generated using the SHAP framework. It summarizes the distribution of SHAP values for each feature across all observations, effectively ranking them by their overall impact on the model output. Absolute SHAP values, regardless of sign, indicate the input variable significance, with larger values corresponding to stronger influence. Additionally, the plot describes the directionality of feature effects: red signifies features that increase the predicted outcome, while blue represents negatively associated features.

The investigation of UFP concentrations utilizing conformal predictions with an XGBoost base estimator, complemented by SHAP for feature importance, shows the complex interplay between various predictors and UFP levels. This analysis not only sheds light on the primary contributors to UFP proliferation but also on their intricate dynamics within environmental systems.

Emissions as Leading Features: The SHAP summary plot reveals the significant influence of emission sources, particularly of black carbon, CO, and NO_x, on the yearly concentrations of UFPs at the global scale. Black carbon, largely released into the air as UFPs, emerges as the most critical predictor, illustrating its direct association with combustion processes as primary UFP sources (Junkermann and Hacker, 2022; Ungeheuer et al., 2022; Garcia Marlès et al., 2023). This direct linkage is consistent with the physical understanding that combustion processes, whether from vehicles, industrial activities, and to a lesser extent biomass burning, release a substantial amount of black carbon into the atmosphere. These particles not only contribute to air quality degradation but also have significant climate implications due to their ability to absorb sunlight, warm the atmosphere and cool the surface (Eckhardt et al., 2023).

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

Open Access Earth System Science Data Discussions

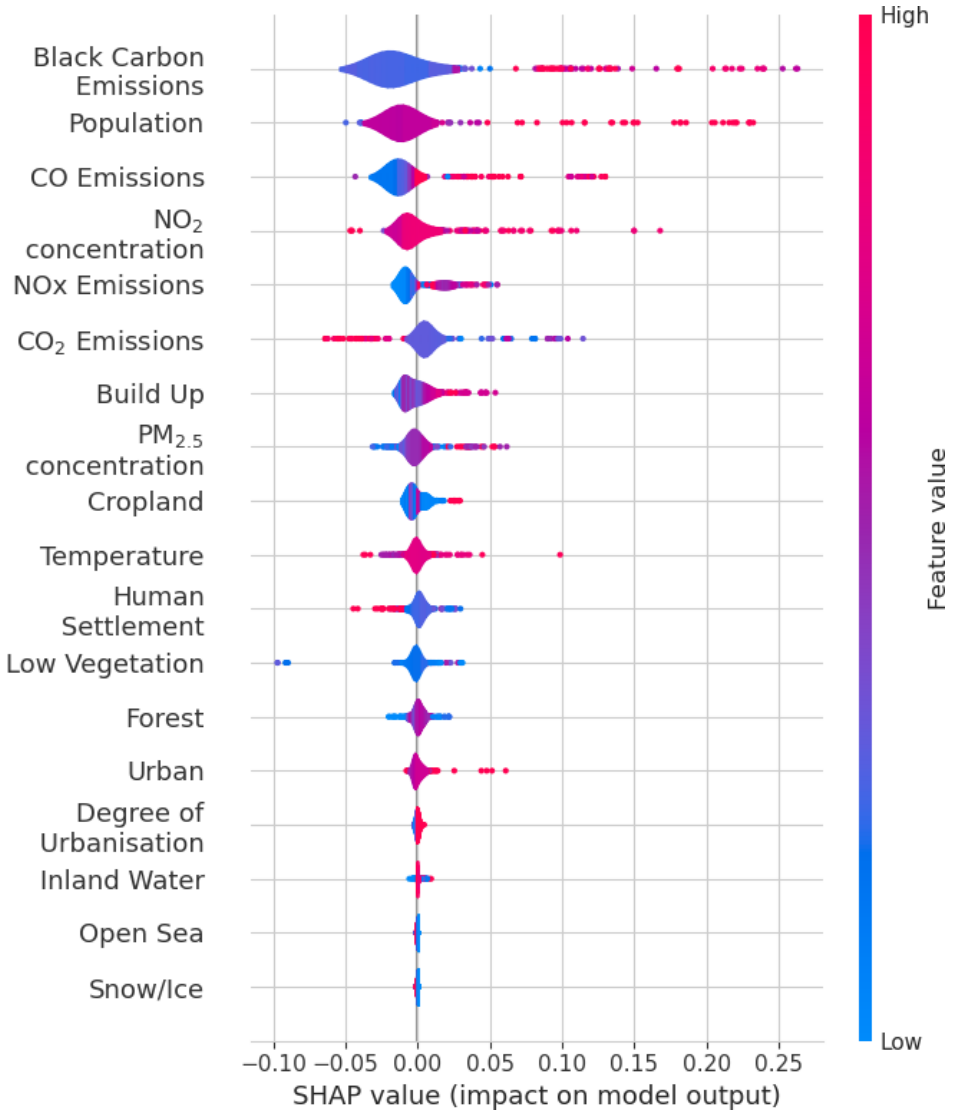

Figure 4. Feature importance for the UFP prediction model based on SHAP summary analysis. Variables are ranked in descending order of their absolute SHAP value, indicating their overall contribution to the model predictions. Colour coding represents the directionality of feature effects: red signifies higher feature values associated with increased UFP predictability, while blue indicates a negative association.

The dual impact of black carbon, exhibiting both a strong positive effect on UFP levels and a negative contribution at relatively low concentrations, suggests a nonlinear relationship between emissions and concentrations, most significantly impacting UFP levels close to the emission location.

Similarly, CO and NO_x are identified as having a substantial positive coincidence on UFP concentrations, though their influence is somewhat less than of black carbon. These gases are also closely associated with combustion processes, serving as precursors to a range of secondary pollutants, including particulate matter. The presence of NO_x, in particular, plays a crucial

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

role in atmospheric reactions leading to the formation of secondary particulate matter, also by enhancing the oxidation of gases which could form particles. The finding that NO_x can negatively impact UFPs at lower concentrations may reflect a similar interplay between emissions, atmospheric chemistry and particle dynamics as black carbon. It underscores the importance of NO_x in both the formation of secondary inorganic and organic aerosols and affecting oxidation processes that induce particle nucleation.

Population Density: Population ranks as the second most influential feature, with its impact being most pronounced at very high UFP levels. This finding aligns with the expectation that higher population densities are associated with increased traffic and industrial emissions, leading to elevated UFP levels. The strong correlation between population density and UFP concentrations emphasizes the role of human activities in both the cause and exposure to air pollution.

Anthropogenic and Urban Features: Other anthropogenic features, such as built-up volume, degree of urbanization, and urban land cover class, positively impact the UFP prediction, albeit to a lesser extent than emissions and population. This underscores the relationship between urban infrastructure and land use patterns and air pollution. The negative impact of low built-up volume on model output illustrates that less urbanized areas have lower UFP levels due to reduced or absent direct sources of pollution.

Role of Ambient Pollutants: The positive contributions of NO_2 and, to a lesser degree, $PM_{2.5}$ concentrations to the model output reinforce the understanding that UFP concentrations directly relate to pollution emissions. Often co-emitted with UFPs from similar sources, these parameters serve as proxies for areas with high UFP levels. However, the association between UFPs and $PM_{2.5}$ is confounded by the fact that at high concentrations PM2.5 provides an increased surface area for coagulation with UFPs and condensation of UFP precursor gases. Furthermore, in desert areas the sources of atmospheric mineral dust particles and UFPs are unrelated.

Temperature: Positioned roughly in the middle of the feature importance list, temperature exhibits a net positive impact on predicted UFP concentrations. This finding may reflect the influence of temperature on atmospheric chemistry and the transport of pollutants, including the relationship between a high temperature and low wind speed in high-pressure regimes during summer. Moreover, the volatility of condensable materials that nucleate and grow secondary particles is sensitive to temperature.

Natural Land Covers' Diverse Impacts: Interestingly, land cover classes such as Open Sea, Snow and Ice, and Forest are ranked low in importance but negatively impact the model output when their values are high. This inverse relationship contrasts with the positive impacts observed for anthropogenic features, even suggesting a role of natural ecosystems in potentially mitigating UFP levels. These land covers may represent areas with lower human activity and hence lower sources of UFPs, and additionally influence the dispersion and deposition of particles.

In summary, the SHAP analysis demonstrates the multifaceted interactions between emissions, human activities, environmental factors, and UFP concentrations. Understanding these relationships helps identify key drivers of air pollution and provides indications for targeted interventions aimed at reducing exposure to harmful UFP levels.

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

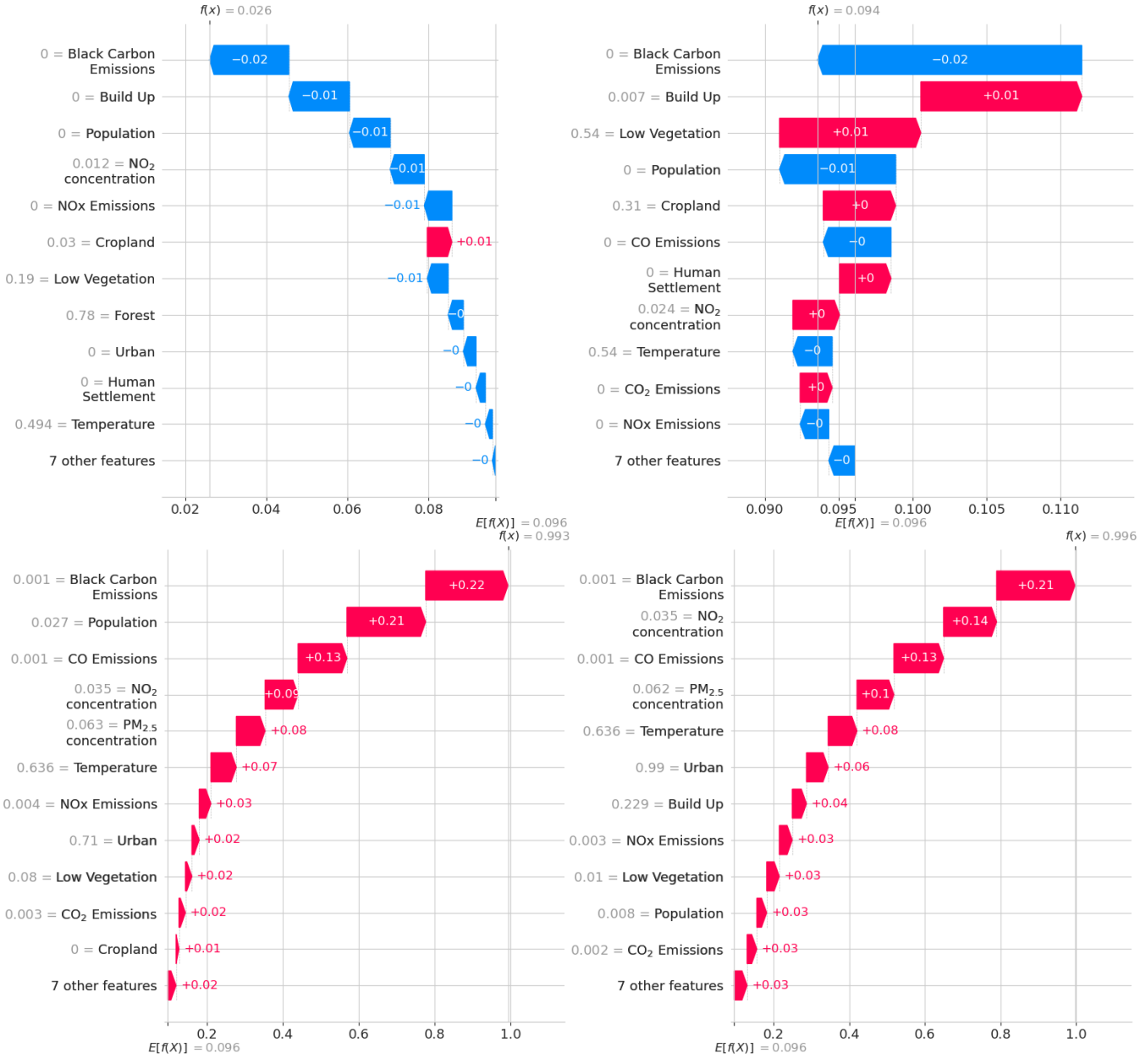

Figure 5. Individual model predictions visualized as waterfall plots. The top left panel shows a prediction for the Oberried region (Germany), where the model predicts relatively low UFP concentrations. The top right panel, for a location at the Diamond Light Source facility in Didcot (UK), shows that the model prediction is close to the expected (mean) values. The bottom SHAP plots show predictions at relatively high UFP levels, one in the city centre of Rome (Italy) and one close to the port in Barcelona (Spain).

Complementing the aforementioned SHAP summary analysis, utilising waterfall SHAP plots for individual predictions by our model provides insightful visualizations of the relative contribution of various predictors. These plots of example

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

locations in Fig. 5 elucidate the interplay of factors influencing UFP levels across distinct locations and conditions, each with unique environmental and anthropogenic characteristics. The SHAP waterfall plots illustrate how the model arrives at specific predictions across a spectrum of ambient UFP concentrations, and offer insights into the influence of various predictors in different regions.

Prediction in Rural/Forested Areas: In the Oberried region of Germany, where a national air quality measurement station (Schauinsland) is located, characterized by rural and forest land cover with negligible urban development or population, the model predicts a low UFP concentration (1200 cm^{-3}). Here, all features except cropland negatively impact the model output. Pollution emissions, including black carbon and NO_x, alongside built-up volume and population, are ranked highly but contribute negatively to UFP predictions. This indicates the model's sensitivity to the absence of anthropogenic activities, consistent with the expectation that in forested areas, without direct pollution sources, UFP levels are relatively low.

Prediction Near Expected Value: The Diamond Light Source facility in Didcot, UK, provides a different scenario. This region, characterized by a significant scientific installation with minimal permanent population but considerable employment, shows a prediction close to the target variable mean (4350 cm^{-3}). The mix of positive and negative feature impacts reflects the unique environmental setup. Notably, black carbon and NO_x emissions negatively influence the UFP prediction, whereas aspects like cropland land cover class, built-up volume, and CO_2 emissions contribute positively. The presence of a large facility with associated human activities amidst an area of low vegetation and limited urban development suggests a complex interplay of factors influencing UFP levels.

High Predictions in Urban Centers: For the highly urbanized and densely populated city center of Rome, Italy, the prediction is relatively high (45600 cm^{-3}), with traffic exhausts contributing heavily to UFP levels. All features, particularly population and combustion-related emissions (black carbon, CO, NO_x and CO_2), contribute positively, underscoring the impact of urban activity and traffic emissions on UFP concentrations. Similarly, for an area near the port of Barcelona, Spain, the prediction is high (45800 cm^{-3}), with emissions and PM2.5 and NO2 concentrations being the most influential positive contributors. The presence of the port and residential areas highlights the significant role of human activities and shipping emissions in UFP pollution.

3.4 Sources of uncertainty

This study applies a novel data science-based methodology to predict UFP concentrations on a global scale at high resolution, yet, ML models need to account for sources of uncertainty, discussed in this section:

- A substantial portion of about 40% of our dataset comprised direct measurements of size distributions, from which we derived UFP data. The remainder of the dataset relies on measurements of PNC, utilized as a proxy of UFP concentrations. Although UFPs are generally understood to dominate PNC by number, this methodology may introduce bias into the model.

- The study of UFPs and their health implications is a burgeoning field within environmental and health sciences. The scarcity of ground stations equipped to measure UFPs, especially over extended time periods, limits the volume and spa-

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

tial coverage of available data. This paucity of data restricts our understanding of UFP distributions and concentrations, particularly in under-monitored low- and middle-income countries.

- Furthermore, the limited dataset also limited the choice of ML models. With a more extensive UFP dataset it would be feasible to employ more highly sophisticated models, such as neural networks that process spatial data directly, rather than relying on the tabular data format used in this study. Such models could potentially offer additional insights into UFP concentrations and their determinants, which will be a future prospect when UFP measurements are included in routine air quality monitoring.

- Data for countries with strong pollution emissions and anticipated elevated and widespread exposure to UFP concentrations, such as India and China, are becoming available but are still scarce. The paucity of data from these regions poses a challenge to the robustness and accuracy of our model predictions. Given the growing industrial activity and urbanization in many areas, there is a need to increase the spatial and temporal coverage of UFP measurements and share them in open-access data repositories.

4 Summary & Conclusions

In this research, we conducted an exhaustive exploration of various data sources, e.g., Earth observations, aimed at characterizing exposure to UFPs, drawing from open-access databases and academic literature. This enabled us to construct a robust ML model, informed by an extensive suite of features, including anthropogenic activities associated with pollution emissions, meteorological conditions and land use characteristics to predict UFP concentrations across the globe at 1 km spatial resolution. Adopting the XGBoost regression model incorporated in a conformal prediction framework was central to our methodology, offering predictive capabilities and uncertainty analysis.

A notable advancement introduced in our study is the integration of explainability of the model results, allowing for an in-depth understanding of the relevance of various features and the model's predictive skill. This level of transparency is essential to raise confidence in the output among researchers and decision-makers pursuing air quality and public health improvements. The result of our study is a dataset that maps global UFP concentrations from 2010 to 2019 on a global scale, which can support public and environmental health research and possibly other applications. This dataset and methodology can offer new insights into the spatial and temporal dynamics of air pollution and innovate informed public health strategies and air quality management policies.

5 Code and data availability

Code to reproduce the results presented in this paper is available in the public github repository https://github.com/pantelisgeor/Ultrafine-Particles. The code provided is written in Python and the workload can be executed by running a series of bash scripts, as documented in the repository description. The code is provided under an MIT lisence, which allows for users to freely use and modify the code.

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

The data produced in this study are available at https://doi.org/10.17617/3.YK9I4B (Georgiades and Pozzer, 2024). Data is provided in NetCDF format, with a yearly temporal resolution and 1 km spatial resolution. Two variables are included in each NetCDF file, UFP (the model prediction for UFP concentration) and CI (UFP $\pm$ CI provides the 95% confidence interval). The datasets are published under a CC BY 4.0 license.

Appendix A: Supplementary Information

Class	Description	Mapped class	Mapped class name
111	Closed forest, evergreen needle leaf	1	Forest
112	Closed forest, evergreen, broad leaf		
113	Closed forest, deciduous needle forest		
114	Closed forest, deciduous broad leaf		
115	Closed forest, mixed		
116	Closed forest, unknown		
121	Open forest, evergreen, needle leaf		
122	Open forest, evergreen broad leaf		
123	Open forest, deciduous broad leaf		
124	Open forest, deciduous broad leaf		
125	Open forest, mixed		
126	Open forest, unknown		
20	Shrubs	2	Low Vegetation
30	Herbaceous vegetation		
60	Bare/sparse vegetation		
100	Moss and lichen		
80	Permament water bodies	3	Inland water
90	Herbaceous wetland		
40	Cultivated and managed vegetation/agriculture (cropland)	4	Cropland
50	Urban/built up	5	Urban
70	Snow and Ice	6	Snow and Ice
200	Open Sea	7	Open Sea

Table A1. Mapping of the CGLS land cover maps to the seven semantically similar classes used in this study.

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

Author contributions. PG wrote the original draft of the manuscript, software and performed formal analysis. All authors were involved in conceptualization of the study, PG, MK and TC performed data curation, PG, MK, TC and MN developed the methodology and investigation procedures. All the authors were involved in reviewing and editing.

Competing interests. The authors declare that that no competing interests are present.

Acknowledgements. We thank the Cyprus Institute's High-Performance Computing Facility for supporting the computational and storage needs of this study. This research was supported by the EMME-CARE project, which received funding from the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 856612 and matching co-funding from the Government of Cyprus. This work has received European Union funding through the European High-Performance Computing Joint Undertaking (JU) and Germany, Bulgaria, Austria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, Greece, Hungary, Ireland, Italy, Lithuania, Latvia, Poland, Portugal, Romania, Slovenia, Spain, Sweden, France, Netherlands, Belgium, Luxembourg, Slovakia, Norway, Tu¨rkiye, Republic of North Macedonia, Iceland, Montenegro, Serbia under grant agreement No. 101101903. This research was also supported by the PREVENT project that has received funding from the European Union's Horizon Europe Research and Innovation Program under Grant Agreement No. 101081276.

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

References

Aalto, P., Hämeri, K., Paatero, P., Kulmala, M., Bellander, T., Berglind, N., Bouso, L., Castaño-Vinyals, G., Sunyer, J., Cattani, G., Marconi, A., Cyrys, J., von Klot, S., Peters, A., Zetzsche, K., Lanki, T., Pekkanen, J., Nyberg, F., Sjövall, B., and Forastiere, F.: Aerosol particle number concentration measurements in five european cities using tsi-3022 condensation particle counter over a three-year period during health effects of air pollution on susceptible subpopulations, Journal of the Air and Waste Management Association, 55, 1064–1076, https://doi.org/10.1080/10473289.2005.10464702, 2005.

Alemayehu, Y. A., Asfaw, S. L., and Terfie, T. A.: Exposure to urban particulate matter and its association with human health risks, Environmental Science and Pollution Research, 27, 27 491–27 506, https://doi.org/10.1007/s11356-020-09132-1, 2020.

Anenberg, S. C., Mohegh, A., Goldberg, D. L., Kerr, G. H., Brauer, M., Burkart, K., Hystad, P., Larkin, A., Wozniak, S., and Lamsal, L.: Long-term trends in urban NO2 concentrations and associated paediatric asthma incidence: estimates from global datasets, The Lancet Planetary Health, 6, e49–e58, https://doi.org/10.1016/s2542-5196(21)00255-2, 2022.

Ashiotis, G., Georgiades, P., Christoudias, T., and Nicolaou, M. A.: Toward Explainable and Transferable Deep Downscaling of Atmospheric Pollutants, IEEE Geoscience and Remote Sensing Letters, 20, 1–5, https://doi.org/10.1109/lgrs.2023.3329710, 2023.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J.: Predictive inference with the jackknife+, The Annals of Statistics, 49, https://doi.org/10.1214/20-aos1965, 2021.

Buchhorn, M., Lesiv, M., Tsendbazar, N.-E., Herold, M., Bertels, L., and Smets, B.: Copernicus Global Land Cover Layers—Collection 2, Remote Sensing, 12, 1044, https://doi.org/10.3390/rs12061044, 2020.

Budholiya, K., Shrivastava, S. K., and Sharma, V.: An optimized XGBoost based diagnostic system for effective prediction of heart disease, Journal of King Saud University - Computer and Information Sciences, 34, 4514–4523, https://doi.org/10.1016/j.jksuci.2020.10.013, 2022.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, ACM, https://doi.org/10.1145/2939672.2939785, 2016.

Crouse, D. L., Peters, P. A., van Donkelaar, A., Goldberg, M. S., Villeneuve, P. J., Brion, O., Khan, S., Atari, D. O., Jerrett, M., Pope, C. A., Brauer, M., Brook, J. R., Martin, R. V., Stieb, D., and Burnett, R. T.: Risk of Nonaccidental and Cardiovascular Mortality in Relation to Long-term Exposure to Low Concentrations of Fine Particulate Matter: A Canadian National-Level Cohort Study, Environmental Health Perspectives, 120, 708–714, https://doi.org/10.1289/ehp.1104049, 2012.

Denier van der Gon, H., Gauss, M., Granier, C., Arellano, S., Benedictow, A., Darras, S., Dellaert, S., Guevara, M., Jalkanen, J.-P., Krueger, K., Kuenen, J., Liaskoni, M., Liousse, C., Markova, J., Prieto Perez, A., Quack, B., Simpson, D., Sindelarova, K., and Soulie, A.: Documentation of CAMS emission inventory products, https://doi.org/10.24380/Q2SI-TI6I, 2023.

Donaldson, K., Stone, V., Clouter, A., Renwick, L., and MacNee, W.: Ultrafine particles, Occupational and Environmental Medicine, 58, 211–216, https://doi.org/10.1136/oem.58.3.211, 2001.

Eckhardt, S., Pisso, I., Evangeliou, N., Zwaaftink, C. G., Plach, A., McConnell, J. R., Sigl, M., Ruppel, M., Zdanowicz, C., Lim, S., Chellman, N., Opel, T., Meyer, H., Steffensen, J. P., Schwikowski, M., and Stohl, A.: Revised historical Northern Hemisphere black carbon emissions based on inverse modeling of ice core records, Nature Communications, 14, https://doi.org/10.1038/s41467-022-35660-0, 2023.

Fontana, M., Zeni, G., and Vantini, S.: Conformal Prediction: a Unified Review of Theory and New Challenges, https://doi.org/10.48550/ARXIV.2005.07972, 2020.

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

Garcia Marlès, M., Alastuey, A., Querol, X., and Hopke, P. K.: Source apportionment of ultrafine particle size distributions in urban Europe, https://doi.org/10.5194/egusphere-egu23-15235, 2023.

GBD 2021 Risk Factors Collaborators: Global burden and strength of evidence for 88 risk factors in 204 countries and 811 subnational locations, 1990-2021: a systematic analysis for the Global Burden of Disease Study 2021, Lancet, 403, 2162–2203, 2024.

Georgiades, P. and Pozzer, A.: Mapping Atmospheric Ultrafine Particles from the Global to the Local Scale., https://doi.org/10.17617/3.YK9I4B, 2024.

Hanberry, B. B.: Urban Land Expansion and Decreased Urban Sprawl at Global, National, and City Scales during 2000 to 2020, Ecosystem Health and Sustainability, 9, https://doi.org/10.34133/ehs.0074, 2023.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., D., S., Simmons, A., Soci, C., Dee, D., and Thépaut, J.-N.: ERA5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), https://doi.org/10.24381/cds.adbb2d47, 2023.

Johansson, U., Boström, H., Löfström, T., and Linusson, H.: Regression conformal prediction with random forests, Machine Learning, 97, 155–176, https://doi.org/10.1007/s10994-014-5453-0, 2014.

Jones, R. R., Hoek, G., Fisher, J. A., Hasheminassab, S., Wang, D., Ward, M. H., Sioutas, C., Vermeulen, R., and Silverman, D. T.: Land use regression models for ultrafine particles, fine particles, and black carbon in Southern California, Science of the Total Environment, 699, 134 234, https://doi.org/10.1016/j.scitotenv.2019.134234, 2020.

Junkermann, W. and Hacker, J.: Unprecedented levels of ultrafine particles, major sources, and the hydrological cycle, Scientific Reports, 12, https://doi.org/10.1038/s41598-022-11500-5, 2022.

Kaur, J., Jhamaria, C., Tiwari, S., and Singh Bisht, D.: Seasonal Variation of Ultrafine Particulate Matter (PM1) and Its Correlation with Meteorological Factors and Planetary Boundary Layer in A Semi-Arid Region, Nature Environment and Pollution Technology, 21, 589–597, https://doi.org/10.46488/nept.2022.v21i02.017, 2022.

Kim, B., Xu, C., and Barber, R. F.: Predictive Inference Is Free with the Jackknife+-after-Bootstrap, 2020.

Kim, K. H., Kabir, E., and Kabir, S.: A review on the human health impact of airborne particulate matter, https://doi.org/10.1016/j.envint.2014.10.005, 2015.

Kivaranovic, D., Ristl, R., Posch, M., and Leeb, H.: Conformal prediction intervals for the individual treatment effect, https://doi.org/10.48550/ARXIV.2006.01474, 2020.

Kohl, M., Lelieveld, J., Chowdhury, S., Ehrhart, S., Sharma, D., Cheng, Y., Tripathi, S. N., Sebastian, M., Pandithurai, G., Wang, H., and Pozzer, A.: Numerical simulation and evaluation of global ultrafine particle concentrations at the Earth's surface, Atmospheric Chemistry and Physics, 23, 13 191–13 215, https://doi.org/10.5194/acp-23-13191-2023, 2023.

Kwon, H.-S., Ryu, M. H., and Carlsten, C.: Ultrafine particles: unique physicochemical properties relevant to health and disease, Experimental & Molecular Medicine, 52, 318–328, https://doi.org/10.1038/s12276-020-0405-1, 2020.

Lelieveld, J., Evans, J. S., Fnais, M., Giannadaki, D., and Pozzer, A.: The contribution of outdoor air pollution sources to premature mortality on a global scale, Nature, 525, 367–371, https://doi.org/10.1038/nature15371, 2015.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems, edited by Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., vol. 30, Curran Associates, Inc., https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf, 2017.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence, 2, 2522–5839, 2020.

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

Marval, J. and Tronville, P.: Ultrafine particles: A review about their health effects, presence, generation, and measurement in indoor environments, Building and Environment, 216, 108 992, https://doi.org/10.1016/j.buildenv.2022.108992, 2022.

Moore, A. and Bell, M.: XGBoost, A Novel Explainable AI Technique, in the Prediction of Myocardial Infarction: A UK Biobank Cohort Study, Clinical Medicine Insights: Cardiology, 16, 117954682211 336, https://doi.org/10.1177/11795468221133611, 2022.

Moreno-Ríos, A. L., Tejeda-Benítez, L. P., and Bustillo-Lecompte, C. F.: Sources, characteristics, toxicity, and control of ultrafine particles: An overview, Geoscience Frontiers, 13, 101 147, https://doi.org/10.1016/j.gsf.2021.101147, 2022.

Mukherjee, A. and Agrawal, M.: World air particulate matter: sources, distribution and health effects, https://doi.org/10.1007/s10311-017-0611-9, 2017.

Münzel, T., Hahad, O., Sørensen, M., Lelieveld, J., Duerr, G. D., Nieuwenhuijsen, M., and Daiber, A.: Environmental risk factors and cardiovascular diseases: a comprehensive expert review, Cardiovascular Research, 118, 2880–2902, https://doi.org/10.1093/cvr/cvab316, 2021.

Nohara, Y., Matsumoto, K., Soejima, H., and Nakashima, N.: Explanation of Machine Learning Models Using Shapley Additive Explanation and Application for Real Data in Hospital, https://doi.org/10.1016/j.cmpb.2021.106584, 2021.

Ohlwein, S., Kappeler, R., Kutlar Joss, M., Künzli, N., and Hoffmann, B.: Health effects of ultrafine particles: a systematic literature review update of epidemiological evidence, International Journal of Public Health, 64, 547–559, https://doi.org/10.1007/s00038-019-01202-7, 2019.

Olsen, Y., Karottki, D. G., Jensen, D. M., Bekö, G., Kjeldsen, B. U., Clausen, G., Hersoug, L.-G., Holst, G. J., Wierzbicka, A., Sigsgaard, T., et al.: Vascular and lung function related to ultrafine and fine particles exposure assessed by personal and indoor monitoring: a cross-sectional study, Environmental Health, 13, 1–10, 2014.

Pesaresi, M. and Politis, P.: GHS-BUILT-V R2023A - GHS built-up volume grids derived from joint assessment of Sentinel2, Landsat, and global DEM data, multitemporal (1975-2030), https://doi.org/10.2905/AB2F107A-03CD-47A3-85E5-139D8EC63283, 2023.

Pezoa, R., Salinas, L., and Torres, C.: Explainability of High Energy Physics events classification using SHAP, Journal of Physics: Conference Series, 2438, 012 082, https://doi.org/10.1088/1742-6596/2438/1/012082, 2023.

Pieters, N., Koppen, G., Van Poppel, M., De Prins, S., Cox, B., Dons, E., Nelen, V., Panis, L. I., Plusquin, M., Schoeters, G., et al.: Blood pressure and same-day exposure to air pollution at school: associations with nano-sized to coarse PM in children, Environmental health perspectives, 123, 737–742, 2015.

Pozzer, A., Anenberg, S., Dey, S., Haines, A., Lelieveld, J., and Chowdhury, S.: Mortality attributable to ambient air pollution: A review of global estimates, GeoHealth, 7, e2022GH000 711, 2023.

Presto, A. A., Saha, P. K., and Robinson, A. L.: Past, present, and future of ultrafine particle exposures in North America, Atmospheric Environment: X, 10, https://doi.org/10.1016/j.aeaoa.2021.100109, 2021.

Saha, P. K., Hankey, S., Marshall, J. D., Robinson, A. L., and Presto, A. A.: High-Spatial-Resolution Estimates of Ultrafine Particle Concentrations across the Continental United States, Environmental Science and Technology, 55, 10 320–10 331, https://doi.org/10.1021/acs.est.1c03237, 2021.

Schiavina, M., Freire, S., and MacManus, K.: GHS-POP R2023A - GHS population grid multitemporal (1975-2030), https://doi.org/10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE, 2023a.

Schiavina, M., Melchiorri, M., and Pesaresi, M.: GHS-SMOD R2023A - GHS settlement layers, application of the Degree of Urbanisation methodology (stage I) to GHS-POP R2023A and GHS-BUILT-S R2023A, multitemporal (1975-2030), https://doi.org/10.2905/A0DF7A6F-49DE-46EA-9BDE-563437A6E2BA, 2023b.

https://doi.org/10.5194/essd-2024-314
Preprint. Discussion started: 7 August 2024

Schraufnagel, D. E.: The health effects of ultrafine particles, Experimental & Molecular Medicine, 52, 311–317, https://doi.org/10.1038/s12276-020-0403-3, 2020.

Trechera, P., Garcia-Marlès, M., Alaustey, A., and Querol, X.: Phenomenology of ultrafine particle concentrations and size distribution across urban Europe, https://doi.org/10.5194/egusphere-egu23-16079, 2023.

Ungeheuer, F., Caudillo, L., Ditas, F., Simon, M., van Pinxteren, D., Kılıç, D., Rose, D., Jacobi, S., Kürten, A., Curtius, J., and Vogel, A. L.: Nucleation of jet engine oil vapours is a large source of aviation-related ultrafine particles, Communications Earth &; Environment, 3, https://doi.org/10.1038/s43247-022-00653-w, 2022.

van Donkelaar, A., Hammer, M. S., Bindle, L., Brauer, M., Brook, J. R., Garay, M. J., Hsu, N. C., Kalashnikova, O. V., Kahn, R. A., Lee, C., Levy, R. C., Lyapustin, A., Sayer, A. M., and Martin, R. V.: Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty, Environmental Science & Technology, 55, 15 287–15 300, https://doi.org/10.1021/acs.est.1c05309, 2021.

Weichenthal, S., Bai, L., Hatzopoulou, M., Van Ryswyk, K., Kwong, J. C., Jerrett, M., van Donkelaar, A., Martin, R. V., Burnett, R. T., Lu, H., and Chen, H.: Long-term exposure to ambient ultrafine particles and respiratory disease incidence in in Toronto, Canada: a cohort study, Environmental Health, 16, https://doi.org/10.1186/s12940-017-0276-7, 2017.

World Health Organization: WHO global air quality guidelines: particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, Genève, Switzerland, 2021.

WorldPop: Global 1km Population, https://doi.org/10.5258/SOTON/WP00647, 2018.

Zhao, H., He, Y., and Shen, J.: Effects of Temperature on Electrostatic Precipitators of Fine Particles and SO3, Aerosol and Air Quality Research, 18, 2906–2911, https://doi.org/10.4209/aaqr.2018.05.0196, 2018.