https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

AEROSOL Research Discussions

Global fields of daily accumulation-mode particle number concentrations using in situ observations, reanalysis data and machine learning

Aino Ovaska[1], Elio Rauth[2,3], Daniel Holmberg[2], Paulo Artaxo[4], John Backman[5], Benjamin Bergmans[6], Don Collins[7], Marco Aurélio Franco[8], Shahzad Gani[1,9], Roy M. Harrison[10], Rakesh K. Hooda[5], Tareq Hussein[1,11], Antti-Pekka Hyvärinen[5], Kerneels Jaars[12], Adam Kristensson[13], Markku Kulmala[1,14,15,16], Lauri Laakso[5,12], Ari Laaksonen[5,17], Nikolaos Mihalopoulos[18], Colin O'Dowd[19], Jakub Ondracek[20], Tuukka Petäjä[1], Kristina Plauškaitė[21], Mira Pöhlker[22], Ximeng Qi[15,16], Peter Tunved[23], Ville Vakkari[5,12], Alfred Wiedensohler[22], Kai Puolamäki[1,2], Tuomo Nieminen[1], Veli-Matti Kerminen[1], Victoria A. Sinclair[1], and Pauli Paasonen[1]

[1]Institute for Atmospheric and Earth System Research (INAR)/Physics, University of Helsinki, Finland
[2]Department of Computer Science, University of Helsinki, Finland
[3]Department of Remote Sensing, Institute of Geography and Geology, University of Würzburg, Germany
[4]Institute of Physics, University of São Paulo, Brazil
[5]Finnish Meteorological Institute (FMI), Erik Palménin aukio 1, 00560, Helsinki, Finland
[6]Institut Scientifique de Service Public (ISSeP), Liege, Belgium
[7]University of California, Riverside, U.S.
[8]Institute of Astronomy, Geophysics and Atmospheric Sciences, University of São Paulo, Brazil
[9]Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, New Delhi, India
[10]School of Geography, Earth & Environmental Sciences, University of Birmingham, United Kingdom
[11]University of Jordan, School of Science, Department of Physics, Environmental and Atmospheric Research Laboratory (EARL), Amman, 11942 Jordan
[12]Atmospheric Chemistry Research Group, Chemical Resource Beneficiation, North-West University, Potchefstroom, South Africa
[13]Department of Physics, Lund University, Sweden
[14]Aerosol and Haze Laboratory, Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing University of Chemical Technology, Beijing 100029, China
[15]Nanjing-Helsinki Institute in Atmospheric and Earth System Sciences, Nanjing University, Nanjing, China
[16]School of Atmospheric Sciences, Nanjing University, Nanjing, China
[17]Department of Technical Physics, University of Eastern Finland
[18]Department of Chemistry, University of Crete, Greece
[19]School of Physics, National University of Ireland, Galway, Ireland
[20]Research Group of Aerosol Chemistry and Physics, Institute of Chemical Process Fundamentals of the CAS, Prague, Czech Republic
[21]Center for Physical Sciences and Technology (FTMC), Vilnius, Lithuania
[22]Leibniz Institute for Tropospheric Research (TROPOS), Leipzig, Germany
[23]Department of Environmental Science, Stockholm University, Sweden

Correspondence: Aino Ovaska (aino.ovaska@helsinki.fi) and Pauli Paasonen (pauli.paasonen@helsinki.fi)

Abstract. Accurate global estimates of accumulation-mode particle number concentrations (N_{100}) are essential for understanding aerosol–cloud interactions, their climate effects, and improving Earth System Models. However, traditional methods

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

relying on sparse in situ measurements lack comprehensive coverage, and indirect satellite retrievals have limited sensitivity in the relevant size range. To overcome these challenges, we apply machine learning (ML) techniques— multiple linear regression (MLR) and eXtreme Gradient Boosting (XGB)—to generate daily global N_{100} fields, using in situ measurements as target variables and reanalysis data from Copernicus Atmosphere Monitoring Service (CAMS) and ERA5 as predictor variables. Our cross-validation showed that ML models captured N_{100} concentrations well in environments well-represented in the training set, with over 70 % of daily estimates within a factor of 1.5 of observations. However, performance declines in underrepresented regions and conditions, such as clean and remote environments, underscoring the need for more diverse observations. The most important predictors for N_{100} in the ML models were aerosol-phase sulphate and gas-phase ammonia concentrations, followed by carbon monoxide and sulfur dioxide. Although black carbon and organic matter showed the highest feature importance values, their opposing signs in the MLR model coefficients suggest their effects largely offset each other's contribution to the N_{100} estimate. By directly linking estimates to in situ measurements, our ML approach provides valuable insights into the global distribution of N_{100} and serves as a complementary tool for evaluating Earth System Model outputs and advancing the understanding of aerosol processes and their role in the climate system.

1 Introduction

Accumulation-mode particles are aerosol particles ranging from 100 to 1000 nm in diameter. They can be emitted directly in this size range from various natural and anthropogenic sources or form through the growth of particles either emitted in smaller sizes or formed by atmospheric new particle formation (e.g. Morawska et al., 1999). In the atmosphere, accumulation mode particles play a critical role in the climate due to their influence on cloud properties and interaction with atmospheric radiation (Forster et al., 2021).

Cloud formation occurs when an air mass becomes supersaturated, leading to the condensation of water vapor on aerosol particles known as cloud condensation nuclei (CCN), forming cloud droplets (Boucher et al., 2013). Whether a particle can act as CCN at a given supersaturation depends on its composition and size (e.g. McFiggans et al., 2006; Andreae and Rosenfeld, 2008). Particles around 100 nm in diameter are generally large enough to activate as CCN under typical atmospheric conditions regardless of their chemical composition (Dusek et al., 2006; Kerminen et al., 2012; Pöhlker et al., 2021), making the number concentrations of accumulation-mode particles a good estimate for CCN-active particles. Aerosol particles can influence the radiative budget both in direct and indirect ways. The number concentration of CCN-active particles affects the cloud's properties, for example cloud albedo, cloud liquid water path, cloud lifetime, and precipitation properties of clouds (e.g. Twomey, 1977; Albrecht, 1989; Forster et al., 2021; Stier et al., 2024). Additionally, because aerosol particles alter transmittance of radiation in the atmosphere, they can modify the atmospheric temperature profile, impacting the evaporation and condensation processes in the clouds (Forster et al., 2021). Due to the complexity of these interactions, aerosol-cloud interactions remain the largest source of uncertainty in the radiative forcing estimates and future climate projections (Forster et al., 2021).

Understanding the global distribution of accumulation-mode particle number concentrations is essential for improving our understanding of CCN and therefore aerosol-cloud interactions. For example, to reliably assess aerosol effects on clouds,

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

AEROSOL Research Discussions

global CCN concentrations need to be captured within a factor of 1.5 of their true values (Rosenfeld et al., 2014). However, obtaining such accuracy with measurements on a global scale is challenging. Although in situ measurements of both CCN and accumulation mode particle number concentrations are available and crucial for understanding spatial variation, they have limited spatial and temporal coverage (Rosenfeld et al., 2014; Schmale et al., 2018). As a result, global observations rely heavily on satellite remote sensing, which introduces its own set of challenges (e.g. Rosenfeld et al., 2014; Bellouin et al., 2020; Quaas et al., 2020). For example, satellites cannot directly observe the aerosol particle number concentrations. Instead, they often rely on indirect retrievals, like radiation extinction-related variables such as aerosol optical depth (AOD) or aerosol index (AI). Inferring number concentrations from these retrievals is challenging because they relate to the entire columnar burden of particles in the atmosphere and are sensitive to other variables, including relative humidity and aerosol particle size. Moreover, satellites cannot detect aerosol loadings beneath clouds, making it difficult to obtain data under the conditions where these measurements would be most needed.

Accumulation-mode particle and CCN number concentrations also pose challenges for Earth System Models (ESMs). Accurately modeling aerosol growth from small particles to the accumulation-mode size range requires detailed numerical descriptions of complex aerosol dynamics within ESMs (Blichner et al., 2021). This task is both challenging and computationally expensive, leading to simplified physical representations in ESMs, adversely affecting their accuracy. Many ESMs employ bulk-mass aerosol schemes without direct particle number concentration calculations (e.g., Yu et al., 2022). If particle number size distributions are represented in the ESMs, they are typically described with modal aerosol schemes, where distributions are represented by several log-normal modes (e.g. Mulcahy et al., 2020; Blichner et al., 2021; Noije et al., 2021). However, this method involves a priori assumptions about the size distribution that often inaccurately reflect the true size distributions and thus alter the flow of particles growing from one mode to another (Blichner et al., 2021; Bergman et al., 2012; Korhola et al., 2014). These issues can be avoided by using sectional schemes, where size distributions are represented by size bins, but these are more computationally expensive (Blichner et al., 2021).

Given the challenges of directly measuring accumulation-mode particle and CCN concentrations, as well as the limitations of the ESMs, there is a clear need to develop alternative estimation methods. One such method is the recent work by Block et al. (2024), who derived global CCN concentrations using aerosol mass mixing ratios from CAMS reanalysis data (CAMS data is discussed further in Sect. 3.2). These aerosol mass concentrations, constrained by satellite-retrieved AOD, were converted into aerosol number size distributions based on estimated size distributions for each aerosol species. They then applied modified kappa-Köhler theory to calculate the number of particles that activate into CCN at specific supersaturation levels. Their approach provides valuable insight into global CCN concentrations at different supersaturations, constrained by the satellite observations assimilated into CAMS reanalysis data. However, it does not incorporate direct CCN or particle number measurements and relies solely on CAMS reanalysis data.

Nair and Yu (2020) presented an alternative approach utilizing machine learning (ML) to estimate CCN concentrations. They selected 46 sites across the globe and employed a chemical transport model to calculate CCN concentrations along with various predictors, including aerosol, chemical, and meteorological variables at these locations. This dataset formed the basis for training a Random Forest Regression Model, which was then evaluated using CCN observations from the Southern Great

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

Plains (USA) measurement station. Although the method relied primarily on modeled CCN concentrations and predictors, it demonstrated the potential of ML techniques for estimating aerosol number concentrations. A follow-up study utilized a similar approach and estimated particle number concentrations (diameters 1.2-120 nm) using Random Forest Regression Model (Yu et al., 2022).

Another machine learning application, that has gained popularity in atmospheric sciences in recent years, is extending observations from measurement stations to larger geographic areas. This method has been quite commonly employed for estimating PM2.5 and other air pollutant concentrations across local and regional scales (e.g., Ma et al., 2019; Di et al., 2019; Kim et al., 2021; Wang et al., 2022, 2023; Yu et al., 2023). Some methods focus on extrapolating measurements using solely the target measurements from measurement stations with no additional predictors. For example, Ma et al. (2019) utilized a neural-network-based spatial-temporal extrapolation method to estimate PM2.5 concentrations in the state of Washington (USA). However, in most cases, the ML models are trained to estimate the concentrations based on a range of widely available variables, including other air quality measurements, meteorological data, satellite retrievals, geographical and land use information, reanalysis datasets, and outputs from chemical transport models.

In this study, we employ ML techniques to bridge the gap between localized in situ measurements of accumulation-mode particle concentrations and the global scale. We train two ML models – a multiple linear regression model and an eXtreme Gradient Boosted model (described in Sect. 2.1) - using in situ measurements of N_{100} as the target variable and reanalysis variables from the CAMS and ERA5 datasets as predictors (described in Sect. 3). These models generate daily number concentration fields for particles with dry diameters larger than 100 nm (N_{100}). Sect. 4 details our methods for training the ML models and assessing the model performance both at the measurement stations and outside of them. Once trained, we use these ML models to generate daily global N_{100} fields N_{100} for 2013. We also investigate the reliability of the global ML models across different regions based on the influence predictor variables have on the models and how the MLR and XGB model fields differ.

2 Background on machine learning methods

This section contains a brief overview of the methods and the two different ML models we used in this study to estimate N_{100}. Further reading on these methods can be found for example in Kuhn and Johnson (2013). The more detailed description of how we applied these techniques is in Sect. 4.

2.1 ML models

Multiple Linear Regression (MLR) is a simple yet effective method that extends ordinary least squares regression to model the relationship between multiple predictor variables and a single target variable (e.g., Kuhn and Johnson, 2013). It assumes a linear relationship between the set of predictors and the target. The MLR model finds a linear equation consisting of coefficient terms for each predictor variable and a constant term (intercept) to minimize the sum of squared residuals between the predicted and observed values.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

AEROSOL Research Discussions

eXtreme Gradient Boosting (XGB) combines a tree-based ensemble method with gradient boosting (Chen and Guestrin, 2016). In simple terms, XGB trains sequentially weak predictive estimators (decision trees) that, at each step, aim to correct the errors of the previous estimators. The final estimate is calculated as the sum of the decision tree estimates. The number of trees can typically be between 100 and 1000. XGB is used both for regression and classification tasks. Here, we used it for regression with squared error as the loss function.

We chose these two ML models because they complement each other well. MLR is a simple, interpretable model that provides insights into the relationships between predictors and the target variable through the coefficients. It also can extrapolate beyond the range of values in the training data, at least if the relationship with the target variable and the covariates is linear. In contrast, XGB is well-suited for complex, non-linear data and interactions but is more computation-intensive and difficult to interpret. XGB is also more limited in its ability to extrapolate beyond the range of values in the training data, as it predicts constant values far outside the training data. By using both MLR and XGB, we can compare two fundamentally different ML methods. The differences in the estimates produced by the ML models may shed light on the global ML model performance, which is otherwise difficult to assess.

2.2 ML training and evaluation process

2.2.1 Training, validation and holdout sets

A typical supervised learning process, such as regression discussed here, involves two main steps: model training and performance evaluation. A portion of the full dataset, called the training set, is used to train the model. During training, the model learns from both the target and predictor variables in the training set, adjusting its internal parameters to capture their relationship.

Once the model is trained, its performance is evaluated with a portion of the full dataset that is separate from the training set. In the testing phase, the trained model receives only the predictor variables and generates estimates for the target variable. These estimates are then compared against the observed target values to assess model performance. To prevent data leakage and ensure reliable model performance assessment, the datasets used for training and testing the model must remain independent.

Testing the final version of the ML model is commonly done with a holdout set. The holdout set is separated from the training set at the beginning of the analysis and is reserved solely for testing the final version. The allocation of data between training and holdout sets depends on the specific application. This includes how much data is assigned to each set and which data points are selected for training versus testing. When data is limited, allocation must be done carefully to ensure that both sets remain representative. In some cases, it may be preferable to forgo data splitting and train the model on all available data. In these cases, resampling methods such as cross-validation (CV) can be used to evaluate model performance using only the training data.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

AEROSOL Research Discussions

2.2.2 K-Fold Cross-Validation

In CV, the original training set is further divided into smaller groups: a new training set and a validation set, which is now used to evaluate the model performance. We used two types of CV, k-fold CV and spatial CV.

K-fold CV involves dividing the original training set into k groups. One group serves as the validation set, while the remaining groups form the new training set. The model undergoes training and testing iteratively, rotating through each group. The process yields k performance values, and the average of these values is utilized to evaluate the model's performance. The benefit of using CV is that each data point can be used both to train the model and to test its performance while maintaining the separation between the sets to assure reliability.

Given the spatial structure of our dataset, we complemented traditional k-fold CV with spatial CV. In spatial CV, folds are defined based on geographical information (e.g., Cho et al., 2020; Beigaitė et al., 2022)—in our case, by measurement station. Because data from the same location is autocorrelated, including a station's data in both the training and validation sets can lead to overly optimistic performance estimates. Spatial CV mitigates this issue by ensuring greater independence between folds because target station's own data is not used in training.

2.2.3 Model optimization

CV is also used for model optimization, a step prior to training the final model. This phase involves fine-tuning the model to enhance its performance on the specific task. In our case, optimization included feature selection and hyperparameter tuning.

Feature selection refers to selecting a subset of predictor variables (also known as features) for the ML mode. If the dataset contains predictor variables that correlate with each other, having multiple variables with similar information is redundant. It can also cause overfitting, where the model becomes too tailored to the training data and performs poorly on unseen data, making the models less generalizable. The best practice is selecting only the relevant variables.

Hyperparameters (HPs) are user-defined parameters that control the complexity of the model. Increasing complexity can improve the performance on the training data, but it also increases the risk of overfitting. Tuning HPs is essential to find the right balance between model complexity and generalization ability. MLR does not require HPs in its basic form, while XGB involves several important HPs, such as the number of trees, tree depth, learning rate, and regularization parameters (XGBoost Developers, 2022). To optimize the HPs, we employed CV, where different combinations of HPs were evaluated to identify the combination of hyperparameters that yielded the best performance across several validation folds.

2.2.4 Model performance metric

For the performance metric, we used root mean squared error (RMSE) between the log10-transformed observed N_{100} values and the log10-transformed estimated N_{100} values ($RMSE_{log10}$). We used log10-transformed N_{100} values in our analysis because we were interested in capturing the correct order of magnitude rather than exact N_{100} values. Additionally, RMSE is scale-dependent resulting in higher errors for higher N_{100} values. Log10-transformation mitigates this issue.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

The $RMSE_{log10}$ calculated using the training set is referred to as training error, and the $RMSE_{log10}$ calculated with a separate holdout set is referred to as a testing error. A low $RMSE_{log10}$ indicates good performance, whereas higher values indicate poorer performance. In this study, we considered the model performance with $RMSE_{log10}$ below 0.2 to be excellent (at least 70 % of the estimated values were within a factor of 1.5 from the observed values, i.e., between the observed value divided by a factor 1.5 and the observed value multiplied with a factor 1.5), below 0.3 good (at least 50 % of the estimated values were within a factor of 1.5 from the observed values) and above 0.3 poor (below 50 % of the estimated values were within a factor of 1.5 from the observed values).

2.2.5 Feature importance

Interpreting the ML model involves assessing the importance of each variable (also known as features). The estimation of feature importance differs between MLR and XGB models. In MLR models, importance is determined by the coefficients of the variables. When variables have a similar range of values or are scaled, the absolute value of a coefficient indicates its importance, and the sign (positive or negative) shows whether an increase in the variable leads to an increase or decrease in N_{100}. In contrast, XGB models do not have a straightforward method for estimating variable importance; instead, they provide various approaches (XGBoost Developers, 2022). We used the gain method, which evaluates importance based on the accuracy improvement in a branch when a variable is included (XGBoost Developers, 2022).

3 Data description and processing

3.1 Measured N_{100} (target variable)

The dataset contained ground-level in situ N_{100} measurements from 35 measurement stations worldwide (Fig. 1). Depending on the station, the measurements were performed with either a Differential Mobility Particle Sizer (DMPS) (Aalto et al., 2001) or a Scanning Mobility Particle Sizer (SMPS) (Wiedensohler et al., 2012). The dataset contained sub-hourly N_{100} calculated from the number concentration of particles between 100 nm and the upper limit of the measurement instrument, which varied between 400 nm and 1000 nm (Table 1). Because the number concentration of accumulation-mode particles is typically dominated by particles with diameters well below 400 nm (Leinonen et al., 2022), it is very unlikely that the differing upper limits have a notable impact on our results. Further description of each station, the measurement instrument used, and references are in Table 1.

We separated the measurements into training and holdout sets based on temporal division (discussed further in Sect. 4.1). The training set contained observations from 2003 to 2019, with the specific measurement periods and data availability varying across stations (Fig. 2). The shortest available time series spanned 201 days, while the longest extended over 6 182 days, altogether 49 490 data points. The holdout set contained N_{100} measurements for 2020-2022. For this time period, we had data from fewer stations, covering only a subset of European stations with altogether 9 587 data points. The data availability of this testing dataset can be seen in Fig. S1.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

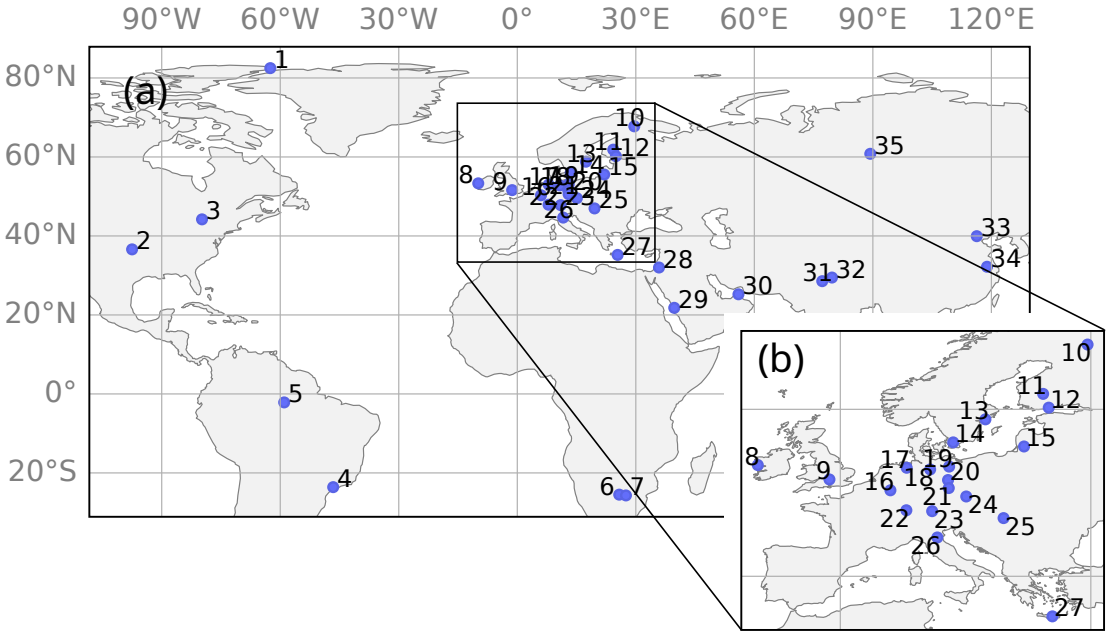

Figure 1. Map of measurement stations. Panel a) shows the map and panel b) zoom-in to Europe. The numbers refer to stations as listed in Table 1.

We processed the N_{100} measurement data by first ensuring all timestamps were in UTC time and then calculating daily averages. To address outliers likely caused by measurement errors, we removed values outside three standard deviations from the station's mean. The observed N_{100} concentrations ranged from a few particles per cm^{-3} to tens of thousands of particles per cm^{-3}. Because the N_{100} concentrations show roughly a lognormal distribution and our aim was to capture the correct order of magnitude rather than exact N_{100} values, we employed log10-transformation for N_{100}.

3.2 Reanalysis data (predictor variables)

Reanalysis data is a gridded dataset created by assimilating observations from various sources, such as in situ measurements and satellite retrievals, into a numerical weather prediction model. In this study, we used reanalysis variables collected from the Copernicus Atmosphere Monitoring Service (CAMS) "CAMS global reanalysis (EAC4)"-dataset (Inness et al., 2019a, b) and "ERA5 hourly data on single levels from 1940 to present"-dataset (Hersbach et al., 2023). Both datasets are generated by the European Centre for Medium-Range Weather Forecasts using the Integrated Forecasting System (IFS) model for numerical weather prediction. CAMS provides global datasets for past atmospheric composition with 3-hourly time resolution and 0.75x0.75-degree spatial resolution. ERA5 offers global datasets for numerous atmospheric variables at hourly time resolution and 0.25x0.25-degree spatial resolution.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

Table 1. List of measurement stations included in this study and their information, including the station number used to identify stations in the figures (No), station full name and country, the abbreviation used in the text, coordinates (latitude, longitude) and altitude above sea level in meters, station environment type, instrumentation, maximum diameter of the measurement, and references. The measurements were conducted either with differential mobility particle sizer (DMPS) or scanning mobility particle sizer (SMPS). Superscripts: a - coastal site, b - mountain site, c - dataset reference Nieminen et al. (2018), d - dataset reference Birmili et al. (2016), e - dataset reference same as measurement site reference.

No	Station	Country	Abbreviation	Latitude	Longitude	Altitude (m a.s.l.)	Environment Type	Instrument	Max. diameter (nm)	Site and Dataset Description
1	Alert	Canada	ALE	82.492	-62.508	75	Polar	smps	500	Leaitch et al. (2013)[c]
2	Southern Great Plains	United States	SGP	36.6	-97.5	300	Rural	smps	750	Marinescu et al. (2019)[c]
3	Egbert	Canada	EGB	44.2	-79.8	251	Rural	smps	420	Pierce et al. (2014)[c]
4	São Paulo	Brazil	SAO	-23.6	-46.6	750	Urban	dmps	800	Backman et al. (2012)[c]
5	Amazonas	Brazil	AMA	-2.146	-59.006	130	Remote	smps	430	Andreae et al. (2015)[e]
6	Botsalano	South Africa	BOT	-25.5	25.8	1400	Rural	dmps	844	Vakkari et al. (2013)[c]
7	Marikana	South Africa	MAR	-25.7	27.5	1170	Urban	dmps	844	Laakso et al. (2008)[c]
8	Mace Head	Ireland	MHD	53.32	-9.88	10	Remote [a]	smps	470	O'Connor et al. (2008)[c]
9	Harwell	England	HRW	51.6	-1.3	126	Rural	smps	450	Charron et al. (2007)[c]
10	Värriö	Finland	VAR	67.76	29.61	390	Remote	dmps	860	Hari et al. (1994)[c]
11	Hyytiälä	Finland	HYY	61.85	24.29	181	Rural	dmps	1000	Hari and Kulmala (2005)[c]
12	Helsinki	Finland	HEL	60.2	24.96	26	Urban	dmps	1000	Järvi et al. (2009)[c]
13	Aspvreten	Sweden	ASP	58.8	17.38	25	Rural	dmps	400	Tunved and Ström (2019)[c]
14	Vavihill	Sweden	VHL	56.04	13.52	172	Rural	dmps	900	Kristensson et al. (2008)[c]
15	Preila	Lithuania	PRL	55.55	22.0	5	Rural	smps	840	Mordas et al. (2016)[c]
16	Vielsalm	Belgium	VIE	50.3	6.0	496	Rural	smps	800	ACTRIS (2024)
17	Bösel	Germany	BSL	53.0	7.95	17	Rural	smps	800	Asmi et al. (2011)[d]
18	Waldhof	Germany	WAL	52.8	10.76	75	Rural	smps	800	UBA (2013)[d]
19	Neuglobsow	Germany	NEU	53.14	13.03	70	Rural	smps	800	UBA (2013)[d]
20	Melpitz	Germany	MLP	51.53	12.9	87	Rural	dmps	800	Engler et al. (2007)[d]

Continued on next page

AEROSOL Research Discussions

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

Continued from previous page

No	Station	Country	Abbreviation	Latitude	Longitude	Altitude (m a.s.l.)	Environment Type	Instrument	Max. diameter (nm)	Site and Dataset Description
21	Annaberg-Buchholz	Germany	ABZ	50.57	12.99	545	Urban	smps	800	Schladitz et al. (2015)[d]
22	Schauinsland	Germany	SCH	47.91	7.91	1205	Rural[b]	smps	800	UBA (2013)[d]
23	Hohenpeissenberg	Germany	HPB	47.8	11.0	988	Rural[b]	smps	800	Birmili et al. (2003)[d]
24	Košetice	Czech Republic	KCE	49.56	15.8	534	Rural	smps	800	Zíková and Ždímal (2013)[e]
25	K-Puszta	Hungary	KPZ	46.97	19.55	125	Rural	dmps	1000	Yli-Juuti et al. (2009)[c]
26	Po Valley	Italy	POV	44.65	11.62	10	Rural	dmps	600	Hamed et al. (2007)[c]
27	Finokalia	Greece (Krete)	FKL	35.2	25.4	250	Remote[a]	smps	848	Kalivitis et al. (2015)[c]
28	Amman	Jordan	AMM	32.01	35.87	1000	Urban	smps	420	Hussein et al. (2019)[e]
29	Hada al Sham	Saudi Arabia	HAD	21.802	39.729	500	Rural	dmps	850	Lihavainen et al. (2016)[e]
30	United Arab Emirates	United Arab Emirates	UAE	25.23	55.98	165	Rural	dmps	800	Kesti et al. (2022)[e]
31	Delhi	India	DEL	28.55	77.19	225	Urban	smps	560	Gani et al. (2020)[e]
32	Mukteshwar	India	MUK	29.47	79.65	2180	Remote	dmps	950	Hooda et al. (2018)[c]
33	Beijing	China	BEI	39.94	116.3	20	Urban	dmps	760	Liu et al. (2020)[e]
34	Nanjing	China	NAN	32.1	118.9	40	Urban	dmps	800	Qi et al. (2015)[e]
35	Zotino	Russia	ZOT	60.8	89.35	114	Remote	dmps	835	Heintzenberg et al. (2011)

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

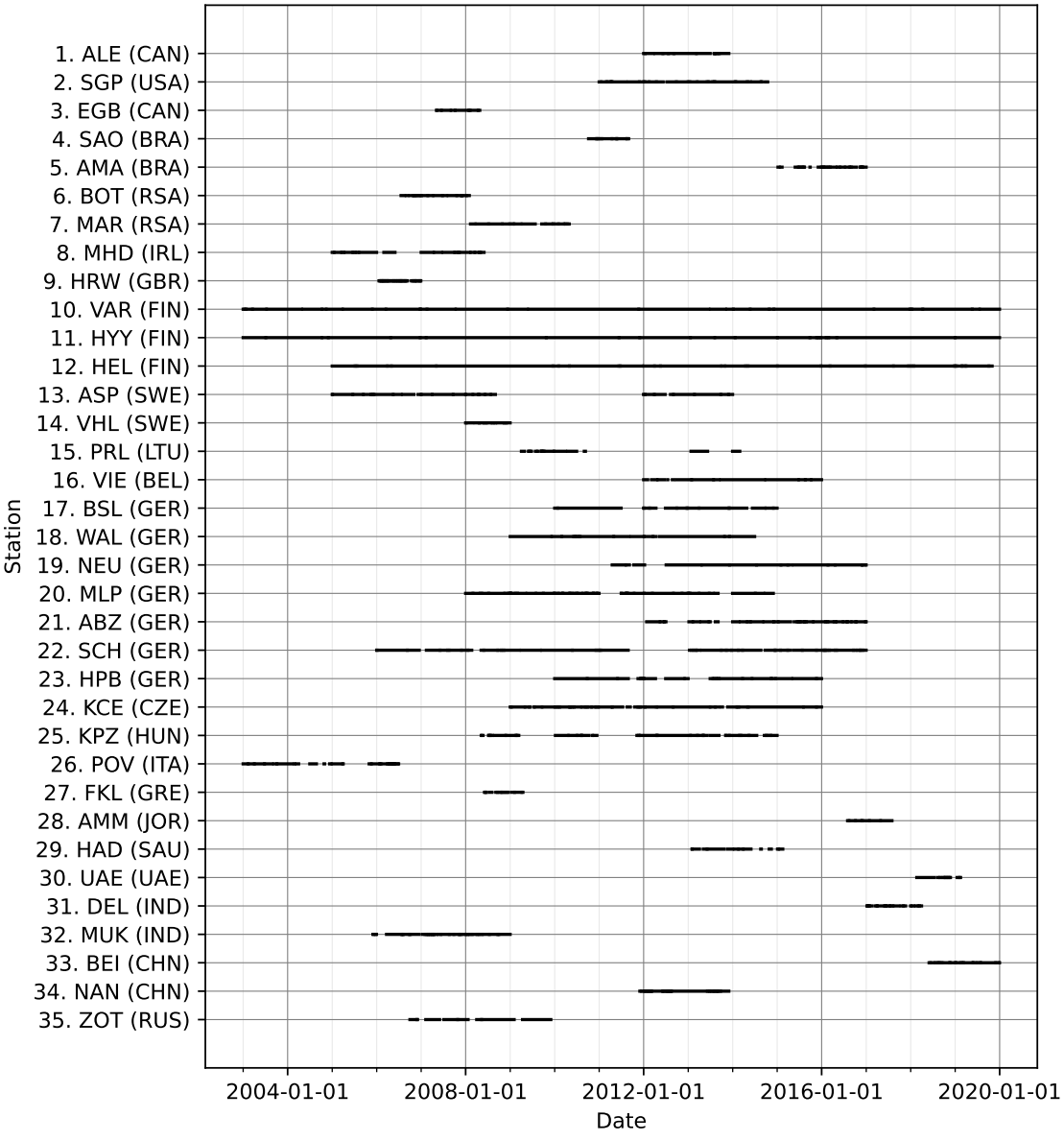

Figure 2. The temporal data availability of N_{100} measurements at different stations in the training set. The station numbers and abbreviations correspond to Table 1.

From CAMS and ERA5 datasets, we selected reanalysis variables known to either directly or indirectly influence the formation, growth, losses or dilution of aerosol particles. Most variables were sourced from the CAMS dataset, while boundary layer height (BLH) was obtained from the ERA5 dataset. The list of reanalysis variables used as predictors is provided in Table 2.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

AEROSOL Research Discussions

Table 2. Information on the variables used in model training and testing. The table lists variable names, variable abbreviations, variable units, model level of reanalysis data if applicable, and whether the variable was log10-transformed. N_{100} was obtained from measurements. The other variables were from reanalysis data, with boundary layer height from ERA5 dataset and other reanalysis variables from CAMS reanalysis data. Wind speed and relative humidity were calculated from CAMS variables. The reanalysis variables contained some single-level variables, but most of the variables were multi-level variables, which we downloaded for model level 60, which is 10 m above ground under standard atmospheric conditions.

Variable Name	**Abbreviation**	**Unit**	**Model level**	**$\log_{10}$-transformation**
Number concentration of particles larger than 100 nm	N_{100}	cm^{-3}	-	yes
Hydrophilic organic matter aerosol mixing ratio	$OM_{h.phil.}$	$kgkg^{-1}$	60	yes
Hydrophobic organic matter aerosol mixing ratio	$OM_{h.phob.}$	$kgkg^{-1}$	60	yes
Hydrophilic black carbon aerosol mixing ratio	$BC_{h.phil.}$	$kgkg^{-1}$	60	yes
Hydrophobic black carbon aerosol mixing ratio	$BC_{h.phob.}$	$kgkg^{-1}$	60	yes
Sulphate aerosol mixing ratio	Sulphate	$kgkg^{-1}$	60	yes
Dust aerosol (0.03 - 0.55 µm) mixing ratio	Dust	$kgkg^{-1}$	60	yes
Sea salt aerosol (0.03 - 0.5 µm) mixing ratio	Sea salt	$kgkg^{-1}$	60	yes
Carbon monoxide mixing ratio	CO	$kgkg^{-1}$	60	yes
Sulphur dioxide mixing ratio	SO_2	$kgkg^{-1}$	60	yes
Ammonia mixing ratio	NH_3	$kgkg^{-1}$	60	yes
Nitrogen monoxide mixing ratio	NO	$kgkg^{-1}$	60	yes
Nitrogen dioxide mixing ratio	NO_2	$kgkg^{-1}$	60	yes
Isoprene mixing ratio	C_5H_8	$kgkg^{-1}$	60	yes
Terpenes mixing ratio	$C_{10}H_{16}$	$kgkg^{-1}$	60	yes
Boundary Layer Height	BLH	m	Single level	no
Specific rain water content mixing ratio	SRWC	$kgkg^{-1}$	60	yes
Air temperature at 2 m height	T	K	Single level	no
Dew point temperature at 2 m height	T_d	K	Single level	no
10-m u-component of wind	U	ms^{-1}	Single level	no
10-m v-component of wind	V	ms^{-1}	Single level	no
10-m wind speed	Wind speed	ms^{-1}	-	yes
2-m relative humidity	RH	%	-	no

Reanalysis variables served as predictors in the training and holdout sets. We matched the reanalysis data to the N_{100} measurements by selecting data from the grid-cells containing the measurement stations and including only days with observations. For the training set, we selected observations from 2003-2019 period and, for the holdout set, observations from 2020-2022 period. However, it should be noted that average conditions within a grid-cell (up to 80 km) may sometimes fail to represent

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

the single-point measurements due to sub-grid scale variability in emission sources, meteorology, and topography. For example, if a measurement station is located near strong sources or within a limited high-emission area, such as a city, the grid-cell average in the reanalysis data may underestimate concentrations due to dilution over a larger area.

In addition to the training and holdout sets, we used reanalysis data as input for generating the global N_{100} fields, retrieving the reanalysis data covering the whole globe for 2013. We chose this year because it had the best availability of observational data for assessing model performance.

We first adjusted the 0.25x0.25-degree resolution of the ERA5 dataset to match the 0.75x0.75-degree resolution of the CAMS dataset by calculating grid-cell averages that correspond to CAMS data grid size. We then calculated daily averages for the variables. Additionally, we derived two variables from CAMS data: 2-m relative humidity (RH) and 10-m wind speed (WS). RH was computed from the dewpoint temperature and air temperature at 2 m height using Alduchov and Eskridge (1996) approximation for saturation vapor pressure. WS was calculated from the 10-m u-component and 10-m v-component of wind.

Finally, we normalized the reanalysis variables that followed lognormal distribution by log10-transforming them (Table 2). Some of the variables had minimum values at zero, and before log10-transforming, we replaced these with the next smallest value of the variable. Additionally, if the log10-transformed value was very low compared to the rest of the values (e.g., 10^{-27}), we shifted the minimum value to 10^{-17}. This increase of the minimum value was necessary because we noticed that in situations when the predictor values were extremely low compared to typical predictor values, it led the MLR model to generate unphysically low N_{100} estimates.

4 Designing and applying training and testing procedures

In this section, we detail how the ML methods described in Sect. 2 were applied in our training and testing process for the ML models. The primary aim of this study was to train global ML models using limited observational data to estimate N_{100} concentrations in areas without measurements. The challenge is not training the global ML models but assessing their performance and reliability outside the measurement stations, which cannot be done with our limited holdout set. Therefore, we designed a methodology incorporating cross-validation and intermediate ML model versions. Although we developed this approach for our specific data, it can be applied to other scientific questions in atmospheric and other fields of science, with similar challenges related to limited spatial observations.

4.1 Training and holdout sets and their limitations

As described in Sect. 3, we allocated the training and holdout sets based on temporal selection, using data from 2003-2019 for training the models and 2020-2022 for assessing model performance. This division was chosen because 84 % of measurements in our dataset were collected during 2003-2019. Training ML models that can be applied globally required a training set that represented diverse environments and meteorological conditions, ideally covering several seasonal cycles at each location to provide reliable analysis. However, as is often the case in atmospheric sciences, most stations did not have long observational series. Given these constraints, we prioritized the training set robustness over the holdout set representativeness and

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

chose to include only stations with 2020-2022 data for the holdout set. This approach also allowed us to investigate temporal extrapolation, where we assessed model performance at the measurement stations but outside the time period used in training.

The presented train-test split had certain limitations. In addition to excluding many of our measurement stations, the holdout set could not assess global ML model performance outside the locations used for training. This drawback was crucial, because our goal was not only to estimate N_{100} at the measurement stations (temporal interpolation and extrapolation), but also to evaluate how well the ML models could predict values in completely new locations (spatial extrapolation). To properly assess spatial extrapolation, we would need a holdout set containing additional stations with sufficiently long time series from different environments. However, long time series of particle number size distributions are not widely available, particularly outside Europe. Therefore, ensuring a wide variety of measurement stations in both training and holdout sets is challenging, and datasets from any additional measurement stations would also improve the training set.

Table 3. Summary of the different intermediate model setups and their train-test splits.

Model setup	Purpose	Train-test split for cross-validation
XGB HP tuning (Sect. 4.3.4)	XGB HP tuning	Spatial train-test split: Target station data used as the validation set and the data from all other sites as training data.
Single-station models (Sect. 4.4.1)	Testing if ML models with reanalysis data can capture observed N_{100}	Temporal train-test split: Data divided into 4-week increments: the first 2 weeks used in the training set, out of the last 2 weeks 3 first and 3 last days discarded and 8 days in between used in the validation set. Rotation of 4-week periods to start from different week of the month.
Station-excluded models (Sect. 4.4.2 and Sect. S3)	Cross-validation: Estimating how the global ML models may perform in environments and conditions outside the existing measurement stations	Spatial train-test split: Target station data used as the validation set and the data from all other sites as training data. For analysis that required a comparable number of data points from all stations, the validation set contained 200 data points with 50 data points sampled per season (Sect. S3). For comparing against station-included models, the validation set included the same days as in station-included models below.
Station-included models (Sect. 4.4.3)	Examining and illustrating how much of the model uncertainty at the target station was linked to the availability of training data in roughly similar environment or meteorological conditions to the target station	Combination of spatial and temporal train-test split: The training data from other stations like in station-excluded models combined with 2 weeks out of 4 weeks target station data as in single-station models. Validation set 8 days out of four weeks target station data as in single-station models.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

AEROSOL Research Discussions

4.2 Intermediate models for inferring global model performance

To address the challenges our dataset posed on training and testing the ML models, we employed CV, which allowed us to maximize data usage by utilizing each data point for both training and testing while maintaining separation between the sets in each CV round. As a result, this method could be applied to all stations, regardless of data length. However, utilizing CV had two main limitations.

First, because CV involves evaluating ML models on the same data used for model optimization, it may overestimate the model performance. To investigate this potential bias, we compared CV performance (training error) with holdout set performance (testing error) at stations where holdout set was available.

Second, for training the final global ML models, we wanted to maximize the training set representation by using all available data from 2003-2019. This approach precluded the direct use of CV for evaluating the final global ML models. To address this, we calculated testing errors for stations with available holdout sets, but for the other stations, we relied on an alternative strategy. We trained several intermediate models and assessed their performance with CV to infer global ML model performance. Although using separate ML models for generating estimates and assessing their performance was not ideal, this method utilized our limited data more effectively than reserving either portions of each station's data or entire station datasets for testing.

We constructed several intermediate models with different setups and corresponding CV train-validation splits (Table 3). The first setup involved single-station models, which we trained and tested using only station-specific data. These provided a simple baseline performance analysis for what our method could achieve. The second setup consisted of station-excluded models, where we utilized spatial CV. We trained station-excluded models with all stations except the target station, which acted as the validation set. This approach provided insight into model performance in locations without measurements. The third setup, station-included models, was similar to the station-excluded models but included a portion of the target station's data in the training set, allowing a comparative analysis against station-excluded models. Additionally, for illustration purposes, we constructed modified versions of the station-included and station-excluded models to generate a time series for 2013. We discuss the different intermediate model setups in more detail in Sect. 4.4.

We structured the model training and testing procedures into three main parts. First, we defined the training and testing procedures, including data sampling, scaling, feature selection, and hyperparameter tuning, to ensure consistency and reliability across all ML models (Sect 4.3). Second, we conducted CV analyses for the intermediate models (Sect 4.4). Finally, we trained the global ML models, assessed feature importance, and produced estimates for 2013 (Sect 4.5).

4.3 Model optimization and training and validation procedures

4.3.1 Train-validation splits for cross-validation

The first step in the analysis was formulating the CV procedures for the intermediate models and determining how to sample and process the training and validation sets to ensure balanced contribution from all stations. We modified the conventional k-

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

fold CV method and devised two main variations for splitting the data into training and validation sets. We used these variations and their combinations when training and testing the intermediate models (Table 3).

The first variation, spatial train-validation split used for spatial CV, treated each measurement station as a group. One station was excluded from the training set, and the model performance was tested on this excluded (target) station. This version was used to construct the station-excluded models (Table 3).

The second variation employed a temporal train-validation split to ensure that the seasonal cycle was represented both in training and testing. Here, each station's data was divided into four increments, with two weeks allocated to the training set, three days discarded, eight days assigned to the validation set, and another three days discarded. Although discarding days reduced the data availability, it minimized autocorrelation between the training and validation sets, preventing overestimated performance. We typically repeated this process four times, rotating the weeks in the sets.

4.3.2 Balancing training set

The train-validation splits allowed us to assess the model performance while maintaining representation from all selected stations. However, the data length varied between the stations, with the shortest measurement series covering 201 days (about 6 and a half months) whereas the longest spanned 6182 days (about 17 years) (Fig. 2). As a result, the training sets contained a different number of days from different stations. Training the models without addressing this imbalance could bias the global ML models towards stations with longer time series. To address the issue, we implemented a weight that was inversely proportional to the number of data points in the station. Data points from stations with longer measurement series were assigned a lower weight and shorter series a higher weight so that all stations had equal influence during training. While this approach sacrificed some benefits of longer measurement series, it preserved all information from these longer datasets and was therefore preferable to sampling only a subset and discarding the rest.

Additionally, most of our stations were situated in Europe (Fig. 1), prompting us to investigate if this Eurocentricity could produce bias in our ML models. We trained models with three different station selection schemes and used cross-validation with spatial train-validation split to evaluate their performance at stations outside of Europe. The station selection schemes were 1) using all stations, 2) sampling a subset of the European stations, and 3) downweighing the data points from European stations. We separately investigated how the selection scheme affected model performance at European stations and non-European stations. The analysis revealed that using all stations yielded comparable model performance as the two other methods for both European and non-European stations. Training the models with data from all stations in the training set even resulted in better median $\mathrm{RMSE}_{\log 10}$, though the improvement was not statistically significant (not shown). Thus, we decided to incorporate data from all stations into our training set.

4.3.3 Feature scaling and selection

An essential part of training the ML models involved processing the predictor variables (features). The variables had different units, and their values differed by several orders of magnitude. Such discrepancies can pose a challenge for ML models, potentially affecting their performance (e.g., Kuhn and Johnson, 2013). Additionally, assessing feature importance with MLR

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

coefficients requires the variables to be scaled. To address this, we centered and scaled the variables - subtracting the mean and dividing by the standard deviation - using a scaling function fitted on the weighted training data. We applied this scaling to both the training and validation or holdout sets.

We also explored different feature selection approaches but ultimately included all variables in our analysis. We investigated how the model performance was affected by selecting only the most important variables, using only a certain type of variable (aerosol variables, meteorological and gas variables), or combining the strongly correlating variables together. However, reducing the number of variables decreased the model performance, likely because all variables were relevant to at least some of the measurement stations. We confirmed, using adjusted R-squared, that including all variables did not artificially inflate model performance due to the larger number of predictors (not shown). As conclusion, we chose to include all variables in our analysis.

4.3.4 Hyperparameter tuning

After establishing the other training and testing procedures, we focused on tuning hyperparameters (HPs) for the XGB model. We used the spatial train-validation split method for cross-validation to ensure the tuned HPs generalized across all stations (Table 3). One station, Schauinsland, Germany (SCH), was excluded from HP tuning due to its frequent positioning above the boundary layer during winter (Birmili et al., 2016). Based on the results, we manually selected parameter combinations that produced strong average $RMSE_{log10}$ across cross-validation rounds. When multiple parameter sets performed well, we chose the ones that minimized training time.

One of the hyperparameters we tuned was $n_{estimators}$, which sets the number of estimators, and, consequently, training rounds during the model training. Even though we tuned this variable, we also chose to use early stopping to avoid overfitting and save computing resources (e.g., Kuhn and Johnson, 2013). Early stopping evaluates model performance after each training round using a validation set and halts training if no improvement is observed after a set number of iterations. In our case, $RMSE_{log10}$ was used as the error metric, and training was stopped if performance did not improve after 10 consecutive rounds.

Throughout our analysis we used one set of tuned HPs. Originally, we formulated the training and testing procedures with default HPs. After deciding the procedure for training the final global ML models (detailed in Sect 4.5.), we tuned the HPs to align with the final training configuration. The final set of HPs can be found in Table S1. We then revisited the training and testing formulations described above to ensure the initial conclusions remained valid.

4.4 Assessing model performance with intermediate models

Once we had established the ML model training and testing procedures, we trained and tested the intermediate models and used the results to investigate the model behavior and performance.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

4.4.1 Single-station models

As outlined in Sect 4.2, our first intermediate model setup involved training single-station models for each individual station (Table 3). These models provided insight into how well ML models trained specifically for one station could predict N_{100} at that location, a simpler task compared to modeling global N_{100} variations. We trained and tested the single-station models using CV with the temporal train-validation split: two weeks from each month were allocated to the training set and eight days to the validation set, repeated four times with different days rotated in the sets (Table 3). For consistency, we scaled the variables, and for XGB, we applied early stopping and the tuned global HPs.

Although we considered tuning HPs for individual single-station models, we found that using globally tuned HPs was sufficient. For instance, when evaluating the performance of a single-station model for Alert, Canada (ALE)—a station with unique characteristics because it is located in very clean polar environment—results showed minimal improvement when using station-specific HPs (not shown). To conserve computational resources, we chose to use the global HP set across all single-station models.

Following cross-validation, we analyzed the results to assess the models' performance at each station and between the MLR and XGB models. To verify the reliability of our results, which CV may overestimate, we also evaluated single station model performance using the holdout set for the stations where it was available.

4.4.2 Station-excluded models

The second intermediate model setup involved station-excluded models, designed to evaluate global model performance in stations not represented in the training data (Table 3). This step was essential for estimating how well the global models could perform in areas without measurements. We employed the spatial train-test split for CV, testing the models using all available data from the target station while training them with data from all other stations. To ensure balanced contributions from each training station, we applied weighting to the data points. We scaled the variables and, for XGB, used tuned hyperparameters and early stopping. We conducted analysis for both MLR and XGB and compared their performance at each station. Additionally, for stations with available holdout sets, we evaluated the performance of the station-excluded models using these sets.

4.4.3 Station-included and station-excluded model comparison

To assess the impact of including data from the target environment in the training set, we constructed station-included models (Table 3). For CV, we used a combination of spatial and temporal train-validation splits. The training set comprised data from all other stations, along with two weeks per month of data from the target station, while the validation set contained eight days per month from the target station. As with the standard temporal train-validation split, we conducted four CV rounds. We also scaled the variables and used tuned HPs and early stopping for XGB.

Additionally, to enable direct comparison between the station-included and station-excluded models, we created a modified version of the station-excluded models. As before, the training set contained data from all stations except the target station, and the validation set included only data from the target station. However, in this version, the validation set was further restricted

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

to include only the days that matched the temporal validation set (Table 3). This process involved four CV rounds, rotating through different validation sets. The scaling, HPs and early stopping were applied as before.

4.4.4 Time series analysis

As the final step in assessing model performance, we analyzed the time series generated by the station-excluded and station-included models, comparing them to the observed N_{100} time series. The goal was to demonstrate the potential performance of the final global ML models, both at the measurement stations and in areas without measurements. This analysis was conducted for 2013, as it had the most comprehensive data availability across different stations.

To generate the estimated N_{100} for this analysis, we followed a procedure similar to the original station-excluded and station-included models (Table 3), with one key modification: the validation sets contained only data from 2013. For the station-excluded models, this involved still using the spatial train-validation split, but with the validation set restricted to 2013 data. Similarly, for the station-included models, we continued to apply the combined spatial and temporal train-validation split, but the validation set consisted solely of 2013 data. However, unlike in the previous steps, we did not use cross-validation rounds for the station-included models; instead, we used the first two weeks of each month as the training set. For both setups, we scaled the data using the scaling function trained on the training sets, and for XGB, we applied the tuned HPs and early stopping.

With the station-excluded and station-included models trained and the corresponding validation sets defined, we generated estimates for 2013. The station-excluded models produced continuous time series, while the station-included models generated time series with only an eight-day period for each month, as determined by the validation set. After generating the estimated N_{100} time series, we compared them to the observed measurements.

4.5 Global ML models and N_{100} fields

In the final part of the analysis, we proceeded to train the global ML models, analyze their feature importance, and generate global N_{100} fields for 2013.

We trained the final global ML models with a training set containing all stations, all available data points, and all variables. As before, the data points in the training set were weighed to ensure an equal contribution from all stations to the model training. We also fitted the scaling function with the training data, scaled the variables with it, and saved it to be used as scaler when generating the global N_{100} fields for 2013. Once the training set was processed, we proceeded to train the global MLR model (MLR_{global}). For the global XGB model (XGB_{global}), we followed the same procedure, except with the addition of the tuned hyperparameters and early stopping. Here it should be noted that in principle early stopping requires separate training and validation sets to evaluate when the model performance plateaus. However, given that the global ML model training did not have a train-test split, we instead used the training set for evaluation. Early stopping caused the model training to interrupt after around 425 training rounds (compared to 900 from our HP tuning), potentially earlier than it would have occurred with separate sets. Nevertheless, because we had utilized early stopping in the previous analyses to mitigate overfitting and save computing resources, we continued to implement it here.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

Once we had trained the MLR_{global} and XGB_{global}, we proceeded to analyze how different variables contributed to these models using model given feature importance.

Finally, to generate the global N_{100} fields for 2013, we utilized the 2013 global reanalysis dataset. After scaling the dataset using the previously fitted scaler, we provided it as input for the MLR_{global} and XGB_{global} models and generated daily N_{100} fields for 2013.

We investigated the global ML models' performance both at measurement stations and in areas without measurements. At the measurement stations, we evaluated the global ML models using the holdout set for stations with N_{100} data available between 2020 and 2022. In areas without measurements, we compared the MLR_{global} and XGB_{global} fields. We calculated the $RMSE_{log10}$ between these estimates for each grid-cell, and if the error value was large, it indicated that the models generated very different estimates for that region, meaning at least one must be inaccurate. Conversely, we could assume the estimates were more reliable if the models produced similar results. However, even when the models yielded similar results, we could not be certain that the estimates were close to the true N_{100} without actual measurements from those locations. For example, if our reanalysis dataset contained a bias in a particular region, both models could produce similar but erroneous results. Since the comparison between MLR_{global} and XGB_{global} fields provided only a rough error estimate, we attempted to develop a more sophisticated method for assessing global performance. However, this effort did not yield results.

5 Results and discussion

5.1 Assessing intermediate model performance

5.1.1 Single-station model performance

The training errors for the single-station models are shown in Fig. 3. While generating these models was not the primary goal of this study, they provided a simpler setting to evaluate our method and identify potential challenges. Many single-station models achieved $RMSE_{log10}$ values below 0.2, and almost all remained under 0.3, indicating that model performance was generally excellent or good.

Testing errors for stations with data from 2020–2022 are shown in Fig. 4. As expected, testing errors were slightly higher than training errors, but the overall conclusions remained consistent. These results demonstrate that estimating N_{100} using ML models and reanalysis data is feasible. However, at some stations (e.g., Harwell, United Kingdom and Preila, Lithuania), model performance was inadequate ($RMSE_{log10} > 0.3$), which we discuss further in Sect. 5.3

5.1.2 Assessing station-excluded and station-included model performance

We first looked at the performance of the station-excluded models, which were trained separately for each station. Fig. 5 depicts the station-excluded N_{100} estimates against the observed N_{100} for all the stations, when no data from the target station was included in the training set. In practice, this means that for each station the estimated N_{100} was produced with a different model and different validation set, and results are presented in one figure. In contrast to the other instances where we used

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

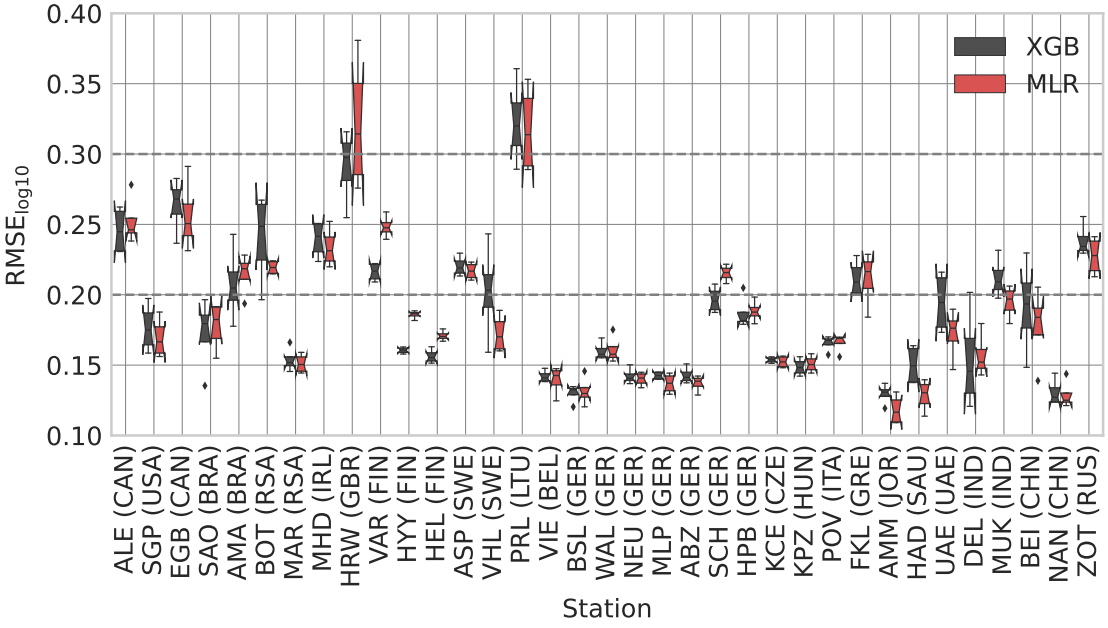

Figure 3. Comparison between training errors (RMSE calculated for log10-transformed concentrations) of the single-station models for one station) with XGB and MLR machine learning models. The boxes and whiskers indicate the variation caused by selecting different train-test splits. The boxes show the quartiles and whiskers show the 1.5 interquartile range of the lower and upper quartile. Data points outside these are considered outliers and marked with individual markers. Additionally notches in the boxplots indicate the confidence interval of the median. If the notches of two boxes do not overlap, it indicates that the medians are statistically significantly different at 5 % significance level.

the station-excluded models, here the estimates were not generated for all the available data from the target station (Table 3). Instead, we used only around 200 days to have a comparable number of data points from all stations in the validation sets. The sampling method for these 200 days is explained in Sect. S3.

Figure 5 provides a rough indicator of how the global ML models would perform in locations not directly represented in the training set. Looking first at the MLR result (Fig. 5a), even when each station had been excluded from the training set, the station-excluded MLR models could produce the range of observed N_{100} values from below 10 cm^{-3} to over 10^4 cm^{-3}. However, the station-excluded models still struggled with replicating the observations at the low concentrations, and in general, 54 % of daily estimates and 15 out of 35 station median estimates fell outside the factor of 1.5 from observations.

For the station-excluded XGB models (Fig. 5b), the station medians were better captured, with only 9 station medians falling outside the 1.5-factor limit. The daily values were also captured slightly better, though still 48 % fell outside the factor of 1.5. The station-excluded XGB models also failed to reproduce extreme values: they could not produce values below 25 cm^{-3}, systematically underestimated values above around 5000 cm^{-3}, and could not produce values above 10^4 cm^{-3}. Overall, these

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

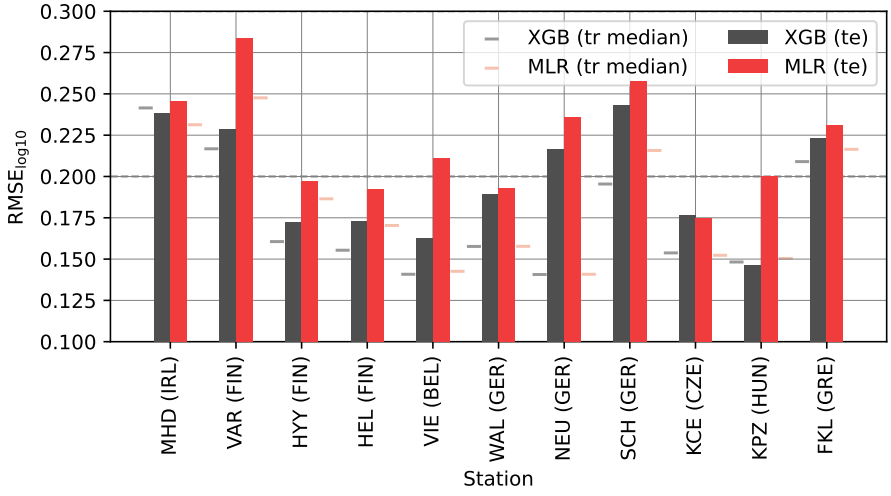

Figure 4. Comparison between the testing error (RMSE calculated for log10-transformed concentrations) of the single-station ML models (ML models trained with data from one station) at each station with XGB and MLR machine learning models. The bars show the testing error for both models, and the lines indicate the median training errors corresponding to Figure 3.

results show that the XGB models tend to be slightly more precise and replicate the median values better, but MLR models are better at extrapolating to low and high concentrations, though they still struggle to capture extreme values.

Next, we analyzed in more detail the MLR and XGB station-excluded performance at different stations. To ensure that the training error analysis (Fig. 6) was reliable, we first compared the training and testing errors against each other at the stations that had data after 2020 (Fig. S2). Because the target station had been left out of the training set in the station-excluded models, the main difference between the training and testing errors was that the training error was calculated with observations before 2020 and testing errors with observations after 2020, whereas the data before 2020 had also been used to optimize the ML models. Figure S2 showed that for station-excluded models, the difference in training and testing errors was small, and we felt confident in drawing conclusions from the training error.

In terms of the training error (Fig. 6), XGB had typically lower or equally good $RMSE_{log10}$ values than MLR, indicating better performance, though there were also stations where MLR performed better. Figure 6 also shows that the station-excluded performance varied depending on the station. The European stations typically had good or even excellent performance, probably because the N_{100}, different emissions, and meteorological conditions at many of the European stations were quite similar to each other. Even when the target station was left out from the training of the station-excluded model, there would still be at least one similar station in the training set. Conversely, stations with poor station-excluded performance might correspond to environments that did not have representation in the training set if the station was excluded from training.

To investigate further this variation in performance, we analyzed the station-included models' performance and compared them against the station-excluded models' performance (Fig. S3). For the stations with excellent station-excluded performance ($RMSE_{log10}$<0.2), we noticed that the differences between the station-included and station-excluded model $RMSE_{log10}$ were

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

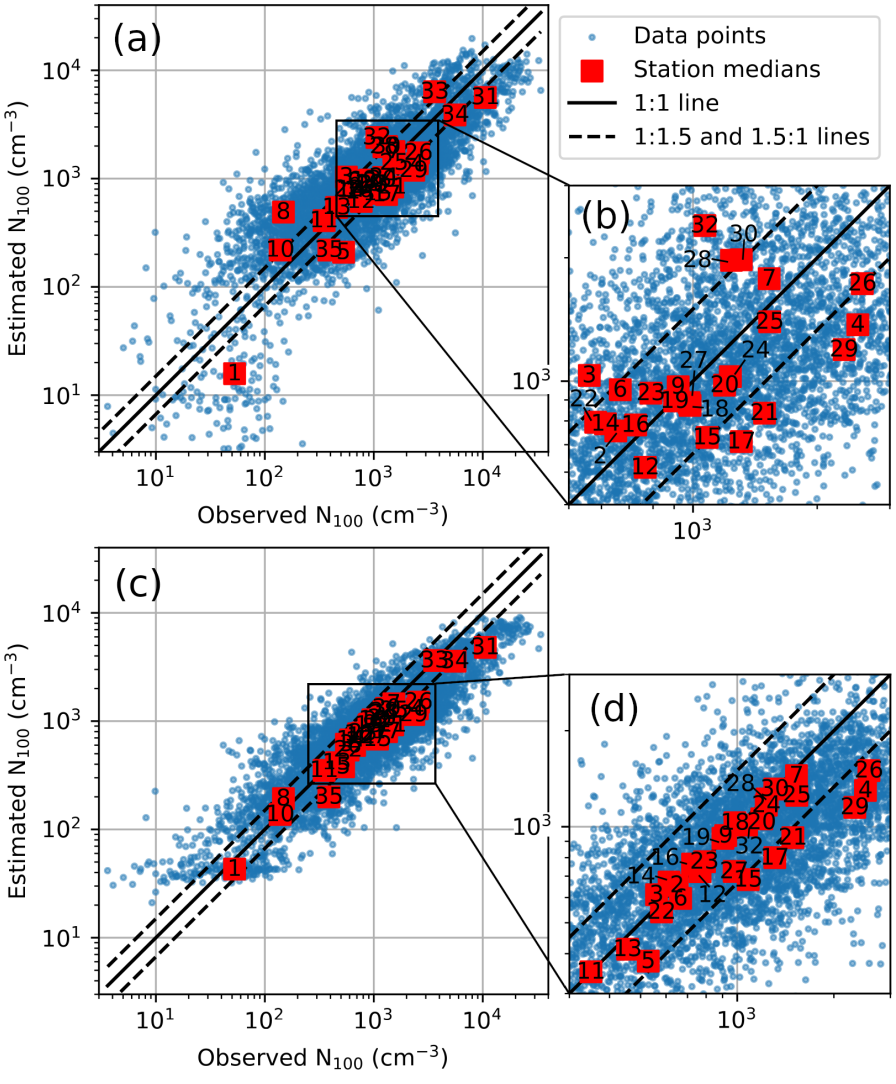

Figure 5. Comparison between observed and estimated N_{100}. The sampling of the data points shown in this figure are explained in Sect. S3. Panel a) shows the result for station-excluded MLR-models and panel b) shows a zoom-in. Panels c)-d) show the result and zoom-in for station-excluded XGB-models. The daily values are indicated in blue and station medians in red. The station medians are additionally marked with numbers which indicate the station as listed in Table 1.

small (below 0.01). This supports our interpretation that for many European sites (Vielsalm, Belgium; Waldhof, Germany; Neuglobsow, Germany and Melpitz, Germany for both MLR and XGB models and Vavihill, Sweden and Košetice, Czech Republic only for XGB model) and some other stations (Southern Great Planes, USA; Amman, Jordan, and Marikana, South

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

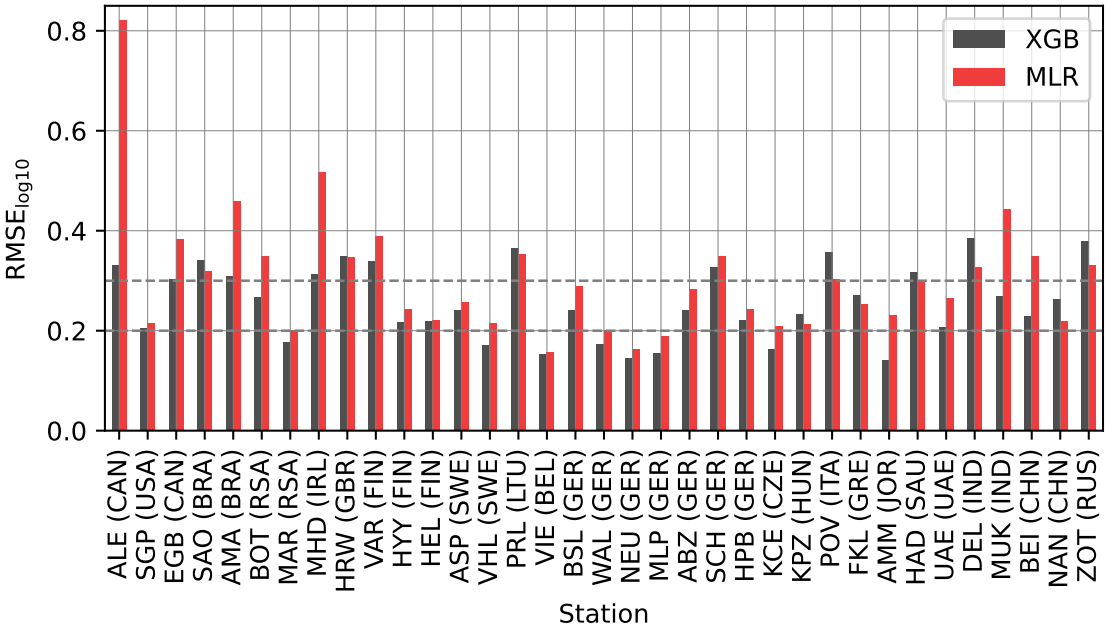

Figure 6. Comparison between station-excluded MLR and XGB model performance (RMSE calculated for log10-transformed concentrations) at each station.

Africa for XGB model), it did not matter whether the station had been excluded from the training, because the other stations could still represent the excluded station during training.

Conversely, for many stations, the station-included models produced clearly better results than station-excluded models (Fig. S3). For the XGB model these stations include Delhi, India; Hada al Sham, Saudi Arabia; São Paulo, Brazil; Po Valley, Italy; Zotino, Russia; Amazonas, Brazil; Nanjing, China; Värriö, Finland and Alert, Canada. The better performance confirms that these stations have some unique characteristics, and without their contribution, the XGB model could not capture the type of environment they represented. For example, when Delhi, India, which has the highest N_{100} in our dataset, was excluded from the training set, the models could not replicate the high N_{100} values. This led to underestimation and poor performance at the station (not shown).

The MLR models were less sensitive to whether station-specific was included in training compared to the XGB models (Fig. S3). Because MLR uses linear predictor functions, adding a small number of new data points does not always affect the model performance, resulting in smaller differences between the station-included and station-excluded model versions. In contrast, any new data in the XGB models can alter the tree structure, affecting model performance. However, this also increases the risk of overfitting, which may reduce the XGB model's ability to generalize outside the measurement stations.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

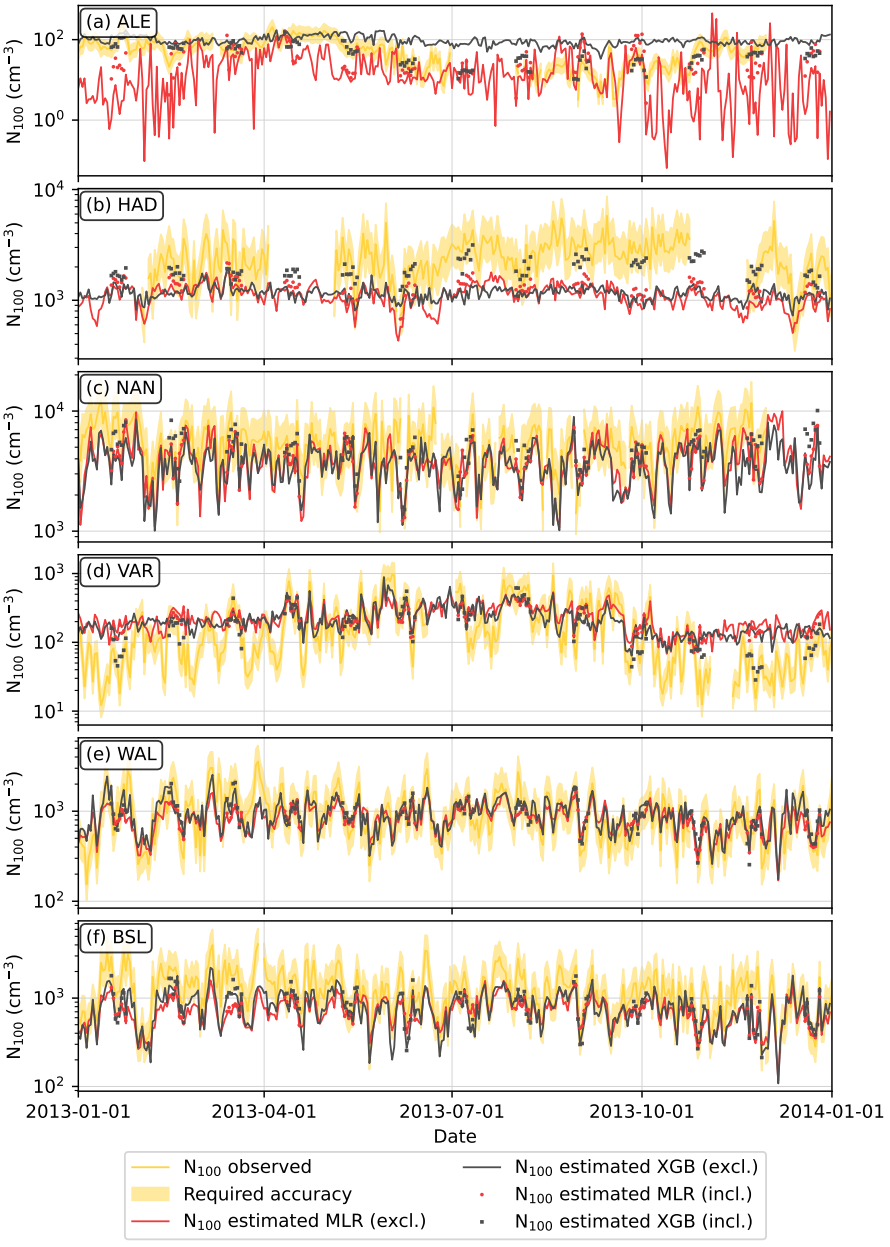

Figure 7. Comparison between observed and model estimated N_{100} time series for 2013 at selected stations. The required accuracy is within factor of 1.5 from the observations based on Rosenfeld et al. (2014). Station-excluded estimates and station-included estimates are described in (Table 3). Panels show the results for stations a) Alert, Canada (ALE) b) Hada al Sham, United Arab Emirates (HAD) c) Nanjing, China (NAN) d) Värriö, Finland (VAR) e) Waldhof, Germany (WAL) f) Bösel, Germany (BSL).

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

5.1.3 Time series

Figure 7 compares the observed N_{100} time series in 2013 to the estimated N_{100} time series produced with the station-excluded and station-included models. The comparison allowed for a better understanding of the ML model behavior outside the measurement stations.

In our dataset, Alert, Canada (ALE) was the sole representative of the extremely clean polar regions (Fig. 7a). When ALE was excluded from the training, neither model performed well at that location, demonstrating the challenge of missing environmental types in the training set. However, when ALE was included in the training, the models, especially XGB produced better estimates.

In Hada al Sham, Saudi Arabia (HAD), both station-excluded models underestimated N_{100}, whereas among the station-included models, the MLR model showed some improvement and the XGB model improved noticeably (Fig. 7b).

Nanjing, China (NAN) N_{100} estimates were captured well, though they were mildly underestimated with the station-excluded models and MLR station-included model (Fig. 7c). The station-included XGB model produced slightly better results. It is possible that specific environmental characteristics in Nanjing contribute to underestimation when using reanalysis data to estimate N_{100}.

In Värriö, Finland (VAR), the models performed well during summer, but the station-excluded models overestimated the low concentrations during winter (Fig. 7d). While the MLR station-included model yielded similar results, the XGB station-included model successfully captured the winter periods as well.

Waldhof, Germany (WAL), a typical European station, was well represented by other stations in the dataset (Fig. 7e). Consequently, even when WAL was excluded from training, the estimated N_{100} time series still aligned closely with the observations. Including data from Waldhof in the training set did not enhance the results. In contrast, in Bösel, Germany (BSL) (Fig. 7f)—another central European station —both the station-included and station-excluded models systematically underestimated N_{100}, although the daily variations were captured well. Birmili et al. (2016) noted that the total particle concentrations in Bösel were higher than at the other rural German sites.

5.2 Global ML models and N_{100} fields

5.2.1 Feature importance

Moving on to the final global ML models, Figure 8 shows the importance of different features in MLR_{global} and XGB_{global}. The two most important variables in both ML models were the black carbon aerosol (BC) mixing ratio and the organic matter aerosol (OM) mixing ratio. In MLR_{global}, these variables were hydrophilic, whereas in XGB_{global}, they were hydrophobic. However, we should not conclude that these variables were truly the most important ones. Due to the underlying dynamics of the CAMS dataset, BC and OM mixing ratios were highly correlated (Fig. S4), as were the hydrophilic and hydrophobic mixing ratios (not shown). Such strong correlations between variables can pose challenges for ML models (e.g., Kuhn and Johnson, 2013). In the MLR_{global} model, we observed an unexpected result: instead of assigning positive coefficients to both variables, it assigned a high positive coefficient to the hydrophilic BC mixing ratio while giving the hydrophilic OM mixing

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

AEROSOL Research Discussions

ratio approximately equally high negative coefficient. This suggests that the MLR_{global} model may have overestimated the influence of hydrophilic BC and then counterbalanced this by assigning a negative coefficient to hydrophilic OM. Typically, their combined effect on N_{100} was quite small. However, if the BC and OM mixing ratios are less closely linked in certain locations or during certain time periods, this imbalance could significantly affect the predicted N_{100} concentrations. To explore this further, we analyzed their relationship in Sect. S5 (Fig. S5).

Aside from the BC and OM mixing ratios, the most important variables influencing the ML models were sulphate aerosol, ammonia, carbon monoxide, and sulfur dioxide mixing ratios followed by temperature (Fig. 8). Since most of these variables are primarily associated with anthropogenic sources, it is unsurprising that in the MLR_{global} model, they exhibited a positive relationship with N_{100} concentrations, meaning that an increase in their concentrations led to an increase in N_{100}.

In contrast, the variables more linked to the natural processes tended to have lower importance and showed both positive and negative coefficients in the MLR_{global} model. Some coefficients aligned directly with expected physical processes. For example, the relationship between specific rainwater content (SRWC) and N_{100} is negative because rain removes aerosol particles from the air. Similarly, the negative coefficient for boundary layer height (BLH) reflects how a larger daily mean BLH dilutes N_{100} by mixing it into a larger volume of air.

Additionally, there were variables that have physically meaningful coefficients, but the interpretation is more nuanced, such as the sea salt aerosol mixing ratio. A higher concentration of sea salt aerosol should result in a higher N_{100} concentration. However, because higher sea salt aerosol concentration often coincides with the arrival of clean marine airmasses, MLR_{global} interprets the relationship to be negative. This is a meaningful interpretation over continental areas, but over oceans (which were not represented in our training set), it would fail to capture the true relationship between sea salt aerosol concentration and N_{100}. Moreover, high sea salt concentration in the sub-0.5 μm size is probably accompanied by high supermicron sea salt aerosol concentration, which gives little additional primary CCN but may substantially suppress secondary CCN formation via acting as a sink for low-volatile vapors and sub-CCN sized particles. A similar phenomenon can explain the negative coefficient of the sub-0.55 μm dust aerosol.

Finally, there were variables for which the MLR coefficients might not be able to capture the physical processes. One of these was temperature, which may have complex relationship with N_{100} depending on locations. For example, in many parts of the world, temperature can be associated with increased volatile organic compound (VOC) emissions, which leads to a larger number of aerosol particles growing to the accumulation-mode size range Paasonen et al. (2013). This effect has a strong correlation with isoprene (C_5H_8) and terpene ($C_{10}H_{16}$) emissions, and MLR_{global} may struggle with variables with strong correlations. However, a negative coefficient assigned for C_5H_8 mixing ratios but positive for $C_{10}H_{16}$ mixing ratio and temperature agrees with several studies suggesting that isoprene likely inhibits the secondary aerosol formation and growth of particles to N_{100} sizes (Lee et al., 2016; Heinritzi et al., 2020). Additionally, natural VOC emissions may be suppressed during the hottest days in many environments. On the opposite side of the temperature spectrum, cold temperatures can also lead to higher N_{100} concentrations due to heating-related residential biomass combustion, which consequently increases aerosol and aerosol precursor emissions. The MLR_{global} cannot capture these complex patterns directly but may attempt to do it

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

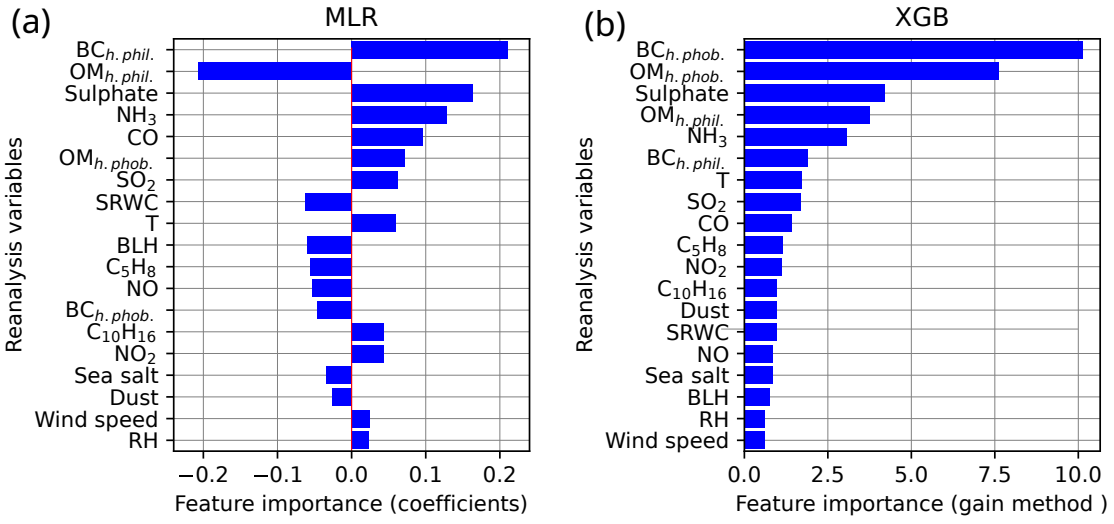

Figure 8. The global ML model feature importance in descending order. Panel a) shows MLR_{global} model feature importance based on MLR coefficients. Panel b) shows XGB_{global} feature importance based on gain method.

indirectly via correlating variables. This may also explain other counterintuitive coefficient values, such as NO2 having a positive coefficient and NO negative coefficient.

5.2.2 The global N_{100} fields

Figure 9 shows the annual mean N_{100} fields in 2013, calculated by averaging the daily N_{100} estimates - Fig. 9a for MLR_{global} and Fig. 9b for XGB_{global}. Both models estimated the highest N_{100} in South Asia and East Asia and the lowest N_{100} in remote locations such as polar areas and deserts.

The comparison between MLR_{global} and XGB_{global} N_{100} fields for 2013 is shown in Figure 9c. Overall, the ML models produced similar values across most continental areas, particularly in large parts of Europe and North America, though the XGB model generally yielded slightly higher estimates. Additionally, the results agreed well ($RMSE_{log10}$<0.15) near most measurement stations as well as in more densely populated areas (Smith, 2017, 2023) even in regions without in situ measurements. This pattern is evident in the most populated areas in the Middle East, Southern Siberia and Central Asia. In South America and Africa, the model agreement was also better in the more populated regions. However, the limited number of measurement stations in these continents may affect the result, because not all populated regions showed strong agreement between the models. A similar trend was observed in South and East Asia. While these regions are overall very densely populated, only the most highly populated areas exhibited strong agreement, which may also relate to the distribution of measurement stations. Although the agreement between models does not confirm accuracy against measurements, it suggests consistency between the models. This consistency is likely because these regions are well-represented in the model training, either directly through a nearby station or indirectly because most of the stations in our dataset are located in anthropogenically influenced areas.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

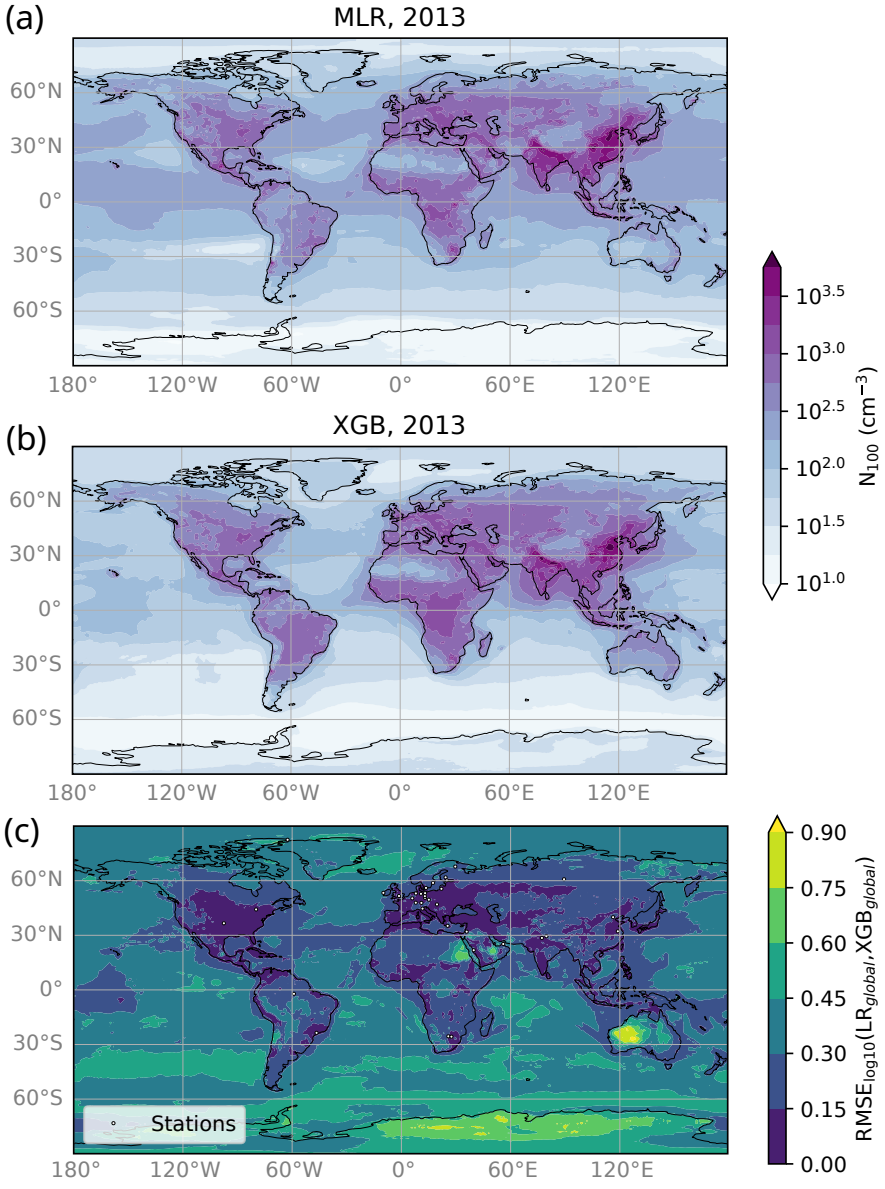

Figure 9. The estimated annual average N_{100} for 2013 a) with MLR_{global} model, and b) with XGB_{global} model. The models were trained with all available measurement data and all measurement stations. Panel c) shows the comparison of estimated daily N_{100} for 2013 from MLR and XGB models, where the color scale shows the root mean squared error between the log10-transformed N_{100} estimates. The smaller the RMSE value the better the models agree. RMSE values below 0.3 indicate that the models agree well and below 0.15 that the models agree very well.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

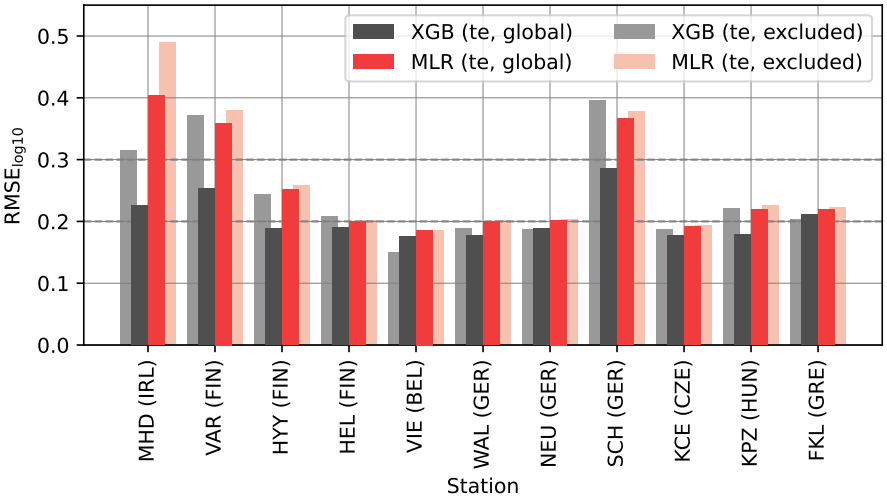

Figure 10. The comparison between MLR_{global} and XGB_{global} testing errors (RMSE calculated for log10-transformed concentrations) and station-excluded model testing errors (corresponding to Figure 6) for the stations that had N_{100} measurements for 2020-2022.

The models diverged in several regions (Fig. 9c), particularly over remote or clean continental environments such as Antarctica, the Australian deserts, the eastern Sahara Desert, and parts of the Middle East ($RMSE_{log10}$ > 0.60). In the latter two regions and some mid-latitude marine regions, the difference appeared to stem from low NH3 values (Fig. S6), which led MLR_{global} to generate lower N_{100} estimates. Other continental areas with smaller but still considerable discrepancies (0.30<$RMSE_{log10}$<0.45) included parts of South America, particularly the Amazon rainforest, the Congo rainforest, and some regions in Africa, including the Kalahari Desert, where the MLR_{global} model consistently predicted lower N_{100} values than XGB_{global}. Additionally, there were some hotspots where MLR_{global} produced clearly higher N_{100} estimates compared to XGB_{global}. The divergence likely stems from different responses to anthropogenic variables in the MLR_{global} and XGB_{global} models. While the anthropogenic variables were important in both models, the linear relationship between the variables and N_{100} in the MLR model seems to cause underestimation in low N_{100} values — common in clean or remote environments. XGB model did not exhibit this behavior possibly due to its nonlinear nature. However, in some locations, the lower N_{100} estimates from the MLR model appear more accurate than those from XGB. For example, in Alert, the station-excluded MLR model captured certain low N_{100} values better than the station-excluded XGB model.

Notable differences emerged also over the oceans (Fig. 9c), which are poorly represented in our training set. In these regions, MLR_{global} typically produced much higher N_{100} estimates than XGB_{global}. However, the models showed better agreement in continental outflow areas, such as the North-Western Pacific Ocean and major shipping routes, likely due to their anthropogenic influence, which makes them better represented in the model training.

The testing errors for MLR_{global} and XGB_{global} models at stations with 2020-2022 observations are shown in Fig. 10. These global ML model errors aligned with previous analyses, such as the station-excluded testing errors (also in Fig. 10),

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

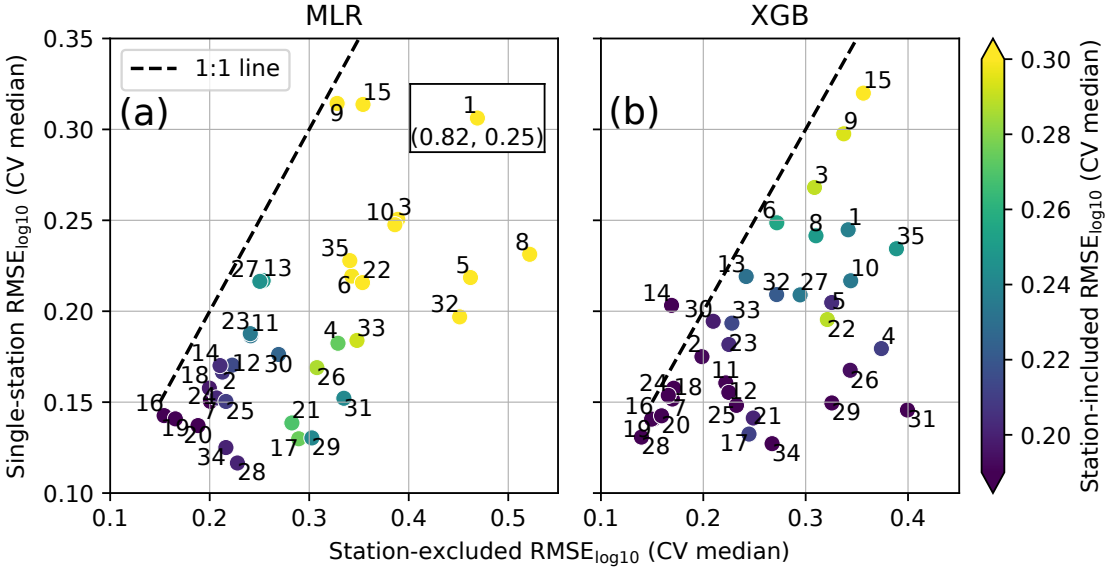

Figure 11. The comparison between the medians of cross-validation (CV) results from single-station and station-excluded models colored with results from station-included model, for a) MLR and b) XGB. The numbers correspond to stations as listed in Table 1. In panel a) one data point (Alert, Canada, 1) was outside figure limits and is indicated separately on the figure with coordinates.

with performance varying by location and the XGB model generally outperforming MLR_{global}. The most notable differences between global and station-excluded model performance occurred at Mace Head (Ireland), Värriö (Finland) and Schauinsland (Germany). In all these locations the global XGB models performed better. This improvement was likely due to the frequent low concentrations at these stations, which are challenging to capture without training representation from the target station. In Mace Head these low concentrations were related to clean airmasses coming from the ocean, in Värriö they were associated with clean winter periods, in Schauinsland to times when the measurement station was above the boundary layer.

5.3 Interpreting results from different ML models

By evaluating model performance across the intermediate models (single-station models, station-included models, and station-excluded models) and global models, we identified three cases where our models struggled to capture N_{100} accurately. Figure 11 presents a comparison of the $RMSE_{log10}$ medians from the CV analyses for single-station models, station-included models, and station-excluded models (the version directly comparable to station-included models, as shown in Table 3).

Firstly, our models struggled with capturing N_{100} at certain stations, even when using single-station models (Fig. 11). While the single-station estimates performed well at most stations, two stations had poor performance ($RMSE_{log10}$>0.3). Additionally, at stations with $RMSE_{log10}$ values between 0.2 and 0.3, certain conditions or characteristics may still be difficult for the single-station models to capture, lowering the performance, even though overall the $RMSE_{log10}$ values are acceptable. Notably, the

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

stations with high single-station $RMSE_{log10}$ often continued to exhibit lower performance in the other intermediate models, suggesting that these locations are inherently difficult to capture with our method (Fig. 11).

Several factors may explain these difficulties. Our dataset may lack key reanalysis variables necessary for accurately estimating N_{100} in these environments. Reanalysis data may also contain biases or struggle to resolve sub-grid scale processes crucial for N_{100} estimates. Additionally, the nonlinear interactions between predictor variables and N_{100} may not be fully captured by our ML models, either due to inherent model constraints or the limited size of the training dataset. Among our datasets, both stations where single-station model $RMSE_{log10}$ exceeded 0.3 (Harwell, United Kingdom and Preila, Lithuania) had relatively short measurement time series. Furthermore, (Xian et al., 2024) reported that CAMS reanalysis AOD differs from AERONET AOD in areas near Preila. Their observation suggests that there may be persistent sub-grid scale variability in aerosol concentrations around the site, which could be contributing to model inaccuracies.

The second challenge our ML models faced was a decline in station-included model performance compared to the single-station models. While we expected some decrease due to the added complexity of incorporating multiple locations, the station-included performance declined notably at some stations. The decline was particularly evident in the MLR models, where at 13 stations the station-included model $RMSE_{log10}$ values were over 1.5 times higher than single-station model $RMSE_{log10}$ values (Fig. S7a). For the XGB models, the station-included performance was notably worse than single-station performance in Schauinsland (Germany), and possibly in Bösel (Germany) and Annaberg-Buchholz (Germany) (Fig. S7b). This weaker performance may arise from variable-N_{100} interactions that differ from other stations. Since the models—especially the MLR model—struggle to capture conflicting variable-N_{100} relationships, stations with unique interactions relative to the rest of the dataset tend to experience the largest performance decline from single-station models to station-included models.

Conversely, at least in one station (Vavihill, Sweden), the station-included XGB model outperformed the single-station XGB model (Fig. S7b). One explanation for this improvement is that Vavihill has a relatively short measurement series in our dataset, which limited the single-station performance. However, in station-included models, Vavihill's data may be supplemented by other similar stations in our dataset, improving the performance.

Thirdly, our models struggled in locations that were not well-represented in our training data. While single-station and station-included models, which incorporated station-specific data, generally captured N_{100} at least moderately well ($RMSE_{log10}$<0.3), station-excluded models performed notably worse at certain sites—even when the station-included performance was excellent ($RMSE_{log10}$<0.2) (discussed in more detail in Sect. 5.1.2). Figure 11 illustrates this pattern especially for the XGB models: when the XGB station-included performance was excellent ($RMSE_{log10}$<0.2), the station-excluded performance varied widely, ranging from excellent ($RMSE_{log10}$<0.2) to poor ($RMSE_{log10}$>0.3). If both station-included and the station-excluded performances were excellent, it indicated that N_{100} in these stations could be captured well even without their own data in the training set because similar stations in our dataset provided sufficient representation. Conversely, as station-excluded $RMSE_{log10}$ increased, it suggested that incorporating station-specific data became increasingly important for accurate estimates. This effect was particularly notable in environments with high N_{100} concentrations compared to the other stations.

Our cross-validation indicated that our training set best represents European urban or rural environments influenced by human activity and similar anthropogenically influenced environments. Even when evaluating the global ML models with a

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

holdout set containing data from 2020-2022 (Fig. 10), the models performed well in capturing N_{100} at the European stations. The global model comparison gives similar results, showing that the models tend to agree in Europe but also in other populated areas.

5.4 ML model limitations

While our results demonstrate promising performance across many environments, the findings from Sect. 5.3 highlight that model accuracy depends strongly on the availability and representativeness of training data. In other words, different limitations in the N_{100} measurements and reanalysis data cause limitations in the ML models. While Sect. 5.3. touched upon these issues, here we discuss them further.

For the N_{100} measurements, the main challenge is data availability. To train ML models that capture diverse environments and meteorological conditions, we require a broad dataset that covers a wide range of locations and time periods. Ideally, we would have at least five years of data from each station. This would allow for the division of data into training, validation, and holdout sets, with at least one full seasonal cycle in each set and multiple cycles in the training set. Since environmental conditions and aerosol concentrations vary between years, such a dataset would enable ML models to generalize better and learn from a broader range of conditions, resulting in more robust estimates. Unfortunately, our current dataset lacks full seasonal coverage at some stations, which makes it harder for the ML models to accurately capture station-specific and global trends. This emphasizes the need for continuous long-term observations.

Another challenge with in situ measurements is potential measurement errors that may remain after filtering. These errors can propagate into the ML models, affecting overall accuracy. Additionally, because our method relies on ground-level N_{100} measurements, our ML models can generate only ground-level N_{100} estimates. However, for many applications knowing the vertical profile of N_{100} would be important. For example, CCN concentrations are particularly important near or above the cloud base (Quaas et al., 2020), especially in cases where the cloud base is decoupled from surface conditions (Su et al., 2024).

Regarding reanalysis data, challenges stem from various biases inherent in the CAMS and ERA5 reanalysis datasets. Block et al. (2024) provide a comprehensive discussion of the biases affecting CAMS aerosol variables, including uncertainties in polar regions due to limited satellite retrievals, omissions such as volcanic activity, and specific volcano-related biases around locations like Mauna Loa (Hawaii, USA) and Altzomoni (Mexico) —both of which appear as hotspots in our MLR model results. CAMS also does not model nitrate aerosol mixing ratios and represents hydrophilic and hydrophobic BC and OM mixing ratios with simplified partitioning based on emission fractions and a conversion rate over time (see Block et al. (2024) and the references therein). Although we did not explore these biases for CAMS gas concentrations and meteorological variables, similar issues are likely present, potentially introducing some errors into our global N_{100} fields.

Moreover, integrating gridded reanalysis data with single-point N_{100} measurements can introduce challenges at the stations located in grid-cells with sub-grid scale variability in emission sources, meteorology, and topography. Because reanalysis data represents grid-cell averages, it may not capture the true predictor variable concentrations at the measurement site, leading to biases in the model's learned relationships. This discrepancy may partly explain the poor performance observed at some stations, even when using single-station models.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

AEROSOL Research Discussions

6 Summary and Conclusions

Observation based data on global accumulation mode particle number concentrations (N_{100}) are essential for assessing global CCN concentrations and their climate impacts as well as for evaluating Earth System Models. According to Rosenfeld et al. (2014), reducing uncertainties in aerosol-cloud interactions requires capturing global CCN concentrations within a factor of 1.5 of true values. In this study, we developed a method for generating global N_{100} fields using a combination of in situ measurements, reanalysis data, and machine learning. For evaluating ML model performance at measurement stations and outside of them, we applied cross-validation to several intermediate models. We also trained global ML models on all available data and generated daily global N_{100} fields for 2013.

We found that at least in a simple setting, such as estimating the N_{100} at a specific location with the single-station models, our method yields good results. This is especially true for the XGB model. However, some stations were more challenging to capture, possibly due to an insufficient number of data points, missing crucial reanalysis variables, or inadequate representation of sub-grid scale variability in concentrations and other reanalysis data biases. Additionally, ML models—particularly the MLR model—may struggle to capture the nonlinear interactions between N_{100} and the reanalysis variables at these stations. The stations where single-station models struggled remained challenging for all types of intermediate models.

Addressing these limitations is challenging, but future work could explore incorporating additional variables. For example, accounting for a station's position relative to the top of the boundary layer, which might help improve ML model performance in high-altitude environments by allowing models to recognize when stations are above it. Additionally, refining the grid-selection scheme could improve accuracy at stations where sub-grid scale variability causes the reanalysis data to misrepresent local conditions. Comparing observed concentrations of key predictor variables with their reanalysis counterparts can help identify discrepancies. If significant differences emerge, selecting a nearby grid-cell that better represents the measurement station—such as choosing a land-only grid-cell instead of one that includes both land and ocean—may enhance model performance.

Our primary approach for evaluating ML model performance in areas without observations was cross-validation using station-excluded models. For each station, we trained an ML model without station-specific data and assessed how well the model reproduced the station's observations. The analysis of these station-excluded models revealed that model performance largely depended on whether the training set contained stations with similar characteristics. This analysis suggests that our global ML models can generalize beyond measurement stations if the environments or conditions resemble the stations in our training set.

For the final global ML models, we investigated feature importance and model interpretation in more detail. Both global ML models identified sulphate aerosol and ammonia, carbon monoxide and sulfur dioxide mixing ratios as the most important variables. BC and OM mixing ratios were also indicated as important, though their combined contribution was likely minor. We used the feature importance to interpret some of the model behavior of the MLR model. We noticed that some variables, such as sea salt aerosol, were represented in ways that do not apply universally across locations and conditions, potentially impacting ML model performance.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

AEROSOL Research Discussions

The comparison between MLR_{global} and XGB_{global} fields for 2013 revealed that the models agreed better in Europe, North America, and many other densely populated and anthropogenically influenced regions, including the most densely populated areas in South America, Africa, Middle East, Southern Siberia, South and East Asia. These areas were likely better represented in the training data, making the ML models potentially more reliable in those regions, though we cannot be certain. Conversely, the ML models showed greater disagreement in remote areas—such as deserts, polar regions, rainforests, and oceans—suggesting these environments may be more challenging for the models to capture. Our analysis did not indicate whether MLR or XGB model, if either, performed better in these regions.

Overall, both the MLR and XGB models have their advantages and disadvantages, and our analysis could not definitively determine which model should be used for generating global N_{100} fields. XGB generally performed better and was able to capture N_{100} also in some unique conditions where the MLR model could not. However, in many locations, the MLR model produced equally good results. Additionally, MLR is less prone to overfitting and can produce better estimates when operating outside the range of N_{100} values in the training set. The MLR model also offers greater interpretability, as its variable coefficients can help identify areas where the model is likely to fail.

Our approach produces valuable results, even though our estimates did not fully meet the accuracy threshold suggested by Rosenfeld et al. (2014). At locations outside the training set, only 9 out of 35 stations had at least one ML model with $RMSE_{log10}$ values below 0.2, meaning that in most locations, fewer than 70 % of daily N_{100} concentration estimates fell within factor of 1.5 of observations. Still, our method provides useful insights and enables global N_{100} estimation where direct observations are unavailable. It complements other observation-based methods, such as satellite-derived approaches or the method outlined in Block et al. (2024) and can be used to evaluate purely model-driven results. A key advantage of our method is that it is directly constrained with in situ measurements of N_{100} rather than relying solely on observations via data assimilation. Although our global N_{100} fields were produced for 2013, the global N_{100} time series can be extended to any period covered by CAMS data (currently 2003-2023). Moreover, this methodology could be applied to estimate other atmospheric variables with available in situ measurements and corresponding reanalysis data.

Further evaluating the performance and reliability of the global MLR and XGB models in different environments and conditions will require additional data. We hope future collaborations will provide access to a wider measurement dataset, including data from stations not currently included in this analysis and more data from stations already part of the study. Although adding new data from measurement stations does not provide a global reliability estimate, it will allow us to assess model performance in new environments and conditions with unseen data. With the larger measurement data set, it would be beneficial and straightforward to retrain the global ML models with the method described in this study. We could also explore using shorter datasets, such as measurement campaign data, for testing the models. While these datasets are too short for model training, they could enrich the holdout set by introducing environments that lack long-term measurements.

Data availability. The measured N_{100} dataset is freely available at https://doi.org/10.5281/zenodo.15222674

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

Author contributions. Conceptualization: AO, PP and TN
Methodology: AO, PP, ER, DH and KP
Software: AO, ER and DH
Formal analysis: AO, ER and DH
Funding acquisition: AO, PP, TP, MK and TN
Investigation (incl. experiments): all co-authors
Data curation: AO, PA, JB, BB, DC, MAF, SG, RMH, RKH, TH, AH, KJ, AK, MK, LL, AL, NM, COD, JO, TP, KPŠ, MP, XQ, PT, VV and AW
Project administration: PP
Supervision: PP, VAS, VMK
Visualization: AO
Writing – original draft: AO
Writing – review & editing: AO, PP, VAS and VMK with all other co-authors

Competing interests. Some authors are members of the editorial board of journal AR.

Acknowledgements. CAMS global reanalysis (EAC4)-dataset (Inness et al., 2019) was downloaded from the Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS) (https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4?tab=overview). Contains modified Copernicus Atmosphere Monitoring Service information 2021. ERA5 hourly data on single levels from 1940 to present-dataset (Hersbach et al., 2023) was downloaded from Copernicus Climate Data Store (2021). Contains modified Copernicus Climate Change Service information 2021. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.
Some of the data included in the analysis were collected at the Southern Great Plains (SGP) site of the Atmospheric Radiation Measurement (ARM) user facility, a U.S. Department of Energy (DOE) Office of Science user facility managed by the Biological and Environmental Research program.
We gratefully acknowledge Richard Leaitch for providing the observation data from Alert and Egbert.
We acknowledge CSC – IT Centre for Science, Finland, for providing computational resources.
ChatGPT has been used for editing the language and code.
Aino Ovaska acknowledges funding from the Doctoral Programme in Atmospheric Sciences at the University of Helsinki (ATM-DP).

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

References

Aalto, P., Hämeri, K., Becker, E., Weber, R., Salm, J., Mäkelä, J. M., Hoell, C., O'dowd, C. D., Hansson, H.-C., Väkevä, M., Koponen, I. K., Buzorius, G., and Kulmala, M.: Physical characterization of aerosol particles during nucleation events, Tellus B: Chemical and Physical Meteorology, 53, 344–358, https://doi.org/10.3402/tellusb.v53i4.17127, 2001.

ACTRIS: Vielsalm - ACTRIS NF labelling, https://actris-nf-labelling.out.ocp.fmi.fi/facility/6, 2024.

Albrecht, B. A.: Aerosols, Cloud Microphysics, and Fractional Cloudiness, Science, 245, 1227–1230, https://doi.org/https://doi.org/10.1126/science.245.4923.1227, 1989.

Alduchov, O. A. and Eskridge, R. E.: Improved Magnus form approximation of saturation vapor pressure, Journal of Applied Meteorology and Climatology, 35, 601–609, 1996.

Andreae, M. O. and Rosenfeld, D.: Aerosol–cloud–precipitation interactions. Part 1. The nature and sources of cloud-active aerosols, Earth-Science Reviews, 89, 13–41, https://doi.org/10.1016/j.earscirev.2008.03.001, 2008.

Andreae, M. O., Acevedo, O. C., Araùjo, A., Artaxo, P., Barbosa, C. G., Barbosa, H. M., Brito, J., Carbone, S., Chi, X., Cintra, B. B., Silva, N. F. D., Dias, N. L., Dias-Júnior, C. Q., Ditas, F., Ditz, R., Godoi, A. F., Godoi, R. H., Heimann, M., Hoffmann, T., Kesselmeier, J., Könemann, T., Krüger, M. L., Lavric, J. V., Manzi, A. O., Lopes, A. P., Martins, D. L., Mikhailov, E. F., Moran-Zuloaga, D., Nelson, B. W., Nölscher, A. C., Nogueira, D. S., Piedade, M. T., Pöhlker, C., Pöschl, U., Quesada, C. A., Rizzo, L. V., Ro, C. U., Ruckteschler, N., Sá, L. D., Sá, M. D. O., Sales, C. B., Santos, R. M. D., Saturno, J., Schöngart, J., Sörgel, M., Souza, C. M. D., Souza, R. A. D., Su, H., Targhetta, N., Tóta, J., Trebs, I., Trumbore, S., Eijck, A. V., Walter, D., Wang, Z., Weber, B., Williams, J., Winderlich, J., Wittmann, F., Wolff, S., and Yáñez-Serrano, A. M.: The Amazon Tall Tower Observatory (ATTO): Overview of pilot measurements on ecosystem ecology, meteorology, trace gases, and aerosols, Atmospheric Chemistry and Physics, 15, 10 723–10 776, https://doi.org/10.5194/acp-15-10723-2015, 2015.

Asmi, A., Wiedensohler, A., Laj, P., Fjaeraa, A. M., Sellegri, K., Birmili, W., Weingartner, E., Baltensperger, U., Zdimal, V., Zikova, N., Putaud, J. P., Marinoni, A., Tunved, P., Hansson, H. C., Fiebig, M., Kivekäs, N., Lihavainen, H., Asmi, E., Ulevicius, V., Aalto, P. P., Swietlicki, E., Kristensson, A., Mihalopoulos, N., Kalivitis, N., Kalapov, I., Kiss, G., Leeuw, G. D., Henzing, B., Harrison, R. M., Beddows, D., O'Dowd, C., Jennings, S. G., Flentje, H., Weinhold, K., Meinhardt, F., Ries, L., and Kulmala, M.: Number size distributions and seasonality of submicron particles in Europe 2008-2009, Atmospheric Chemistry and Physics, 11, 5505–5538, https://doi.org/10.5194/acp-11-5505-2011, 2011.

Backman, J., Rizzo, L. V., Hakala, J., Nieminen, T., Manninen, H. E., Morais, F., Aalto, P. P., Siivola, E., Carbone, S., Hillamo, R., Artaxo, P., Virkkula, A., Petäjä, T., and Kulmala, M.: On the diurnal cycle of urban aerosols, black carbon and the occurrence of new particle formation events in springtime São Paulo, Brazil, Atmospheric Chemistry and Physics, 12, 11 733–11 751, https://doi.org/10.5194/acp-12-11733-2012, 2012.

Beigaitė, R., Mechenich, M., and Žliobaitė, I.: Spatial Cross-Validation for Globally Distributed Data, in: Discovery Science. DS 2022, edited by Pascal, P. and Ienco, D., vol. 13601 of *Lecture Notes in Computer Science*, pp. 127–140, Springer, Cham, https://doi.org/10.1007/978-3-031-18840-4_10, 2022.

Bellouin, N., Quaas, J., Gryspeerdt, E., Kinne, S., Stier, P., Watson-Parris, D., Boucher, O., Carslaw, K. S., Christensen, M., Daniau, A. L., Dufresne, J. L., Feingold, G., Fiedler, S., Forster, P., Gettelman, A., Haywood, J. M., Lohmann, U., Malavelle, F., Mauritsen, T., McCoy, D. T., Myhre, G., Mülmenstädt, J., Neubauer, D., Possner, A., Rugenstein, M., Sato, Y., Schulz, M., Schwartz, S. E., Sourdeval,

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

O., Storelvmo, T., Toll, V., Winker, D., and Stevens, B.: Bounding Global Aerosol Radiative Forcing of Climate Change, Reviews of Geophysics, 58, 1–45, https://doi.org/10.1029/2019RG000660, 2020.

Bergman, T., Kerminen, V. M., Korhonen, H., Lehtinen, K. J., Makkonen, R., Arola, A., Mielonen, T., Romakkaniemi, S., Kulmala, M., and Kokkola, H.: Evaluation of the sectional aerosol microphysics module SALSA implementation in ECHAM5-HAM aerosol-climate model, Geoscientific Model Development, 5, 845–868, https://doi.org/10.5194/gmd-5-845-2012, 2012.

Birmili, W., Berresheim, H., Plass-Dülmer, C., Elste, T., Gilge, S., Wiedensohler, A., and Uhrner, U.: Atmospheric Chemistry and Physics The Hohenpeissenberg aerosol formation experiment (HAFEX): a long-term study including size-resolved aerosol, H2SO4 , OH, and monoterpenes measurements, Atmos. Chem. Phys, 3, 361–376, www.atmos-chem-phys.org/acp/3/361/, 2003.

Birmili, W., Weinhold, K., Rasch, F., Sonntag, A., Sun, J., Merkel, M., Wiedensohler, A., Bastian, S., Schladitz, A., Löschau, G., Cyrys, J., Pitz, M., Gu, J., Kusch, T., Flentje, H., Quass, U., Kaminski, H., Kuhlbusch, T. A. J., Meinhardt, F., Schwerin, A., Bath, O., Ries, L., Wirtz, K., and Fiebig, M.: Long-term observations of tropospheric particle number size distributions and equivalent black carbon mass concentrations in the German Ultrafine Aerosol Network (GUAN), Earth system science data, 8, 355–382, https://doi.org/10.5194/essd-8-355-2016, 2016.

Blichner, S. M., Sporre, M. K., Makkonen, R., and Berntsen, T. K.: Implementing a sectional scheme for early aerosol growth from new particle formation in the Norwegian Earth System Model v2: Comparison to observations and climate impacts, Geoscientific Model Development, 14, 3335–3359, https://doi.org/10.5194/gmd-14-3335-2021, 2021.

Block, K., Haghighatnasab, M., Partridge, D. G., Stier, P., and Quaas, J.: Cloud condensation nuclei concentrations derived from the CAMS reanalysis, Earth System Science Data, 16, 443–470, https://doi.org/10.5194/essd-16-443-2024, 2024.

Boucher, O., Randall, D., Artaxo, P., Bretherton, C., Feingold, G., Forster, P., Kerminen, V.-M., Kondo, Y., Liao, H., Lohmann, U., Rasch, P., Satheesh, S., Sherwood, S., Stevens, B., and Zhang, X.: Clouds and Aerosols, in: Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

Charron, A., Birmili, W., and Harrison, R. M.: Factors influencing new particle formation at the rural site, Harwell, United Kingdom, Journal of Geophysical Research Atmospheres, 112, https://doi.org/10.1029/2007JD008425, 2007.

Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Cho, D., Yoo, C., Im, J., and Cha, D. H.: Comparative Assessment of Various Machine Learning-Based Bias Correction Methods for Numerical Weather Prediction Model Forecasts of Extreme Air Temperatures in Urban Areas, Earth and Space Science, 7, https://doi.org/10.1029/2019EA000740, 2020.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M. B., Choirat, C., Koutrakis, P., Lyapustin, A., Wang, Y., Mickley, L. J., and Schwartz, J.: An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution, Environment International, 130, https://doi.org/10.1016/j.envint.2019.104909, 2019.

Dusek, U., Frank, G. P., Hildebrandt, L., Curtius, J., Schneider, J., Walter, S., Chand, D., Drewnick, F., Hings, S., Jung, D., Borrmann, S., and Andreae, M. O.: Size Matters More Than Chemistry for Cloud-Nucleating Ability of Aerosol Particles, Science, 312, 1375–1378, https://doi.org/10.1126/science.1125261, 2006.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

Engler, C., Rose, D., Wehner, B., Wiedensohler, A., Brüggemann, E., Brüggemann, B., Gnauk, T., Spindler, G., Tuch, T., and Birmili, W.: Size distributions of non-volatile particle residuals (D p <800 nm) at a rural site in Germany and relation to air mass origin, Atmos. Chem. Phys, 7, 5785–5802, www.atmos-chem-phys.net/7/5785/2007/, 2007.

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D., Mauritsen, T., Palmer, M., Watanabe, M., Wild, M., and Zhang, H.: Earth's energy budget, climate feedbacks, and climate sensitivity, in: Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, edited by Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., p. 923–1054, Cambridge University Press, https://doi.org/doi: 10.1017/9781009157896.009, 2021.

Gani, S., Bhandari, S., Patel, K., Seraj, S., Soni, P., Arub, Z., Habib, G., Ruiz, L. H., and Apte, J. S.: Particle number concentrations and size distribution in a polluted megacity: The Delhi Aerosol Supersite study, Atmospheric Chemistry and Physics, 20, 8533–8549, https://doi.org/10.5194/acp-20-8533-2020, 2020.

Hamed, A., Joutsensaari, J., Mikkonen, S., Sogacheva, L., Maso, M. D., Kulmala, M., Cavalli, F., Fuzzi, S., Facchini, M. C., Decesari, S., Mircea, M., Lehtinen, K. E. J., and Laaksonen, A.: Atmospheric Chemistry and Physics Nucleation and growth of new particles in Po Valley, Italy, Atmos. Chem. Phys, 7, 355–376, https://doi.org/https://doi.org/10.5194/acp-7-355-2007, 2007.

Hari, P. and Kulmala, M.: Station for Measuring Ecosystem-Atmosphere Relations (SMEAR II), Boreal Environment Research, 10, 315–322, 2005.

Hari, P., Kulmala, M., Pohja, T., Lahti, T., Siivola, E., Palva, E., Aalto, P., Hameri, K., Vesala, T., Luoma, S., and Pulliainen, E.: Air pollution in eastern Lapland: challenge for an environmental measurement station, Silva Fennica, 28, 29–39, https://doi.org/https://doi.org/10.14214/sf.a9160, 1994.

Heinritzi, M., Dada, L., Simon, M., Stolzenburg, D., Wagner, A. C., Fischer, L., Ahonen, L. R., Amanatidis, S., Baalbaki, R., Baccarini, A., Bauer, P. S., Baumgartner, B., Bianchi, F., Brilke, S., Chen, D., Chiu, R., Dias, A., Dommen, J., Duplissy, J., Finkenzeller, H., Frege, C., Fuchs, C., Garmash, O., Gordon, H., Granzin, M., Haddad, I. E., He, X., Helm, J., Hofbauer, V., Hoyle, C. R., Kangasluoma, J., Keber, T., Kim, C., Kürten, A., Lamkaddam, H., Laurila, T. M., Lampilahti, J., Lee, C. P., Lehtipalo, K., Leiminger, M., Mai, H., Makhmutov, V., Manninen, H. E., Marten, R., Mathot, S., Mauldin, R. L., Mentler, B., Molteni, U., Müller, T., Nie, W., Nieminen, T., Onnela, A., Partoll, E., Passananti, M., Petäjä, T., Pfeifer, J., Pospisilova, V., Quéléver, L. L., Rissanen, M. P., Rose, C., Schobesberger, S., Scholz, W., Scholze, K., Sipile, M., Steiner, G., Stozhkov, Y., Tauber, C., Tham, Y. J., Vazquez-Pufleau, M., Virtanen, A., Vogel, A. L., Volkamer, R., Wagner, R., Mingyi, W., Lena, W., Daniela, W., Xiao, M., Yan, C., Ye, P., Zha, Q., Zhou, X., Amorim, A., Baltensperger, U., Hansel, A., Kulmala, M., Tome, A., Winkler, P. M., Worsnop, D. R., Donahue, N. M., Kirkby, J., and Curtius, J.: Molecular understanding of the suppression of new-particle formation by isoprene, Atmospheric Chemistry and Physics, 20, 11 809–11 821, https://doi.org/10.5194/acp-20-11809-2020, 2020.

Heintzenberg, J., Birmili, W., Otto, R., Andreae, M. O., Mayer, J.-C., Chi, X., and Panov, A.: Aerosol particle number size distributions and particulate light absorption at the ZOTTO tall tower (Siberia), 2006–2009, Atmospheric Chemistry and Physics, 11, 8703–8719, https://doi.org/10.5194/acp-11-8703-2011, 2011.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Sabater, J. M., Nicolas, J., Peubey, C., Radu, R., Rozum, I., Schepers, D., 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, accessed: 2023-03-14, 2023.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

Hooda, R. K., Kivekäs, N., O'Connor, E. J., Coen, M. C., Pietikäinen, J. P., Vakkari, V., Backman, J., Henriksson, S. V., Asmi, E., Komppula, M., Korhonen, H., Hyvärinen, A. P., and Lihavainen, H.: Driving Factors of Aerosol Properties Over the Foothills of Central Himalayas Based on 8.5 Years Continuous Measurements, Journal of Geophysical Research: Atmospheres, 123, 13,421–13,442, https://doi.org/10.1029/2018JD029744, 2018.

Hussein, T., Dada, L., Hakala, S., Petäjä, T., and Kulmala, M.: Urban aerosol particle size characterization in Eastern Mediterranean Conditions, Atmosphere, 10, https://doi.org/10.3390/atmos10110710, 2019.

Inness, A., Ades, M., Agusti-Panareda, A., Barre, J., Benedictow, A., Blechschmidt, A.-M., Dominguez, J. J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., Razinger, M., Remy, S., Schulz, M., and Suttie, M.: The CAMS reanalysis of atmospheric composition, Atmospheric chemistry and physics, 19, 3515–3556, https://doi.org/10.5194/acp-19-3515-2019, 2019a.

Inness, A., Ades, M., Agustí-Panareda, A., Barré, J., Benedictow, A., Blechschmidt, A., Dominguez, J., Engelen, R., Eskes, H., Flemming, J., Huijnen, V., Jones, L., Kipling, Z., Massart, S., Parrington, M., Peuch, V.-H., M., R., Remy, S., Schulz, M., and Suttie, M.: CAMS global reanalysis (EAC4), Copernicus atmosphere monitoring service (CAMS) atmosphere data store (ADS), https://ads.atmosphere.copernicus.eu/datasets/cams-global-reanalysis-eac4?tab=overview, accessed: 2023-03-14, 2019b.

Järvi, L., Hannuniemi, H., Hussein, T., Junninen, H., Aalto, P. P., Hillamo, R., Mäkelä, T., Keronen, P., Siivola, E., Vesala, T., and Kulmala, M.: The urban measurement station SMEAR III: Continuous monitoring of air pollution and surface-atmosphere interactions in Helsinki, Finland, Boreal environment research, 14, 86–109, 2009.

Kalivitis, N., Kerminen, V. M., Kouvarakis, G., Stavroulas, I., Bougiatioti, A., Nenes, A., Manninen, H. E., Petäjä, T., Kulmala, M., and Mihalopoulos, N.: Atmospheric new particle formation as a source of CCN in the eastern Mediterranean marine boundary layer, Atmospheric Chemistry and Physics, 15, 9203–9215, https://doi.org/10.5194/acp-15-9203-2015, 2015.

Kerminen, V. M., Paramonov, M., Anttila, T., Riipinen, I., Fountoukis, C., Korhonen, H., Asmi, E., Laakso, L., Lihavainen, H., Swietlicki, E., Svenningsson, B., Asmi, A., Pandis, S. N., Kulmala, M., and Petäjä, T.: Cloud condensation nuclei production associated with atmospheric nucleation: A synthesis based on existing literature and new results, Atmospheric Chemistry and Physics, 12, 12 037–12 059, https://doi.org/10.5194/acp-12-12037-2012, 2012.

Kesti, J., Backman, J., O'Connor, E. J., Hirsikko, A., Asmi, E., Aurela, M., Makkonen, U., Filioglou, M., Komppula, M., Korhonen, H., and Lihavainen, H.: Aerosol particle characteristics measured in the United Arab Emirates and their response to mixing in the boundary layer, Atmospheric Chemistry and Physics, 22, 481–503, https://doi.org/10.5194/acp-22-481-2022, 2022.

Kim, M., Brunner, D., and Kuhlmann, G.: Importance of satellite observations for high-resolution mapping of near-surface NO2 by machine learning, Remote Sensing of Environment, 264, https://doi.org/10.1016/j.rse.2021.112573, 2021.

Korhola, T., Kokkola, H., Korhonen, H., Laaksonen, A., Lehtinen, K. E., and Romakkaniemi, S.: Reallocation in modal aerosol models: Impacts on predicting aerosol radiative effects, Geoscientific Model Development, 7, 161–174, https://doi.org/10.5194/gmd-7-161-2014, 2014.

Kristensson, A., Maso, M. D., Swietlicki, E., Hussein, T., Zhou, J., Kerminen, V. M., and Kulmala, M.: Characterization of new particle formation events at a background site in southern Sweden: Relation to air mass history, Tellus, Series B: Chemical and Physical Meteorology, 60 B, 330–344, https://doi.org/10.1111/j.1600-0889.2008.00345.x, 2008.

Kuhn, M. and Johnson, K.: Applied Predictive Modeling, Springer Science+Business Media, 1 edn., 2013.

Laakso, L., Laakso, H., Aalto, P. P., Keronen, P., Petäjä, T., Nieminen, T., Pohja, T., Siivola, E., Kulmala, M., Kgabi, N., Molefe, M., Mabaso, D., Phalatse, D., Pienaar, K., and Kerminen, V.-M.: Basic characteristics of atmospheric particles, trace gases and meteorology in

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

a relatively clean Southern African Savannah environment, Atmos. Chem. Phys, 8, 4823–4839, www.atmos-chem-phys.net/8/4823/2008/, 2008.

Leaitch, W. R., Sharma, S., Huang, L., Toom-Sauntry, D., Chivulescu, A., Macdonald, A. M., Salzen, K. V., Pierce, J. R., Bertram, A. K., Schroder, J. C., Shantz, N. C., Chang, R. Y., and Norman, A. L.: Dimethyl sulfide control of the clean summertime Arctic aerosol and cloud, Elementa, 1, https://doi.org/10.12952/journal.elementa.000017, 2013.

Lee, S. H., Uin, J., Guenther, A. B., de Gouw, J. A., Yu, F., Nadykto, A. B., Herb, J., Ng, N. L., Koss, A., Brune, W. H., Baumann, K., Kanawade, V. P., Keutsch, F. N., Nenes, A., Olsen, K., Goldstein, A., and Ouyang, Q.: Isoprene suppression of new particle formation: Potential mechanisms and implications, Journal of Geophysical Research, 121, 14 621–14 635, https://doi.org/10.1002/2016JD024844, 2016.

Leinonen, V., Kokkola, H., Yli-Juuti, T., Mielonen, T., Kühn, T., Nieminen, T., Heikkinen, S., Miinalainen, T., Bergman, T., Carslaw, K., Decesari, S., Fiebig, M., Hussein, T., Kivekäs, N., Krejci, R., Kulmala, M., Leskinen, A., Massling, A., Mihalopoulos, N., Mulcahy, J. P., Noe, S. M., Noije, T. V., O'connor, F. M., O'dowd, C., Olivie, D., Pernov, J. B., Petäjä, T., Øyvind Seland, Schulz, M., Scott, C. E., Skov, H., Swietlicki, E., Tuch, T., Wiedensohler, A., Virtanen, A., and Mikkonen, S.: Comparison of particle number size distribution trends in ground measurements and climate models, Atmospheric Chemistry and Physics, 22, 12 873–12 905, https://doi.org/10.5194/acp-22-12873-2022, 2022.

Lihavainen, H., Alghamdi, M. A., Hyvärinen, A. P., Hussein, T., Aaltonen, V., Abdelmaksoud, A. S., Al-Jeelani, H., Almazroui, M., Almehmadi, F. M., Zawad, F. M. A., Hakala, J., Khoder, M., Neitola, K., Petäjä, T., Shabbaj, I. I., and Hämeri, K.: Aerosols physical properties at Hada Al Sham, western Saudi Arabia, Atmospheric Environment, 135, 109–117, https://doi.org/10.1016/j.atmosenv.2016.04.001, 2016.

Liu, Y., Yan, C., Feng, Z., Zheng, F., Fan, X., Zhang, Y., Li, C., Zhou, Y., Lin, Z., Guo, Y., Zhang, Y., Ma, L., Zhou, W., Liu, Z., Dada, L., Dällenbach, K., Kontkanen, J., Cai, R., Chan, T., Chu, B., Du, W., Yao, L., Wang, Y., Cai, J., Kangasluoma, J., Kokkonen, T., Kujansuu, J., Rusanen, A., Deng, C., Fu, Y., Yin, R., Li, X., Lu, Y., Liu, Y., Lian, C., Yang, D., Wang, W., Ge, M., Wang, Y., Worsnop, D. R., Junninen, H., He, H., Kerminen, V. M., Zheng, J., Wang, L., Jiang, J., Petäjä, T., Bianchi, F., and Kulmala, M.: Continuous and comprehensive atmospheric observations in Beijing: a station to understand the complex urban atmospheric environment, Big Earth Data, 4, 295–321, https://doi.org/10.1080/20964471.2020.1798707, 2020.

Ma, J., Ding, Y., Cheng, J. C., Jiang, F., and Wan, Z.: A temporal-spatial interpolation and extrapolation method based on geographic Long Short-Term Memory neural network for PM2.5, Journal of Cleaner Production, 237, https://doi.org/10.1016/j.jclepro.2019.117729, 2019.

Marinescu, P. J., Levin, E. J., Collins, D., Kreidenweis, S. M., and Heever, S. C. V. D.: Quantifying aerosol size distributions and their temporal variability in the Southern Great Plains, USA, Atmospheric Chemistry and Physics, 19, 11 985–12 006, https://doi.org/10.5194/acp-19-11985-2019, 2019.

McFiggans, G., Artaxo, P., Baltensperger, U., Coe, H., Facchini, M. C., Feingold, G., Fuzzi, S., Gysel, M., Laaksonen, A., Lohmann, U., Mentel, T. F., Murphy, D. M., D., C. O., Snider, J. R., and Weingartner, E.: The effect of physical and chemical aerosol properties on warm cloud droplet activation, Atmospheric chemistry and physics, 6, 2593–2649, https://doi.org/10.5194/acp-6-2593-2006, 2006.

Morawska, L., Thomas, S., Jamriska, M., and Johnson, G.: The modality of particle size distributions of environmental aerosols, Atmospheric Environment, 33, 4401–4411, 1999.

Mordas, G., Plauškaite, K., Prokopčiuk, N., Dudoitis, V., Bozzetti, C., and Ulevicius, V.: Observation of new particle formation on Curonian Spit located between continental Europe and Scandinavia, Journal of Aerosol Science, 97, 38–55, https://doi.org/10.1016/j.jaerosci.2016.03.002, 2016.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

Mulcahy, J. P., Johnson, C., Jones, C. G., Povey, A. C., Scott, C. E., Sellar, A., Turnock, S. T., Woodhouse, M. T., Abraham, N. L., Andrews, M. B., Bellouin, N., Browse, J., Carslaw, K. S., Dalvi, M., Folberth, G. A., Glover, M., Grosvenor, D. P., Hardacre, C., Hill, R., Johnson, B., Jones, A., Kipling, Z., Mann, G., Mollard, J., O'Connor, F. M., Palmiéri, J., Reddington, C., Rumbold, S. T., Richardson, M., Schutgens, N. A., Stier, P., Stringer, M., Tang, Y., Walton, J., Woodward, S., and Yool, A.: Description and evaluation of aerosol in UKESM1 and HadGEM3-GC3.1 CMIP6 historical simulations, Geoscientific Model Development, 13, 6383–6423, https://doi.org/10.5194/gmd-13-6383-2020, 2020.

Nair, A. A. and Yu, F.: Using machine learning to derive cloud condensation nuclei number concentrations from commonly available measurements, Atmospheric chemistry and physics, 20, 12 853–12 869, https://doi.org/10.5194/acp-20-12853-2020, 2020.

Nieminen, T., Kerminen, V.-M., Petaja, T., Aalto, P. P., Arshinov, M., Asmi, E., Baltensperger, U., Beddows, D. C. S., Beukes, J. P., Collins, D., Ding, A., Harrison, R. M., Henzing, B., Hooda, R., Hu, M., Horrak, U., Kivekas, N., Komsaare, K., Krejci, R., Kristensson, A., Laakso, L., Laaksonen, A., Leaitch, W. R., Lihavainen, H., Mihalopoulos, N., Nemeth, Z., Nie, W., O'Dowd, C., Salma, I., Sellegri, K., Svenningsson, B., Swietlicki, E., Tunved, P., Ulevicius, V., Vakkari, V., Vana, M., Wiedensohler, A., Wu, Z., Virtanen, A., and Kulmala, M.: Global analysis of continental boundary layer new particle formation based on long-term measurements, Atmospheric chemistry and physics, 18, 14 737–14 756, https://doi.org/10.5194/acp-18-14737-2018, 2018.

Noije, T. V., Bergman, T., Sager, P. L., O'Donnell, D., Makkonen, R., Gonçalves-Ageitos, M., Döscher, R., Fladrich, U., Hardenberg, J. V., Keskinen, J. P., Korhonen, H., Laakso, A., Myriokefalitakis, S., Ollinaho, P., Garciá-Pando, C. P., Reerink, T., Schrödner, R., Wyser, K., and Yang, S.: EC-Earth3-AerChem: A global climate model with interactive aerosols and atmospheric chemistry participating in CMIP6, Geoscientific Model Development, 14, 5637–5668, https://doi.org/10.5194/gmd-14-5637-2021, 2021.

O'Connor, T. C., Jennings, S. G., and O'Dowd, C. D.: Highlights of fifty years of atmospheric aerosol research at Mace Head, Atmospheric Research, 90, 338–355, https://doi.org/10.1016/J.ATMOSRES.2008.08.014, 2008.

Paasonen, P., Asmi, A., Petäjä, T., Kajos, M. K., Äijälä, M., Junninen, H., Holst, T., Abbatt, J. P. D., Arneth, A., Birmili, W., Gon, H. D. V. D., Hamed, A., Hoffer, A., Laakso, L., Laaksonen, A., Leaitch, W. R., Plass-dülmer, C., Pryor, S. C., Räisänen, P., Swietlicki, E., Wiedensohler, A., Worsnop, D. R., matti Kerminen, V., and Kulmala, M.: Warming-induced increase in aerosol number concentration likely to moderate climate change, Nature Geoscience, 6, 438–442, https://doi.org/10.1038/ngeo1800, 2013.

Pierce, J. R., Westervelt, D. M., Atwood, S. A., Barnes, E. A., and Leaitch, W. R.: New-particle formation, growth and climate-relevant particle production in egbert, canada: Analysis from 1 year of size-distribution observations, Atmospheric Chemistry and Physics, 14, 8647–8663, https://doi.org/10.5194/acp-14-8647-2014, 2014.

Pöhlker, M. L., Zhang, M., Braga, R. C., Krüger, O. O., Pöschl, U., and Ervens, B.: Aitken mode particles as CCN in aerosol- And updraft-sensitive regimes of cloud droplet formation, Atmospheric Chemistry and Physics, 21, 11 723–11 740, https://doi.org/10.5194/acp-21-11723-2021, 2021.

Qi, X. M., Ding, A. J., Nie, W., Petäjä, T., Kerminen, V. M., Herrmann, E., Xie, Y. N., Zheng, L. F., Manninen, H., Aalto, P., Sun, J. N., Xu, Z. N., Chi, X. G., Huang, X., Boy, M., Virkkula, A., Yang, X. Q., Fu, C. B., and Kulmala, M.: Aerosol size distribution and new particle formation in the western Yangtze River Delta of China: 2 years of measurements at the SORPES station, Atmospheric Chemistry and Physics, 15, 12 445–12 464, https://doi.org/10.5194/acp-15-12445-2015, 2015.

Quaas, J., Arola, A., Cairns, B., Christensen, M., Deneke, H., Ekman, A. M., Feingold, G., Fridlind, A., Gryspeerdt, E., Hasekamp, O., Li, Z., Lipponen, A., Mülmenstädt, J., Nenes, A., Penner, J. E., Rosenfeld, D., Schrödner, R., Sinclair, K., Sourdeval, O., Stier, P., Tesche, M., Diedenhoven, B. V., and Wendisch, M.: Constraining the Twomey effect from satellite observations: Issues and perspectives, Atmospheric Chemistry and Physics, 20, 15 079–15 099, https://doi.org/10.5194/acp-20-15079-2020, 2020.

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

Rosenfeld, D., Andreae, M. O., Asmi, A., Chin, M., Leeuw, G., Donovan, D. P., Kahn, R., Kinne, S., Kivekäs, N., Kulmala, M., Lau, W., Schmidt, K. S., Suni, T., Wagner, T., Wild, M., and Quaas, J.: Global observations of aerosol-cloud-precipitation-climate interactions, Rev. Geophys., 52, 750–808, https://doi.org/https://doi.org/10.1002/2013RG000441, 2014.

Schladitz, A., Leníček, J., Beneš, I., Kováč, M., Skorkovský, J., Soukup, A., Jandlová, J., Poulain, L., Plachá, H., Löschau, G., and Wiedensohler, A.: Air quality in the German-Czech border region: A focus on harmful fractions of PM and ultrafine particles, Atmospheric Environment, 122, 236–249, https://doi.org/10.1016/j.atmosenv.2015.09.044, 2015.

Schmale, J., Henning, S., Decesari, S., Henzing, B., Keskinen, H., Sellegri, K., Ovadnevaite, J., Pöhlker, M., Brito, J., Bougiatioti, A., Kristensson, A., Kalivitis, N., Stavroulas, I., Carbone, S., Jefferson, A., Park, M., Schlag, P., Iwamoto, Y., Aalto, P., Äijälä, M., Bukowiecki, N., Ehn, M., Fröhlich, R., Frumau, A., Herrmann, E., Herrmann, H., Holzinger, R., Kos, G., Kulmala, M., Mihalopoulos, N., Nenes, A., O'Dowd, C., Petäjä, T., Picard, D., Pöhlker, C., Pöschl, U., Poulain, L., Swietlicki, E., Aneae, M., Artaxo, P., Wiedensohler, A., Ogren, J., Matsuki, A., Yum, S. S., Stratmann, F., Baltensperger, U., and Gysel, M.: Long-term cloud condensation nuclei number concentration, particle number size distribution and chemical composition measurements at regionally representative observatories, Atmospheric chemistry and physics, 18, 2853–2881, https://doi.org/10.5194/acp-18-2853-2018, 2018.

Smith, D. A.: Visualising world population density as an interactive multi-scale map using the global human settlement population layer, Journal of Maps, 13, 117–123, https://doi.org/10.1080/17445647.2017.1400476, 2017.

Smith, D. A.: World Population Density 2020, https://luminocity3d.org/WorldPopDen/#3/37.37/49.04, accessed: 2025-03-04, 2023.

Stier, P., van den Heever, S. C., Christensen, M. W., Gryspeerdt, E., Dagan, G., Saleeby, S. M., Bollasina, M., Donner, L., Emanuel, K., Ekman, A. M., Feingold, G., Field, P., Forster, P., Haywood, J., Kahn, R., Koren, I., Kummerow, C., L'Ecuyer, T., Lohmann, U., Ming, Y., Myhre, G., Quaas, J., Rosenfeld, D., Samset, B., Seifert, A., Stephens, G., and Tao, W. K.: Multifaceted aerosol effects on precipitation, Nature Geoscience, 17, 719–732, https://doi.org/10.1038/s41561-024-01482-6, 2024.

Su, T., Li, Z., Henao, N. R., Luan, Q., and Yu, F.: Constraining effects of aerosol-cloud interaction by accounting for coupling between cloud and land surface, Sci. Adv, 10, https://doi.org/10.1126/sciadv.adl5044, 2024.

Tunved, P. and Ström, J.: On the seasonal variation in observed size distributions in northern Europe and their changes with decreasing anthropogenic emissions in Europe: Climatology and trend analysis based on 17 years of data from Aspvreten, Sweden, Atmospheric Chemistry and Physics, 19, 14 849–14 873, https://doi.org/10.5194/acp-19-14849-2019, 2019.

Twomey, S.: The Influence of Pollution on the Shortwave Albedo of Clouds, Journal of the Atmospheric Sciences, 34, 1149–1152, https://doi.org/10.1175/1520-0469(1977)034<1149:TIOPOT>2.0.CO;2, 1977.

UBA: Das Luftmessnetz des Umweltbundesamtes: Langzeitmessungen, Prozessverständnis und Wirkungen ferntransportierter Luftverunreinigungen, Tech. rep., edited by Schleyer, R. and Bieber, E. and Wallasch, M., Umweltbundesamt (UBA), Dessau-Rosslau, 2013.

Vakkari, V., Beukes, J. P., Laakso, H., Mabaso, D., Pienaar, J. J., Kulmala, M., and Laakso, L.: Long-term observations of aerosol size distributions in semi-clean and polluted savannah in South Africa, Atmospheric Chemistry and Physics, 13, 1751–1770, https://doi.org/10.5194/acp-13-1751-2013, 2013.

Wang, S., Huo, Y., Mu, X., Jiang, P., Xun, S., He, B., Wu, W., Liu, L., and Wang, Y.: A High-Performance Convolutional Neural Network for Ground-Level Ozone Estimation in Eastern China, Remote Sensing, 14, https://doi.org/10.3390/rs14071640, 2022.

Wang, Z., Chen, P., Wang, R., An, Z., and Qiu, L.: Estimation of PM2.5 concentrations with high spatiotemporal resolution in Beijing using the ERA5 dataset and machine learning models, Advances in Space Research, 71, https://doi.org/10.1016/j.asr.2022.12.016, 2023.

Wiedensohler, A., Birmili, W., Nowak, A., Sonntag, A., Weinhold, K., Merkel, M., Wehner, B., Tuch, T., Pfeifer, S., Fiebig, M., M., A. F., Asmi, E., Sellegri, K., Depuy, R., Venzac, H., Villani, P., Laj, P., Aalto, P., Ogren, J. A., Swietlicki, E., Williams, P., Roldin, P., Quincey,

https://doi.org/10.5194/ar-2025-18
Preprint. Discussion started: 1 July 2025

P., Hüglin, C., Fierz-Schmidhauser, R., Gysel, M., Weingartner, E., Riccobono, F., Santos, S., Grüning, C., Faloon, K., Beddows, D., Harrison, R., Monahan, C., Jennings, S. G., D., C. O., Marinoni, A., Horn, H. G., Keck, L., Jiang, J., Scheckman, J., McMurry, P. H., Deng, Z., Zhao, C. S., Moerman, M., Henzing, B., de Leeuw, G., Löschau, G., and Bastian, S.: Mobility particle size spectrometers: harmonization of technical standards and data structure to facilitate high quality long-term observations of atmospheric particle number size distributions, Atmospheric measurement techniques, 5, 657–685, https://doi.org/10.5194/amt-5-657-2012, 2012.

XGBoost Developers: XGBoost Documentation, https://xgboost.readthedocs.io/en/stable/index.html, accessed: 2025-04-01, 2022.

Xian, P., Reid, J. S., Ades, M., Benedetti, A., Colarco, P. R., Silva, A. D., Eck, T. F., Flemming, J., Hyer, E. J., Kipling, Z., Rémy, S., Sekiyama, T. T., Tanaka, T., Yumimoto, K., and Zhang, J.: Intercomparison of aerosol optical depths from four reanalyses and their multi-reanalysis consensus, Atmospheric Chemistry and Physics, 24, 6385–6411, https://doi.org/10.5194/acp-24-6385-2024, 2024.

Yli-Juuti, T., Riipinen, I., Aalto, P. P., Nieminen, T., Maenhaut, W., Janssens, I. A., Claeys, M., Salma, I., Ocskay, R., Hoffer, A., Imre, K., and Kulmala, M.: Characteristics of new particle formation events and cluster ions at K-puszta, Hungary, Hungary. Boreal Env. Res, 14, 683–698, 2009.

Yu, F., Luo, G., Nair, A. A., Tsigaridis, K., and Bauer, S. E.: Use of Machine Learning to Reduce Uncertainties in Particle Number Concentration and Aerosol Indirect Radiative Forcing Predicted by Climate Models, Geophysical Research Letters, 49, https://doi.org/10.1029/2022GL098551, 2022.

Yu, W., Ye, T., Zhang, Y., Xu, R., Lei, Y., Chen, Z., Yang, Z., Zhang, Y., Song, J., Yue, X., Li, S., and Guo, Y.: Global estimates of daily ambient fine particulate matter concentrations and unequal spatiotemporal distribution of population exposure: a machine learning modelling study, The Lancet Planetary Health, 7, e209–e218, https://doi.org/10.1016/S2542-5196(23)00008-6, 2023.

Zíková, N. and Ždímal, V.: Long-Term Measurement of Aerosol Number Size Distributions at Rural Background Station Košetice, Aerosol and Air Quality Research, 13, 1464–1474, https://doi.org/10.4209/aaqr.2013.02.0056, 2013.