Environmental Science: Atmospheres

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Environ. Sci.: Atmos.*, 2025, **5**, 367

Estimating the atmospheric aerosol number size distribution using deep learning

Yusheng Wu, *[a] Martha Arbayani Zaidan,[ab] Runlong Cai,†[a] Jonathan Duplissy,[ac] Magdalena Okuljar,‡[a] Katrianne Lehtipalo,[ad] Tuukka Petäjä[a] and Juha Kangasluoma[a]

The submicron aerosol number size distribution significantly impacts human health, air quality, weather, and climate. However, its measurement requires sophisticated and expensive instrumentation that demands substantial maintenance efforts, leading to limited data availability. To tackle this challenge, we developed estimation models using advanced deep learning algorithms to estimate the aerosol number size distribution based on trace gas concentrations, meteorological parameters, and total aerosol number concentration. These models were trained and validated with 15 years of ambient data from three distinct environments, and data from a fourth station were exclusively used for testing. Our estimative models successfully replicated the trends in the test data, capturing the temporal variations of particles ranging from approximately 10–500 nm, and accurately deriving total number, surface area, and mass concentrations. The model's accuracy for particles below 75 nm is limited without the inclusion of total particle number concentration as training input, highlighting the importance of this parameter for capturing the dynamics of smaller particles. The reliance on total particle number concentration, a parameter not routinely measured at all in air quality monitoring sites, as a key input for accurate estimation of smaller particles presents a practical challenge for broader application of the models. Our models demonstrated a robust generalization capability, offering valuable data for health assessments, regional pollution studies, and climate modeling. The estimation models developed in this work are representative of ambient conditions in Finland, but the methodology in general can be applied in broader regions.

Received 8th September 2024
Accepted 3rd February 2025

DOI: 10.1039/d4ea00127c

rsc.li/esatmospheres

Environmental significance

Aerosol particles, especially submicron particles, play a critical role in air quality, climate, and human health. However, traditional measurements of aerosol number size distributions are limited by expensive, high-maintenance instruments. This study addresses these limitations by using deep learning models to predict aerosol particle size distributions from widely available air quality data. By offering a reliable and accessible method to estimate critical environmental data, the study facilitates better health assessments and pollution management. The approach also broadens access to data that can support climate modeling, particularly in regions lacking the resources for continuous physical monitoring.

1 Introduction

Ambient aerosol particles significantly impact human health, air quality, weather, and climate. Submicron particles, in particular, contribute significantly due to their higher number concentration compared to larger particles, despite generally having lower mass concentrations. The established link between particulate matter exposure and adverse health effects underscores the particular relevance of submicron particles to pulmonary diseases, cancer, and mortality.[1,2] Due to their small size, these particles can penetrate deep into the respiratory tract and enter the bloodstream, significantly increasing health risks.[1–3] The deposition locations of submicron particles within the human body are heavily influenced by their diffusion coefficient and accumulation capacity, both of which are determined by particle size.[4,5] In ambient air, freshly nucleated and emitted particles must grow to a certain size to significantly reduce visibility or become activated as cloud condensation nuclei (CCN) under specific humidity conditions.[6] Accordingly,

[a]*Institute for Atmospheric and Earth System Research (INAR)/Physics, Faculty of Science, University of Helsinki, Finland. E-mail: yusheng.wu@helsinki.fi*
[b]*Department of Computer Science, Faculty of Science, University of Helsinki, Finland*
[c]*Helsinki Institute of Physics, HIP, Faculty of Science, University of Helsinki, Finland*
[d]*Finnish Meteorological Institute, Helsinki, Finland*
† Present address: Shanghai Key Laboratory of Atmospheric Particle Pollution and Prevention (LAP3), Department of Environmental Science & Engineering, Fudan University, Shanghai, China.
‡ Present address: International Laboratory for Air Quality and Health, School of Earth and Atmospheric Sciences, Queensland University of Technology, Brisbane, Australia.

the impact of ambient particles is not only dependent on their total number concentration but also on their size. These particles play a crucial role in both local and global climate.[6–8]

In order to understand the impacts of submicron particles, comprehensive measurements have been carried out worldwide. For example, continuous long-term observation of atmospheric variables, including ambient particles, has been performed as the key method for gaining a comprehensive understanding of the interactions between humans, nature, and the atmosphere.[9] Additionally, air quality monitoring stations have been equipped with instruments to measure total particle number concentration and particle size distributions, which is crucial for understanding the health impacts of submicron particles, enhancing air quality assessments by capturing detailed data on particle size and count, identifying pollution sources, ensuring compliance with emerging regulations, and supporting environmental research on atmospheric chemistry and climate change.[10–12]

Unfortunately, collecting long-term particle size distribution (typically represented as $dN/d\log D_p$, indicating the particle number concentration across size bins normalized by using the logarithmic span of particle diameters) data over a large spatial scale poses a significant challenge due to the high cost of instrumentation and the substantial maintenance workload required. Advanced particle measurement instruments are often expensive, necessitating considerable financial investment for widespread deployment. Additionally, these instruments require regular calibration and maintenance to ensure accurate data collection, which adds to the operational burden.[13,14] As a result, many regions may lack comprehensive particle size distribution data, limiting the ability to fully understand and mitigate the impacts of particulate matter on public health and the environment.[15]

Machine learning is a promising tool for addressing gaps in particle size distribution data, particularly in the context of aerosol physical properties. In the era of artificial intelligence and big data, data mining and machine learning technologies have significantly advanced atmospheric science by enabling more sophisticated data processing and analysis. These technologies have broad applications in understanding aerosol properties, including new-particle formation (NPF), but are not limited to this phenomenon. For instance, in Hyytiälä, Finland, clustering and classification methods have been used to investigate the relationship between the formation and growth of new particles and environmental variables, such as relative humidity and the condensation sink of gaseous precursors.[16,17] Similarly, in the Po Valley, Italy, discriminant analysis has been employed to classify nucleation events, identifying relative humidity, O_3, and radiation as significant factors influencing NPF.[18] A multivariate non-linear mixed-effect model has further demonstrated that relative humidity and O_3 are major predictors of NPF across multiple European sites, including the Po Valley, Melpitz in Germany, and Hohenpeissenberg in Germany.[19] Beyond NPF, machine learning approaches such as mutual information have been effectively utilized to explore non-linear associations between atmospheric variables and various aerosol phenomena.[17,20] Deep learning techniques, including image identification and Bayesian classification methods, have successfully classified NPF events, demonstrating the versatility of these tools in atmospheric research.[21,22] Furthermore, the use of remote sensing generates substantial image data, which are well-suited for analysis through deep learning and other machine learning algorithms. For example, a transfer learning-based method has been applied to study temporal changes in dust properties,[23] while other studies have demonstrated the potential of machine learning in remote sensing applications by providing technical tutorials and showcasing specific architectures, such as a pre-trained AlexNet with pyramid pooling for image scene classification.[24,25] Machine learning algorithms have also been used to develop virtual sensors for estimating the concentration of atmospheric variables, showcasing their ability to model aerosol characteristics without the need for extensive traditional measurement infrastructures.[26,27] A study[28] explores the use of random forest techniques to gain quantitative insights into the impact of air mass history and coastal conditions on the formation and growth of nucleation mode particles in the atmosphere. Another recent study[29] presents a convolutional neural network-based approach to identify new particle formation events from longitudinal global particle number size distribution data, providing a valuable tool for understanding new atmospheric particle formation processes. However, despite these advancements, the application of machine learning to estimate time series data of aerosol particles remains relatively uncommon. This is primarily due to the limited availability of sufficient training data and the complexities involved in hyperparameter tuning, which pose significant challenges in achieving accurate estimation models.

In this paper, we demonstrate that the particle number size distribution can be accurately estimated using data from routine air quality measurements. Estimative models for the aerosol number size distribution are developed based on pre-processed ambient trace gas concentrations, meteorological conditions, and aerosol number concentration. We utilize recurrent neural networks (RNNs) to build these models and systematically tune hyperparameters using an automated machine learning (AutoML) approach.

2 Materials and methods

2.1 Measurement data

All data used in this study come from stations located in Finland and are summarized in Table 1. Data from three stations for measuring atmosphere-ecosystem relations (SMEAR stations)[30] are used to train models for estimating ambient particle size distributions in different ways. These stations are chosen to represent the typical ambient environments in Finland. SMEAR I is a subarctic forest remote station in northern Finland. SMEAR II is a boreal forest regional station in southern Finland.[31] SMEAR III is located in an urban environment in Helsinki, southern Finland.[32] Data from a fourth station, Qvidja, are used as a test set. The model's performance was evaluated only for the year 2019 at the Qvidja station due to limited data availability at that site. Qvidja is located in a coastal

Table 1 Summary of input data used in the study

Station	Location	Environment	Data type	Variables measured	Years
SMEAR I	Northern Finland	Subarctic forest	Meteorological, trace gas, and particle data	Wind speed, wind direction, temperature, relative humidity, pressure, radiation, NO_x, SO_2, CO, O_3, N_{tot}, and particle size distribution (DMPS and APS)	2005–2019 (training)
SMEAR II	Southern Finland	Boreal forest	Meteorological, trace gas, and particle data	Wind speed, wind direction, temperature, relative humidity, pressure, radiation, NO_x, SO_2, CO, O_3, N_{tot}, and particle size distribution (DMPS and APS)	2005–2019 (training)
SMEAR III	Helsinki, Southern Finland	Urban environment	Meteorological, trace gas, and particle data	Wind speed, wind direction, temperature, relative humidity, pressure, radiation, NO_x, SO_2, CO, O_3, N_{tot}, and particle size distribution (DMPS and APS)	2005–2019 (training)
Qvidja	Southwestern Finland	Coastal agriculture	Meteorological, trace gas, and particle data	Wind speed, wind direction, temperature, relative humidity, pressure, radiation, NO_x, SO_2, CO, O_3, and particle size distribution (DMPS and APS) (no N_{tot})	2019 (test)
Test data	SMEAR Stations (I, II, and III)	Various environments	Meteorological, trace gas, and particle data	Same as the respective training data	2020 (test)

agriculture environment in southwestern Finland.[33] Data used as training model inputs include meteorological data (wind speed, wind direction, temperature, relative humidity, pressure, and radiation), trace gas data (NO_x, SO_2, CO, and O_3), and total particle number concentration (N_{tot}) measured by using a condensation particle counter (CPC). N_{tot} is included in selected models to assess its importance, because N_{tot} is not measured in all air quality monitoring stations. The target, measured particle size distribution data, is used to train and test models by comparing them with model outputs. Particle size distribution data are measured by using a differential mobility particle sizer (DMPS) and an aerodynamic particle sizer (APS). The DMPS measured particles in the size range of 3–1000 nm, while the APS measured particles in the aerodynamic diameter size range of 0.53–20 μm, and since the measurement ranges of the DMPS and APS overlap, the DMPS was used to determine the number concentrations of particles up to 700 nm, while APS data were applied for particle sizes exceeding 700 nm.[34,35] Data from 2005 to 2019 from SMEAR stations are used as the training set. Data for 2020 from SMEAR stations and 2019 from Qvidja are the test set.

2.2 Data preprocessing

Data cleaning and synchronisation: the raw data from the SMEAR stations and the Qvidja station need to be checked for errors and inconsistencies. This may involve identifying and correcting outliers, handling missing data, and ensuring that the data from different instruments are properly aligned in time. The cleaned data are then synchronised to a 10-minute time resolution to create a consistent time series for analysis.

Interpolation of missing values: gaps in the data, where measurements are missing for periods up to 6 hours, are filled through interpolation. The threshold reflects a balanced approach for optimizing both data quality and data availability. Linear interpolation is used, which estimates the missing values based on the values of the nearest available data points. This ensures that the time series is complete and continuous, which is important for the model to learn temporal patterns effectively.

Handling of negative size distribution values: negative values in the particle size distribution data are not physically meaningful and are likely due to measurement errors or instrument noise. These negative values are replaced with a small positive number (10^{-5}). This value is chosen to be small enough to minimise its impact on the model training process while still allowing the data to be visualised on a logarithmic scale.

Extraction of temporal features: the day of the year and hour of the day are extracted from the timestamps of the data and added as new features. This allows the model to learn seasonal and diurnal patterns in the data. For example, the model can learn that particle concentrations tend to be higher during certain times of the year or day.

Normalisation of input features: all input features, including meteorological parameters, trace gas concentrations, and particle number concentration, are normalised by removing the mean and scaling to unit variance. This ensures that all features have a similar range and distribution, which can improve the performance and stability of the neural network. Normalisation can prevent features with larger scales from dominating the learning process and can help the optimisation algorithm converge faster.

2.3 Recurrent neural networks

Recurrent Neural Networks (RNNs) are employed to build estimation models due to their ability to handle sequential data

Table 2 Abbreviations of models and estimations

Input station and variables	Testing station	Model	Estimation
SMEAR I: met + gas	SMEAR I	SM1Train-MetGas	SM1Train-SM1Test-MetGas
SMEAR I: met + gas + N_{tot}	SMEAR I	SM1Train-MetGasNtot	SM1Train-SM1Test-MetGasNtot
SMEAR II: met + gas	SMEAR II	SM2Train-MetGas	SM2Train-SM2Test-MetGas
SMEAR II: met + gas + N_{tot}	SMEAR II	SM2Train-MetGasNtot	SM2Train-SM2Test-MetGasNtot
SMEAR III: met + gas	SMEAR III	SM3Train-MetGas	SM3Train-SM3Test-MetGas
SMEAR III: met + gas + N_{tot}	SMEAR III	SM3Train-MetGasNtot	SM3Train-SM3Test-MetGasNtot
ALL SMEAR: met + gas	SMEAR I	AllTrain-MetGas	AllTrain-SM1Test-MetGas
ALL SMEAR: met + gas + N_{tot}	SMEAR I	AllTrain-MetGasNtot	AllTrain-SM1Test-MetGasNtot
ALL SMEAR: met + gas	SMEAR II	AllTrain-MetGas	AllTrain-SM2Test-MetGas
ALL SMEAR: met + gas + N_{tot}	SMEAR II	AllTrain-MetGasNtot	AllTrain-SM2Test-MetGasNtot
ALL SMEAR: met + gas	SMEAR III	AllTrain-MetGas	AllTrain-SM3Test-MetGas
ALL SMEAR: met + gas + N_{tot}	SMEAR III	AllTrain-MetGasNtot	AllTrain-SM3Test-MetGasNtot
ALL SMEAR: met + gas	Qvidja	AllTrain-MetGas	AllTrain-QvidjaTest-MetGas

effectively. While many algorithms are available for modeling, including linear regression (with or without regularization), tree-based ensemble methods like random forests, support vector regression, and deep learning algorithms like convolutional neural networks, atmospheric processes are often highly non-linear. Therefore, linear algorithms are less suitable for our purposes. RNNs are chosen for their inherent capability to model time series data by retaining memory of previous inputs, which allows them to learn from past information during training.

RNNs are particularly well-suited for this task because of their ability to capture temporal dependencies in sequential data. Unlike other deep learning architectures, such as convolutional neural networks, RNNs have an internal memory that allows them to learn from past information and use it to estimate future events. This is essential for accurately modelling atmospheric processes, which are often influenced by historical conditions. RNNs inherently reduce the reliance on extensive feature engineering by automatically learning and extracting relevant features from sequential data. Their ability to capture complex dependencies, temporal patterns, and contextual information within the input data enables them to effectively process and model intricate relationships without requiring additional manual feature augmentation.[36,37] By utilising RNNs, our model can effectively learn the complex relationships between meteorological variables, trace gas concentrations, and particle size distribution over time, leading to more accurate estimations.

We use Mean Squared Error (MSE) as the loss function because it effectively measures the average squared difference between estimated and actual values, providing a clear indication of model accuracy and helping to minimize estimation errors by penalizing larger deviations more significantly.

In RNNs, hyperparameter tuning is a crucial step in training models. Traditionally, with many parameters to optimize, long training times, and multiple folds to prevent information leakage, this process can be cumbersome. We utilize Optuna with default settings for model tuning. Optuna is a popular AutoML tool based on Bayesian methods.[38] The hyperparameters are tuned as follows: the number of layers in the long short-term memory (LSTM) network is set to 1 or 2 to balance model complexity and training efficiency; the batch size ranges from 32 to 256 to manage computational resources and training stability; the number of units per layer ranges from 8 to 128 to capture varying levels of feature representation; the output size varies from 32 to 256 to accommodate different estimation requirements; and the dropout ratio ranges from 0 to 0.5 to prevent overfitting while maintaining model generalization. The hyperparameter space includes three gradient descent optimizers: Root Mean Squared Propagation (RMSprop) with a learning rate from 10^{-5} to 10^{-1}, decay from 0.85 to 0.99, and momentum from 10^{-5} to 10^{-1}; Adaptive Moment Estimation (Adam) with a learning rate from 10^{-5} to 10^{-1}; and Stochastic Gradient Descent (SGD) with a learning rate from 10^{-5} to 10^{-1} and momentum from 10^{-5} to 10^{-1}. Tuning each model required approximately two hours on a single GPU machine.

To examine the interactions between variables from different stations, we train models using data from individual stations as well as a combined dataset from all three stations to capture both station-specific and overall patterns. In total, we build 8 models and generate 13 estimations (see Table 2).

2.4 Random forest

A random forest model is constructed to assess feature importance, which measures the contribution of each feature to the model's estimative performance.[39] Unlike RNNs, which can leverage GPU hardware acceleration, random forest models treat each particle size bin independently and do not benefit from such acceleration, resulting in higher computational costs. Consequently, only three years of SMEAR II data (2015–2017) are used for training the random forest model. Cross-validation with random grid search is employed to optimize the model's hyperparameters.

3 Results

3.1 RNN model evaluation

Fig. 1 provides a detailed comparison between the estimated and measured aerosol number size distributions. As shown in

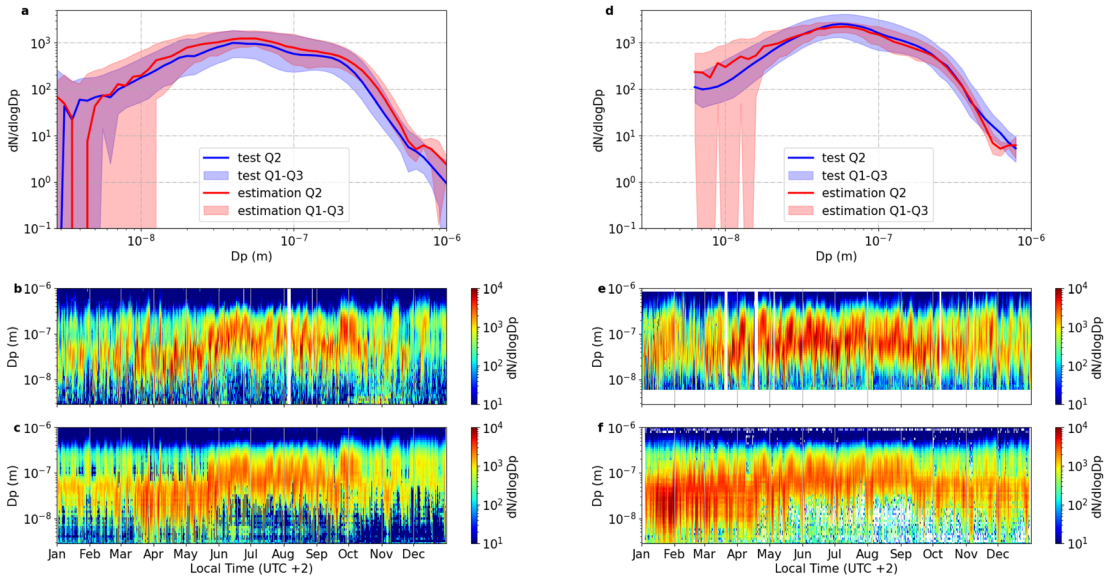

Fig. 1 Comparison of estimations of SMEAR II (AllTrain-SM2Test-MetGas) with the test set. (a) Solid lines represent medians, and shaded areas indicate the 1st and 3rd quartiles. (b) Measured time series from SMEAR II, 2020. (c) Estimated time series for SMEAR II, 2020. Comparison of estimations of Qvidja (AllTrain-QvidjaTest-MetGas) with the test set. (d) Solid lines represent medians, and shaded areas indicate the 1st and 3rd quartiles. (e) Measured time series from SMEAR II, 2020. (f) Estimated time series for SMEAR II, 2020.

Fig. 1a and d, the plots illustrate the distribution of particle number concentrations ($dN/d\log D_p$) against particle diameter from SMEAR II and Qvidja. The solid lines represent the median values (Q2) of the measured and estimated data, while the shaded regions indicate the interquartile range, bound by the first (Q1) and third quartiles (Q3). This visualization highlights how closely the model's estimations align with the actual measurements, providing insight into the accuracy and reliability of the estimative model, particularly across different particle sizes. As shown in Fig. 1b and c, the time series of the particle size distribution is shown, with particle diameter plotted against time and the data colored by the corresponding $dN/d\log D_p$ values. Fig. 1b and e present the observed data, while Fig. 1c and f depict the model's estimations. These figures allow for a temporal comparison, demonstrating the model's capability to replicate not only the overall distribution but also the temporal evolution of particle concentrations across different sizes. One notable aspect is the model's underperformance in estimating particles smaller than 10 nm. This is largely due to the limitations of the Condensation Particle Counter (CPC) used in the measurements, which has a lower detection limit of 10 nm. As a result, the model does not have sufficient training data for these smaller particles, leading to discrepancies in this size range. Despite this limitation, the model performs robustly for larger particles, successfully capturing variations in particle size distribution. This capability is crucial for understanding the temporal dynamics of aerosol concentrations and their implications for environmental and health-related studies. The observations at Qvidja (Fig. 1e) indicate low values from January to March and high values from April to September, while the model (Fig. 1f) does not capture the trend well. This discrepancy likely arises because the Qvidja station is located in a coastal agricultural environment, which differs from the environments of the three SMEAR stations (subarctic forest, boreal forest, and urban) used for training the model. The training data may not fully capture the unique processes influencing the aerosol size distribution at the coastal Qvidja site. Furthermore, the Qvidja data were only used for testing and were limited to the year 2019 due to the availability of data from that station. This is the only year for which a complete and usable dataset is available for model evaluation. This limitation might also contribute to the observed discrepancies.

Fig. 2 also highlights the distinct patterns in estimative accuracy across different environmental settings, as evidenced by the variation in coefficient of determination (R^2) values. R^2 is a statistical measure indicating the proportion of the variance in the dependent variable that is predictable from the independent variables, and is used to evaluate the models' goodness of fit. Negative R^2 values, observed in some cases, suggest that the model's estimations, for those specific instances, do not accurately capture the variations in the observed data and perform worse than simply estimating the mean value. The convex shape of the R^2 curves across all test sets suggests that the estimative models exhibit the highest accuracy in the intermediate particle size range (10–300 nm), while estimations for smaller (<10 nm) and larger (>300 nm) particles are less accurate. This pattern likely stems from the relatively higher abundance and stability of mid-sized particles, making them easier for the models to predict. Notably, the inclusion of the total particle number concentration (N_{tot}) as an input variable significantly enhances the model's performance for smaller particles, particularly those below 75 nm. This improvement is evident in the higher R^2 values observed in the models that

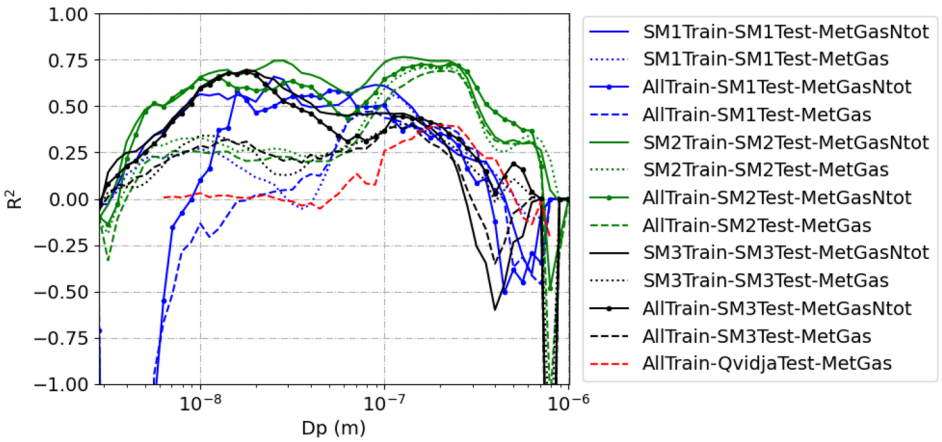

Fig. 2 Performance of models trained on variant data sets, evaluated on variant test sets.

include N_{tot}, indicating that N_{tot} provides critical information that helps capture the dynamics of smaller particle formation and growth processes. While the inclusion of N_{tot} significantly improves the model's performance, particularly for smaller particles, it's essential to acknowledge that N_{tot} measurements are not standard at most monitoring sites. This limitation restricts the model's applicability to sites with available N_{tot} data, highlighting the need for either wider implementation of N_{tot} measurements or alternative approaches to capture the dynamics of smaller particles in the absence of these data. For the SMEAR II and III stations, the model trained on data from all stations performs comparably to those trained on station-specific data, reflecting the similarity in environmental conditions and particle formation processes between these sites. However, the performance for SMEAR I is noticeably lower, likely due to the unique subarctic environment at this station, which differs significantly from the other locations. This discrepancy underscores the importance of tailoring estimative models to the specific characteristics of the station's environment. In the case of the Qvidja station, which was not included in the training set, the model trained on all available data still shows reasonable performance within the 100–400 nm size range. This suggests that while the general patterns of particle size distribution can be captured even in previously unseen environments, the model's estimative power diminishes outside the trained size range, particularly for very small and very large particles. This highlights the need for more diverse training data to improve the generalizability of the models across different environments.

3.2 Derived variables and atmospheric relevance

One of the reasons that particle size distribution is useful is that many important variables can be derived from it. The three commonly used ones are particle number concentration, surface area, and volume (by assuming a constant density, volume can be converted into mass concentration). The number concentration of particles (or ultrafine particles, UFPs) is often used as a single parameter in atmospheric health studies.[40–42] Particle surface area is also highlighted in aerosol related health studies because the aerosol reactive surface area may be one determining factor driving the adverse health effects.[43–45] Particle mass concentration has been used as an important air pollution index for a long time. We use the whole size range of 3–1000 nm for evaluating the derived variables from the estimated particle size distribution. Fig. 3 presents a comparison between the estimated and measured values of three essential variables derived from the particle number size distribution: total particle number concentration (Fig. 3a), surface area (Fig. 3b), and volume (Fig. 3c). The results demonstrate a strong correlation between the estimated and measured values for total number concentration and surface area, with the data points closely aligning along the diagonal, indicating that the model accurately captures these variables. However, the estimation for total particle volume exhibits a broader spread (Fig. 3c), suggesting that the model has greater difficulty accurately representing the contributions of larger particles, which are less frequent but significantly impact volume. There are differences in performance between stations. For instance, at SMEAR III, the correlation for total particle number is stronger than for total surface area and volume. This could be due to the presence of a larger number of bigger particles at this site, which are challenging to model. Overall, the model demonstrates robust performance, particularly in estimating number concentration and surface area, underscoring its potential utility in atmospheric studies.

3.3 Feature importance analysis

Fig. 4 shows the feature importance analysis conducted using the random forest model based on 3 years of SMEAR II data. The *y*-axis represents the different input features used in the model, such as N_{tot} (total particle number concentration), CO, and various meteorological variables (wind components WX and WY, temperature, relative humidity, *etc.*). The *x*-axis displays the importance score for each feature, indicating how much each variable contributes to the estimation of particle size distributions across different size bins. In this figure, the importance of N_{tot} is clearly highlighted, showing a significant impact on estimating particle sizes within the 10–100 nm range. This

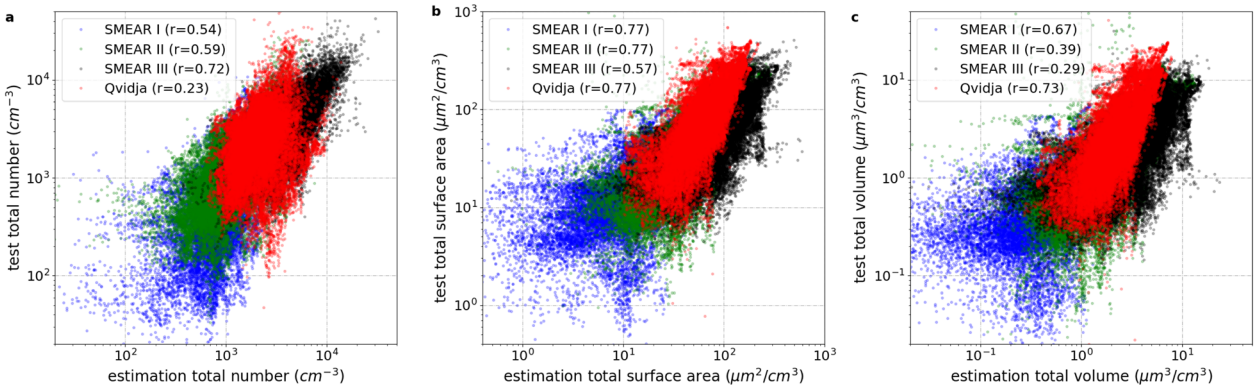

Fig. 3 Comparison of estimated *versus* test data for variables derived from the particle number size distribution across multiple sites. (a) Total number concentration (cm^{-3}), (b) total surface area (μm^2 cm^{-3}), and (c) total volume (μm^3 cm^{-3}). Each point on the plot represents a paired comparison between the model's estimation and the corresponding measured value. The color of the points indicates the station where the data was collected.

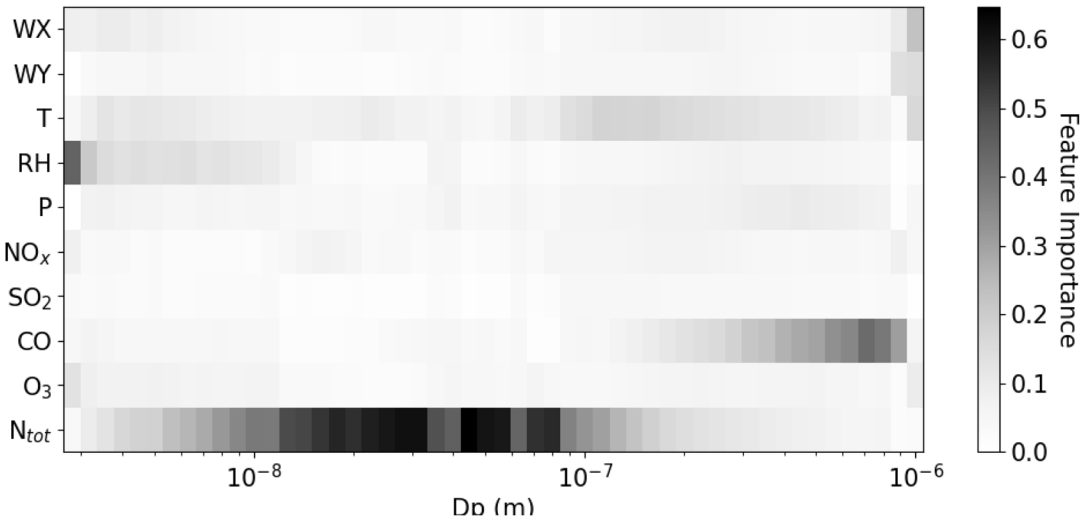

Fig. 4 Random forest feature importance (WX: north component of horizontal wind speed and WY: east component of horizontal wind speed).

suggests that N_{tot} is the most influential feature in determining the concentration of particles in this size range. CO also shows some estimative power, particularly for particles in the 300–1000 nm size range, although its influence is less pronounced compared to that of N_{tot}. The other features, including meteorological parameters like wind speed (WX and WY), temperature, and humidity, exhibit relatively lower importance, indicating that while they contribute to the model, their impact is not as substantial as that of N_{tot} and CO. To ensure that the feature importance values are representative, we averaged the importance scores across all size bins. This provides a more comprehensive view of which features consistently contribute to the model's estimative accuracy.

3.4 Exploring new particle formation events

A new particle formation (NPF) event refers to the process by which gas-phase molecules nucleate and grow within the atmosphere. Fig. 5 focuses on estimating these NPF events at the SMEAR II station, highlighting the performance of models trained with and without N_{tot} as an input variable. The figure compares the measured and estimated particle size distributions over two specific days: a non-event day (2020-06-16) and an event day (2020-05-02). In each subplot, the *x*-axis represents time, while the *y*-axis denotes particle diameter in meters. The color gradient illustrates particle number concentration ($dN/d\log D_p$), with warmer colors indicating higher concentrations. Subplots (a) and (e) show the measured particle size distributions for the non-event and event days, respectively. On the event day (subplot (e)), a distinctive "banana" shape appears at particle sizes below 30 nm, signaling the occurrence of a new particle formation event. Subplots (b) and (f) present estimations from a model trained without N_{tot}. This model fails to accurately capture the particle formation and growth dynamics on the event day, missing the characteristic "banana" shape entirely. This indicates the model's inability to predict the burst of particles smaller than 30 nm during an NPF event when N_{tot} is excluded. In contrast, subplots (c) and (g) display estimations

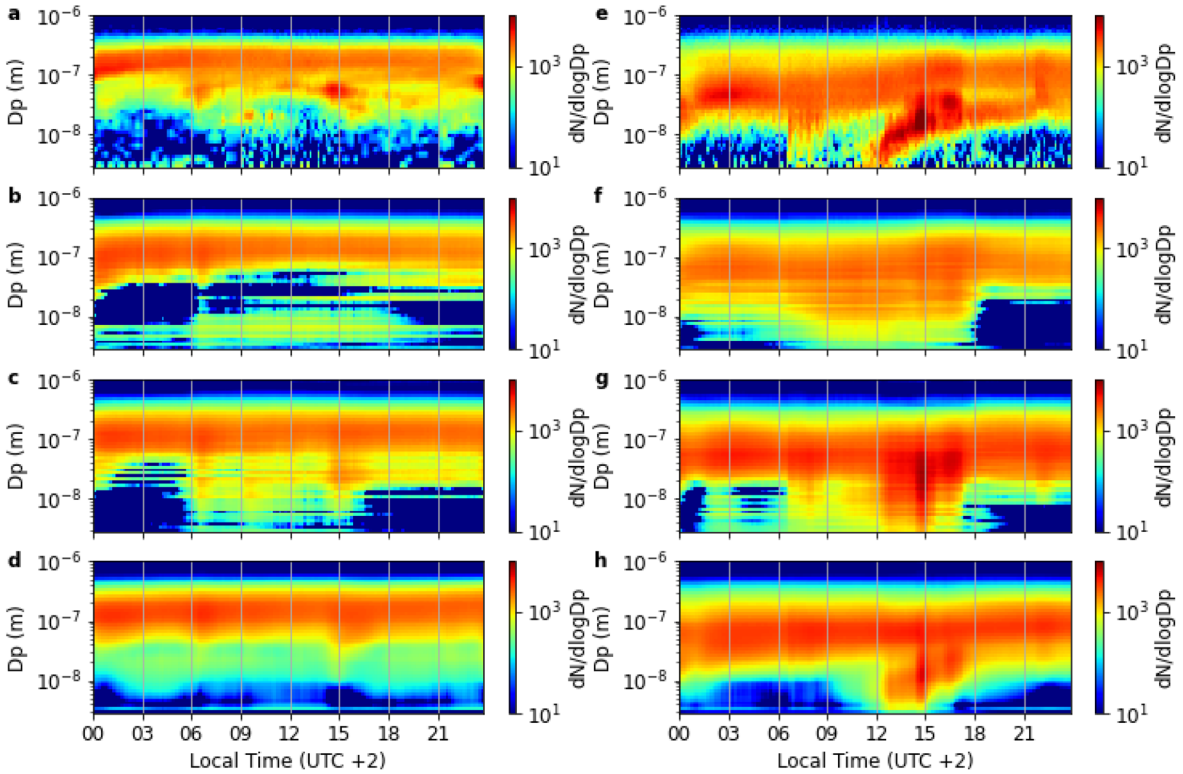

Fig. 5 Examples of the non-event day (2020-06-16. (a) Test. (b) AllTrain-SM2Test-MetGas. (c) AllTrain-SM2Test-MetGasNtot. (d) SM2Train-SM2Test-MetGasNtot) and event day (2020-05-02. (e) Test. (f) AllTrain-SM2Test-MetGas. (g) AllTrain-SM2Test-MetGasNtot. (h) SM2Train-SM2Test-MetGasNtot). The color gradient represents the particle size number distribution ($dN/d\log D_p$).

from a model trained with N_{tot}. These estimations successfully capture the burst of particles smaller than 30 nm on the event day, accurately reproducing the "banana" shape observed in the measured data. This demonstrates the importance of including N_{tot} as an input variable for accurately modeling particle formation dynamics. Furthermore, subplot (d) provides an even more detailed estimation for the event day using a model specifically trained for the SMEAR II station with N_{tot} included. This model not only captures the particle formation event but also offers a more precise depiction of the particle growth dynamics compared to the generalized model shown in subplot (g). The discrepancy in capturing the small event observed at 6–9 am local time between the generalized and specified models could be attributed to several factors. The generalized model, trained on data from all three SMEAR stations, may have learned a broader range of atmospheric conditions and particle formation patterns, allowing it to detect subtle variations that the specified model, trained solely on SMEAR II data, might miss. Additionally, the small event might be associated with specific local meteorological conditions or source contributions that are less pronounced at SMEAR II compared to the other stations, contributing to the specified model's inability to capture it. Further investigation is needed to fully understand the reasons behind this difference in model performance. These comparisons underscore the critical role of including N_{tot} as an input variable in models estimating new particle formation events. The detailed analysis of feature importance and the performance comparison across different models clearly demonstrate that incorporating N_{tot} significantly enhances the accuracy of particle size distribution estimations, particularly for smaller particles during NPF events. Since the model's performance for particles below 10 nm is limited, the NPF event case study may primarily reflect the model's ability to capture overall trends and patterns of particle formation and growth, rather than providing precise quantification of particle concentrations in the sub-10 nm range.

4 Discussion

The most significant aspect the models have in common is that they work better for medium particle sizes (about 10–300 nm), while the performance for smaller and larger sizes is not as good. One possible reason is the inherent data quality of the training set. For sub-10 nm particles, the primary reason is that the N_{tot} is measured with a CPC that detects particles only larger than about 10 nm and does not reflect variations in sub-10 nm particle concentrations. The concentration of large-size particles in the atmosphere is relatively low, leading to a larger relative uncertainty in the measurements. Small particles, mainly introduced to the atmosphere through the new particle formation process, depend strongly on gas-phase precursors such as sulfuric acid, ammonia, and other low volatile compounds, which are not included as inputs in the models. On the other hand, large particles are often affected by seasonal

and local events, such as pollen and dust, which are sporadic and not explicitly included in the model input. The importance of N_{tot} as an input for accurate estimation of small particles is evident in Fig. 2, 4, and 5. The model trained with N_{tot} can provide more detailed information than that without N_{tot} (Fig. 5). Therefore, if cost-effective CPCs that can be added to routine observations are developed, it will greatly promote research on the generation and evolution of atmospheric particles.

In terms of the accuracy of the estimated results, the model including N_{tot} as an input can predict the new particle formation event, but it is not sufficient for more detailed quantitative analysis, such as calculating the new particle formation rate and growth rate. The reason that models cannot predict particles smaller than 100 nm in Qvidja (Fig. 2) may be related to the marine agricultural environment, which has a fast nucleation and growth rate locally, and these particle formation mechanisms are not present at the other stations. Fig. 3a shows that the correlation of the derived number concentrations of the test and estimation is weak because particles smaller than 100 nm contribute significantly to the total particle number concentrations.

The model's performance was found to be most accurate within the 10–500 nm particle size range, which can be attributed to the availability of more extensive and reliable data for these particle sizes. The limitations of the instruments used to measure particle size distribution, particularly for particles smaller than 10 nm and larger than 500 nm, likely contributed to the reduced accuracy in these ranges. To improve the model's ability to predict the full spectrum of particle sizes, future research should focus on incorporating data from instruments with wider detection ranges and enhanced sensitivity, particularly for the smallest and largest particle sizes.

The methods of this research can still be improved in the future. Use of longer data sets may train better models. When we selected the features, we did not use $PM_{2.5}$. This is because the concentration of $PM_{2.5}$ in the atmosphere in Finland is relatively low and close to the detection limit of the instruments, so the measurement results fluctuate significantly. $PM_{2.5}$, as a common measurement parameter, is likely to be used as a feature for model input in the future, especially in urban environments. This study is based on the understanding that N_{tot}, reflecting the total number of particles present at a given time, directly influences their distribution across different size ranges at the same time point. However, future studies could explore using lagged N_{tot} values, representing past measurements, as an alternative approach to better capture the temporal evolution of particle size distribution. While our current analysis focused primarily on diurnal and seasonal trends, future studies could investigate the models' performance in estimating weekly variations, particularly in environments with distinct weekly patterns in anthropogenic emissions.[46] The stability of the boundary layer, which can significantly influence vertical mixing and aerosol transport processes, was not explicitly considered in the current model. Incorporating parameters reflecting boundary layer stability, such as mixing layer height or stability indices, could potentially enhance the model's ability to capture the temporal variations in aerosol size distribution, particularly during periods of strong diurnal changes in atmospheric stability.[47] As our models are trained exclusively on data from Finland, their applicability might be limited to geographical regions with comparable meteorological conditions, such as similar temperature ranges, relative humidity levels, and wind patterns, as well as similar atmospheric composition, including concentrations of trace gases and pre-existing aerosols. To assess the model's generalizability, testing with data from different environments, particularly those with distinct meteorological and atmospheric characteristics, is crucial. It is foreseeable that if data from stations of different environment types are used to train the model, a model with better performance and generalization will be trained. It is possible to train models for different environments; for instance, models trained specifically for urban environments may be used for aerosol particle exposure estimates.

Transfer learning using certain time series neural network models is worth trying, instead of training a model from scratch.[48] It is worth mentioning that ensemble learning can integrate multiple models into one, demonstrating strong estimative performance in many fields,[49] and it can also be a direction to improve models in the future.[50] This study primarily highlights the estimative capabilities of RNN models; however, a thorough evaluation of their computational efficiency compared to that of traditional models remains an important area for future exploration. Finally, causal inference has flourished in the past two decades and is considered one of the key directions for data science.[51,52] If the tools of causal inference are applied to follow-up studies, the interpretability of the model will be improved, potentially guiding traditional research based on physical and chemical analyses.

5 Conclusion

Overall, the estimated results successfully capture the trends and patterns of the measured data, as evidenced in Fig. 1. This indicates that the models are effective in replicating the general behaviors of aerosol number size distributions. However, the tendency of the trained models to be overfitted to specific data source stations suggests a need for further refinement. By incorporating data from a broader range of measuring sites, these models can be generalized to predict aerosol distributions across more diverse geographical areas. Additionally, due to the higher quality of sampling data in the medium particle size range, the models demonstrated superior estimating accuracy for this range, achieving R^2 values around 0.5, compared to 0.2 for small and large size ranges. This disparity highlights the importance of data quality in model performance. The conversion of estimated results into number concentration, surface area concentration, and volume concentration of particulate matter further validates the model's estimation capabilities, with correlation coefficients mostly ranging from 0.5 to 0.8. These results underscore the model's potential to effectively predict different particulate matter characteristics, which is critical for applications in environmental health and air quality

management. The feature importance analysis using the random forest model revealed that total concentration and CO are the most significant features, indicating that these factors are vital for accurate estimations. Looking ahead, there are several avenues for enhancing this research. Future improvements could involve using a larger sample size, incorporating additional features, expanding the hyperparameter space, experimenting with more algorithms, and integrating data from various sites. Furthermore, exploring model integration and applying causal inference methods could enhance model robustness and interpretability, ultimately leading to more accurate and generalizable estimates in diverse environments.

Data availability

Data are available from the SMEAR database: **https://smear.avaa.csc.fi/**. No new data were generated.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was funded by the Research Council of Finland (325656, 364223, 356134, 346370, and 355330), Business Finland (6909/31/2022), and a University of Helsinki three year grant (75284132).

Notes and references

1 A. Zanobetti, M. Franklin, P. Koutrakis and J. Schwartz, *Environ. Health*, 2009, **8**, 1–12.
2 R. McConnell, T. Islam, K. Shankardass, M. Jerrett, F. Lurmann, F. Gilliland, J. Gauderman, E. Avol, N. Künzli, L. Yao, J. Peters and K. Berhane, *Environ. Health Perspect.*, 2010, **118**, 1021–1026.
3 U. Gehring, A. H. Wijga, M. Brauer, P. Fischer, J. C. de Jongste, M. Kerkhof, M. Oldenwening, H. A. Smit and B. Brunekreef, *Am. J. Respir. Crit. Care Med.*, 2010, **181**, 596–603.
4 R. J. Delfino, C. Sioutas and S. Malik, *Environ. Health Perspect.*, 2005, **113**, 934–946.
5 E. V. Bräuner, L. Forchhammer, P. Møller, J. Simonsen, M. Glasius, P. Wåhlin, O. Raaschou-Nielsen and S. Loft, *Environ. Health Perspect.*, 2007, **115**, 1177–1182.
6 J. H. Seinfeld and S. N. Pandis, *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*, 2016.
7 V.-M. Kerminen, M. Paramonov, T. Anttila, I. Riipinen, C. Fountoukis, H. Korhonen, E. Asmi, L. Laakso, H. Lihavainen, E. Swietlicki, B. Svenningsson, A. Asmi, S. N. Pandis, M. Kulmala and T. Petäjä, *Atmos. Chem. Phys.*, 2012, **12**, 12037–12059.
8 Y. Shen, A. Virkkula, A. Ding, K. Luoma, H. Keskinen, P. P. Aalto, X. Chi, X. Qi, W. Nie, X. Huang, T. Petäjä, M. Kulmala and V.-M. Kerminen, *Atmos. Chem. Phys.*, 2019, **19**, 15483–15502.
9 M. Kulmala, *Nature*, 2018, **553**, 21–23.
10 K. Donaldson, V. Stone, P. Gilmour, D. Brown and W. MacNee, *Philos. Trans. R. Soc. London, Ser. A Philos. Trans.: Math., Phys. Eng. Sci.*, 2000, **358**, 2741–2749.
11 K.-H. Ahn and B. Y. H. Liu, *J. Aerosol Sci.*, 1990, **21**, 263–275.
12 R. S. Sokhi, N. Moussiopoulos, A. Baklanov, J. Bartzis, I. Coll, S. Finardi, R. Friedrich, C. Geels, T. Grönholm, T. Halenka, M. Ketzel, A. Maragkidou, V. Matthias, J. Moldanova, L. Ntziachristos, K. Schäfer, P. Suppan, G. Tsegas, G. Carmichael, V. Franco, S. Hanna, J.-P. Jalkanen, G. J. M. Velders and J. Kukkonen, *Atmos. Chem. Phys.*, 2022, **22**, 4615–4703.
13 P. H. McMurry, *Atmos. Environ.*, 2000, **34**, 1959–1999.
14 M. A. Zaidan, L. Dada, M. A. Alghamdi, H. Al-Jeelani, H. Lihavainen, A. Hyvärinen and T. Hussein, *Appl. Sci.*, 2019, **9**, 4475.
15 P. L. Fung, M. A. Zaidan, O. Surakhi, S. Tarkoma, T. Petäjä and T. Hussein, *Atmos. Meas. Tech.*, 2021, **14**, 5535–5554.
16 S. Hyvönen, H. Junninen, L. Laakso, M. Dal Maso, T. Grönholm, B. Bonn, P. Keronen, P. Aalto, V. Hiltunen, T. Pohja, S. Launiainen, P. Hari, H. Mannila and M. Kulmala, *Atmos. Chem. Phys.*, 2005, **5**, 3345–3356.
17 M. A. Zaidan, V. Haapasilta, R. Relan, P. Paasonen, V.-M. Kerminen, H. Junninen, M. Kulmala and A. S. Foster, *Atmos. Chem. Phys.*, 2018, **18**, 12699–12714.
18 S. Mikkonen, K. E. Lehtinen, A. Hamed, J. Joutsensaari, M. Facchini and A. Laaksonen, *Atmos. Chem. Phys.*, 2006, **6**, 5549–5557.
19 S. Mikkonen, H. Korhonen, S. Romakkaniemi, J. Smith, J. Joutsensaari, K. Lehtinen, A. Hamed, T. Breider, W. Birmili, G. Spindler, C. Plass-Duelmer, M. Facchini and A. Laaksonen, *Geosci. Model Dev.*, 2011, **4**, 1–13.
20 P. Laarne, E. Amnell, M. A. Zaidan, S. Mikkonen and T. Nieminen, *Atmosphere*, 2022, **13**, 1046.
21 J. Joutsensaari, M. Ozon, T. Nieminen, S. Mikkonen, T. Lähivaara, S. Decesari, M. C. Facchini, A. Laaksonen and K. E. Lehtinen, *Atmos. Chem. Phys.*, 2018, **18**, 9597–9615.
22 M. Zaidan, V. Haapasilta, R. Relan, H. Junninen, P. Aalto, M. Kulmala, L. Laurson and A. Foster, *Tellus B: Chem. Phys. Meteorol.*, 2018, **70**, 1–10.
23 Y. Ma, W. Gong and F. Mao, *J. Quant. Spectrosc. Radiat. Transfer*, 2015, **153**, 119–130.
24 L. Zhang, L. Zhang and B. Du, *IEEE Geosci. Remote Sens. Mag.*, 2016, **4**, 22–40.
25 X. Han, Y. Zhong, L. Cao and L. Zhang, *Remote Sens.*, 2017, **9**, 848.
26 M. A. Zaidan, D. Wraith, B. E. Boor and T. Hussein, *Appl. Sci.*, 2019, **9**, 4976.
27 M. A. Zaidan, N. H. Motlagh, P. L. Fung, D. Lu, H. Timonen, J. Kuula, J. V. Niemi, S. Tarkoma, T. Petäjä, M. Kulmala and T. Hussein, *IEEE Sens. J.*, 2020, **20**, 13638–13652.
28 C. Yang, H. Dong, Y. Chen, L. Xu, G. Chen, X. Fan, Y. Wang, Y. J. Tham, Z. Lin, M. Li, *et al.*, *Environ. Sci. Technol.*, 2023, **58**, 1187–1198.
29 S. Kecorius, L. Madueño, M. Lovric, N. Racic, M. Schwarz, J. Cyrys, J. A. Casquero-Vera, L. Alados-Arboledas, S. Conil, J. Sciare, *et al.*, *Sci. Data*, 2024, **11**, 1239.

30 P. Hari, E. Nikinmaa, T. Pohja, E. Siivola, J. Bäck, T. Vesala and M. Kulmala, *Physical and Physiological Forest Ecology*, Springer Netherlands, pp. , pp. 471–487.
31 P. Hari and M. Kulmala, *Environ. Res.*, 2005, **10**, 315–322.
32 J. Järvi, H. Hannuniemi, T. Hussein, H. Junninen, P. P. Aalto, R. Hillamo, T. Mäkelä, P. Keronen, E. Siivola, T. Vesala and M. Kulmala, *Boreal Environ. Res.*, 2009, **14**, 86.
33 M. Olin, M. Okuljar, M. P. Rissanen, J. Kalliokoski, J. Shen, L. Dada, M. Lampimäki, Y. Wu, A. Lohila, J. Duplissy, M. Sipilä, T. Petäjä, M. Kulmala and M. Dal Maso, *Atmos. Chem. Phys. Discuss.*, 2022, **2022**, 1–26.
34 M. Dal Maso, M. Kulmala, I. Riipinen, R. Wagner, T. Hussein, P. P. Aalto and K. E. J. Lehtinen, *Boreal Environ. Res.*, 2025, **10**, 323–336.
35 A. Virkkula, J. Backman, P. Aalto, M. Hulkkonen, L. Riuttanen, T. Nieminen, M. Dal Maso, L. Sogacheva, G. De Leeuw and M. Kulmala, *Atmos. Chem. Phys.*, 2011, **11**, 4445–4468.
36 F. Shaheen, B. Verma and M. Asafuddoula, *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2016, pp. 1–8.
37 H. Sheil, O. Rana and R. Reilly, *arXiv*, 2018, preprint, arXiv:1807.08207, DOI: **10.48550/arXiv.1807.08207**.
38 T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.
39 K. J. Archer and R. V. Kimes, *Comput. Stat. Data Anal.*, 2008, **52**, 2249–2260.
40 M. W. Frampton, M. J. Utell, W. Zareba, G. Oberdörster, C. Cox, L. S. Huang, P. E. Morrow, F. Lee, D. Chalupa, L. M. Frasier, D. Speers and J. Stewart, *Research Report*, Health Effects Institute, 2004, pp. 1–47.
41 M. Strak, H. Boogaard, K. Meliefste, M. Oldenwening, M. Zuurbier, B. Brunekreef and G. Hoek, *Occup. Environ. Med.*, 2010, **67**, 118–124.
42 S. Weichenthal, R. Kulka, A. Dubeau, C. Martin, D. Wang and R. Dales, *Environ. Health Perspect.*, 2011, **119**, 1373–1378.
43 R. J. Delfino, C. Sioutas and S. Malik, *Environ. Health Perspect.*, 2005, **113**, 934–946.
44 G. Oberdörster, E. Oberdörster and J. Oberdörster, *Environ. Health Perspect.*, 2005, **113**, 823–839.
45 X. Meng, Y. Ma, R. Chen, Z. Zhou, B. Chen and H. Kan, *Environ. Health Perspect.*, 2013, **121**, 1174–1178.
46 A. Georgoulias and K. Kourtidis, *Atmos. Chem. Phys.*, 2011, **11**, 4611–4632.
47 H. Luo, L. Dong, Y. Chen, Y. Zhao, D. Zhao, M. Huang, D. Ding, J. Liao, T. Ma, M. Hu, et al., *Atmos. Chem. Phys.*, 2022, **22**, 2507–2524.
48 K. Weiss, T. M. Khoshgoftaar and D. Wang, *J. Big Data*, 2016, **3**, 1–40.
49 O. Sagi and L. Rokach, *Wiley Interdiscip. Rev.: Data Min. Knowl. Discovery*, 2018, **8**, e1249.
50 M. A. Zaidan, N. H. Motlagh, P. L. Fung, A. S. Khalaf, Y. Matsumi, A. Ding, S. Tarkoma, T. Petäjä, M. Kulmala and T. Hussein, *IEEE Trans. Ind. Inf.*, 2023, **19**, 1366–1379.
51 J. Pearl, *Stat. Surv.*, 2009, **3**, 96–146.
52 J. Pearl, *Causality: Objectives and Assessment*, 2010, pp. 39–58.