



Machine learning-based downscaling of aerosol size distributions from a global climate model

Antti Vartiainen^{1,2}, Santtu Mikkonen^{1,3}, Ville Leinonen⁴, Tuukka Petäjä⁵, Alfred Wiedensohler⁶, Thomas Kühn⁷, and Tuuli Miinalainen⁸

¹Department of Technical Physics, University of Eastern Finland, Kuopio, 70211, Finland

²Advanced Computing Facility, CSC – IT Center for Science Ltd, Espoo, 02150, Finland

³Department of Environmental and Biological Sciences, University of Eastern Finland, Kuopio, 70211, Finland

⁴Aerosol Physics Laboratory, Tampere University, Tampere, 33014, Finland

⁵Institute for Atmospheric and Earth System Research (INAR)/Physics, Faculty of Science, University of Helsinki, Helsinki, 00014, Finland

⁶Leibniz Institute for Tropospheric Research (TROPOS), Leipzig, 04318, Germany

⁷Weather and Climate Impact Research, Finnish Meteorological Institute, Helsinki, 00560, Finland

⁸Space and Earth Observation Centre, Finnish Meteorological Institute, Helsinki, 00560, Finland

Correspondence: Antti Vartiainen (antti.vartiainen@uef.fi)

Received: 19 February 2025 – Discussion started: 24 March 2025

Revised: 8 August 2025 – Accepted: 2 September 2025 – Published: 24 October 2025

Abstract. Air pollution, particularly exposure to ultrafine particles (UFPs) with diameters below 100 nm, poses significant health risks, yet their spatial and temporal variability complicates impact assessments. This study explores the potential of machine learning (ML) techniques in enhancing the accuracy of a global aerosol-climate model's outputs through statistical downscaling to better represent observed data at specific sites. Specifically, the study focuses on the particle number size distributions from the global aerosol-climate model ECHAM-HAMMOZ. The coarse horizontal resolution of ECHAM-HAMMOZ (approx. 200 km) makes modeling sub-gridscale phenomena, such as UFP concentrations, highly challenging. Data from three European measurement stations (Helsinki, Leipzig, and Melpitz) were used as target of downscaling, covering nucleation, Aitken, and accumulation particle size ranges during years 2016–2018. Six different ML methods (Random Forest, XGBoost, Neural Networks, Support Vector Machine, Gaussian Process Regression and Generalized Linear Model) were employed, with hyperparameter optimization and feature selection integrated for model improvement. A separate ML model was trained for each of the sites and size ranges. Results showed a notable improvement in prediction accuracy for all particle sizes compared to the original global model outputs, particu-

ularly for the accumulation subrange. Challenges remained particularly in downscaling the nucleation subrange, likely due to its high variability and the discrepancy in spatial scale between the climate model representation and the underlying processes. Additionally, the study revealed that the choice of downscaling method requires careful consideration of spatial and temporal dimensions as well as the characteristics of the target variable, as different particle size ranges or variables in other studies may necessitate tailored approaches. The study demonstrates the feasibility of ML-based downscaling for enhancing air quality assessments. This approach could support future epidemiological studies and inform policies on pollutant exposure. Future integration of ML models dynamically into global climate model frameworks could further refine climate predictions and health impact studies.

1 Introduction

Air pollution is considered one of the leading global health risks, in terms of both associated premature deaths and disability (GBD 2019 Risk Factors Collaborators, 2020). Fine particulate matter (PM_{2.5}, particle diameter < 2.5 µm) has been found to be especially harmful; a recent report from

the Global Burden of Disease study identifies it as the most important environmental health risk factor (McDuffie et al., 2021). Although PM_{2.5} has undergone extensive study, exposure to and health impacts of smaller particle sizes, such as ultrafine particles (UFPs) with diameters below 100 nm, remain less well understood (Fuzzi et al., 2015; Vogli et al., 2023). Different sized particles contribute to different aspects of the ambient particle concentration – UFPs mainly control the concentrations in terms of number, while coarser particles control the concentrations in terms of mass (PM_{2.5}). The size of aerosol particles influences, for example, their deposition in the human respiratory tract and their reactive surface area, which in turn can affect their potential to cause health problems (Kreyling et al., 2004; Schraufnagel, 2020). According to epidemiological and toxicological studies, UFPs can more easily enter the alveoli in the lungs, and from there reach other organs (Kreyling et al., 2004; Schraufnagel, 2020). Compared to larger particles, they can thus potentially contribute to, for example, diabetes (Bai et al., 2018), cancer (Pagano et al., 1996), and ischemic cardiovascular disease (Downward et al., 2018; Li et al., 2017; Ostro et al., 2015) more strongly. However, due to their high spatial and temporal variability, estimating exposure to UFPs is challenging, leading to uncertain or even conflicting conclusions regarding their health impacts (Vachon et al., 2024a; Schraufnagel, 2020). Currently, both the World Health Organization (World Health Organization, 2021) and the European Union (European Council, 2008) provide guidelines on safe exposure limits for PM_{2.5} and PM₁₀ (diameter < 10 µm), but no such limits exist for UFP. Indeed, according to a recent review of the topic (Schraufnagel, 2020), UFPs are, in many ways, “at the frontier of air pollution research”.

As the availability of exposure data limits the potential to conduct epidemiological UFP studies (Vachon et al., 2024a), various approaches have been used to gain more information on UFP number concentrations. To study exposure to pollutants, observations from a scarce network of sites have typically been expanded to cover the study area, such as a city, through methods like land use regression (LUR) (Venuta et al., 2024; Amini et al., 2024; Wolf et al., 2017) or interpolation (Jung et al., 2023). Sometimes, more measurements are done in relatively short campaigns to improve the spatial coverage (Vogli et al., 2023; Downward et al., 2018), or satellite-based observations are added as inputs to LUR models to more accurately represent spatial or temporal variability of pollutants (Zani et al., 2020; Jung et al., 2023; Stafoggia et al., 2019). However, most such studies are focused on pollutants other than UFPs (Lin et al., 2022). In recent years, machine learning (ML) has also been a common tool in improving the accuracy of LUR models, often outperforming traditional statistical methods (Vachon et al., 2024b).

While improved spatial characterization of present-day air quality is valuable for understanding its health implications, predicting how air quality may evolve in the future is also important. Actions taken to mitigate climate change might

significantly affect the emissions of pollutants or their precursors, thus hindering the prediction ability of LUR models. Furthermore, since these models are purely descriptive and not integrated with physics-based tools, they cannot be used for studying air quality under varying emission scenarios.

Besides the said statistical methods, local-scale air quality is commonly represented using deterministic models that simulate, for example, the emissions, transport, and transformation of pollutants (Sofiev et al., 2006; Johansson et al., 2022). As the inputs of these models can in principle be modified based on the climate change scenario of interest, they could be suitable for long-term air quality prediction. However, many physics-based models simulate only gaseous pollutants such as NO_x (Pepe et al., 2016) and O₃ (Sharma et al., 2013); aerosols, if supported, may be limited to larger particle sizes (Friberg et al., 2017), omitting UFP. Additionally, running simulations can be computationally expensive and always requires boundary conditions from global climate models, further increasing computational costs. The physics-based air quality models are also not ideal for all sites, as accounting for urban infrastructure or complex terrain requires detailed information about local topography, which is often either unavailable or not accurately captured by local-scale models. (Hinestroza-Ramirez et al., 2023).

Compared to air quality models, global-scale climate models generally produce more output variables, potentially also containing size-resolved representations of aerosols. Simulating long-term global changes in aerosol concentrations is possible with climate models, as they incorporate a broader range of atmospheric processes and feedback mechanisms compared to regional climate models. Since global-scale models are already necessary to generate boundary conditions for regional models, using them directly for generating air quality estimates might seem practical. However, for local-scale air quality estimation, the resolution of current climate models is far too coarse, typically ranging from tens to hundreds of kilometers horizontally and tens to hundreds of meters vertically (Turnock et al., 2020). Particularly for UFPs, the challenge arises from their number concentrations being governed by processes such as primary emissions and secondary formation and growth, which occur both in multiple scales. The initial cluster formation occurs in sub-grid spatial scales and it is highly spatially variable (Dada et al., 2023), including a contribution from traffic as well (Rönkkö et al., 2017), while the growth to Aitken and accumulation mode sizes (see Sect. 3) takes place in synoptical scale (Petäjä et al., 2022). All of this makes the particle size distribution of the nucleation mode highly variable in space and time.

An approach known as downscaling can be applied to the low-resolution outputs of global climate models, with the aim of improving their accuracy in the local scale. In this context, the nested approach of initializing regional climate models with boundary conditions from global simulations is called dynamical downscaling

(Maraun and Widmann, 2018). Another technique, statistical downscaling, instead aims to find a statistical dependence between coarse-resolution outputs and local observations of the quantity of interest. This dependence can later be used for output correction as a post-processing step. The benefit of statistical downscaling lies in its computational efficiency compared to the computationally much more costly dynamical downscaling (Xu et al., 2020). Most of the literature on downscaling focuses on correcting meteorological variables such as temperature (Li et al., 2020; Goyal et al., 2011; Kim and Villarini, 2024) and precipitation (Xu et al., 2020; Vandal et al., 2017; Sachindra et al., 2018). Some recent studies have applied statistical downscaling to air quality variables (Miinalainen et al., 2023; Gouldsbrough et al., 2024; Ivatt and Evans, 2020) but only a few to UFP number concentrations (Kohl et al., 2023). Although the statistical methods for downscaling have typically been simple bias corrections or linear regressions (Maraun and Widmann, 2018), many downscaling studies from the past few years have instead utilized ML methods with promising results (Xu et al., 2020; Sachindra et al., 2018; Miinalainen et al., 2023; Gouldsbrough et al., 2024). To our knowledge, however, none so far has applied ML methods to UFP downscaling.

In this study, using various ML methods, we have down-scaled aerosol particle number size distributions produced by a global aerosol-climate model to better match observations from three measurement stations. We used data from two urban stations (Helsinki, Finland and Leipzig, Germany) and from one rural background measurement station (Melpitz, Germany). The size distribution was represented by three size ranges, the so-called nucleation, Aitken, and accumulation subranges. We have opted to avoid calling these subranges “modes”, as the subranges do not exactly match the conventional mode definitions due to limitations in the size resolution of the climate model representation. Based on these subranges, we categorized the simulated and observed daily average particle number concentrations (PNCs). All three subranges overlap with the UFP size range (< 100 nm), with the nucleation subrange being barely above molecular cluster size and a part of the accumulation subrange exceeding 100 nm. A broad selection of ML methods, six in total, was used and is further described in Sect. 4.4. Moreover, ML model hyperparameters were optimized and feature selection performed, to obtain optimal model configurations for the task. Finally, a glimpse into the inner workings of the black box ML models was provided by a game theoretical method, SHAP, that aims to explain the usage of features by these models. The objective of the study is to act as a proof of concept, showcasing the potential of ML methods in improving predicted PNCs in different parts of the UFP size range. Additionally, comparing the performance of ML methods can help determine whether some of them could be particularly recommended. Ultimately, such ML-based downscaling of UFP could serve to study air quality simultaneously with other aspects of climate change, helping

improve policy-making by accounting for more diverse consequences. Past air quality datasets could also be expanded using this method for studying UFP health effects. A new air quality directive of the European Union (European Council, 2024) includes a mandate to measure UFP concentrations throughout urban and rural supersites in Europe. In the next decade, this is expected to enhance the availability of UFP data for health studies. Our approach could provide additional insights into UFP number concentrations already before the new measurements are implemented.

2 Global aerosol-climate model simulation

We conducted the global climate model simulation with the aerosol-climate model ECHAM6.3-HAM2.3 (ECHAM-HAMMOZ) (Schultz et al., 2018; Tegen et al., 2019; Neubauer et al., 2019). ECHAM-HAMMOZ includes the general circulation model ECHAM (Stevens et al., 2013), the aerosol module HAM (Tegen et al., 2019; Neubauer et al., 2019) and also the chemistry module MOZART (Schultz et al., 2018). We use the aerosol microphysics module SALSA2.0 (Kokkola et al., 2018), which discretizes the aerosol size distribution into ten size classes and treats a soluble and an insoluble sub-population separately. A more detailed description of the SALSA module is presented in Kokkola et al. (2018). The size classes range from 3 nm to 10 μ m, from which we have selected the seven smallest classes (3 to 700 nm) as a basis of the nucleation, Aitken, and accumulation subranges that will constitute the target variables of our study, which are compared against measurements (see Sect. 3 for more details). Additionally, all ten size classes are included as input variables in the ML models, along with the other simulated variables on which the downscaling is based (see Table S1 in the Supplement). In the aerosol module HAM, the aerosol compounds treated are black carbon (BC), sulfate (SU), organic aerosol (OA), sea salt (SS), and mineral dust (DU). The grid resolution that was used in our simulation was T63L47, corresponding approximately $1.9^\circ \times 1.9^\circ$ horizontal resolution. The grid extends vertically to 0.01 hPa (~ 80 km) and there are 47 vertical hybrid layers. The layer nearest to surface has a height of approximately 65 m. The chemistry module MOZART was not included in our setup, but instead a simplified scheme for sulfur chemistry was used (Feichter et al., 1996; Stier et al., 2005).

The simulation was performed using prescribed sea surface temperature and sea ice cover from data from the Atmospheric Model Intercomparison Project (AMIP) of the Program for Climate Model Diagnosis and Intercomparison (PCDMI) (Taylor et al., 2012). In addition, large-scale meteorological fields, wind and surface pressure, were nudged towards ERA5 reanalysis data (Hersbach et al., 2020, 2017). Temperature and free static energy were allowed to evolve freely (Zhang et al., 2014). The anthropogenic aerosol

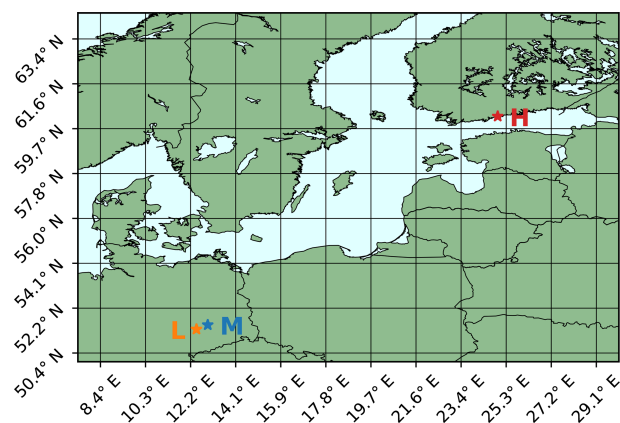


Figure 1. Measurement site locations marked on the map of north-eastern Europe. M, L, and H refer to the Melpitz, Leipzig, and Helsinki sites, respectively. The grid lines show the coarse output resolution of ECHAM-HAMMOZ. As can be seen, Leipzig and Melpitz are located in the same grid cell. Coordinates have been rounded to one decimal place.

emissions were prescribed as monthly averages from the ECLIPSE V6b emission inventory (Stohl et al., 2015; IIASA, 2024). A more detailed description of the aerosol emission input fields can be found from Miinalainen et al. (2023).

3 Aerosol measurements

Measurement data used in this study consist of atmospheric measurements from three stations (Fig. 1), spanning years 2016–2018. Specifically, particle number size distribution data measured with the DMPS/SMPS (Differential/Scanning Mobility Particle Sizer) instruments were obtained from the stations. The data are openly available from SmartSMEAR (SmartSMEAR, 2025; Junninen et al., 2009, Helsinki) and EBAS databases (EBAS, 2025, Leipzig and Melpitz). These sites were chosen because of the relatively long and continuous time series of aerosol size distributions available. In this section, we focus on providing short descriptions of the stations' locations and regional representativeness, relevant for interpreting the results. More detailed information about the measuring stations can be found in the references of their short descriptions in the paragraphs below.

The size distribution data have been reformulated into number concentrations of three subranges, nucleation (< 7.7 nm), Aitken (7.7–50 nm), and accumulation (50–700 nm). These size ranges were selected to correspond to the SALSA bins 1a1 for nucleation, 1a2–1a3 for Aitken, and 2a1–2a4 for accumulation. The three subranges that constitute the target variables of the study were formed by summation over the relevant bins. The hourly number concentrations of the three subranges have then been averaged to one-day time resolution. These data processing steps were performed by Leinonen et al. (2022), and are further dis-

cussed there. However, in our study, days with fewer than eight hours of measurements were also removed from the dataset to ensure representativeness. No other conversions or postprocessing steps were performed. Daily averaged number concentrations of the subranges, which act as the target variables in our study, are presented together with simulated ECHAM-HAMMOZ concentrations in Fig. 2.

The Helsinki station ($60^{\circ}12' \text{ N } 24^{\circ}58' \text{ E}$) is situated in Helsinki, Finland (see Hussein et al., 2008, and Järvi et al., 2009, for details). Helsinki is the largest city in Finland (approx. 675 000 inhabitants), and more than 1.5 million inhabitants live in the Helsinki metropolitan area. The station is classified as an urban station. In Helsinki, the particle number size distribution is measured with DMPS (TSI), with size range of 3 nm to $1 \mu\text{m}$. All three subranges were available from the site.

The Melpitz station ($51^{\circ}32' \text{ N } 12^{\circ}54' \text{ E}$) is situated in Germany, in the southwest of the small town of Torgau (approx. 20 000 inhabitants), immediately west of the village of Melpitz (Hamed et al., 2010). It is classified as a rural background station (Birmili et al., 2016). Particle number size distribution is measured with SMPS (TSI) with size range of 5 to 800 nm. All three subranges were also available from the Melpitz station.

The Leipzig station ($51^{\circ}21' \text{ N } 12^{\circ}26' \text{ E}$) is situated in the city of Leipzig, Germany (approx. 590 000 inhabitants), about 4 km east of the city center (more detailed description in Birmili et al., 2016). It is classified as an urban background station. Aitken and accumulation subrange number concentrations are available from this site. The measured particle size range was between 10 and 800 nm (using DMPS), and for this reason, the number concentration of the nucleation subrange is not available.

4 Analysis methods

4.1 Downscaling workflow

Statistical downscaling, in this application, refers specifically to finding and utilizing a relationship between the large number of ECHAM-HAMMOZ output variables and the observed particle number concentrations (PNCs) at the three sites with the aid of ML methods. If the ML models can learn this dependence, they no longer need the measurement data to function, but can predict the PNC based purely on the climate simulation. In other words, the outputs of the ECHAM simulation constitute the inputs of the ML models, and the output of the ML models is the number concentration of a specific size subrange. The ML output is also referred to as the target variable, while the input variables are commonly called features. The inputs are properly introduced in Sect. 4.2 and listed exhaustively in Table S1. The downscaling is site-specific, that is, the ML models are trained separately for each station and size subrange.

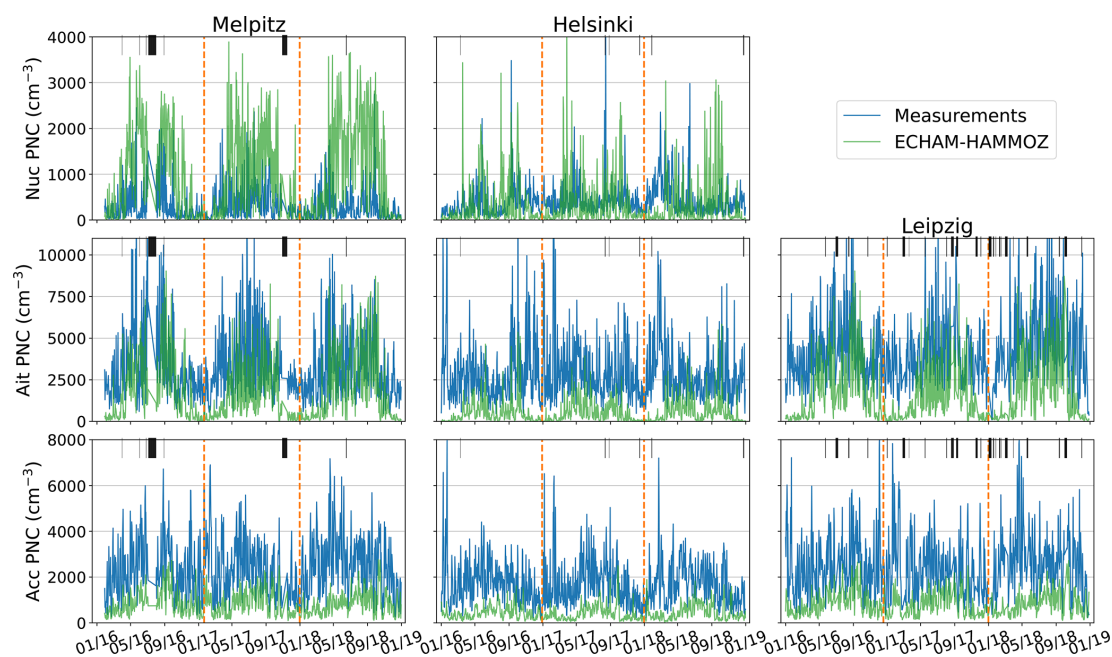


Figure 2. Daily averages of measured and simulated PNCs per subrange of particle size distribution (rows) at each of the three sites (columns), from 2016 to 2018. The black bars represent missing data. Partitioning of the data into training, validation, and testing subsets is shown by the dashed vertical lines. Acc, Ait, and Nuc refer to the accumulation, Aitken, and nucleation subranges, respectively.

As the first step of the ML pipeline, the dataset consisting of simulated features (i.e., the simulation outputs of ECHAM-HAMMOZ) and measured PNC was split into three subsets, known as training, validation, and testing sets (see Sect. 4.2). Next, the number of input variables was reduced through feature selection (Sect. 4.3), the hyperparameters of the models were tuned (Sect. 4.5), and the models were trained, while evaluating the models' performance at each step where necessary. In ML terminology, hyperparameters refer to the tunable parameters of the ML algorithms. This is in contrast to the internal parameters, such as linear model coefficients or neural network weights, which are not controlled by the user but automatically selected by the algorithms. After the optimized, trained models were obtained, further analysis and comparison was done.

On each iteration step of the hyperparameter optimization, an ML model with particular hyperparameter values was fit to the training data, and the trained model was then tested on the validation data to obtain a measure of its goodness-of-fit. The purpose of this common approach is to avoid overfitting the model to the training data, which would impair its generalizability. A separate validation set was used instead of k -fold cross-validation to avoid temporal leakage of information, where future data is used for training and past for testing (Fraga et al., 2023). Rolling variants of k -fold cross-validation that retain the ordering of the data were considered, but initial tests showed poor generalization across folds, possibly due to the seasonality inherent in the data. Thus, k -fold cross-validation was not used in this study.

When the optimization was finished, the combination of hyperparameters leading to the best validation performance was selected, and the ML model with this configuration was retrained on the combined training and validation subsets. Then, the model was applied to the testing subset to evaluate its performance on completely unseen data. This was done because there could be a slight, optimistic bias in the validation score when the hyperparameters have been selected to be optimal for the validation subset. Because some of the ML methods (RF and the NNs, see Sect. 4.4) utilize randomness as part of their algorithms, and therefore depend on the initialization of the random number generator, the retraining part was repeated with 50 different seed numbers. In the results, the mean performance is shown for these methods. Finally, the SHAP method (see Sect. 4.6) was applied to the trained models to investigate their use of input variables.

4.2 Data preparation

For each station, we used ECHAM-HAMMOZ data from only one ECHAM grid cell, which contained the station coordinates and altitude. The ECHAM-HAMMOZ data with a vertical dimension were interpolated to correspond to the station altitude. This was done by utilizing the CDO command line tool (Schulzweida, 2023), and by using the nearest layer to the surface as well as the second lowest layer. Thus, the global scale simulation data were matched with the point measurements. Both the observed and simulated data were then averaged to daily resolution. We selected 93 vari-

ables directly from ECHAM-HAMMOZ data as input variables, covering meteorology, aerosol composition and size distribution. Additional variables were created to replace or complement the existing simulated variables. The u and v components of wind were transformed into two cyclic directional components (north-south, $wind_{ns}$ and east-west, $wind_{ew}$, both varying between -1 and 1) and one variable for the absolute magnitude of the wind vector ($wind_{speed}$). Finally, time of the year was represented by $time_{ws}$ (winter–summer variability) and $time_{sa}$ (spring–autumn variability). These were also formed cyclically (and thus vary between -1 and 1) to avoid a large difference between the value of the last day of the year and the first of the next year, which occurs when using linear time and can disrupt the ML models. In practice, transforming a variable into two cyclic components was done by applying sinusoidal functions (sine and cosine) to the proportion of the variable's current value relative to its maximum (for example, $time_{ws} = \cos([\text{day number}] / [\text{days in year}])$). In our naming convention for all four cyclic features, the positive direction is referenced first (i.e., $wind_{ns} = 1$ means northerly wind and $time_{ws} = 1$ means New Year). All in all, these changes resulted in 100 input variables to be considered for the ML models (see Table S1). These variables were chosen for their potential relevance for aerosol formation, transport or removal, or because they represent properties of the particle size distribution in the climate model.

The three years of data were split into three subsets used for training, validation, and testing of the ML models. Because of missing data, not all subsets could cover a full year (See Fig. 2). The last year of data (2018) was reserved for testing to prioritize the completeness of the test results, while the rest of the data were split equally between training and validation. The difference in the number of samples between the subsets therefore depended on the measurement site. As atmospheric phenomena often show strong seasonality, it was deemed beneficial to have data from throughout the year in all subsets, even if increasing the size of the training subsets could also have been useful.

As most of the ML methods used in this study benefit from feature normalization, the input data were normalized to zero mean and one standard deviation, computed from the training set. No other preprocessing techniques were applied. For all parts of the ML procedure that utilize randomness, the random number generator was initialized with an arbitrary seed number (1024858913).

4.3 Feature selection

Reducing the number of input variables by removing redundant or less impactful ones can improve the performance of ML models, as well as mitigate unnecessary computational costs. Simpler models are also easier to interpret. Therefore, a feature selection scheme was applied before feeding the simulation data into the ML models. A typical way to drop

redundant variables is to see if some of the inputs correlate strongly, and only retain non-correlating ones (up to some threshold value). Dependence between each feature and the target variable, on the other hand, can be an indicator of the feature's relevance. Each criterion alone, however, provides only a partial view and could result either in an ineffective selection or the removal of important information. In addition, as feature selection is still often performed manually rather than in a data-driven manner, these issues could be further exacerbated by human error.

In this study, these ideas were combined to take both redundancy and relevance into account through a data-driven approach. First, a threshold for redundancy (hereafter red_thresh) was selected based on the methods described in Sect. 4.5.1. For each feature, the number of high-correlation pairs (correlations larger than red_thresh) that included the feature was counted. The feature participating in the largest number of such pairs was then dropped. This was repeated until no pair of features exceeded the threshold. In case the number of high-correlation cases was equal for two or more features, the magnitudes of the correlations were compared, and the one with the larger sum of correlations was dropped. If only two highly correlated features remain and the sums are thus equal, we have removed the one that appears earlier in the column order of the input data. After this, another threshold was set for relevance (later rel_thresh), and each feature whose correlation with the target fell below this threshold was also dropped. Both relevance and redundancy were measured by Spearman's correlation coefficient to account for nonlinear dependencies.

The two threshold values were optimized along the model-specific hyperparameters for each ML method and dataset to ensure optimal choice of features. Our approach was inspired by the more commonly used minimum redundancy – maximum relevance (mRMR) method (Ding and Peng, 2005), which we also initially compared to both our approach and a more typical correlation-based selection procedure (not presented here). As further motivation for our approach, it is worth noting that mRMR does not offer a mechanism to adjust the relative weighting of relevance and redundancy, which could be beneficial in certain cases.

Naturally, removing any feature during selection entails a trade-off: the model loses some amount of information, even if that information appears redundant or insignificant in the training data. While feature selection is based on observed correlations and redundancy, it is still possible that a removed feature could improve predictions in future scenarios – particularly for out-of-distribution data. To mitigate this risk, we combined feature selection and model-specific hyperparameter optimization using cross-validation on a holdout validation set, allowing the process to account for generalization performance. Additionally, as a robustness check, we repeated the model training using all available input features (i.e., without applying feature selection), to compare perfor-

mance and ensure that potentially valuable information was not systematically excluded.

4.4 Machine learning methods

In this study, the downscaling task is performed using six different statistical methods. Unless mentioned otherwise, all methods were implemented by the Python package *scikit-learn* version 1.1.1 (Pedregosa et al., 2011). The implementations feature a varying number of hyperparameters, some of which were optimized (listed in Tables S2–S8), while others were either left to their default values or given some other constant value. If some non-default constant value was used, it is mentioned in this section. For more detailed information on the effects of the hyperparameters, readers are referred to the documentation of the methods.

Random Forest (RF) and XGBoost (further abbreviated XGB in figures) are based on different approaches to an ensemble of decision trees. RF takes advantage of randomness to reduce the dependence between individual trees in the ensemble, thereby reducing the ensemble's total variance (Breiman, 2001; Hastie et al., 2009). XGBoost, belonging to the class of gradient boosting methods, generally builds smaller trees with less initial variance and aims to reduce the total bias of the ensemble by sequentially adding trees that correct the errors made by the preceding trees (Friedman, 2001). Unlike the other five methods, XGBoost was implemented by the standalone XGBoost library version 2.0.3 for Python (Chen and Guestrin, 2016). The `RandomForestRegressor` function from *scikit-learn* was used for RF.

Neural networks (NN), in their most basic form, are made of layers of interconnected nodes that each produce a linear combination of the incoming signals, which is then transformed by a non-linear activation function (Alpaydin, 2014). The first layer is composed of the inputs, while the last layer produces the output(s). The layers between them are referred to as hidden layers. Such simple feedforward NNs are also known as multilayer perceptrons (MLP). The *scikit-learn* function `MLPRegressor` was used as our NN, and two different versions were trained separately: one with one hidden layer (NN1), and another with two hidden layers (NN2). These model architectures were also considered distinct from the point of view of our comparison, increasing the number of methods in the results section from six to seven. Before optimization, two hyperparameters were given constant values based on preliminary tests: batch size was set to 32 and solver to “Adam”.

Instead of fitting a complex non-linear function to the training data, the Support Vector Machine (SVM) transforms the data into a higher dimension, where it then fits a linear model to it (Cortes and Vapnik, 1995). In practice, this computationally demanding coordinate transformation can be replaced by a kernel operation by choosing suitable basis functions (Alpaydin, 2014). Additionally, if a data point's distance to the fitted hyperplane were smaller than a specified

amount, the point would be ignored by the fit. This way, the model's tolerance to minor errors can be controlled (Alpaydin, 2014). A function called SVR from *scikit-learn* was used to implement the SVM model. Before optimization, the upper limit for solver iterations (*max_iter*) was set to 10 000, as some unsuitable hyperparameter combinations could cause the iteration to become stuck. Concurrently, the SVM's cache size hyperparameter was increased from the default 200 to 1000 MB to avoid issues with insufficient memory.

Gaussian Processes (GP) take a Bayesian approach to ML by conditioning a prior distribution, again represented by a kernel function, on the training data. The mean of the resulting posterior process can then be used as a prediction (Rasmussen and Williams, 2005). `GaussianProcessRegressor` from *scikit-learn* was used for this study. Most of its hyperparameters were set before the optimization, leaving *alpha* as the only optimizable parameter. The number of restarts was set to nine (meaning ten runs in total), and *normalize_y* was set to True, as recommended for zero-mean, unit-variance priors in the documentation. Additionally, *copy_X_train* was set to False, as the training inputs are not changed during the optimization and thus do not need to be saved. An RBF (Radial Basis Function) kernel with length scale bounds (1×10^{-10} , 1×10^2) was selected as the covariance function of the GP. The length scale of the kernel is optimized internally by `GaussianProcessRegressor`, and not as part of the hyperparameter optimization procedure.

The sixth method in the comparison was the Generalized Linear Model (GLM). It generalizes, and improves upon, linear regression by allowing a non-Gaussian error distribution, and enabling a nonlinear relationship between the inputs and the target through a so-called link function (McCullagh and Nelder, 1989). Nevertheless, GLM does not utilize interactions between inputs unless they are explicitly defined, making it considerably simpler compared to the other methods. Because of its relative simplicity, GLM is not always considered a pure ML method. In *scikit-learn*, GLM is implemented by the `TweedieRegressor` function.

4.5 Hyperparameter optimization

4.5.1 Optimization methods

Finding the hyperparameter values that result in a model configuration with the highest predictive performance can be seen as an optimization problem, where the *objective function* to be optimized takes the hyperparameters as inputs and produces as output some measure of the goodness-of-fit of the corresponding ML model. Each evaluation of the objective therefore involves training an ML model and testing it against observations, which can make a brute force search through hyperparameter combinations extremely slow. To minimize the number of evaluations, the Bayesian optimization (BO) approach aims to approximate the expensive-to-evaluate objective through a surrogate function, such as a

Gaussian Process (Brochu et al., 2010). The surrogate function is updated every time a new point is evaluated, and can be used to strategically select the next point either in a region of uncertainty (favoring exploration) or closer to previously found extrema (favoring exploitation). An *acquisition function* determines which points should be evaluated, and can often be tuned to balance the trade-off between exploration and exploitation.

In this study, a BO algorithm from the Python package `bayesian-optimization` version 1.4.3 was used to search for optimal values of the hyperparameters and feature selection thresholds (Nogueira, 2014). This implementation uses a GP as the surrogate function. The default kernel for the GP is the Matérn kernel, which has a parameter ν controlling the smoothness of the sampled functions. Another tunable parameter of the optimizer is the noise level α of the GP itself. For the acquisition function, the package's default option is the Upper Confidence Bound (UCB) function

$$\text{UCB}(\mathbf{x}) = \mu(\mathbf{x}) + \kappa \sigma(\mathbf{x}) \quad (1)$$

where κ controls how much weight should be given to the posterior's standard deviation $\sigma(\mathbf{x})$ relative to its mean $\mu(\mathbf{x})$ at some point $\mathbf{x}$ of the hyperparameter space. That is, a higher κ favors exploration, focusing the search on regions of higher uncertainty. We have generally used the default settings for both the acquisition function and the GP, apart from some customization that is described next.

Many of the hyperparameters in ML models are either integer-valued (e.g. number of estimators in an ensemble) or categorical (e.g. choice of activation function in NNs), while the GP of the BO algorithm utilized in this study only supports optimization of hyperparameters with continuous values. A common solution to this is to take the point suggested by the acquisition function, and either round the hyperparameters to the closest integer or one-hot-encode the categorical ones before evaluating the objective, depending on which one is needed. As demonstrated by Garrido-Merchán and Hernández-Lobato (2020), this approach causes the GP to ignore that an interval around an integer becomes known when one point is evaluated in the interval, as all values in that interval are rounded to the same integer value. This can lead to unnecessary evaluations and thus slow down the iteration. In the worst case, the algorithm can even become stuck on one point. Therefore, the authors propose that the transformation (i.e., rounding and encoding) of the hyperparameters should be done inside the kernel function, so that the acquisition function gains accurate information about the posterior when evaluating a new point. We have applied this approach to the default Matérn kernel, keeping it otherwise unchanged.

As the range of the hyperparameters can be wide and the general location of the optimum can be uncertain, it can be useful to optimize some hyperparameters logarithmically. This is not supported by the BO package by default, but it is easy to implement by transforming, at the beginning of the

objective function, the hyperparameter x in question to 10^x , effectively optimizing the value of the exponent. This transformation was applied to many hyperparameters in almost all ML models, and is also indicated differently in Tables S2–S8.

In addition to BO, the optimization of the hyperparameters was also done using a randomized search (RS), which would be expected to perform worse, as long as the BO iteration proceeds properly. As there are multiple parameters to tune for the optimizer itself that can significantly affect its performance (Snoek et al., 2012), a suboptimal selection could potentially make the BO method inferior to a purely random procedure. In our application, where a large number of models are optimized, it would be highly impractical to inspect every model individually to make sure the BO iteration has succeeded, especially with the limited options for visualization available. Visualizing aspects of the optimization process can make it easier to verify that the parameter space has been thoroughly explored and that fitting the GP has been successful. Due to these limitations, both BO and RS were used. While the models and datasets of this study are relatively small and thus a complex method like BO may not lead to major computational gains, the same methodology could be applied in future studies with more computationally intensive problems. These could be, for example, downscaling of a longer or higher resolution time series, training a single model on data from a large number of sites, or using large deep learning models.

4.5.2 Optimization procedure

The selected optimization method, either BO or RS, was executed for 300 iterations. In the case of BO, the first 30 points were also sampled randomly to have sufficient data for the acquisition function to operate on. Another case, called “pure BO” in the results section, was run without sampling these initial points. As for the parameters of the optimizer itself, the ν of the Matérn kernel was set to 1.5 (making the samples from the GP once differentiable), while the α of the GP was set to 1×10^{-2} when the model had categorical hyperparameters, and left to the default 1×10^{-6} otherwise. Three options (1, 2.5 and 10) were tried for the κ parameter of the acquisition function to account for different needs for exploration and exploitation.

In addition to these five cases (RS, pure BO, and BO with three different values of κ), two more cases were formed by not optimizing the feature selection (FS) parameters as part of either RS or BO (with $\kappa = 2.5$). Hence, all 100 input variables were included in the models. These two cases are called “RS & no FS” and “No FS”, respectively. It should be noted that without feature selection, the only hyperparameter of the GP model is α , and therefore it would not make much sense to use BO to optimize it, as it is also based on fitting a GP. In this case, RS was used instead, meaning that “RS & no FS” and “No FS” refer to the same procedure when GP is concerned.

4.6 Evaluation of model performance

To assess model performance, we employed five complementary evaluation metrics. In addition to the commonly used mean absolute error (MAE), root mean squared error (RMSE), and Pearson's correlation coefficient (r), we also used the coefficient of determination (ρ^2) and the scaled MAE (sMAE), defined as:

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

$$\text{sMAE}(y, \hat{y}) = \frac{\text{MAE}}{\bar{y}} = \frac{1}{\bar{y}} \cdot \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (3)$$

Here, y denotes the vector of observed values, $\hat{y}$ the predicted values, and $\bar{y}$ their mean.

The coefficient of determination, ρ^2 , measures the proportion of variance in the observations explained by the model. While often referred to as R^2 , we use ρ^2 to emphasize that in non-linear models this value can be negative, unlike in unconstrained linear regression where R^2 is bounded between 0 and 1. A perfect match between predictions and observations yields $\rho^2 = 1$.

The scaled MAE (sMAE), applied e.g. in Mikkonen et al. (2020), normalizes MAE by the mean of the observed values. This allows for better comparability across datasets of differing magnitudes, such as the particle size distribution sub-ranges examined here. As the mean PNC is strictly positive, sMAE remains well-defined throughout.

RMSE and MAE are expressed in the same units as the target variable (1 cm^{-3} in this case), with RMSE penalizing large errors more heavily. Pearson's r , in turn, quantifies the association between predicted and observed values and is insensitive to the scale of errors, making it useful for assessing rank consistency.

For model selection during hyperparameter optimization, ρ^2 was used as the primary criterion of goodness-of-fit.

In addition to numerical evaluation, we applied a game-theoretical interpretation method, SHAP (SHapley Additive exPlanations), to assess the influence of individual input features on model predictions (Lundberg and Lee, 2017; Molnar, 2022). SHAP assigns a contribution value to each feature per prediction, indicating the direction and magnitude of its effect relative to the average prediction. Positive (negative) SHAP values imply that the feature increased (decreased) the model's output.

We used the `shap` Python package (v0.40.0) with the permutation explainer applied to the test data (Lundberg and Lee, 2017). The results were aggregated across all test days and visualized using beeswarm plots, which show both the distribution and strength of feature effects. In these plots, a consistent increase (or decrease) in SHAP values with rising feature values suggests a positive (or negative) association with the target variable.

The SHAP analysis helps identify which input variables the models depend on most and whether these dependencies align with known atmospheric processes. This interpretability is especially valuable when evaluating black-box models such as neural networks or ensemble methods.

5 Results and discussion

5.1 Comparison of ML models

In Fig. 3, we present a comparison of all seven downscaling methods across the eight datasets. The performance of the methods varied depending on the dataset: all methods were among the best in at least one of the datasets, but most of them also failed in some cases, yielding ρ^2 s close to, or even less than, zero. Only RF and GP showcased stable performance, as they never resulted in a ρ^2 less than 0.1, and were never among the worst performing methods. On average, XGBoost had the highest ρ^2 (0.263), followed by SVM (0.250). XGBoost was also the best method for four out of the eight datasets. It only failed in the nucleation dataset of Helsinki, where it had a lower ρ^2 than any other model (see also Table S4 for hyperparameters, some of which are atypical). However, this dataset turned out to be difficult for all methods, as none of them were able to reach a ρ^2 above 0.15. Generally, the differences between methods were smaller than the differences between datasets, and in many cases, multiple methods were nearly equal in performance. Only some datasets had one method that clearly outperformed the others; this was XGBoost in the nucleation and accumulation datasets of Melpitz and in the Aitken dataset of Helsinki, and GLM in the accumulation dataset of Leipzig. Additionally, XGBoost and SVM were the two best methods for all sub-ranges from Melpitz, indicating some commonality between these datasets.

Overall, other ML methods have a slight advantage over GLM, as its average ρ^2 is the lowest across datasets (0.176). There is, however, strong variance in its performance, as it is among the best methods in both Leipzig's accumulation dataset and Helsinki's nucleation dataset, but among the worst in the six remaining datasets. In three of the six datasets, it is strictly the weakest, and in the other three, only two methods (NN2 and SVM) perform slightly worse. Particularly, the previously mentioned RF and GP were never outperformed by GLM, except in Leipzig's accumulation dataset. Moreover, GLM results in negative ρ^2 (-0.126) in the Aitken dataset of Helsinki, a drastic difference to all other methods.

For RF and the two neural networks, a mean ρ^2 from 50 different initializations is shown in both the table and the graph in Fig. 3. The magnitude of the 2σ confidence intervals, given in parentheses, indicates that randomness had a relatively minor effect on the performance of these models, except for those models that performed poorly to begin

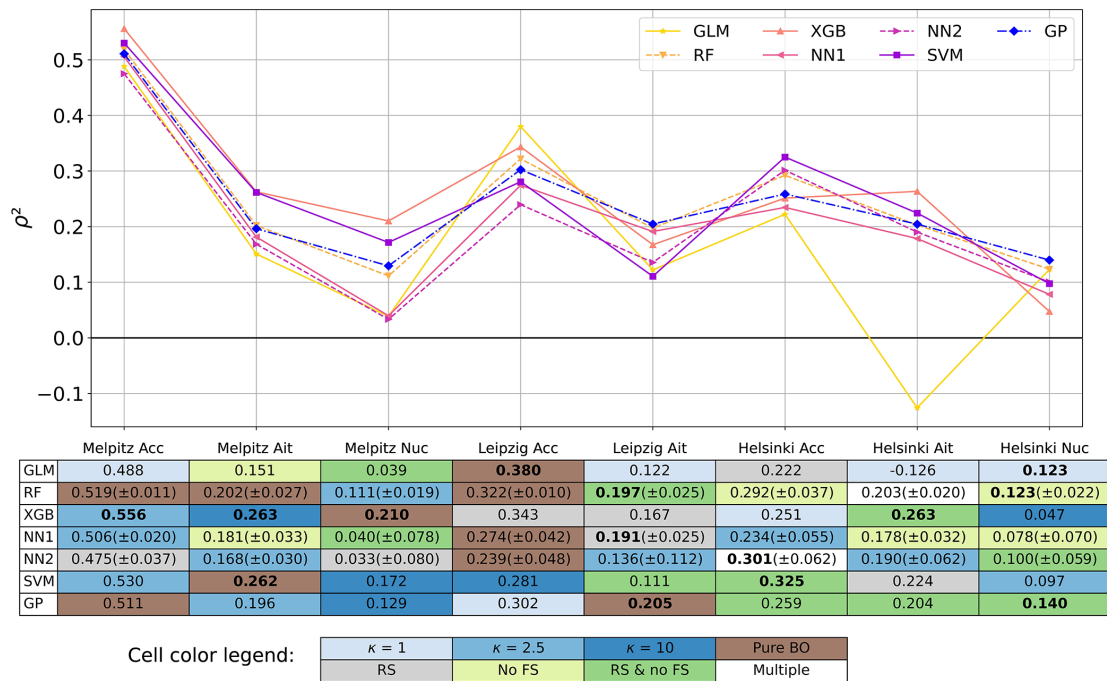


Figure 3. Test set performances (ρ^2) of the optimized models for all eight datasets. For the methods that are affected by randomness, the 2σ confidence intervals computed from 50 different initializations are also shown in the table. The background colors in the table represent the optimization method used. The methods and the abbreviations are explained in Sect. 4.5.1 and 4.5.2. In the cases where multiple optimization methods produced the exact same result, the background is left blank. These cases were Helsinki Ait (Pure BO, $\kappa = 1$, and $\kappa = 2.5$) and Helsinki Acc ($\kappa = 1$, $\kappa = 2.5$, $\kappa = 10$, and RS). The downscaling method(s) that achieved the highest ρ^2 for a given dataset are shown in bold (differences of less than 0.025 are disregarded).

with. It is also interesting to compare the two variations of the neural network. In all datasets from Helsinki, adding another layer to the neural network was beneficial. The simpler one-layer network yielded better results in all other datasets. This could be linked to a higher complexity in modeling the particle number size distribution in Helsinki compared to the other sites. This complexity may also be reflected in the qualities of the optimized models: the three best models developed for Helsinki's subranges utilize all 100 features, while at least some amount of feature selection was beneficial for all of the other datasets' best models (see Sect. 4.3 for a description of feature selection, and Tables S2–S8 for optimization results). Conversely, the accumulation subrange of Leipzig seems to have been a less complex target for downscaling, as the optimal number of features for it was lower than for other datasets, both when considering the best method (GLM, 19 features) and the average of all methods (28 features). In this case, interactions between features were not needed either, as GLM does not utilize those, unlike the other methods. It is of course possible that having access to more training data or an even wider range of input variables would reveal some interactions that were not found by our current procedure. In that sense, the simplicity of the best model might only indicate that something, like outliers in the

training data, confused the more complex methods while not affecting the linear model to the same extent.

5.2 Downscaling performance

Figure 4 shows the PNC results of the most successful downscaling methods for the test subset (2018) of each of the eight datasets. In all cases, a clear improvement is observed compared to the original subranges simulated by ECHAM-HAMMOZ, both visually and based on the five metrics shown in the figures. XGBoost achieved the highest ρ^2 for all subranges from Melpitz and the Aitken subrange from Helsinki. Gaussian process regression resulted in the best model for Leipzig's Aitken subrange and Helsinki's nucleation subrange, support vector machine for Helsinki's accumulation subrange, and the generalized linear model for Leipzig's accumulation subrange. Generally, the downscaling of the accumulation subrange was most successful, whereas the nucleation subrange seems to have been more difficult to downscale, resulting in relatively low ρ^2 s in both Helsinki and Melpitz. All three downscaled accumulation subranges have higher ρ^2 s, correlations, and sMAEs than any of the other datasets, even though many peaks and troughs are still estimated incorrectly. The downscaling model trained on the accumulation subrange of Melpitz per-

forms best out of the three, producing a ρ^2 of 0.56 and sMAE of 0.24.

The original ECHAM-simulated nucleation subranges differ significantly from the measured ones, likely contributing to the relatively poor performance of the downscaling models for that subrange. In Melpitz, the ρ^2 of the original is lowest among all subranges (-14.53), and the sMAE is highest (3.36). The strongly negative ρ^2 indicates that the large-scale approximation of the size distribution in ECHAM is a poor representation of the nucleation subrange at this site. In wintertime, the simulation represents the measurements reasonably well, but a strong overestimation is apparent from spring to autumn. Similarly in Helsinki, the correlation between the simulated and measured nucleation subranges is almost nonexistent (0.03) before downscaling. The simulation is unable to predict the high peaks in number concentration during spring and early summer, but instead predicts peaks for autumn, when the measured concentrations are relatively low. By downscaling, these differences can be greatly reduced: sMAE drops from 1.01 to 0.56 in Helsinki and from 3.36 to 0.74 in Melpitz, and the previously negligible correlation in Helsinki increases to 0.38. Thus, even though the performance metrics of the downscaled nucleation subrange are worse even when compared to the non-downscaled ECHAM-simulation of the accumulation subrange in Melpitz, the improvements are significant considering the starting point. Additionally, it should be noted that arithmetic means instead of medians were used in the daily averaging to preserve the highly variable nature of the data. Using medians would smooth the time series, which, while possibly improving downscaling results, would also depict the nucleation subrange less realistically.

The representation of new particle formation and nucleation-sized particles is, on many occasions, not sufficient in global climate models (Williamson et al., 2019). This can be due to, for instance, errors in estimating nucleation rates. As a study by Laakso et al. (2022) shows, ECHAM-SALSA tends to favor partitioning of sulfuric acid to the particle phase due to nucleation over condensation, which may lead to overestimation of nucleation subrange particles. Kokkola et al. (2018) compared ECHAM-SALSA number size distributions to observation data, and their results revealed that at some measurement stations, ECHAM-SALSA overestimates the nucleation mode number concentrations. Furthermore, ECHAM-SALSA does not model new particle formation due to nitrates, which may cause differences between modelled and measured nucleation subrange number concentrations. The representation of nucleation-sized aerosols could be enhanced by including a volatility basis set (VBS) scheme (Donahue et al., 2011), which can improve the representation of secondary organic aerosols. In addition to limitations in representing the nucleation mode, other input variables can also contribute to challenges in downscaling. The coarse spatial resolution of global-scale models naturally

limits their ability to accurately capture processes other than just new particle formation.

A strong variability can be seen in both the simulated and measured Aitken subranges at the German sites (Melpitz and Leipzig), although the peaks and troughs match poorly. In winter, the simulated concentrations decrease more than they should. The downscaling methods are generally able to bring the concentrations to a more realistic level, but they fail to capture the true variability in the data. Compared to the subranges simulated by ECHAM-HAMMOZ, the downscaled concentrations no longer fluctuate as rapidly, but instead seem to more carefully follow an average level between the peaks and troughs of the measured time series. This can be acceptable when the main focus of a study is on long-term averages, and not on e.g. maximum daily exposures. Regardless, an improvement compared to the original is seen in all reported metrics. For the Aitken subrange from Helsinki, the original simulated concentration is mostly too low, and has a less drastic summertime variability compared to the German sites. The downscaling by XGBoost fixes the underestimation and brings the variability closer to that of the measurements. Based on the metrics, the results are quite similar to the other Aitken subrange datasets. The increase in the correlation coefficient from 0.12 to 0.53 is largest out of all datasets, and the improvements in sMAE and ρ^2 are also among the largest.

To summarize, downscaling was generally more effective for larger particle sizes than for smaller ones, and for the rural Melpitz site compared to the urban sites. The eight datasets were further examined through statistical tests comparing the means of the training, validation and testing subsets of the measured PNC (not shown). These tests found significant differences between the years for most subranges and sites, amounting to five out of eight cases in total. If the subsets differ substantially, ML models may struggle to generalize from one dataset to another. To potentially reduce the variation between subsets, the temporal dimension of the data could be expanded beyond three years, thereby enlarging each subset. Training the models with more than one year of data, in particular, could enhance generalization performance. Therefore, we recommend collecting more data for future studies, if possible.

We can compare our results to previous studies to place them in a broader context, although no directly comparable studies exist. For example, Ivatt and Evans (2020) trained an XGBoost model to improve the ozone predictions of a chemistry transport model, and achieved an improvement in Pearson's r of 0.36 (from 0.48 to 0.84). This is similar, though in most cases slightly higher, to the improvement achieved by our models. In addition to a different target variable, their higher time resolution and lower spatial resolution (mean of multiple sites) complicate the comparison. XGBoost was also the most successful model in the study by Venuta et al. (2024), which produced spatiotemporal UFP predictions (logarithm of PNC) with a daily time

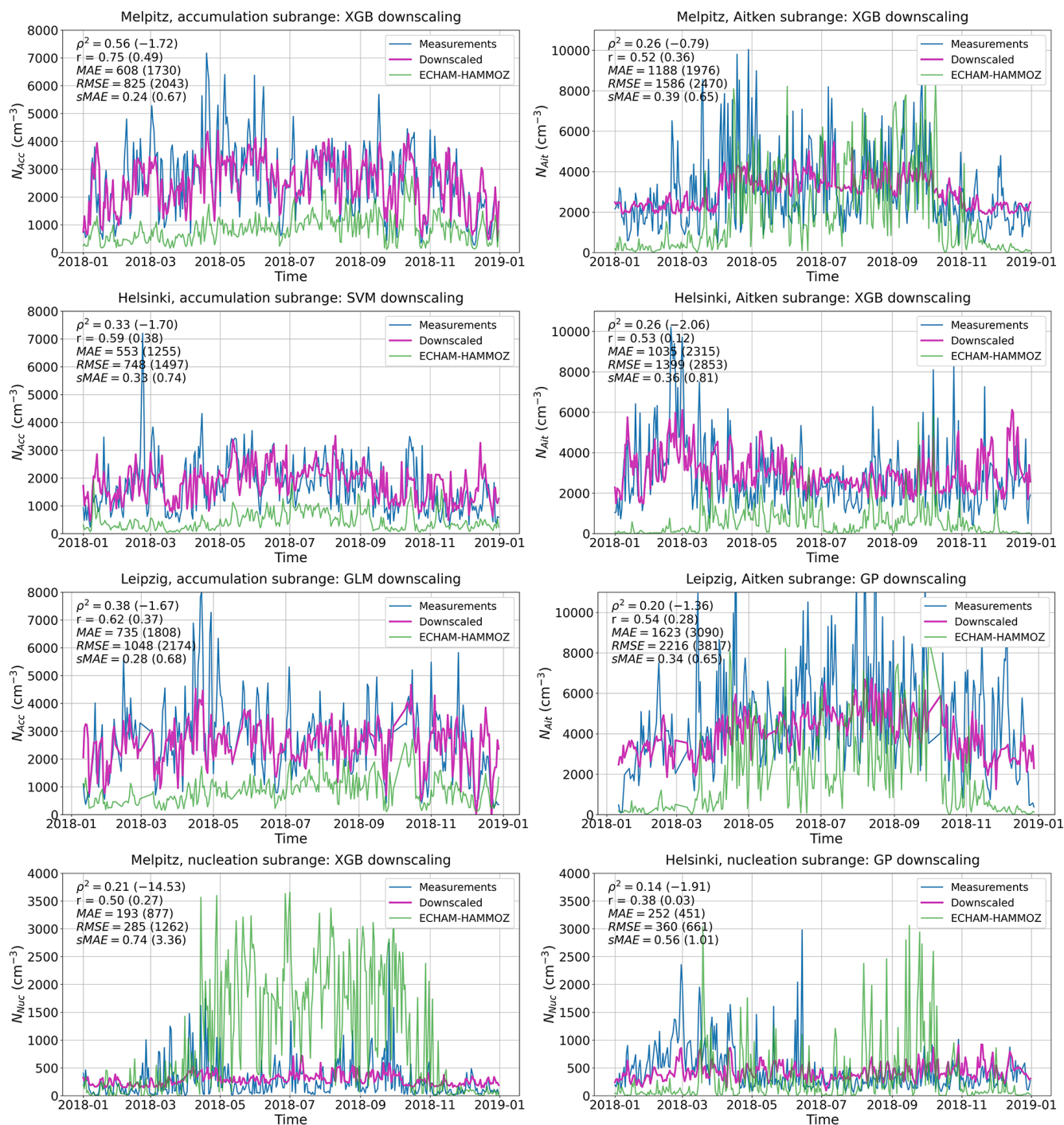


Figure 4. Daily average particle number concentrations per subrange in 2018 (test set), for all three sites. Measurements are shown in blue, ECHAM-HAMMOZ outputs in green, and the results of downscaling by the best model for each dataset (i.e., the bolded cells in Fig. 3) in purple. Goodness-of-fit metrics are reported in the top left corners of each figure, first for the downscaling and then, in parentheses, for ECHAM-HAMMOZ.

resolution. Their ρ^2 of about 0.72 was significantly higher than ours, though a direct comparison is again difficult due to the smoothing effect of the log-transform combined with data trimming they performed. Additionally, they used observational weather data instead of climate simulations to train the models. As our models can in theory be used to predict far into the future and produce non-transformed estimates of UFP concentrations, the seemingly lower metrics are still competitive, especially considering the substantial improvement over the ECHAM predictions.

5.3 Notes on optimization

In the table of Fig. 3, the method of hyperparameter optimization that resulted in the best model is represented by the color of the cell's background. In the cases where the background of the table is white, multiple optimization methods yielded the exact same hyperparameter values and hence also ρ^2 . This means that the hyperparameters resulting in the highest ρ^2 were discovered either during the initial random steps of the iterations (which were now deterministic due to the fixed seed number), or by the convergence of the BO algorithm to the same hyperparameter values during the non-random steps. The latter was the case for the RF model trained on the Aitken data from Helsinki. This makes sense given that one of the equally performing optimization methods was "Pure BO", which didn't utilize random iterations. For the NN2 model trained on Helsinki's accumulation subrange, on the other hand, the optimum was found from among the initially sampled points in all four identically performing cases.

It can be seen that for most models, the BO methods were superior to RS. However, it is surprising that in some cases randomized search (RS) led to higher ρ^2 . The number of iterations for both approaches was the same, and BO searches the parameter space more methodically, so it should have been able to find a better combination of hyperparameter values. In these situations, it is possible that the hyperparameters don't have a clear optimum, and therefore a reasonably good combination can be found randomly. Then, RS could work slightly better than BO purely by chance. Another possibility is that the few alternatives which were tried for the parameters of the optimizer itself (e.g. *kappa*, *alpha*, and *nu*) were suboptimal for that specific model and dataset. Selecting the parameters correctly can be challenging when the number of different models and datasets is large, and when the options for visualization are limited, such as in high-dimensional spaces. Finally, it is possible that the optimization algorithm itself didn't fully work as intended in these cases, or even got stuck without actually converging on a solution, possibly due to the additional complexity in the acquisition function caused by the treatment of discrete-valued and categorical hyperparameters, as mentioned in Nguyen et al. (2020). This problem could be difficult to diagnose in a comprehensive

model comparison study, when every result cannot feasibly be individually inspected.

Simultaneously with the model hyperparameters, feature selection was also optimized through two threshold parameters for redundancy and relevance (see Sect. 4.3 and 4.5.2). For the GP models trained on the datasets of Helsinki, both "No FS" and "RS & no FS" involved using RS and were thus identical, for reasons discussed in Sect. 4.5.2. It is also interesting to note that if optimized correctly, our feature selection method could have resulted in practically no selection (i.e., full set of features) by setting the thresholds for redundancy and relevance to 0.99 and 0, respectively. Therefore, it should theoretically always be equal or superior to the "No FS" cases where all 100 features were used without any selection procedure. However, this might be further complicated by the effect of an increased number of hyperparameters on the capability of the optimization algorithm to find the optimum.

In conclusion, BO can improve the results of hyperparameter tuning relative to a randomized search, but can be significantly affected by the selection of the BO parameters and therefore requires careful analysis of the optimization process. Due to this tradeoff between the simplicity of RS and the (generally) improved optimization performance of BO, BO may be preferable when developing one computationally expensive ML model. However, when the number of models under optimization is large, the interpretability and ease of implementation of RS can make it a more practical choice. Other Python packages that implement similar optimization methods, though not only Bayesian, also exist and could alternatively be utilized. Some examples are Hyperopt (Bergstra et al., 2013), Optuna (Akiba et al., 2019), and SMAC3 (Lindauer et al., 2022).

5.4 Interpreting the models

The SHAP method, described in Sect. 4.6, was used to analyze the features in the ML models. Figure 5 shows a summary of the most important features across all models. The height of the bars relates to how many models were strongly influenced by the corresponding feature, defined by the feature being among the ten highest when ranked by mean absolute SHAP value. For example, the north-south directional component of wind was among the ten most important in 46 models out of the total 56. In general, the wind-related features are seen to be important for the prediction of all subranges of the size distribution, though less so for the nucleation subrange. Solar radiation is also one of the most important variables. ML models for the smallest two size ranges seem to utilize emissions of organic carbon, whereas accumulation subrange is connected to sulfur dioxide (SO₂) and sulfate (SO₄), according to the SHAP values. Interestingly, the feature for geopotential height is mainly used by the Helsinki models.

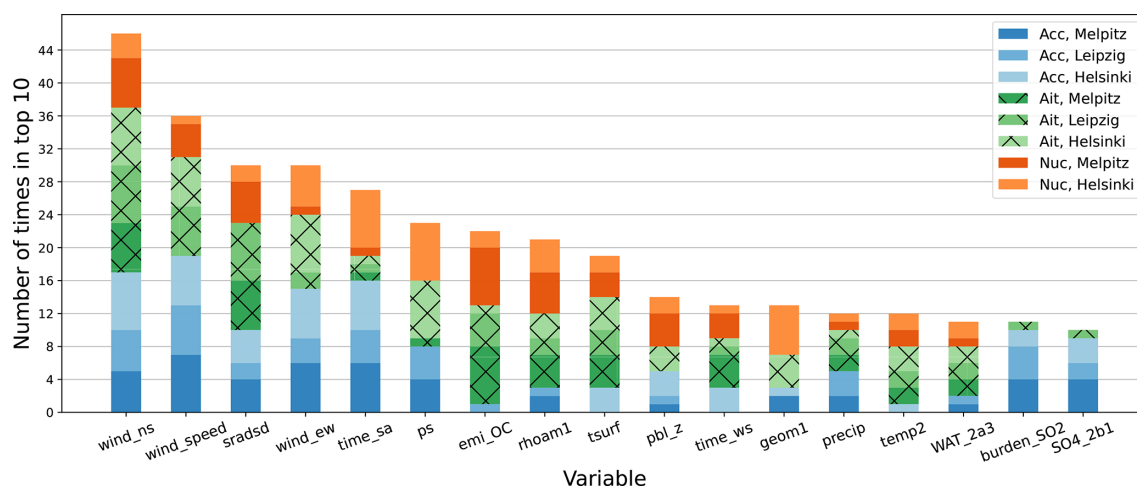


Figure 5. Most important input variables across all eight datasets, measured by mean absolute SHAP values. All seven ML models were analyzed for each dataset. Hence, the upper limit for the height of the bars is 56. Bars with height less than ten are not shown.

The modewise summaries of the SHAP explanations (given in Figs. S1–S3) can be examined for additional insights. Figure S1 shows that also variables related to dust and black carbon, which are not present in the summary figure (Fig. 5), are contributing to many of the ML models for accumulation subrange. Variables used for downscaling Aitken subrange (Fig. S2) do not substantially differ from the ones shown in the summary of Fig. 5. In general, we recommend refraining from using SHAP to interpret weakly performing models, such as most of the ones for nucleation subrange (Fig. S3), as any conclusions made are likely to be misleading.

The best-performing ML models were studied in detail using SHAP (Figs. S4–S15). These model explanations can be compared to experimental studies from the sites to see how well the statistical relationships found by the models correspond to the physical characteristics of the locations. The Helsinki station can be taken as an example. In previous research (Järvi et al., 2009), the surrounding area has been subdivided into three distinct land use sectors, of which the road sector to the southeast has been found to be the largest contributor to the accumulation mode (100–1000 nm), especially during springtime. In addition to the road itself, long-range transport from the east is hypothesized to contribute to this sector's accumulation mode. On the other hand, an increased concentration of ultrafine particles (3–100 nm) has been associated roughly equally with the road sector and the urban sector to the north. The vegetation sector to the west remains a direction of slightly less polluted air throughout the year. These findings are in line with the effects of wind direction in our models: Figures S4–S8 for Helsinki show that the east-west wind component is important in all models, and that its effect is positive (i.e., easterly wind is connected to increased pollutant concentrations). The two best models for Helsinki's accumulation subrange (Figs. S4–S5)

both also include the north-south component, which has a negative effect on PNC. This means that the models predict higher concentrations when wind is blowing from the south. Järvi et al. (2009) point out that ship emissions from the harbor, located approximately in this direction, can affect accumulation mode PNC. Moreover, the springtime increase is also captured by these two models. It is interesting to note that NN2 has almost the same ρ^2 as SVM, despite using far fewer features (18 and 100, respectively; see Table S9).

As the models for Helsinki's nucleation subrange (Figs. S7–S8) are quite weak, and therefore unlikely to capture the relevant effects, we compare the UFP of Järvi et al. (2009) only to our Aitken subrange downscaling model (Fig. S6). The positive effect of northerly wind and the negative effect of temperature on the Aitken subrange PNC seem realistic, as wood combustion in the urban sector is a significant source of pollutants in the area. Importance of the variables boundary layer height and atmospheric pressure might also be related to the same phenomenon.

Likely, some (or even most) of the features shown in the SHAP plots are only deemed important in terms of their contribution to the downscaling because they correlate with some physically relevant quantity, and not because they themselves cause changes in PNC. For example, this is probably the case with the sea salt variable in Figs. S13 and S15, as Melpitz is located nowhere near marine environments. In the ECHAM-HAMMOZ data, PM25_SS is highly correlated with certain variables (num_2a6, num_2a7, WAT_2a6, and WAT_2a7) that might more realistically be connected to PNC in Melpitz, however. Hence, when employing SHAP values to assess feature importance, it is important to note that SHAP explains how specific models operate and is not to be interpreted as a tool for causal inference of physical systems.

In this analysis, it should naturally be recognized that all features originate from a simulation of large-scale climate, and therefore do not necessarily represent the immediate surroundings of the measurement sites. Additionally, SHAP is known to be sensitive to correlated features (Aas et al., 2021), which most of our models include; if accurate explanations of the models are crucial, care should be taken to remove all (even somewhat) correlating features before training or use more robust explanation methods.

5.5 Performance considerations

Although our results indicate that some of the ML methods may on average result in higher goodness-of-fit metrics, there are other aspects that might affect the choice of downscaling method. For example, the training and inference durations for different ML methods can vary differently as a function of the dimensionality of input data. In our study, the computational performance of the methods was not considered important, as the downscaling was done as a post-processing step. However, if the downscaling was included as an online correction in the climate model itself, speed of the method would be critical. We have compared the computational performance of the seven model architectures separately for the training, optimization, and inference steps (Fig. S16). There is significant variation in performance: in terms of training, GLM and SVM are by far the fastest, while the NN architectures take longest to train on average. When applying the models for inference, however, NNs are among the fastest, along with GLM. This might make them preferable in applications where computing time is costly. In the optimization phase, the BO algorithm itself takes relatively long to iterate through, reducing the difference in total duration between most methods; still, NNs are the slowest, though there was large variation depending on the number of NN nodes. Adding a layer to the NNs slowed their training substantially. These findings are naturally only indicative of how the methods perform computationally, and may not apply to datasets of different size. Moreover, parallel computation of the training or inference algorithms can yield additional speedups, which could be another advantage of the methods capable of being parallelized. Of our six methods, RF and NN training can be run in parallel, as the trees in RF can be trained independently, and the NN training can be split into independent batches. Parts of the XGBoost algorithm can also be parallelized, though the trees of the ensemble must still be trained successively (Chen and Guestrin, 2016). Another advantage of NNs is that their structure is ideal for multi-target regression, i.e., the number of target variables can be freely chosen. This way, all three subranges of the size distribution could be downscaled with a single NN. Using the other ML methods, a separate model needs to be trained for each individual output variable.

6 Conclusions

This study provides a proof of concept for using ML methods to improve the site-specific accuracy of aerosol particle number size distributions derived from global-scale climate models. By employing six ML methods, optimized through feature selection and hyperparameter tuning, significant improvements were observed in the simulated particle concentrations, especially for accumulation and Aitken subranges. Among the methods, XGBoost demonstrated, on average, superior performance across various datasets. Despite these advances, the nucleation subrange proved more challenging to downscale, likely due to high spatial variability and limitations in the underlying large-scale climate model outputs, particularly in the processes contributing to new particle formation.

The findings underscore the potential of ML-enhanced downscaling as a computationally efficient alternative to traditional methods, offering robust applications in air quality and epidemiological studies. It was observed that downscaling methods can significantly enhance model accuracy at individual measurement sites. However, the selection of a suitable downscaling method requires precision and depends on the target variable's characteristics, as well as spatial and, assumably, temporal dimensions. For example, while particle size ranges were the focus here, the same methods could be applied to other variables as well. Future research should focus on expanding the geographical scope of measurement data, integrating additional features to capture local-scale variations, and exploring dynamic downscaling during climate simulations. Additionally, deep learning methods specialized for time series regression, such as Long Short-Term Memory NNs (Hochreiter and Schmidhuber, 1997), might further improve the downscaling performance by accounting for temporal dependencies in the observed and simulated data (if training data with no missing values were available). These advancements could enhance the predictive accuracy of particle size distributions in coarse-scale climate models, contributing to better assessments of climate change impacts and health outcomes.

Code and data availability. The ECHAM6-HAMMOZ model is made available to the scientific community under the HAMMOZ Software License Agreement, which defines the conditions under which the model can be used. The license can be retrieved from https://redmine.hammoz.ethz.ch/attachments/291/License_ECHAM-HAMMOZ_June2012.pdf (last access: 20 October 2025). The model data can be reproduced using ECHAM-HAMMOZ model revision 6588 from the repository <https://redmine.hammoz.ethz.ch/projects/hammoz/> (HAMMOZ consortium, 2025, login required).

The Python code files for data processing, machine learning, and plotting are provided in the Zenodo archive at <https://doi.org/10.5281/zenodo.17079843> (Vartiainen, 2025). The same archive also contains the ECHAM-HAMMOZ simulation set-

tings, the trained ML models, as well as the processed simulation and measurement datasets.

The particle number size distribution measurement data are openly available in the EBAS (<https://ebas.nilu.no/>, last access: 17 September 2025) and SmartSMEAR (<https://smear.avaa.csc.fi/>, last access: 17 September 2025) databases.

Supplement. The supplement related to this article is available online at <https://doi.org/10.5194/amt-18-5763-2025-supplement>.

Author contributions. AV: conceptualization; methodology; investigation; validation; formal analysis; writing – original draft; visualization. SM: conceptualization; methodology; supervision; writing – review and editing. VL: methodology; investigation; writing – review and editing. TP: investigation; writing – review and editing; data curation. AW: investigation; writing – review and editing; data curation. TK: conceptualization; writing – review and editing. TM: conceptualization; methodology; supervision; writing – review and editing.

Competing interests. At least one of the (co-)authors is a member of the editorial board of *Atmospheric Measurement Techniques*. The peer-review process was guided by an independent editor, and the authors also have no other competing interests to declare.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements. We are grateful to Dr Leena Järvi from University of Helsinki (INAR/Physics & HELSUS) for administration of the SMEAR III station in Helsinki.

Financial support. This research has been funded by the University of Eastern Finland doctoral program (UEF Doctoral school, 2023) and supported by the following Research Council of Finland (RCoF) grants: Competitive funding to strengthen university research profiles (PROFI) for the University of Eastern Finland (grant nos. 325022 and 352968), The Atmosphere and Climate Competence Center (ACCC) Flagship (grant nos. 337549, 357902, 359340, 337550, 357904, 359341, 359342, and 359343), Flagship of Advanced Mathematics for Sensing Imaging and Modelling (grant no. 359196), RESEMON project (grant nos. 330165 and 337552), and ClimAirPathways (grant no. 355531). Additionally, financial support from University of Helsinki (ACTRIS-HY) and the RI-URBANS project (EU Horizon 2020 grant no. 101036245) is gratefully acknowledged.

Review statement. This paper was edited by Daniel Perez-Ramirez and reviewed by two anonymous referees.

References

- Aas, K., Jullum, M., and Løland, A.: Explaining individual predictions when features are dependent: More accurate approximations to Shapley values, *Artificial Intelligence*, 298, 103502, <https://doi.org/10.1016/j.artint.2021.103502>, 2021.
- Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-Generation Hyperparameter Optimization Framework, in: *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631, <https://doi.org/10.1145/3292500.3330701>, 2019.
- Alpaydin, E.: *Introduction to Machine Learning*, MIT Press, Cambridge, MA, 3rd edn., ISBN 978-0-262-02818-9, 2014.
- Amini, H., Bergmann, M. L., Taghavi Shahri, S. M., Tayebi, S., Cole-Hunter, T., Kerckhoffs, J., Khan, J., Meliefste, K., Lim, Y.-H., Mortensen, L. H., Hertel, O., Reeh, R., Gaarde Nielsen, C., Loft, S., Vermeulen, R., Andersen, Z. J., and Schwartz, J.: Harnessing AI to unmask Copenhagen's invisible air pollutants: A study on three ultrafine particle metrics, *Environmental Pollution*, 346, 123664, <https://doi.org/10.1016/j.envpol.2024.123664>, 2024.
- Bai, L., Chen, H., Hatzopoulou, M., Jerrett, M., Kwong, J., Burnett, R., Donkelaar, A., Copes, R., Martin, R., Van Ryswyk, K., Lu, H., Kopp, A., and Weichenthal, S.: Exposure to Ambient Ultrafine Particles and Nitrogen Dioxide and Incident Hypertension and Diabetes, *Epidemiology*, 29, 1, <https://doi.org/10.1097/EDE.0000000000000798>, 2018.
- Bergstra, J., Yamins, D., and Cox, D.: Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures, in: *Proceedings of the 30th International Conference on Machine Learning*, edited by: Dasgupta, S. and McAllester, D., vol. 28 of *Proceedings of Machine Learning Research*, PMLR, Atlanta, Georgia, USA, 115–123, <https://proceedings.mlr.press/v28/bergstra13.html> (last access: 20 October 2025), 2013.
- 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., Gerwig, H., 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 Syst. Sci. Data*, 8, 355–382, <https://doi.org/10.5194/essd-8-355-2016>, 2016.
- Breiman, L.: Random Forests, *Machine Learning*, 45, 5–32, <https://doi.org/10.1023/A:1010950718922>, 2001.
- Brochu, E., Cora, V. M., and de Freitas, N.: A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning, *arXiv [preprint]*, <https://doi.org/10.48550/arXiv.1012.2599>, 12 December 2010.
- Chen, T. and Guestrin, C.: XGBoost: A Scalable Tree Boosting System, in: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Associa-

- tion for Computing Machinery, 785–794, ISBN 9781450342322, <https://doi.org/10.1145/2939672.2939785>, 2016.
- Cortes, C. and Vapnik, V.: Support-vector networks, *Machine Learning*, 20, 273–297, <https://doi.org/10.1007/BF00994018>, 1995.
- Dada, L., Okuljar, M., Shen, J., Olin, M., Wu, Y., Heimsch, L., Herlin, I., Kankaanrinta, S., Lampimäki, M., Kalliokoski, J., Baalbaki, R., Lohila, A., Petäjä, T., Maso, M. D., Duplissy, J., Kerminen, V.-M., and Kulmala, M.: The synergistic role of sulfuric acid, ammonia and organics in particle formation over an agricultural land, *Environ. Sci.: Atmos.*, 3, 1195–1211, <https://doi.org/10.1039/D3EA00065F>, 2023.
- Ding, C. and Peng, H.: Minimum redundancy feature selection from microarray gene expression data, *Journal of Bioinformatics and Computational Biology*, 3, 185–205, <https://doi.org/10.1142/s0219720005001004>, 2005.
- Donahue, N. M., Epstein, S. A., Pandis, S. N., and Robinson, A. L.: A two-dimensional volatility basis set: 1. organic-aerosol mixing thermodynamics, *Atmos. Chem. Phys.*, 11, 3303–3318, <https://doi.org/10.5194/acp-11-3303-2011>, 2011.
- Downward, G., Nunen, E., Kerckhoffs, J., Vineis, P., Brunekreef, B., Boer, J., Messier, K., Roy, A., Verschuren, W., Schouw, Y., Sluijs, I., Gulliver, J., Hoek, G., and Vermeulen, R.: Long-Term Exposure to Ultrafine Particles and Incidence of Cardiovascular and Cerebrovascular Disease in a Prospective Study of a Dutch Cohort, *Environmental Health Perspectives*, 126, 127007, <https://doi.org/10.1289/EHP3047>, 2018.
- EBAS: EBAS database, <https://ebas.nilu.no/> (last access: 17 September 2025), 2025.
- European Council: On Ambient Air Quality and Cleaner Air for Europe 2008/50/EC, Off. J. Eur. Union, 1, 1–44, <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008L0050&from=en> (last access: 17 September 2025), 2008.
- European Council: On Ambient Air Quality and Cleaner Air for Europe 2024/2881, Off. J. Eur. Union, 1, 1–70, <https://eur-lex.europa.eu/eli/dir/2024/2881/oj/eng> (last access: 17 September 2025), 2024.
- Feichter, J., Kjellström, E., Rodhe, H., Dentener, F., Lelieveld, J., and Roelofs, G.-J.: Simulation of the tropospheric sulfur cycle in a global climate model, *Atmospheric Environment*, 30, 1693–1707, [https://doi.org/10.1016/1352-2310\(95\)00394-0](https://doi.org/10.1016/1352-2310(95)00394-0), 1996.
- Fraga, R. P., Kang, Z., and Axthelm, C. M.: Effect of Machine Learning Cross-validation Algorithms Considering Human Participants and Time-series: Application on Biometric Data Obtained from a Virtual Reality Experiment, *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 67, 2162–2167, <https://doi.org/10.1177/21695067231192258>, 2023.
- Friberg, M. D., Kahn, R. A., Holmes, H. A., Chang, H. H., Sarnat, S. E., Tolbert, P. E., Russell, A. G., and Mulholland, J. A.: Daily ambient air pollution metrics for five cities: Evaluation of data-fusion-based estimates and uncertainties, *Atmospheric Environment*, 158, 36–50, <https://doi.org/10.1016/j.atmosenv.2017.03.022>, 2017.
- Friedman, J. H.: Greedy function approximation: A gradient boosting machine., *The Annals of Statistics*, 29, 1189–1232, <https://doi.org/10.1214/aos/1013203451>, 2001.
- Fuzzi, S., Baltensperger, U., Carslaw, K., Decesari, S., Denier van der Gon, H., Facchini, M. C., Fowler, D., Koren, I., Langford, B., Lohmann, U., Nemitz, E., Pandis, S., Riipinen, I., Rudich, Y., Schaap, M., Slowik, J. G., Spracklen, D. V., Vignati, E., Wild, M., Williams, M., and Gilardoni, S.: Particulate matter, air quality and climate: lessons learned and future needs, *Atmos. Chem. Phys.*, 15, 8217–8299, <https://doi.org/10.5194/acp-15-8217-2015>, 2015.
- Garrido-Merchán, E. C. and Hernández-Lobato, D.: Dealing with categorical and integer-valued variables in Bayesian Optimization with Gaussian processes, *Neurocomputing*, 380, 20–35, <https://doi.org/10.1016/j.neucom.2019.11.004>, 2020.
- GBD 2019 Risk Factors Collaborators: Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019, *The Lancet*, 396, 1223–1249, [https://doi.org/10.1016/S0140-6736\(20\)30752-2](https://doi.org/10.1016/S0140-6736(20)30752-2), 2020.
- Gouldsbrough, L., Hossaini, R., Eastoe, E., Young, P. J., and Vieno, M.: A machine learning approach to downscale EMEP4UK: analysis of UK ozone variability and trends, *Atmos. Chem. Phys.*, 24, 3163–3196, <https://doi.org/10.5194/acp-24-3163-2024>, 2024.
- Goyal, M., Burn, D., and Ojha, C.: Evaluation of machine learning tools as a statistical downscaling tool: Temperatures projections for multi-stations for Thames River Basin, Canada, *Theoretical and Applied Climatology*, 108, 519–534, <https://doi.org/10.1007/s00704-011-0546-1>, 2011.
- Hamed, A., Birmili, W., Joutsensaari, J., Mikkonen, S., Asmi, A., Wehner, B., Spindler, G., Jaatinen, A., Wiedensohler, A., Korhonen, H., Lehtinen, K. E. J., and Laaksonen, A.: Changes in the production rate of secondary aerosol particles in Central Europe in view of decreasing SO₂ emissions between 1996 and 2006, *Atmos. Chem. Phys.*, 10, 1071–1091, <https://doi.org/10.5194/acp-10-1071-2010>, 2010.
- HAMMOZ consortium: ECHAM-HAMMOZ model data and input files, <https://redmine.hammoz.ethz.ch/projects/hammoz/>, last access: 20 October 2025.
- Hastie, T., Tibshirani, R., and Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer New York, ISBN 9780387848587, 2009.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: Complete ERA5 from 1940: Fifth generation of ECMWF atmospheric reanalyses of the global climate: Copernicus Climate Change Service (C3S) Data Store (CDS), <https://doi.org/10.24381/cds.143582cf>, 2017.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vam-

- borg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049, <https://doi.org/10.1002/qj.3803>, 2020.
- Hinestroza-Ramirez, J. E., Lopez-Restrepo, S., Yarce Botero, A., Segers, A., Rendon-Perez, A. M., Isaza-Cadavid, S., Heemink, A., and Quintero, O. L.: Improving Air Pollution Modelling in Complex Terrain with a Coupled WRF–LOTOS–EUROS Approach: A Case Study in Aburrá Valley, Colombia, *Atmosphere*, 14, <https://doi.org/10.3390/atmos14040738>, 2023.
- Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Computation*, 9, 1735–1780, <https://doi.org/10.1162/neco.1997.9.8.1735>, 1997.
- Hussein, T., Martikainen, J., Junninen, H., Sogacheva, L., Wagner, R., Maso, M. D., Riipinen, I., Aalto, P. P., and Kulmala, M.: Observation of regional new particle formation in the urban atmosphere, *Tellus B: Chemical and Physical Meteorology*, 60, 509–521, 2008.
- IIASA: ECLIPSE V6b, <https://iiasa.ac.at/models-tools-data/global-emission-fields-of-air-pollutants-and-ghgs> (last access: 3 July 2024), 2024.
- Ivatt, P. D. and Evans, M. J.: Improving the prediction of an atmospheric chemistry transport model using gradient-boosted regression trees, *Atmos. Chem. Phys.*, 20, 8063–8082, <https://doi.org/10.5194/acp-20-8063-2020>, 2020.
- Järvi, L., Hannuniemi, H., Hussein, T., Junninen, H., Aalto, P., Keronen, P., Kulmala, M., Keronen, P., Hillamo, R., Mäkelä, T., Siivola, E., and Vesala, T.: The urban measurement station SMEAR III: Continuous monitoring of air pollution and surface-atmosphere interactions in Helsinki, Finland, *Boreal Environment Research*, 14, 1797–2469, 2009.
- Johansson, L., Karppinen, A., Kurppa, M., Kousa, A., Niemi, J. V., and Kukkonen, J.: An operational urban air quality model ENFUSER, based on dispersion modelling and data assimilation, *Environmental Modelling & Software*, 156, 105460, <https://doi.org/10.1016/j.envsoft.2022.105460>, 2022.
- Jung, C.-R., Chen, W.-T., Young, L.-H., and Hsiao, T.-C.: A hybrid model for estimating the number concentration of ultrafine particles based on machine learning algorithms in central Taiwan, *Environment International*, 175, 107937, <https://doi.org/10.1016/j.envint.2023.107937>, 2023.
- Junninen, H., Lauri, A., Keronen, P., Aalto, P., Hiltunen, V., Hari, P., and Kulmala, M.: Smart-SMEAR: on-line data exploration and visualization tool for SMEAR stations, *Boreal Environment Research*, 14, 447–457, <https://smear.avaa.csc.fi/> (last access: 17 September 2025), 2009.
- Kim, T. and Villarini, G.: Projected changes in daily precipitation, temperature and wet-bulb temperature across Arizona using statistically downscaled CMIP6 climate models, *International Journal of Climatology*, 44, 1994–2010, <https://doi.org/10.1002/joc.8436>, 2024.
- Kohl, M., Lelieveld, J., Chowdhury, S., Ehrhart, S., Sharma, D., Cheng, Y., Tripathi, S. N., Sebastian, M., Pandithurai, G., Wang, H., and Pozzer, A.: Numerical simulation and evaluation of global ultrafine particle concentrations at the Earth's surface, *Atmos. Chem. Phys.*, 23, 13191–13215, <https://doi.org/10.5194/acp-23-13191-2023>, 2023.
- Kokkola, H., Kühn, T., Laakso, A., Bergman, T., Lehtinen, K. E. J., Mielonen, T., Arola, A., Stadler, S., Korhonen, H., Ferrachat, S., Lohmann, U., Neubauer, D., Tegen, I., Siegenthaler, Le Drian, C., Schultz, M. G., Bey, I., Stier, P., Daskalakis, N., Heald, C. L., and Romakkaniemi, S.: SALSA2.0: The sectional aerosol module of the aerosol–chemistry–climate model ECHAM6.3.0-HAM2.3-MOZ1.0, *Geosci. Model Dev.*, 11, 3833–3863, <https://doi.org/10.5194/gmd-11-3833-2018>, 2018.
- Kreyling, W., Semmler, M., and Möller, W.: Dosimetry and Toxicology of Ultrafine Particles, *Journal of aerosol medicine: the official journal of the International Society for Aerosols in Medicine*, 17, 140–52, <https://doi.org/10.1089/0894268041457147>, 2004.
- Laakso, A., Niemeier, U., Visioni, D., Tilmes, S., and Kokkola, H.: Dependency of the impacts of geoengineering on the stratospheric sulfur injection strategy – Part 1: Intercomparison of modal and sectional aerosol modules, *Atmos. Chem. Phys.*, 22, 93–118, <https://doi.org/10.5194/acp-22-93-2022>, 2022.
- 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., van Noije, T., O'Connor, F. M., O'Dowd, C., Olivie, D., Pernov, J. B., Petäjä, T., Seland, Ø., 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, *Atmos. Chem. Phys.*, 22, 12873–12905, <https://doi.org/10.5194/acp-22-12873-2022>, 2022.
- Li, X., Li, Z., Huang, W., and Zhou, P.: Performance of statistical and machine learning ensembles for daily temperature downscaling, *Theoretical and Applied Climatology*, 140, 571–588, <https://doi.org/10.1007/s00704-020-03098-3>, 2020.
- Li, Y., Lane, K. J., Corlin, L., Patton, A. P., Durant, J. L., Thanikachalam, M., Woodin, M., Wang, M., and Brugge, D.: Association of Long-Term Near-Highway Exposure to Ultrafine Particles with Cardiovascular Diseases, Diabetes and Hypertension, *International Journal of Environmental Research and Public Health*, 14, <https://doi.org/10.3390/ijerph14050461>, 2017.
- Lin, S., Ryan, I., Paul, S., Deng, X., Zhang, W., Luo, G., Dong, G.-H., Nair, A., and Yu, F.: Particle surface area, ultrafine particle number concentration, and cardiovascular hospitalizations, *Environmental Pollution*, 310, 119795, <https://doi.org/10.1016/j.envpol.2022.119795>, 2022.
- Lindauer, M., Eggensperger, K., Feurer, M., Biedenkapp, A., Deng, D., Benjamins, C., Ruhkopf, T., Sass, R., and Hutter, F.: SMAC3: A Versatile Bayesian Optimization Package for Hyperparameter Optimization, *Journal of Machine Learning Research*, 23, 1–9, <http://jmlr.org/papers/v23/21-0888.html> (last access: 20 October 2025), 2022.
- Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, in: *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, Curran Associates Inc., Red Hook, NY, USA, 4768–4777, ISBN 9781510860964, 2017.
- Maraun, D. and Widmann, M.: *Statistical Downscaling and Bias Correction for Climate Research*, Cambridge University Press, <https://doi.org/10.1017/9781107588783>, 2018.
- McCullagh, P. and Nelder, J.: *Generalized Linear Models*, Second Edition, Taylor & Francis, ISBN 9780412317606, 1989.
- McDuffie, E., Martin, R., Yin, H., and Brauer, M.: *Global Burden of Disease from Major Air Pollution Sources (GBD MAPS)*:

- A Global Approach, Research report, Health Effects Institute, 2021, 1–45, PMCID: PMC9501767, PMID: 36148817, 2021.
- Miinalainen, T., Kokkola, H., Lipponen, A., Hyvärinen, A.-P., Soni, V. K., Lehtinen, K. E. J., and Kühn, T.: Assessing the climate and air quality effects of future aerosol mitigation in India using a global climate model combined with statistical downscaling, *Atmos. Chem. Phys.*, 23, 3471–3491, <https://doi.org/10.5194/acp-23-3471-2023>, 2023.
- Mikkonen, S., Németh, Z., Varga, V., Weidinger, T., Leinonen, V., Yli-Juuti, T., and Salma, I.: Decennial time trends and diurnal patterns of particle number concentrations in a central European city between 2008 and 2018, *Atmos. Chem. Phys.*, 20, 12247–12263, <https://doi.org/10.5194/acp-20-12247-2020>, 2020.
- Molnar, C.: Interpretable Machine Learning, 2nd edn., <https://christophm.github.io/interpretable-ml-book> (last access: 20 October 2025), 2022.
- Neubauer, D., Ferrachat, S., Siegenthaler-Le Drian, C., Stier, P., Partridge, D. G., Tegen, I., Bey, I., Stanelle, T., Kokkola, H., and Lohmann, U.: The global aerosol–climate model ECHAM6.3–HAM2.3 – Part 2: Cloud evaluation, aerosol radiative forcing, and climate sensitivity, *Geosci. Model Dev.*, 12, 3609–3639, <https://doi.org/10.5194/gmd-12-3609-2019>, 2019.
- Nguyen, D., Gupta, S., Rana, S., Shilton, A., and Venkatesh, S.: Bayesian Optimization for Categorical and Category-Specific Continuous Inputs, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, <https://doi.org/10.1609/aaai.v34i04.5971>, 2020.
- Nogueira, F.: Bayesian Optimization: Open source constrained global optimization tool for Python, GitHub, <https://github.com/bayesian-optimization/BayesianOptimization> (last access: 17 September 2025), 2014.
- Ostro, B., Hu, J., Goldberg, D., Reynolds, P., Hertz, A., Bernstein, L., and Kleeman, M.: Associations of Mortality with Long-Term Exposures to Fine and Ultrafine Particles, Species and Sources: Results from the California Teachers Study Cohort, *Environmental health perspectives*, 123, <https://doi.org/10.1289/ehp.1408565>, 2015.
- Pagano, P., De Zaiacomo, T., Scarcella, E., Bruni, S., and Calamoscia, M.: Mutagenic Activity of Total and Particle-Sized Fractions of Urban Particulate Matter, *Environmental Science & Technology*, 30, 3512–3516, <https://doi.org/10.1021/es960182q>, 1996.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.: Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 12, 2825–2830, 2011.
- Pepe, N., Pirovano, G., Lonati, G., Balzarini, A., Toppetti, A., Riva, G., and Bedogni, M.: Development and application of a high resolution hybrid modelling system for the evaluation of urban air quality, *Atmospheric Environment*, 141, 297–311, <https://doi.org/10.1016/j.atmosenv.2016.06.071>, 2016.
- Petäjä, T., Tabakova, K., Manninen, A., Ezhova, E., O'Connor, E., Moiseev, D., Sinclair, V. A., Backman, J., Levula, J., Luoma, K., Virkkula, A., Paramonov, M., Rätty, M., Äijälä, M., Heikkinen, L., Ehn, M., Sipilä, M., Yli-Juuti, T., Virtanen, A., Ritsche, M., Hickmon, N., Pulik, G., Rosenfeld, D., Worsnop, D. R., Bäck, J., Kulmala, M., and Kerminen, V.-M.: Influence of biogenic emissions from boreal forests on aerosol–cloud interactions, *Nature Geoscience*, 15, 42–47, <https://doi.org/10.1038/s41561-021-00876-0>, 2022.
- Rasmussen, C. E. and Williams, C. K. I.: *Gaussian Processes for Machine Learning*, The MIT Press, ISBN 9780262256834, <https://doi.org/10.7551/mitpress/3206.001.0001>, 2005.
- Rönkkö, T., Kuuluvainen, H., Karjalainen, P., Keskinen, J., Hillamo, R., Nieminen, J. V., Pirjola, L., Timonen, H. J., Saarikoski, S., Saukko, E., Järvinen, A., Silvennoinen, H., Rostedt, A., Olin, M., Yli-Ojanperä, J., Nousiainen, P., Kousa, A., and Dal Maso, M.: Traffic is a major source of atmospheric nanocluster aerosol, *Proceedings of the National Academy of Sciences*, 114, 7549–7554, <https://doi.org/10.1073/pnas.1700830114>, 2017.
- Sachindra, D., Ahmed, K., Rashid, M. M., Shahid, S., and Perera, B.: Statistical downscaling of precipitation using machine learning techniques, *Atmospheric Research*, 212, 240–258, <https://doi.org/10.1016/j.atmosres.2018.05.022>, 2018.
- Schraufnagel, D.: The health effects of ultrafine particles, *Experimental & Molecular Medicine*, 52, 311–317, <https://doi.org/10.1038/s12276-020-0403-3>, 2020.
- Schultz, M. G., Stadler, S., Schröder, S., Taraborrelli, D., Franco, B., Krefting, J., Henrot, A., Ferrachat, S., Lohmann, U., Neubauer, D., Siegenthaler-Le Drian, C., Wahl, S., Kokkola, H., Kühn, T., Rast, S., Schmidt, H., Stier, P., Kinnison, D., Tyndall, G. S., Orlando, J. J., and Wespes, C.: The chemistry–climate model ECHAM6.3–HAM2.3–MOZ1.0, *Geosci. Model Dev.*, 11, 1695–1723, <https://doi.org/10.5194/gmd-11-1695-2018>, 2018.
- Schulzweida, U.: CDO User Guide, Zenodo, <https://doi.org/10.5281/zenodo.10020800>, 2023.
- Sharma, S., Sharma, P., and Khare, M.: Hybrid modelling approach for effective simulation of reactive pollutants like Ozone, *Atmospheric Environment*, 80, 408–414, <https://doi.org/10.1016/j.atmosenv.2013.08.021>, 2013.
- SmartSMEAR: SmartSMEAR database, <https://smear.avaa.csc.fi/> (last access: 17 September 2025), 2025.
- Snoek, J., Larochelle, H., and Adams, R. P.: Practical Bayesian Optimization of Machine Learning Algorithms, in: *Advances in Neural Information Processing Systems*, edited by: Pereira, F., Burges, C., Bottou, L., and Weinberger, K., vol. 25, Curran Associates, Inc., ISBN 9781627480031, https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf (last access: 20 October 2025), 2012.
- Sofiev, M., Siljamo, P., Valkama, I., Ilvonen, M., and Kukkonen, J.: A dispersion modelling system SILAM and its evaluation against ETEX data, *Atmospheric Environment*, 40, 674–685, <https://doi.org/10.1016/j.atmosenv.2005.09.069>, 2006.
- Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., de' Donato, F., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., Viegi, G., Kloog, I., and Schwartz, J.: Estimation of daily PM₁₀ and PM_{2.5} concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model, *Environment International*, 124, 170–179, <https://doi.org/10.1016/j.envint.2019.01.016>, 2019.
- Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., Block, K., Brokopf, R., Fast, I., Kinne, S., Kornblüth, L., Lohmann, U., Pincus, R., Reichler, T., and Roeckner, E.: Atmospheric component of the MPI-M Earth System Model: ECHAM6, *Jour-*

- nal of Advances in Modeling Earth Systems, 5, 146–172, <https://doi.org/10.1002/jame.20015>, 2013.
- Stier, P., Feichter, J., Kinne, S., Kloster, S., Vignati, E., Wilson, J., Ganzeveld, L., Tegen, I., Werner, M., Balkanski, Y., Schulz, M., Boucher, O., Minikin, A., and Petzold, A.: The aerosol-climate model ECHAM5-HAM, *Atmos. Chem. Phys.*, 5, 1125–1156, <https://doi.org/10.5194/acp-5-1125-2005>, 2005.
- Stohl, A., Aamaas, B., Amann, M., Baker, L. H., Bellouin, N., Bernsten, T. K., Boucher, O., Cherian, R., Collins, W., Daskalakis, N., Dusinska, M., Eckhardt, S., Fuglestedt, J. S., Harju, M., Heyes, C., Hodnebrog, Ø., Hao, J., Im, U., Kanakidou, M., Klimont, Z., Kupiainen, K., Law, K. S., Lund, M. T., Maas, R., MacIntosh, C. R., Myhre, G., Myriokefalitakis, S., Oivié, D., Quaas, J., Quennehen, B., Raut, J.-C., Rumbold, S. T., Samset, B. H., Schulz, M., Seland, Ø., Shine, K. P., Skeie, R. B., Wang, S., Yttri, K. E., and Zhu, T.: Evaluating the climate and air quality impacts of short-lived pollutants, *Atmos. Chem. Phys.*, 15, 10529–10566, <https://doi.org/10.5194/acp-15-10529-2015>, 2015.
- Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An Overview of CMIP5 and the Experiment Design, *Bulletin of the American Meteorological Society*, 93, 485–498, <https://doi.org/10.1175/BAMS-D-11-00094.1>, 2012.
- Tegen, I., Neubauer, D., Ferrachat, S., Siegenthaler-Le Drian, C., Bey, I., Schutgens, N., Stier, P., Watson-Parris, D., Stanelle, T., Schmidt, H., Rast, S., Kokkola, H., Schultz, M., Schroeder, S., Daskalakis, N., Barthel, S., Heinold, B., and Lohmann, U.: The global aerosol–climate model ECHAM6.3–HAM2.3 – Part 1: Aerosol evaluation, *Geosci. Model Dev.*, 12, 1643–1677, <https://doi.org/10.5194/gmd-12-1643-2019>, 2019.
- Turnock, S. T., Allen, R. J., Andrews, M., Bauer, S. E., Deushi, M., Emmons, L., Good, P., Horowitz, L., John, J. G., Michou, M., Nabat, P., Naik, V., Neubauer, D., O'Connor, F. M., Oivié, D., Oshima, N., Schulz, M., Sellar, A., Shim, S., Takemura, T., Tilmes, S., Tsigaridis, K., Wu, T., and Zhang, J.: Historical and future changes in air pollutants from CMIP6 models, *Atmos. Chem. Phys.*, 20, 14547–14579, <https://doi.org/10.5194/acp-20-14547-2020>, 2020.
- Vachon, J., Buteau, S., Liu, Y., Van Ryswyk, K., Hatzopoulou, M., and Smargiassi, A.: Spatial and spatiotemporal modelling of intra-urban ultrafine particles: A comparison of linear, nonlinear, regularized, and machine learning methods, *Science of The Total Environment*, 954, 176523, <https://doi.org/10.1016/j.scitotenv.2024.176523>, 2024a.
- Vachon, J., Kerckhoffs, J., Buteau, S., and Smargiassi, A.: Do machine learning methods improve prediction of ambient air pollutants with high spatial contrast? A systematic review, *Environmental Research*, 262, 119751, <https://doi.org/10.1016/j.envres.2024.119751>, 2024b.
- Vandal, T. J., Kodra, E., and Ganguly, A. R.: Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation, *Theoretical and Applied Climatology*, 137, 557–570, 2017.
- Vartiainen, A.: Machine Learning-Based Downscaling of Aerosol Size Distributions from a Global Climate Model, Zenodo [dataset], <https://doi.org/10.5281/zenodo.17079843>, 2025.
- Venuta, A., Lloyd, M., Ganji, A., Xu, J., Simon, L., Zhang, M., Saeedi, M., Yamanouchi, S., Lavigne, E., Hatzopoulou, M., and Weichenenthal, S.: Predicting within-city spatiotemporal variations in daily median outdoor ultrafine particle number concentrations and size in Montreal and Toronto, Canada, *Environmental Epidemiology*, 8, e323, <https://doi.org/10.1097/EE9.0000000000000323>, 2024.
- Vogli, M., Peters, A., Wolf, K., Thorand, B., Herder, C., Koenig, W., Cyrus, J., Maestri, E., Marmiroli, N., Karasch, S., Zhang, S., and Pickford, R.: Long-term exposure to ambient air pollution and inflammatory response in the KORA study, *Science of The Total Environment*, 912, 169416, <https://doi.org/10.1016/j.scitotenv.2023.169416>, 2023.
- Williamson, C. J., Kupc, A., Axisa, D., Bilsback, K. R., Bui, T., Campuzano-Jost, P., Dollner, M., Froyd, K. D., Hodshire, A. L., Jimenez, J. L., Kodros, J. K., Luo, G., Murphy, D. M., Nault, B. A., Ray, E. A., Weinzierl, B., Wilson, J. C., Yu, F., Yu, P., Pierce, J. R., and Brock, C. A.: A large source of cloud condensation nuclei from new particle formation in the tropics, *Nature*, 574, 399–403, <https://doi.org/10.1038/s41586-019-1638-9>, 2019.
- Wolf, K., Cyrus, J., Harciníková, T., Gu, J., Kusch, T., Hampel, R., Schneider, A., and Peters, A.: Land use regression modeling of ultrafine particles, ozone, nitrogen oxides and markers of particulate matter pollution in Augsburg, Germany, *Science of The Total Environment*, 579, 1531–1540, <https://doi.org/10.1016/j.scitotenv.2016.11.160>, 2017.
- World Health Organization: WHO global air quality guidelines. Particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide, ISBN 9789240034228, 2021.
- Xu, R., Chen, N., Chen, Y., and Chen, Z.: Downscaling and Projection of Multi-CMIP5 Precipitation Using Machine Learning Methods in the Upper Han River Basin, *Advances in Meteorology*, 2020, 8680436, <https://doi.org/10.1155/2020/8680436>, 2020.
- Zani, N. B., Lonati, G., Mead, M. I., Latif, M. T., and Crippa, P.: Long-term satellite-based estimates of air quality and premature mortality in Equatorial Asia through deep neural networks, *Environmental Research Letters*, 15, 104088, <https://doi.org/10.1088/1748-9326/abb733>, 2020.
- Zhang, K., Wan, H., Liu, X., Ghan, S. J., Kooperman, G. J., Ma, P.-L., Rasch, P. J., Neubauer, D., and Lohmann, U.: Technical Note: On the use of nudging for aerosol–climate model intercomparison studies, *Atmos. Chem. Phys.*, 14, 8631–8645, <https://doi.org/10.5194/acp-14-8631-2014>, 2014.