https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

Sensitivity Analysis for a Statistical Prediction Model Utilized for Particle Number Concentrations with a Focus on Submicron Size Fractions

Adeeb Ktaish[1], Nahid Atashi[1], Ola Surakhi[4], Yu Zhao[5], Tianyi Zhao[5], Martha Arbayani Zaidan[1,2], Tuukka Petäjä[1], Tuija Jokinen[1,3], Wei Du[1], Markku Kulmala[1], and Tareq Hussein[1,6]

[1]Institute for Atmospheric and Earth System Research (INAR/Physics), Faculty of Science, University of Helsinki, 00014 Helsinki, Finland
[2]Department of Computer Science, Faculty of Science, University of Helsinki, 00014 Helsinki, Finland
[3]Climate & Atmosphere Research Centre (CARE-C), The Cyprus Institute, 1645 Nicosia, Cyprus
[4]Data Science and Artificial Intelligence, American University of Madaba, Madaba, 11821, Jordan
[5]School of Infrastructure Engineering, Dalian University of Technology, No. 2 Linggong Road, Ganjingzi District, Dalian 116024, China
[6]University of Jordan, School of Science, Department of Physics, Environmental and Atmospheric Research Laboratory (EARL), Amman 11942, Jordan

Correspondence: Adeeb Ktaish (adeeb.ktaish@helsinki.fi) Tareq Hussein (tareq.hussein@helsinki.fi)

Keywords. Bayesian statistical framework; Boreal Forest; Rural; Particle size distribution; Ultrafine; Accumulation

Abstract. Long-term particle number size distribution (PNSD) observations are sparse and costly to maintain, limiting their use for prediction and process attribution. We develop a Bayesian regression with autoregressive (AR) errors and sequential prediction, estimating posteriors via Markov chain Monte Carlo (MCMC). The study uses multi-year data set from 1996 till 2024, hourly observations from three Finnish SMEAR stations—SMEAR I (Värriö), SMEAR II (Hyytiälä), and SMEAR III (Helsinki)—with predictors drawn from routinely measured meteorology and trace gases. The developed model was applied on 5 different scenarios with different aims.Scenario I showed that expanding the historical training period improve or stabilize prediction performance. Scenario II demonstrated that predictive performance depends strongly on the training year. Scenario III showed that adding increasingly old observations provide no additional benefit once recent observations were already included. Scenario IV demonstrated that progressively updating the model with more recent data generally improved predictive performance. Scenario V showed that increasing the length of consecutive training windows generally reduced prediction errors. Addtionally, the model captures both diurnal and seasonal variability and achieves high coefficients of determination (R^2) at 1 h resolution indicating that most observed variability is explained. These results suggest a successful prediction for particle number concentrations. The potential of this approach is to make future forecasting for particle number concentrations for the up coming years.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

1 Introduction

An aerosol is a two–phase system consisting of solid and/or liquid particles suspended in a carrier gas, which is the air (Hinds, 1999). Aerosols can be emitted directly to the atmosphere from primary sources or be formed *in situ* via nucleation-mode of gas–phase species(Kulmala et al., 2022; Thakur et al., 2022; Peltola et al., 2026; Paasonen et al., 2013; Zaidan et al., 2020). New particle formation (NPF) is commonly defined as a sudden burst of sub–nanometer (1–3 nm) particles followed by their subsequent growth(Kulmala et al., 2004; Kerminen et al., 2018; Lee et al., 2019). NPF proceeds in two main stages (Kulmala et al., 2014, 2013; Lehtipalo et al., 2025): (i) the formation of a critical cluster during the transition from vapor to a condensed phase, and (ii) growth of this cluster to larger diameters (typically 2–3 nm and beyond) while competing against losses.

Atmospheric aerosols influence climate through both direct and indirect effects (Kulmala et al., 2005, 2013; Kerminen et al., 2012) . Direct effects arise from scattering and absorption of radiation, while indirectly by modifying cloud properties by acting as cloud condensation nuclei (Kulmala et al., 2005). From a health perspective, ultrafine particles (UFP; $D_p < 100$ nm) dominate number concentrations, deposit efficiently in the respiratory tract, and may translocate to extrapulmonary organs; number–based metrics are therefore important for exposure assessment (Hussein et al., 2015; Ma et al., 2022).

Particle size distribution (PSD) observations are important to exposure and impacts assessments. However, spatially distributed PSD monitoring remains sparse because instruments are expensive to deploy, maintain, and operate as well as require regular calibration (McMurry, 2000; Seinfeld and Pandis, 2016), limiting coverage and continuity. To bridge these observational gaps, researchers increasingly rely on modelling approaches that broadly fall into two classes: (1) Deterministic dispersion and chemical–transport models and (2) statistical models assisted with machine learning (ML) and artificial intelligence (AI). Deterministic dispersion and chemical–transport models explicitly represent sources and sinks but require detailed emissions inventories and substantial computational resources (Seinfeld and Pandis, 2016; Jacob, 1999). In contrast, statistical models learn empirical relationships from observations and can be deployed operationally with modest requirements (Mølgaard et al., 2012). Prior work has used meteorology, traffic activity, or both to estimate particle number concentrations (Hussein et al., 2006; Westmoreland et al., 2007; Liu et al., 2008), while extensive statistical forecasting has focused on gas–phase pollutants (Schlink et al., 2006; Chelani and Devotta, 2006; Hrušt et al., 2009; Cheon et al., 2009; Kumar and Jain, 2010; Moustris et al., 2010; Jaiswal et al., 2018; Wei et al., 2019; Koo et al., 2020; Middya and Roy, 2022; Zaidan et al., 2020). For example, Hussein et al. (2006) modelled daily mean number concentrations using temperature (T) and wind speed (U) with second–order polynomial terms across 36 wind sectors, achieving good performance for daily means but without capturing diurnal variability or providing an operational forecast. Building on this line of work, Mølgaard et al. (2012) incorporated traffic intensity, relative humidity (RH), and temporal factors (time of day, day of week, seasonality) and added an autoregressive component within a Bayesian framework; however, their analysis covered only a short period (2006–2008). In parallel, deep–learning approaches have been proposed to estimate number size distributions or bin–wise concentrations from routine covariates (Wu et al., 2025).

Despite these advances, there remains a notable lack of long time–series, multi–site studies that jointly model or predict nucleation–mode concentrations across contrasting environments. Very few efforts combine multi–year datasets from multiple sites and simultaneously predict nucleation, UFP, and accumulation modes in urban, boreal–forest, and remote settings.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

However, there remains a notable lack of long time series data, PSD records, and multi-site studies that jointly model or predict nucleation-mode concentrations across contrasting environments. As a result, very few studies have combined long-term datasets from multiple sites and simultaneously predicted the nucleation mode in three distinct settings (e.g., urban, boreal forest, and subarctic / remote background).

Here, we address this gap by integrating long-term observations from three environments and developing a statistical framework that leverages routinely measured meteorology and gas-phase pollutants to estimate particle number concentrations with explicit emphasis on the nucleation mode, UFP, and accumulation mode.

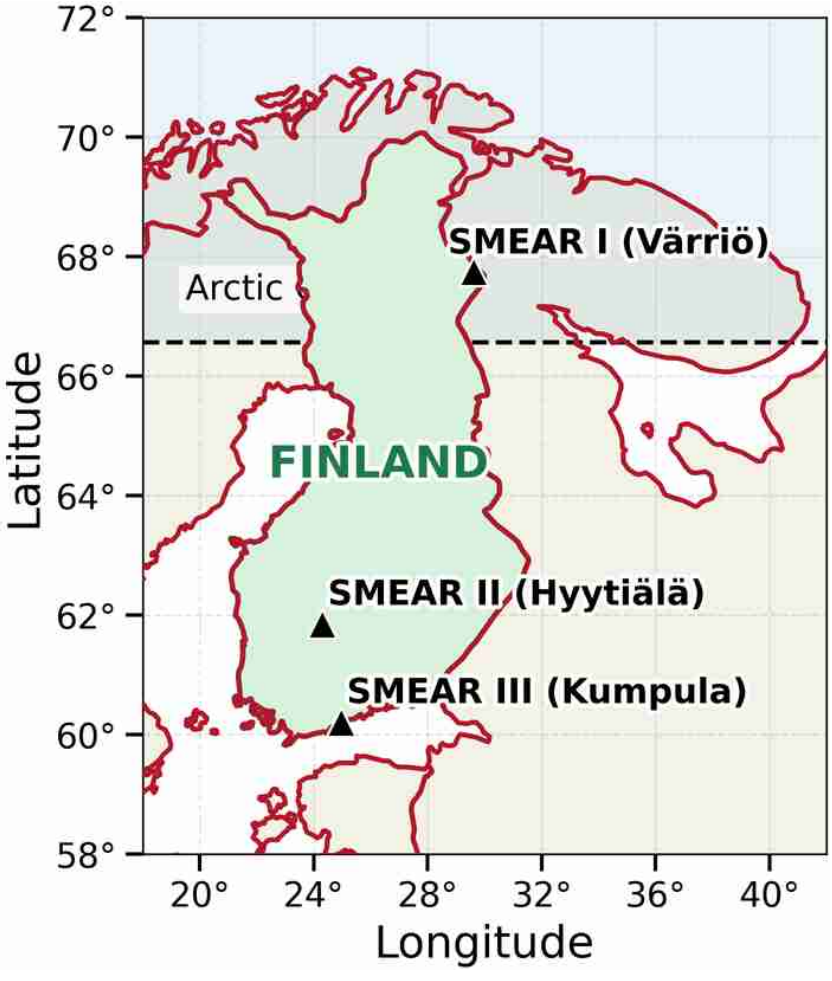


Figure 1. A map of Finland showing the locations of the SMEAR (Station for Measuring Ecosystem-Atmosphere Relations) stations: SMEAR III (Helsinki–Kumpula; urban background), SMEAR II (Hyytiälä; rural boreal forest), and SMEAR I (Värriö; subarctic / remote background site).

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

2 Methodology

2.1 Data base and data handling

The data utilized in this study consists of aerosol particle number size distributions, gases, and meteorological conditions. The data bases were collected from the SMEAR network (Station for Measuring Ecosystem–Atmosphere Relations). The primary purpose of the SMEAR stations is to measure exchanges of momentum, energy, and matter in different environments and to provide continuous long-term measurements of the chemical and physical properties of atmospheric aerosols, gas pollutants, turbulent exchange, and basic meteorology (Junninen et al., 2025). In this study, we focus on three SMEAR stations: Helsinki (SMEAR III), Hyytiälä (SMEAR II), and Värriö (SMEAR I) Fig. 1. SMEAR III is an Urban background site located on the Kumpula campus, about 4 km NNE of Helsinki city centre (Järvi et al., 2009). SMEAR II is a boreal forest site situated in a Scots pine forest near Hyytiälä. SMEAR I is rural background located in Värriö, close to the Russian border and north of the Arctic Circle and represents a remote location with very low local emissions (Järvi et al., 2009).

The quality of the aerosol and meteorological data was checked. Detailed analysis was performed to remove suspicious data. Additionally, both data sets were aggregated to the the hourly average by retaining only values above the 80th percentile. Daily averages were then calculated from the filtered hourly data, followed by monthly and annual means derived from the daily.

2.1.1 Particle number size distributions

The particle number size distributions were measured with a Differential Mobility Particle Sizer (DMPS) at all selected sites. The DMPS technique is based on bipolar charging of aerosols followed by size classification according to electrical mobility in a Diffeential Mobility Analyzer (DMA). Bearing in mind we are dealing here with long-term aerosol data base, it is acceptable to expect differences in the DMPS system setup changing from time to time at the same site as a matter of updates in the system or increasing the capability of the system to cover a wider particle diameter range. Besides that, it is also acceptable that the DMPS systems operated at the selected SMEAR sites slightly varied in setup. However, the DMPS systems were maintained and calibrated regularly to ensure comparability and standardization. In general, the cut-off diameter of the DMPS systems was below 10 nm and the upper diameter range was about 950 nm at most. Detailed information about the DMPS system in the past 25 years can be found in our previous publications about SMEAR stations included in this study (Järvi et al., 2009) (Tunved et al., 2003). All aerosol measurements underwent standardized quality-assurance, laboratory verification, and correction procedures to ensure data validity prior to analysis.

The particle number concentration N was calculated by integrating the particle number size distribution $\frac{dN}{d\log D_p}$ over the relevant diameter range,

$$N = \int n_N^0(D_p)\, d(\log D_p). \tag{1}$$

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

2.1.2 Supporting gaseous and meteorological conditions data-base

The gaseous concentrations and meteorological conditions data were used as input variable for traning and prediction at all sites. The gaseous components included here were: ozone (O_3), nitrogen oxides (NO, NO_2, NO_x), sulfur dioxide (SO_2), and carbon monoxide (CO). The meteorological conditions included: Temperature (T), relative humidity (RH), precipitation (P), solar radiation (SR), and wind speed (WS) and direction (WD). These gaseous and meteorological variables were selected for their high data completeness to minimize imputation and to preserve the contiguous time series required by the auto regressive component. For Helsinki (SMEAR III), meteorological variables were obtained from the Helsinki Region Environmental Services Authority (HSY) because these measurements provided higher temporal coverage.

Although the seasons length vary from southern to Northern Finland, but assumed standard seasons as follows: winter (December–February), spring (March–May), summer (June–August), and autumn (September–November) (Ruuskanen et al., 2003).

The measurement setup varied at the three sites, but the measurement followed a standard protocol and calibration. In Helsinki (SMEAR III) and Hyytiälä (SMEAR II), gaseous components, T, WS and WD were measured at a height about 4.0–4.4 m. In Värriö (SMEAR I), gaseous components were measured at a height about 2 m, T and WS at about 4.4m, and WD at 16m.

2.2 Bayesian Generalized Additive Model (GAM)

We assumed that particle number concentrations in each size fraction (nucleation, ultrafine (UFP), and accumulation) follow a log-normal distribution. Histograms of hourly particle number concentrations N for the different size fractions–nucleation ($D_p < 25$ nm), UFP ($D_p < 100$ nm), and accumulation ($D_p > 100$ nm)–at the three sites (Helsinki, Hyytiälä, and Värriö) are shown in (Fig. 2) and support this assumption.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

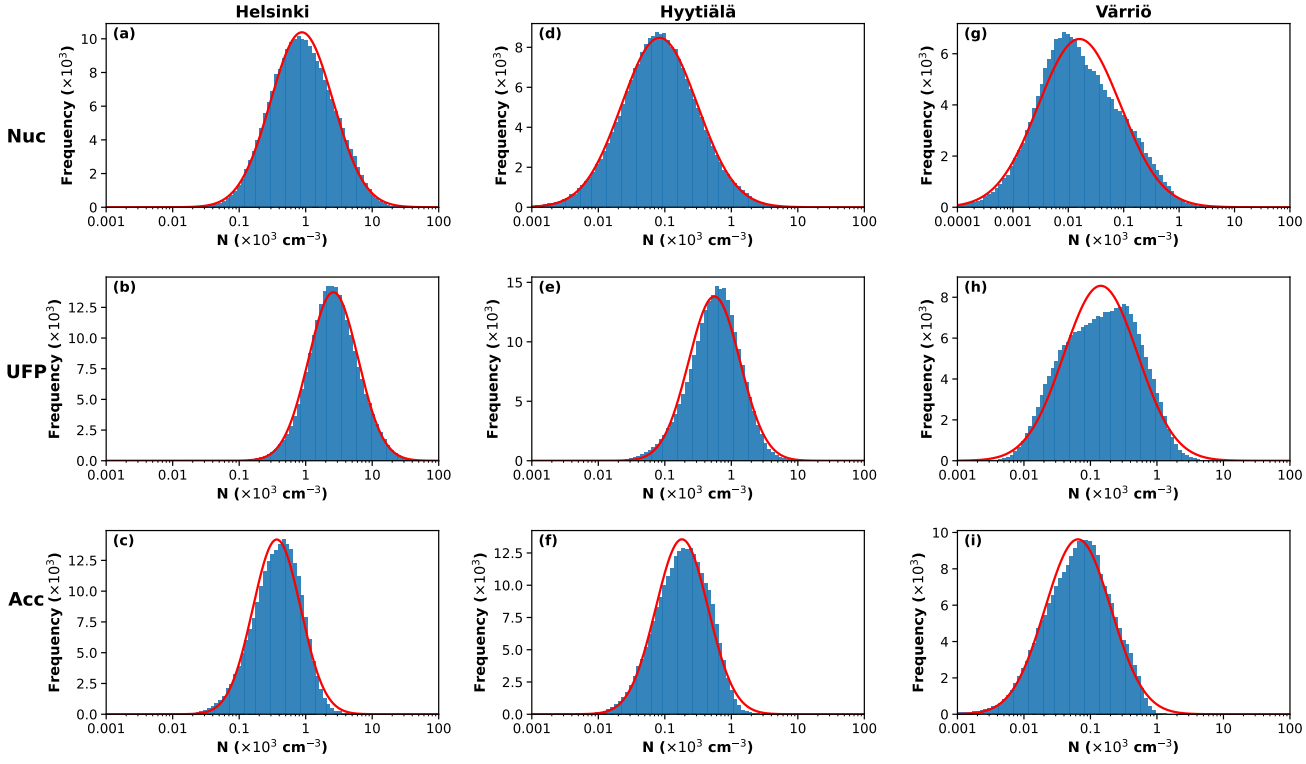


Figure 2. Histograms of hourly particle number concentrations N for the different size fractions: nucleation-mode (10 $\leq$D$_p$ $\leq$ 25 nm),UFP (10 $\leq D_p \leq$ 100 nm), and accumulation-mode (100 $\leq D_p \leq$ 300 nm)in (a–c) Helsinki(SMEAR III), (d–f) Hyytiälä (SMEAR II), and (g–i) Värriö (SMEAR I).

Accordingly, we consider the following mathematical formula for the GAM model:

$$\log N = f(\mathbf{x}) + \varepsilon, \tag{2}$$

where N is the particle number concentration for a certain particle size fractions, $\mathbf{x}$ is the vector of covariates, and ε is an error term assumed to follow a normal distribution.

2.2.1 Bayesian GAM formulation

Following Aldrin and Haff (2005) and Mølgaard et al. (2012), we model the particle number concentration, $\log N$, as linear in its parameters. The covariates include meteorological variables wind speed (U), wind direction (WD), temperature (T), precipitation (P), solar radiation (SR), relative humidity (RH) and trace gases [SO_2, NO_2, CO, and O_3]. The temporal structure is captured by a periodic component $f(t)$ with separate weekday versus weekend/holiday diurnal harmonics, seasonal harmonics, their interactions, and a linear trend. The relationship between particle number concentration and the covariates can be written

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

as

$$\begin{aligned}\log N = {} & C(\phi) + a_1(\phi)U + a_2(\phi)U^2 + b_1(\phi)T_t + b_2(\phi)T^2 + d(\phi_t)\,\mathrm{RH} \\ & + factor * P + factor * SO_2 + factor * \mathrm{NO_2} + factor * \mathrm{CO} + factor * \mathrm{O_3} + factor * \mathrm{SR} \\ & + f(t) + \varepsilon\end{aligned} \tag{3}$$

where $C(\phi)$ is a wind-direction–dependent baseline, $a_1(\phi)$ and $a_2(\phi)$ modulate the linear and quadratic wind-speed effects, $b_1(\phi)$ and $b_2(\phi)$ modulate the linear and quadratic temperature effects, $d(\phi)$ modulates the humidity effect, the β's are regression coefficients for precipitation, radiation, and trace gases, $f(t)$ is the time-dependent component, and ε_t is an error term assumed to follow a normal distribution.

By using the leading terms of a Fourier series in wind direction, we parameterise each of the functions C, a_1, a_2, b_1, b_2, and d with seven parameters. For example,

$$a_1(\phi) = a_{1,1} + \sum_{j=1}^{3} \{a_{1,2j}\sin(j\phi) + a_{1,2j}\cos(j\phi)\}, \tag{4}$$

and analogous expressions are used for $C(\phi)$, $a_2(\phi)$, $b_1(\phi)$, $b_2(\phi)$, and $d(\phi)$ with their own sets of coefficients.

The function of time $f(t)$ is composed of periodic functions of the time of day and the time of year and a linear trend to account for long-term changes. To capture daily variations, periodic functions are applied separately for weekdays and for weekends/holidays. Let t be measured in days, $t_d = 2\pi t$ the time of day, and $t_y = 2\pi t/365.25$ the time of year. Let $I_{\mathrm{off}}(t)$ be an indicator equal to 1 on weekends and holidays and 0 on workdays. Then The function of time $f(t)$ can be written as

$$\begin{aligned} f(t) = {} & 1_{\mathrm{off}}(t)\sum_{j=1}^{2}\{\tau_1 + \tau_2\sin(jt_d) + \tau_{2j+1}\cos(jt_d)\} \\ & + \left[1 - 1_{\mathrm{off}}(t)\right]\sum_{j=1}^{2}\{\tau_{2j+4}\sin(jt_d) + \tau_{2j+5}\cos(jt_d)\} \\ & + \sin(t_y)\sum_{j=1}^{2}\{\tau_{2j+8}\sin(jt_d) + \tau_{2j+9}\cos(jt_d)\} \\ & + \cos(t_y)\sum_{j=1}^{2}\{\tau_{2j+12}\sin(jt_d) + \tau_{2j+13}\cos(jt_d)\} \\ & + \sum_{j=1}^{2}\{\tau_{2j+16}\sin(jt_y) + \tau_{2j+17}\cos(jt_y)\} + \tau_{22}\,t, \end{aligned} \tag{5}$$

where $t_d = 2\pi t$ and $t_y = 2\pi t/365.25$ with t measured in days, $1_{\mathrm{off}}(t)$ is an indicator function equal to 1 on weekends and holidays and 0 on workdays, and the τ's are parameters.

Aldrin and Haff (2005) have mentioned that the error termε is auto-correlated. This problem occurs when we are modeling a time series. To address this, Chib (1993) developed a clear/robust framework to handle regression models with autocorrelated errors. He proposed a model in the following form

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

$$y_t = \mathbf{X}_t^{\mathrm{T}} \boldsymbol{\beta} + \varepsilon_t, \tag{6}$$

where the error term follows an autoregressive structure

$$\varepsilon_t = u_t + \phi_1 \varepsilon_{t-1} + \cdots + \phi_p \varepsilon_{t-p}, \tag{7}$$

with y_t denoting the observation at time t, $\mathbf{X}^{\top}$ the vector of covariates at time t, and $\boldsymbol{\beta}$ the vector of regression coefficients. The parameters $\phi_1, \ldots, \phi_p$ are autoregressive coefficients, and u_t is an independent normally distributed innovation with mean zero and standard deviation σ. This framework can be directly applied to our setting because the regression model in Eq. (3) is linear in its parameters. In our study, we set $y_t = \log N$ and define $\mathbf{X}$ and $\boldsymbol{\beta}$ accordingly.

We applied this autoregressive error formulation at 1 h, 3 h, and daily resolutions. Specifically, we included six autoregressive parameters corresponding to lags of one time step: 6 h, 12 h, 24 h, 2 d, and 1 week. At the 3 h resolution, this corresponds to an $\boldsymbol{\phi}$-vector with 56 components, of which ϕ_1, ϕ_2, ϕ_4, ϕ_8, ϕ_{16}, and ϕ_{56} are allowed to be non-zero. We verified that, with these six non-zero parameters, the innovation process u_t shows no remaining autocorrelation.

2.2.2 Bayesian GAM inference and prediction

This model was implemented in a Bayesian framework. Bayesian inference provides a principled way to combine prior information with observed data. By adhering to the likelihood principle, it yields consistent inferences across different sampling designs and offers intuitive interpretations, such as assigning 95 % probability that the true parameter vector lies within a 95 % credible interval. Let $\boldsymbol{\Theta}$ denote the parameter vector and z a generic set of observations. The conditional probability density of $\boldsymbol{\Theta}$ given z is obtained from Bayes' theorem

$$p(\boldsymbol{\Theta} \mid z) = \frac{p(z \mid \boldsymbol{\Theta})\, p(\boldsymbol{\Theta})}{p(z)}, \tag{8}$$

where $p(z \mid \boldsymbol{\Theta})$ is the conditional density of z given $\boldsymbol{\Theta}$ (the likelihood), $p(\boldsymbol{\Theta})$ is the prior density, and $p(z)$ is the marginal density of z. When z is fixed, $p(z)$ acts as a normalising constant.

To use Bayes' theorem in our model we set $z = \mathbf{y}_L$, where $\mathbf{y}_L$ denotes the learning (training) observations, and we define

$$\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\phi} \\ \sigma \end{pmatrix},$$

with $\boldsymbol{\beta}$ the regression coefficients from Eq. (3), $\boldsymbol{\phi}$ the autoregressive parameters from Eqs. (6)–(7), and σ the innovation standard deviation. The corresponding covariates are collected in X_L. We indicate learning data by the subscript L on both $\mathbf{y}_L$ and X_L. Under a prior that is independent of X_L, Bayes' theorem gives the posterior

$$p(\boldsymbol{\Theta} \mid X_L, \mathbf{y}_L) = \frac{p(\mathbf{y}_L \mid \boldsymbol{\Theta}, X_L)\, p(\boldsymbol{\Theta})}{p(\mathbf{y}_L, X_L)}. \tag{9}$$

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

We assume limited prior information about the parameters. Thus, all reasonable choices of the prior would yield very similar posterior distributions. Following Chib (1993), parameter vectors for which $\sum_j \phi_j > 0.99$ are assigned zero prior probability, ensuring stationarity of the autoregressive error process. For the remainingparameters we use an improper reference prior as suggested by Chib (1993).Forecasts of y_t can be obtained from the posterior distribution. Because the error term in Eq. (7) is autoregressive, forecast distributions depend on previous values of ε_t and must be computed sequentially in chronological order, starting from the first unknown value of y_t. Let n denote the index of the last known observation at the time of forecasting, and let the subscript W denote the last week of known observations $(X_W, \mathbf{y}_W)$. The one-step-ahead forecast distribution for time $n+1$ is

$$p(y_{n+1} \mid X_{n+1}, X_W, \mathbf{y}_W, X_L, \mathbf{y}_L) = \int_{\Theta} p(y_{n+1} \mid \boldsymbol{\Theta}, X_{n+1}, X_W, \mathbf{y}_W)\, p(\boldsymbol{\Theta} \mid X_L, \mathbf{y}_L)\, d\boldsymbol{\Theta}. \tag{10}$$

This equation shows that the integration is over the multidimensional parameter space Θ of $\boldsymbol{\Theta}$. In this study we condition on the forecasted covariates X_{n+1} and do not integrate over their uncertainty. This simplifies the analysis and focuses on the response model, while likely yielding optimistic accuracy relative to fully operational forecasting in which future covariates (e.g. tomorrow's temperature, wind,and gases) are themselves uncertain. All values of ε_t are functions of $\boldsymbol{\Theta}, X_W, \mathbf{y}_W$. In practice, the data sets $(X_W, \mathbf{y}_W)$ and $(X_L, \mathbf{y}_L)$ overlap, although for computational reasons this need not always be the case. Once the forecast distribution of y_{n+1} is determined, the corresponding forecast error ε_{n+1} becomes available. This can then be used in the calculation of the next forecast distribution. The two-step-ahead forecast distribution for time $n+2$ is

$$\begin{aligned} P(y_{n+2} \mid X_{n+2}, X_w, y_w, y_l, X_l) = \int_{\epsilon_{n+1}, \Theta} & P(y_{n+2} \mid \Theta, X_{n+2}, \epsilon_{n+1}, X_w, y_w) \\ & \times P(\epsilon_{n+1} \mid \Theta, X_w, y_w) \times P(\Theta \mid y_l, X_l)\, d\Theta\, d\epsilon_{n+1}. \end{aligned} \tag{11}$$

By iterating this procedure, it is in principle possible to continue making forecasts arbitrarily far into the future. However, it is expected that the uncertainty of the forecasts $y_{n+\ell}$ increases as the forecast horizon ℓ grows.

Monte Carlo (MC) methods were used to evaluate the integrals in Eqs. (10) and (11). In particular, we drew samples $\boldsymbol{\Theta}$ from the posterior distribution $p(\boldsymbol{\Theta} \mid X_L, \mathbf{y}_L)$. Markov chain Monte Carlo (MCMC) algorithms are generic and effective tools for sampling from such posterior distributions. However, they must be used with care because successive samples of the parameter vector are not independent, and the convergence of the Markov chain needs to be monitored. We employed the Gibbs sampler proposed by Chib (1993), which is a specific type of MCMC algorithm tailored to regression models with autoregressive errors. There is a possibility that the algorithm may visit a neighbourhood of the parameter space where the sum of the components of the autoregressive vector $\boldsymbol{\phi}$ is close to one. In such states the conditional distribution of the regression coefficients $\boldsymbol{\beta}$ can be highly uncertain, and the corresponding update may be poorly identified. Moreover, when updating $\boldsymbol{\phi}$, the algorithm may repeatedly select vectors whose elements approximately sum to one. As a consequence, the Markov chain can become trapped in a local maximum of the posterior.

In addition, the algorithm involves matrix inversions that are nearly singular when $\sum_j \phi_j$ is close to one, which may lead to numerical instability. The forecast distribution is calculated when the sample from the posterior is available. When implement-

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



ing the Gibbs sampler, the first 200 iterations were discarded as burn-in. Discarding an initial part of the chain is necessary because the Markov chain must converge to its stationary distribution before the samplesrepresent the target posterior. Furthermore, when fitting the model some of the residuals u_t could not be computed due to missing data in the covariates. These residuals were therefore excluded from the likelihood evaluation.

2.2.3 Training and applying the GAM model

We applied the GAM prediction for different particle size fractions (Nucleation mode ($10 \leq$D$_p \leq 25$ nm), UFP fraction ($10 \leq D_p \leq 100$ nm), and Accumulation mode ($100 \leq D_p \leq 300$ nm) based on hourly ; in other words, training and prediction was performed with the same data set. The GAM model was trained and applied for the three sites based on five scenarios from year 1996 till 2024

Scenario I: training data included all available data-set and prediction was performed for a whole certain year; for example, year 2020 was predicted based on training data during 1996–2024. Accordingly, we performed in total as many as 28 predictions for each site.

Scenario II: training the data included a specific year used to predict another year; for example, year 2005 was predicted based on training data during different individual years (for example year 2020 or 1999 regardless of before or after the predicted year). Accordingly, for each site, we performed 29×29 matrix of predictions (841 in total).

Scenario III: training was based on a group of years (at least two years) and prediction for a following year; for example, year 2024 was predicted from many trainings including different previous years: 2022–2023 (two years), 2021-2023 (three years), 2020–2023 (four years), and so on until 1999–2023 (28 years). Accordingly, the total number of predictions for all years was 378. Note in this prediction scenario, the end year in training was always one year before the predicted year.

Scenario IV: training was based on a group of years (at least one year) and prediction for a later year; for example, year 2024 was predicted from many trainings including different previous years: 1996 (one year only), 1996–1997 (two years), 1996–1998 (three years), 1996–1999 (four years), and so on until 1996–2023 (28 years). Accordingly, the total number of predictions for all years was 351. Note in this prediction scenario, the start year in training was always 1996.

Scenario V: training was based on a group of years (at least 2 year) and prediction for a later year; for example year 2024 was predicted several times (25 predictions) by training two years (namely: 1996–1997, 1997–1998, 1998–1999, and so on until 2022–2023) and then predicted several times (24 predictions) by training three years (namely: 1996–1998, 1997–1999, 1998–2000, and so on until 2021–2023), and so on. In total for year 2024 there were 325 predictions with training years varying from 2 years up to 26 years. Accordingly, the total number of predictions made for all years between 1996-2024 following this scenario was 20475.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

2.2.4 Performance and prediction quality check

The GAM model provides the estimated (expected) value of the particle number concentration at time

$$\hat{Y}_t = \mathbb{E}[Y_t]. \tag{12}$$

The model performance can be assessed by several measures: mean squared error (MSE), mean absolute error (MAE), the coefficient of determination (R^2), Root mean square error (RMSE) and mean percentage error (MPE). They are mathematically written as

$$\mathrm{MSE} = \frac{1}{n}\sum_{m}(Y_m - \hat{Y}_m)^2, \tag{13}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_m - \hat{Y}_m|, \tag{14}$$

$$R^2 = 1 - \frac{\sum_m (Y_m - \hat{Y}_m)^2}{\sum_m (Y_m - \bar{Y})^2} = 1 - \frac{\mathrm{MSE}}{\mathrm{Var}(Y)}\,\frac{n}{n-1}, \tag{15}$$

where Y_m is the true value, $\hat{Y}_m$ is the predicted value, and n is the number of samples.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{16}$$

$$\mathrm{MPE} = \frac{100}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{y_i}\right) \tag{17}$$

The model was first implemented, followed by an evaluation of its prediction performance. We optimised the sequential updating approach presented by Dawid (1984) . As more learning data become available over time, the model prediction is expected to improve by leveraging this accumulated information. The posterior distribution $p(\boldsymbol{\Theta} \mid X, \mathbf{y})$ of the parameters is therefore updated successively. This step could in principle be carried out on a daily or hourly basis, but to reduce computational time and effort we updated $p(\boldsymbol{\Theta} \mid X, \mathbf{y})$ only every 20th day. In our framework we assume that the forecast is issued every afternoon, using particle number measurements available up to noon. The prediction distribution is then calculated for the subsequent two days, from midnight to midnight.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

The R^2, MSE, MAE,RMSE, and MPE were calculated based on the hourly predictions. Additionally, we evaluated the model performance change among different scenarios and within each scenario, the R^2 was calculated and compared as a function of training data length and time gap between training data and predicted year. For instance,

$$\text{Loss} = R^2_{\text{best}} - R^2 \tag{18}$$

where R^2_{best} is the highest R^2 for that training data, and R^2 is the value for a given predicted year. A loss of zero indicates that the year is the best-performing year, while larger values indicate lower performance. Finally, for each training window, the median loss, the 25th percentile, the 75th percentile, and the maximum loss were calculated. The median loss represents the typical reduction in performance relative to the best year, the inter quartile range shows the spread.

3 Results and Discussion

3.1 An overview about the particle number concentrations

Total particle number concentrations N_{tot} in Helsinki (SMEAR III) exhibited a decreasing long-term trend since late 1990s. This decreasing trend is explained by controlled emissions from anthropogenic activities. In general, N_{tot} in Helsinki was dominated by ultrafine particles (UFP; $10 \leq D_p < 100\,\text{nm}$), which also exhibited the same decreasining trend (Fig. 3a and Fig. 4b). N_{tot} and also UFP concentrations were higher in the winter than in the summer reflecting a strong seasonal variation. This was mainly due to increased emissions and variations in meteorological conditions (low temperature and inversion). For example, UFP and N_{tot} concentrations were affected by traffic and heating emissions (traced by NO_2 and CO) that could be trapped beneath shallow boundary layer in the winter. The Nucleation mode (N_{nuc}; $10 \leq D_p \leq 25\,\text{nm}$) concentration also showed similar decreasing trend as those for UFP and N_{tot} (Fig. 4a). N_{nuc} also showed a seasonal variations reflecting the occurrence of high anthropogenic emissions in the winter and high frequency of new particle formation (NPF) in the spring. N_{acc} showed a slightly decreasing trend but also exhibited a seasonal variability (Fig. 4c).

In Hyytiälä (SMEAR II), N_{tot} was generally lower than what was observed in Helsinki but it maintained a rather constant annual variation with a clear seasonal variation (Fig. 3c) . The UFP concentration in Hyytiälä also dominated N_{tot} and interestingly, all particle size fractions maintaned a rather constant annual variation and as expected they exhibited a seasonal variation (Fig. 4). In Värriö (SMEAR I), N_{tot} was generally lower than those in Helsinki and Hyytiälä and all particle size fractions maintained rather constant annual variations with clear seasonal variations (Fig. 3 and Fig. 4).

3.1.1 Particle number concentration variation with meteorological conditions

A critical part in successful prediction is to pre-investigate the relationship between the input parameters and the predicted parameters (here particle number concentration). The input parameters included meteorological parameters and gaseous components. The temperature and wind speed usually have the major part affecting the aerosols concentrations because these

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

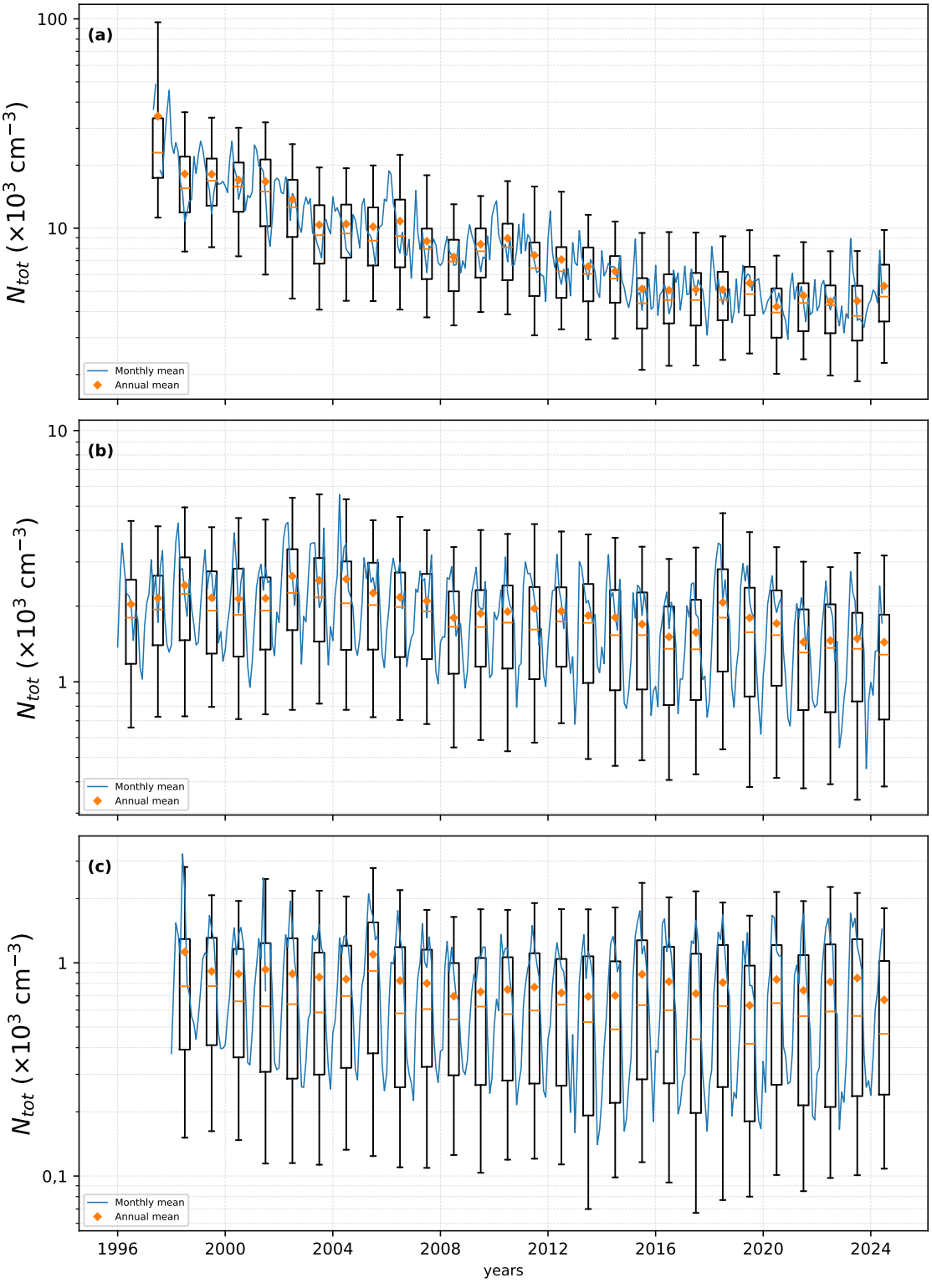


Figure 3. Time series for the total particle number concentration (N_{tot}) in (a) Helsinki (SMEAR III), (b) Hyytiälä (SMEAR II), and (c) Värriö (SMEAR I).

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

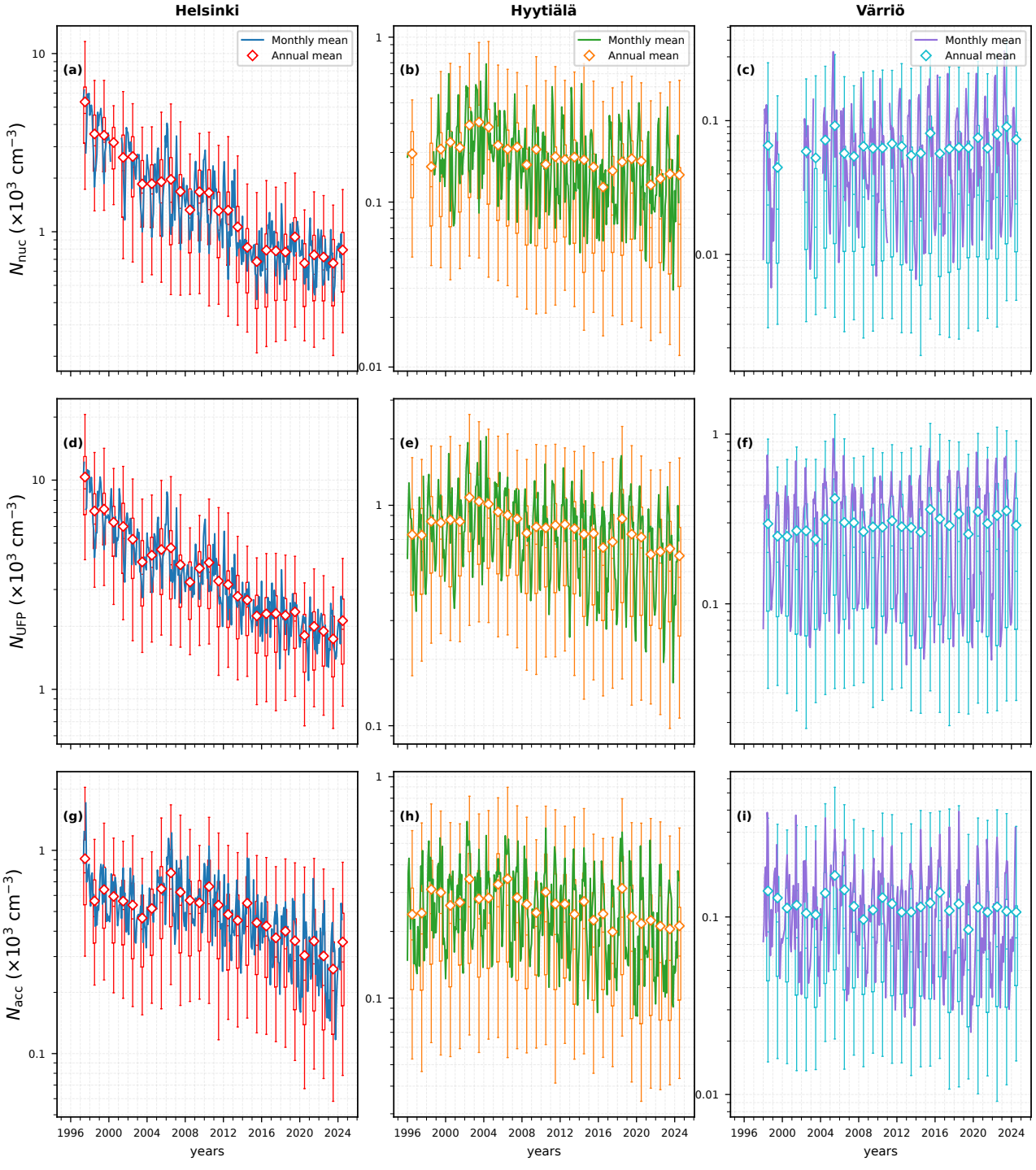


Figure 4. Time series for particle size fraction number concentration in Helsinki (SMEAR III) panels (a) nucleation (10 $\leq$D$_p$ $\leq$ 25 nm), (d-f) UFP (10 $\leq$ D_p $<$ 100 nm), and (g-i) accumulation (100 $\leq$ D_p $\leq$ 300 nm), Hyytiälä (SMEAR II) panels (d) nucleation (10 $\leq$D$_p$ $\leq$ 25 nm), (e) UFP (10 $\leq$ D_p $\leq$ 100 nm), and (f) accumulation (100 $\leq$ D_p $\leq$ 300 nm), and Värriö (SMEAR I) panels (g) nucleation (10 $\leq$D$_p$ $\leq$ 25 nm), (h) UFP (10 $\leq$ D_p $\leq$ 100 nm), and (i) accumulation (100 $\leq$ D_p $\leq$ 300 nm).

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



parameters affect the boundary layer as well as dispersion and dilution of pollutants in the air. The wind direction also play a role in dilution when clean air enters the region and mixes with the local air that contains local air pollution within the boundary layer. Particle number concentrations, in general, have either positive or negative correlation with some gaseous air pollutants. The positive correlation is likely indicating that the aerosol particles and those gaseous pollutants originate from the same source. the negative correlation is yet to be investigated depending on the physicochemical dynamics and gas-to-particle relationship in the air. As an advantage of the positive correlation, prediction of aerosol particle number concentration can be enhanced by including those gaseous components that originate from the sources as those of the particulate matter . For example, it is most likely that UFP as well as nucleation mode particles concentrations have a high correlation coefficient with NO_x, which can be a good predictor for number concentrations; and thus, be included in the model training (Fig. 5).

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

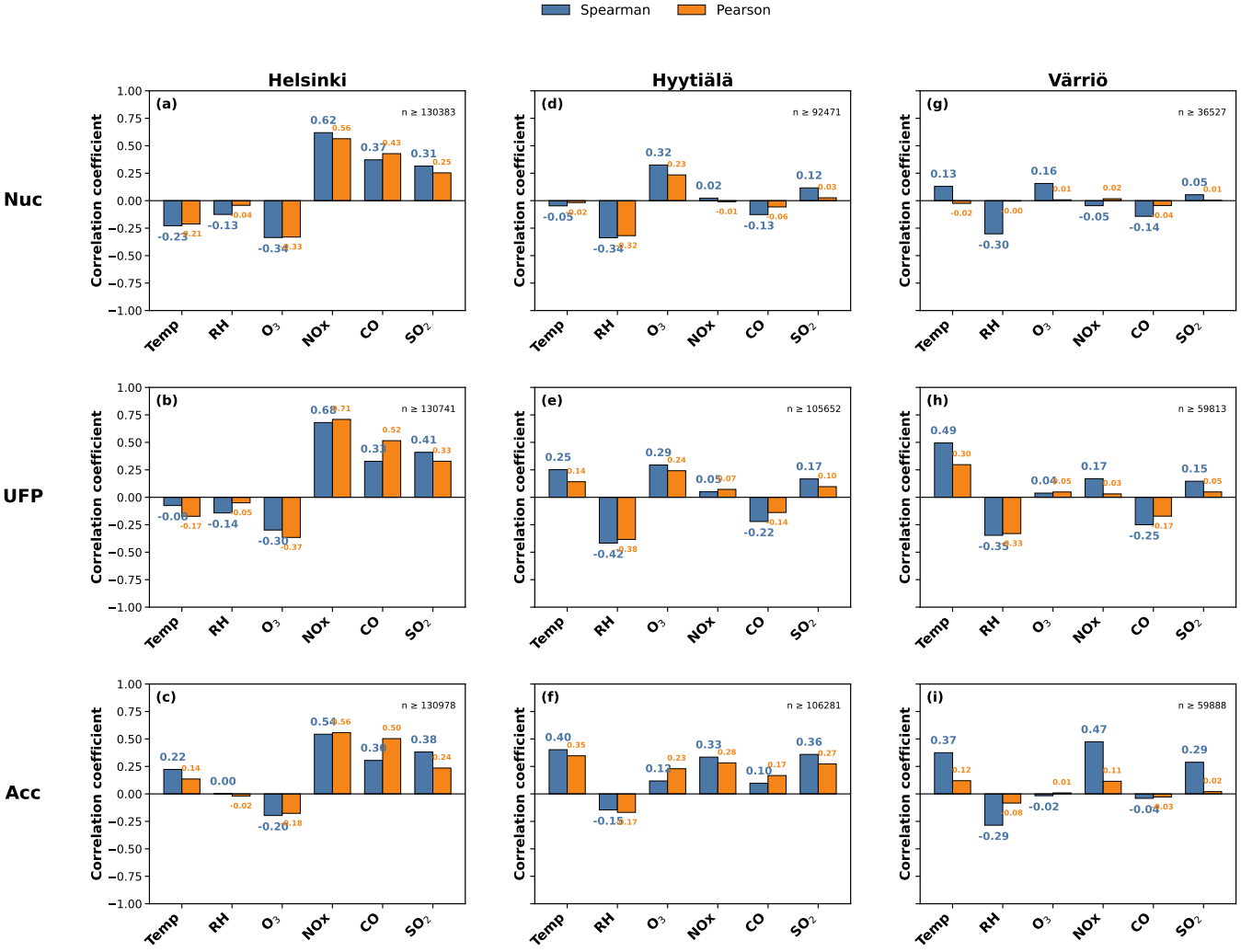


Figure 5. Correlation coefficients between size-fractionated particle number concentrations and selected meteorological and gaseous variables at Helsinki, Hyytiälä, and Värriö. Rows correspond to nucleation-mode particles (N_{nuc}), ultrafine particles (N_{UFP}), and accumulation-mode particles (N_{acc}). Blue bars indicate Spearman correlation coefficients and orange bars indicate Pearson correlation coefficients.

3.1.1.1 The effect of wind speed and direction

In Helsinki (SMEAR III), all particle size fractions exhibited a decreasing trend as wind speed increases showing that ventilation and dilution under stronger winds, reduces the contribution of near-field sources and lower the urban background loading. (Hussein et al., 2006) (Fig. 6)..

In Hyytiälä (SMEAR II), N_{nuc} and UFP exhibited an increasing trend as wind speed increased , which is completely the opposite compared to Helsinki, where as N_{acc} showed a similar pattern this is consistent with stronger winds being associated

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



with better ventilation and reduced accumulation-mode loading which can increase the survival probability of freshly formed clusters into the nucleation-size range (Dal Maso et al., 2005) (Fig. 6).

In Varrio (SMEAR I), all particle size fractions exhibited slightly increasing trend, opposite to Helsinki, suggesting that particle number is strongly influenced by air-mass origin (Vehkamäki et al., 2004) (Fig. 6).

In Helsinki(SMEAR III) particle number concentrations are systematically higher across all size fractions, reflecting the strong influence of urban primary emissions and regional air-mass transport (Järvi et al., 2009). The median concentrations of the three size fractions decreased with increasing wind speed, indicating enhanced ventilation and dilution under stronger winds, which reduce the contribution of near-field sources and lower the urban background loading (Hussein et al., 2006). On the other hand, the wide spread within each wind-speed bin shows that wind speed alone cannot explain the variability, since high concentrations occur across the full wind-speed range. This indicates that short term source variability and changes in boundary-layer mixing also influence the observed concentrations (Fig. 6 a Fig. 6d Fig. 6g)

At Hyytiälä (SMEAR II), concentrations are lower than in Helsinki, reflecting the cleaner background conditions at the boreal forest station (Hari and Kulmala, 2005; Kulmala et al., 2004). N_{acc} decreases with wind speed, whereas N_{nuc} shows a modest increase in the median. This pattern is consistent with stronger winds being associated with better ventilation and reduced accumulation-mode loading which can increase the survival probability of freshly formed clusters into the nucleation-size range (Dal Maso et al., 2005) (Fig. 6 b Fig. 6e Fig. 6h)

At Värriö (SMEAR I) concentrations are the lowest overall, consistent with the remote subarctic setting and weaker local sources (Hari and Kulmala, 2005). The medians for N_{nuc}, N_{UFP} ,and N_{acc} tend to increase with wind speed, suggesting that particle number is strongly influenced by air-mass origin and episodic events (Vehkamäki et al., 2004) (Fig. 6c Fig. 6f Fig. 6i)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

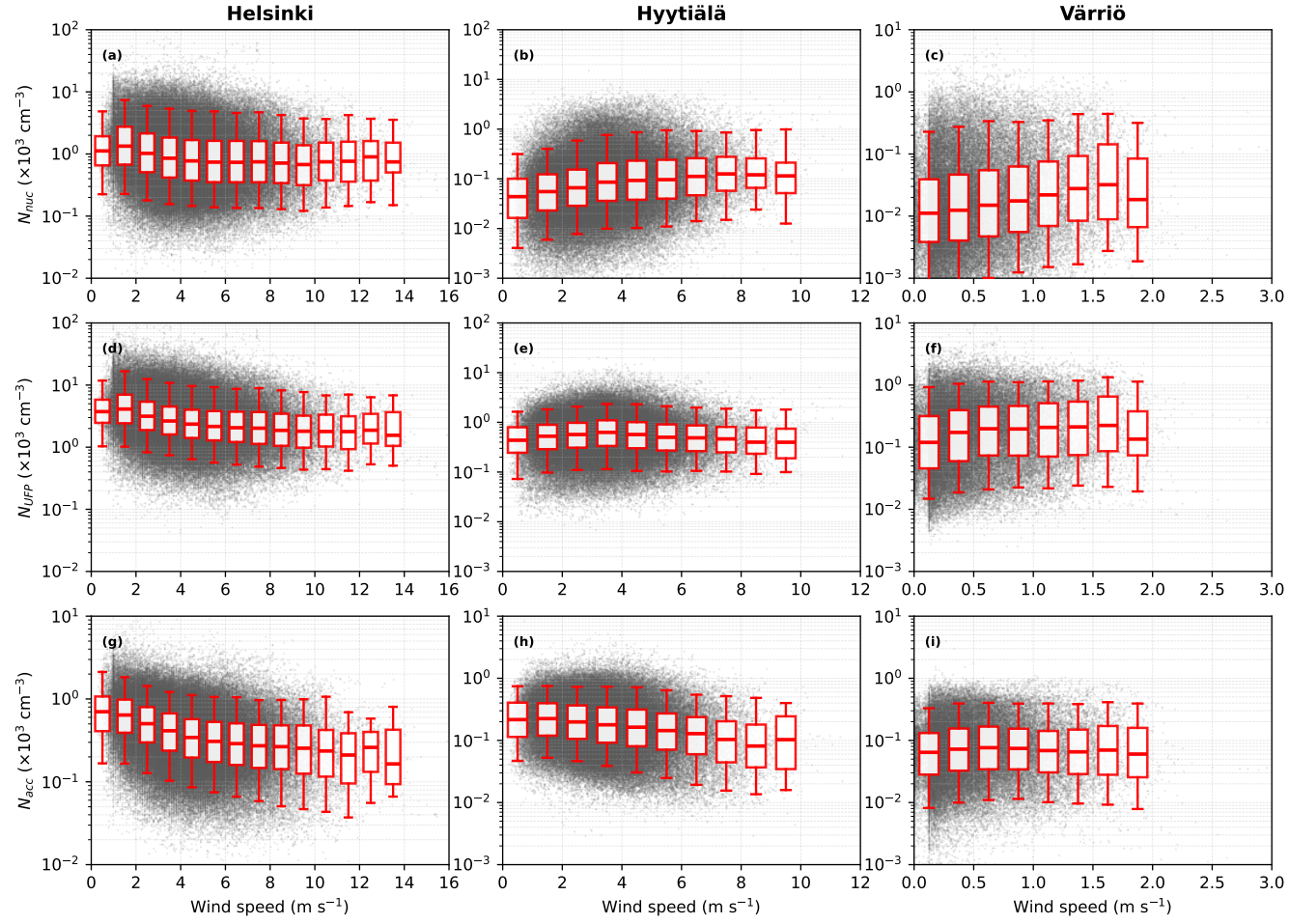


Figure 6. Wind-speed dependence of size-fractionated particle number concentrations at three SMEAR sites. Helsinki (SMEAR III), Hyytiälä (SMEAR II), and Värriö (SMEAR I). Rows show nucleation-mode N_{nuc}, ultrafine N_{UFP}, and accumulation-mode N_{acc} Scatter points are coloured by wind direction (WD) binned into 30° sectors. Red boxplots summarise the distribution of N within wind-speed bins (median and interquartile range

3.1.1.2 The effect of temperature and relative humidity

In Helsinki (SMEAR III), N_{nuc} and UFP exhibit the same trend, concentrations are high at lower temperature and decrease to 0 and then remain constant. This behaviour is consistent with enhanced cold-season primary particle production and reduced dilution under stable, shallow boundary layers in winter(Järvi et al., 2009). N_{acc} how only a weak temperature dependence with a shallow U-shaped pattern (Järvi et al., 2009; Tunved et al., 2003). This response is expected because particles with diameters >100 nm represent a mixture of regional background aerosol and urban contributions, which are less tightly coupled to instantaneous local emissions than smaller modes (Tunved et al., 2003; Okuljar et al., 2021)(Fig. 7).

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



In Hyytiälä (SMEAR II), N_{nuc} and UFP shows a constant value. This is due to concentrations occurred mainly under low to moderate relative humidity rather than under the most humid conditions, suggesting that NPF was favoured during relatively dry and photochemically active periods, whereas high humidity was associated with conditions less favourable for particle formation and survival, and also reflect a balance between NPF production, subsequent growth and removal by coagulation and wet scavenging (Kulmala et al., 2004; Kerminen et al., 2018) (Fig. 7e). N_{acc} increases strongly toward warm temperatures, indicating efficient warm-season growth into CCN-relevant sizes (Hari and Kulmala, 2005; Kerminen et al., 2018). A key driver is the temperature dependence of biogenic VOC emissions and their oxidation to low-volatility products that promote condensational growth and secondary organic aerosol formation (Ehn et al., 2014; Kulmala et al., 2004) (Fig. 7h).

In Värriö (SMEAR I) particle size fractions exhibits different trend compared to Helsinki and Hyytiälä except for N_{acc}. N_{nuc} values shows a constant trend local primary urban sources are largely absent (Kerminen et al., 2018) (Fig. 7).Relative humidity was generally high over a broad temperature range, suggesting that cleaner,drier, and more photochemically favorable air masses supported nucleation more effectively than very humid conditions. for UFP concentrations generally increases with temperature, pointing to enhanced warm-season secondary production and growth under clean conditions (Kulmala et al., 2004; Kerminen et al., 2018) forN_{acc} also increases toward warm temperatures, consistent with summer secondary aerosol formation and growth in a remote background environment ((Ehn et al., 2014; Tunved et al., 2003) (Fig. 7).

3.1.2 Particle number concentration versus gases

3.1.2.1 Variation with respect to NO_x and O_3

In Helsinki (SMEAR III), all particle size fractions exhibited an increasing trend as NO_x increased, this is due to the influence of local and regional anthropogenic sources on particle number concentrations at the sites (Järvi et al., 2009; **?**; Laakso et al., 2003), The highest-NO_x points are associated with the lowest-O_3, supporting a fresh-plume regime where NO rapidly titrates ozone ($NO + O_3 \rightarrow NO_2$) and suppresses O_3 close to sources (Seinfeld and Pandis, 2016) (Fig. 8).

In Hyytiälä (SMEAR II), N_{nuc} and UFP shows a constant median range, as NO_x increases O3 remains constant this is consistent with NPF being frequent under relatively clean, photochemically favourable conditions at a boreal forest site where low condensation sink allows newly formed clusters to survive and grow into the nucleation-mode size range(Kerminen et al., 2018; Dada et al., 2017, 2018). N_{acc}have an increasing trend as NO_x increases indicating that active photochemistry and more aged/regionally mixed air, conditions typically associated with NPF occurrence and growth. (Laakso et al., 2003; **?**) (Fig. 8).

In Värriö (SMEAR I) the all particle size fraction exhibit the same trend compared to Hyytiälä, as NO_x increases the O3 remains constant that consistent with a remote subarctic background site where local primary combustion sources are minimal and variability is dominated by air-mass history and regional-scale processes rather than direct local emissions(Hussein et al., 2009) (Fig. 8).

3.1.2.2 Variation with respect to CO and SO_2

In Helsinki (SMEAR III), all size fractions increase strongly with CO, acting as an indicator of combustion influence and polluted air-mass conditions at an urban background site,indicating that urban/regional pollution conditions, vapour availability,

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

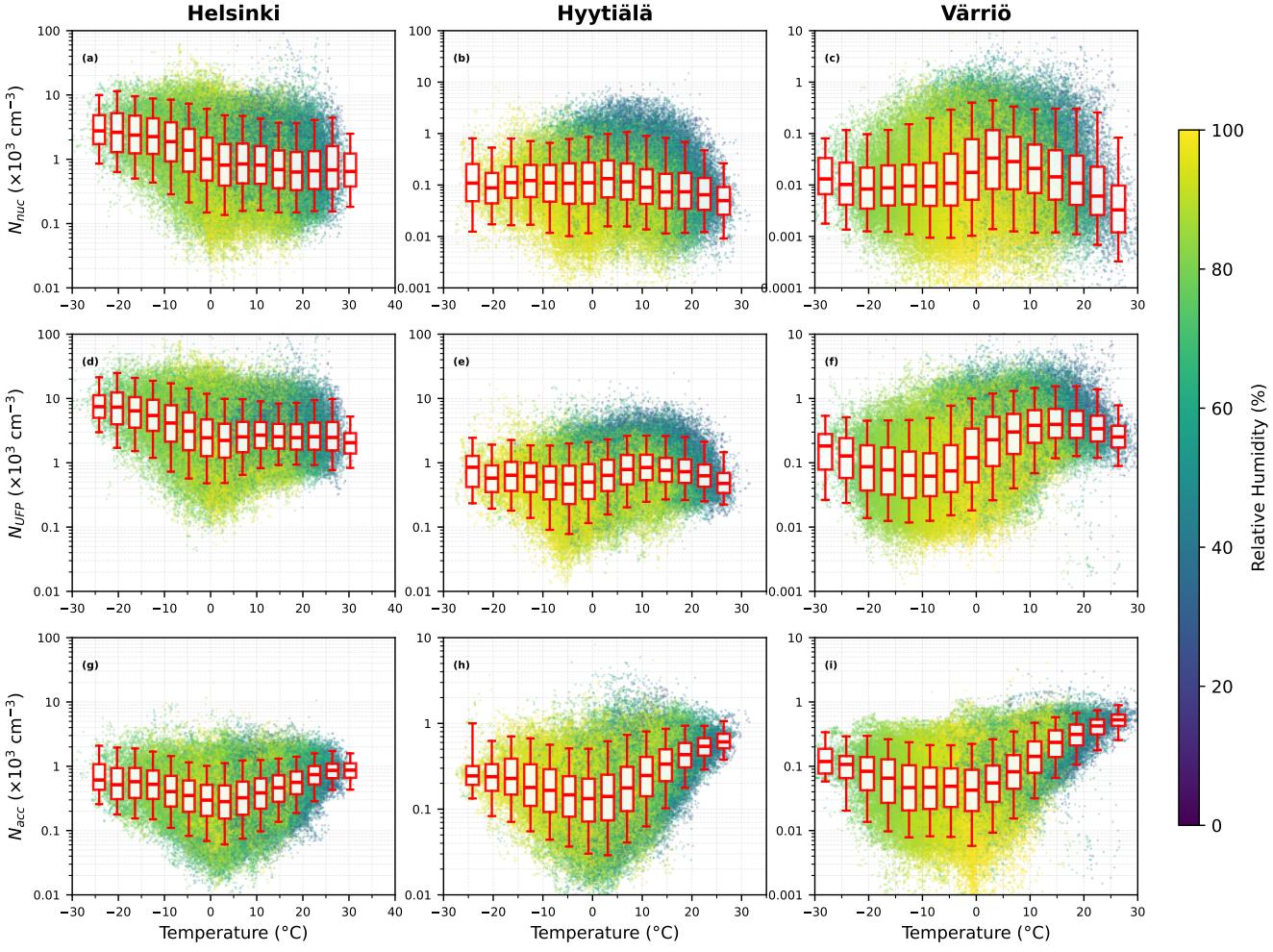


Figure 7. Temperature dependence of nucleation-mode ($10 \leq$D$_p \leq 25$ nm),UFP ($10 \leq D_p < 100$ nm), and accumulation-mode ($100 \leq D_p \leq 300$ nm) particle number concentrations in Helsinki (SMEAR III) Hyytiälä (SMEAR II) Värriö (SMEAR I) Scatter points show hourly data coloured by relative humidity (RH), and red box-and-whisker symbols indicate the median, interquartile range, and 5th–95th percentiles.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

and competing sink effects that can limit the survival of freshly formed clusters under the most polluted conditions (Kerminen et al., 2018; Järvi et al., 2009; Seinfeld and Pandis, 2016; **?**; Laakso et al., 2003). The SO_2 colouring further indicates that the highest-CO conditions are frequently accompanied by elevated SO_2 (Fig. 9).

In Hyytiälä (SMEAR II), N_{nuc}, UFP exhibits a different trend compared to Helsinki indicating that CO alone is not a strong discriminator of nucleation and UFP number concentration at this boreal forest site. This is consistent with N_{nuc} being governed mainly by precursor availability and cluster survival against the condensation and coagulation sink (Kulmala et al., 2004; Kerminen et al., 2018; Dada et al., 2017, 2018). N_{UFP} exhibits at most a weak negative tendency with increasing CO, whereas N_{acc} increases clearly with CO,consistent with transported/aged pollution enhancing the accumulation mode and strengthening the overall sink (Laakso et al., 2003; **?**) (Fig. 9).

3.2 GAM model predictions

3.2.1 Scenario I

In Scenario I, the model is applied to each year, using training data for each target year that includes all available data from the first year through the year immediately preceding the target year.

In Helsinki (SMEAR III), the model exhibits different performance for the particle size fractions. For N_{nuc}, R^2 is relatively high during the early years and declines during the later years. This is consistent with the error metrics: lower values occurred during the early years and then increased toward several later years.

For UFP, R^2 shows consistent patterns throughout all years, where it fluctuates between 0,78 and 0,83. The MAE, MSE, and RMSE show lower values and narrow ranges after the first year, while MPE is almost 0.

for N_{acc}, the R^2 shows an increasing trend across the predicted years. The MAE, MSE, and RMSE generally decline over the same period, with some individual-year fluctuations. MPE also remains comparatively small for most years.

In Hyytiälä (SMEAR II) and Värriö (SMEAR I), a similar pattern was observed across the three particle size fractions. For N_{nuc}, UFP, and N_{acc}, the R^2 improved as additional years were added to the training years, while MAE, RMSE, and MSE decreased and became more stable.

In Hyytiälä (SMEAR II), the R^2 for N_{nuc} drops below 0.4 in 1999 because there was no spring observation in the training dataset. Therefore, the model had not learned the spring pattern before predicting 1999 because there are no values for the nucleation mode in the dataset. In contrast, the high R^2 in 1998 is partly because data were only available in July and from October to December, so the model was evaluated only during those periods.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

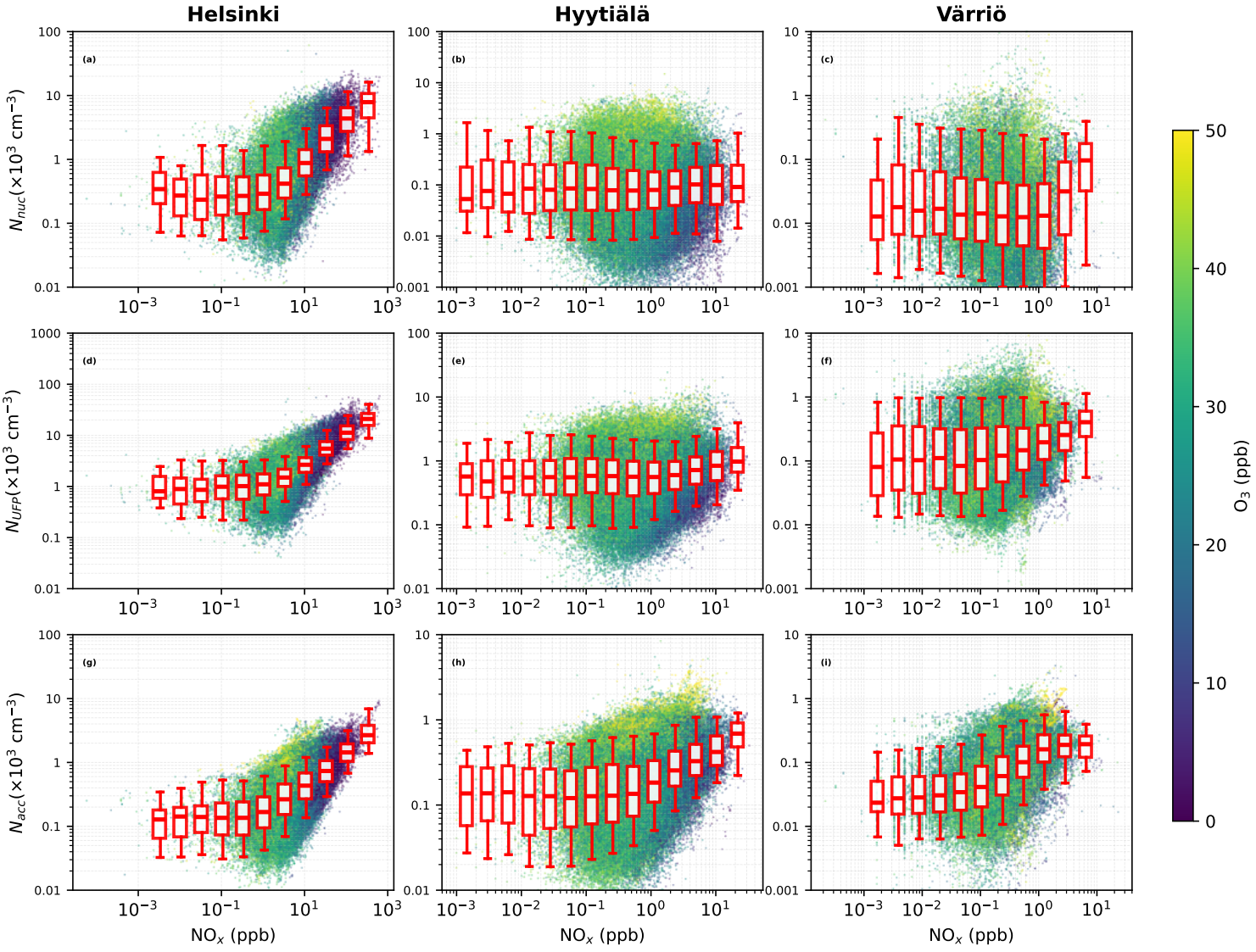


Figure 8. Scatter plot of O_3 and NO_x for nucleation-mode ($10 \leq D_p \leq 25$ nm),UFP ($10 \leq D_p < 100$ nm), and accumulation-mode ($100 \leq D_p \leq 300$ nm) particle number concentrations in Helsinki (SMEAR III), Hyytiälä (SMEAR II), and Värriö (SMEAR I). Scatter points show hourly data coloured by O_3.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

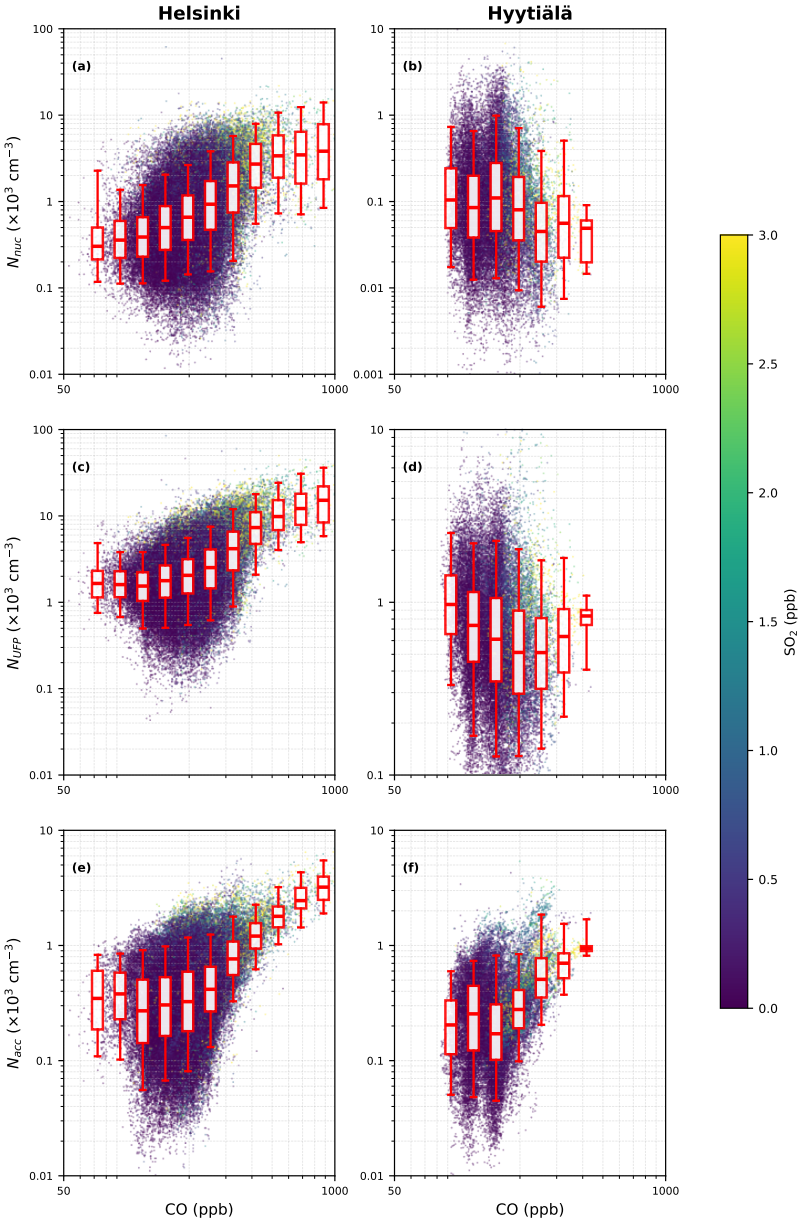


Figure 9. Scatter plots of SO_2 and CO for nucleation-mode ($10 \leq D_p \leq 25$ nm), UFP ($10 \leq D_p < 100$ nm), and accumulation-mode ($100 \leq D_p \leq 300$ nm) particle number concentrations in Helsinki (SMEAR III), Hyytiälä (SMEAR II), and Värriö (SMEAR I). Scatter points show hourly data coloured by SO_2.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

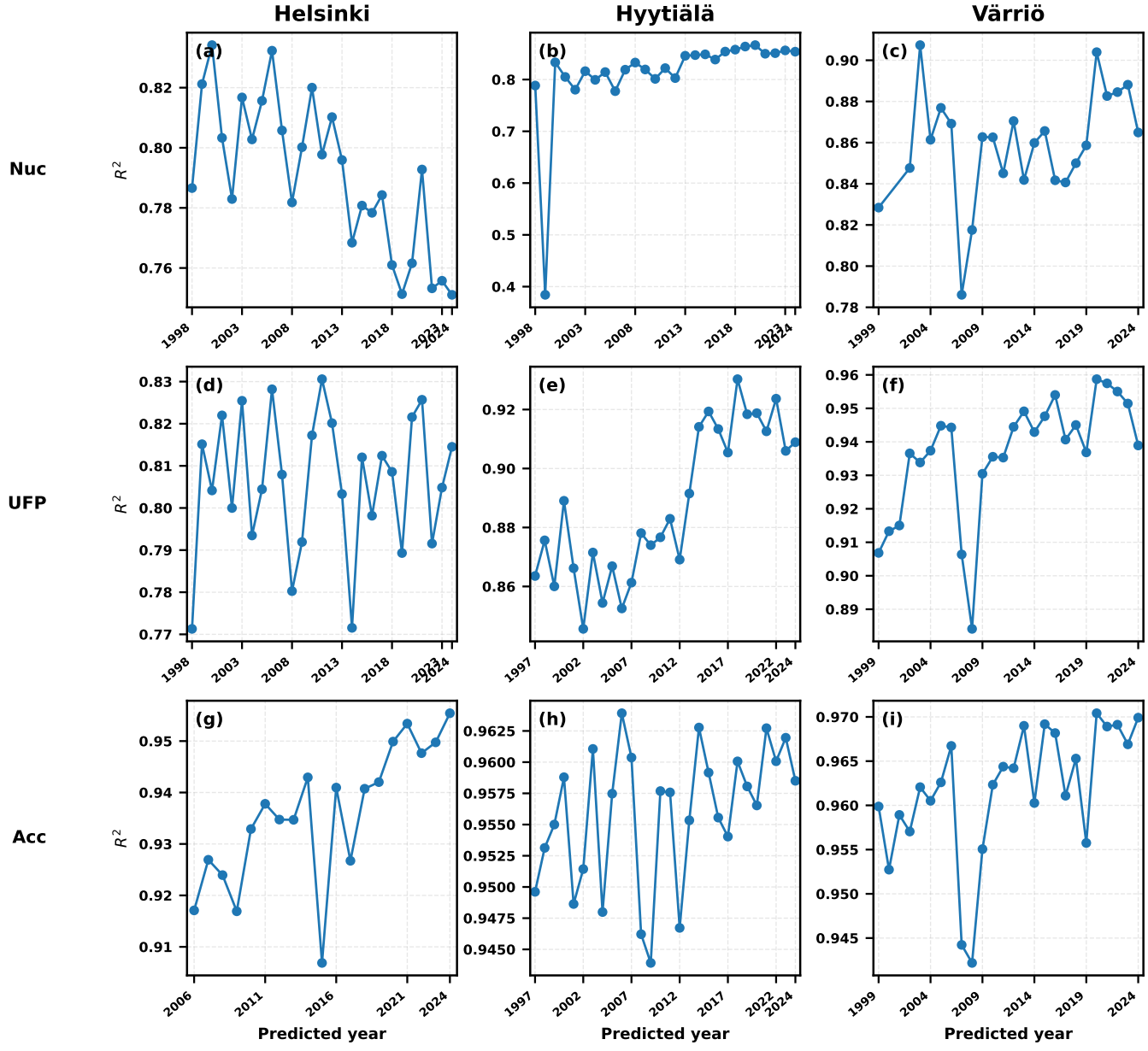


Figure 10. plot showing the R^2 for scenario I in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode ($10 \leq D_p \leq 25$ nm), (d-f) UFP ($10 \leq D_p < 100$ nm), and (g-i) accumulation ($100 \leq D_p \leq 300$ nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

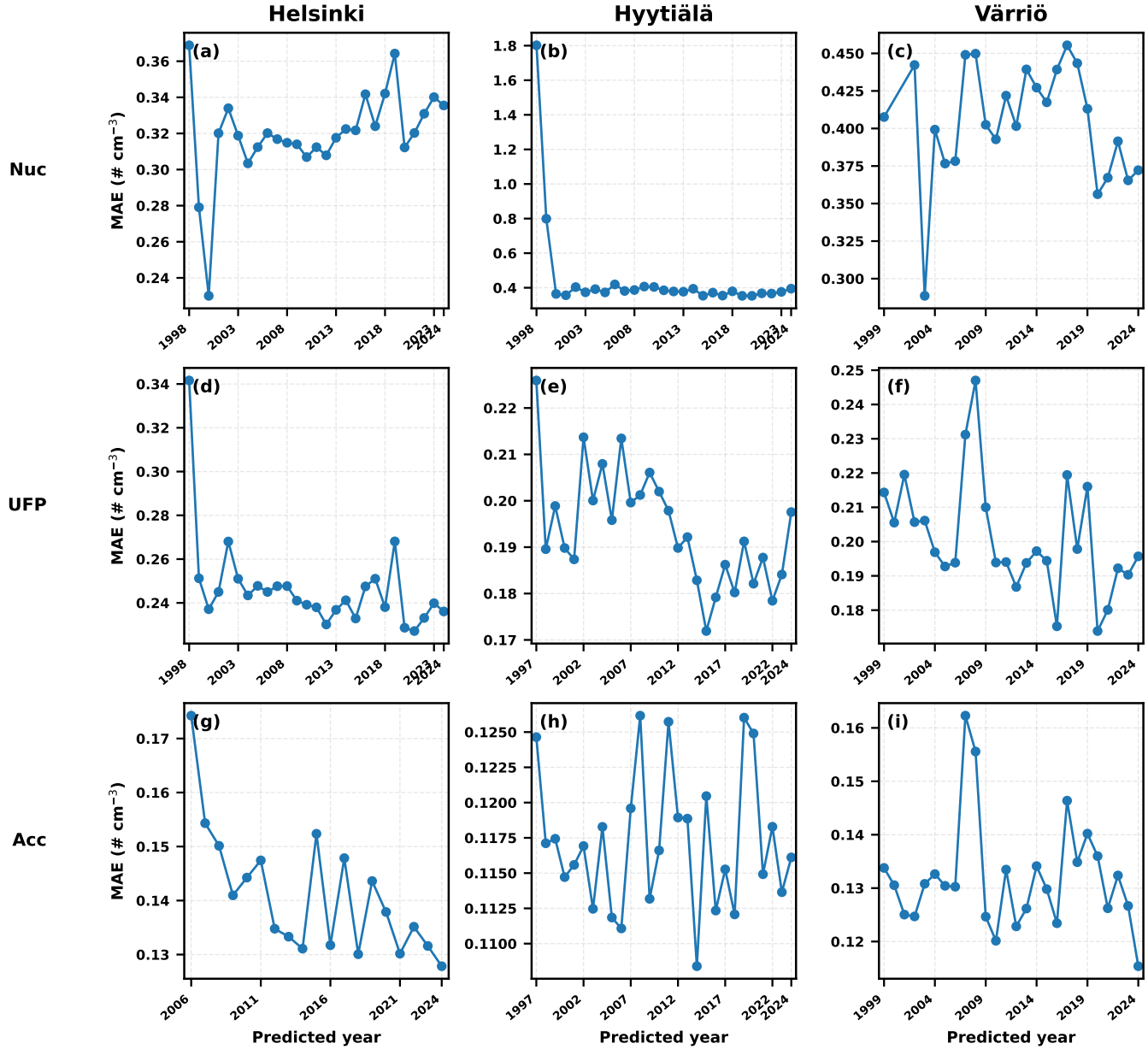


Figure 11. plot showing the Mean Absolute Error (MAE) for Scenario I in Helsinki (SMEAR III), Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode ($10 \leq D_p \leq 25$ nm), (d-f) UFP ($10 \leq D_p \leq 100$ nm), and (g-i) accumulation ($100 \leq D_p \leq 300$ nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

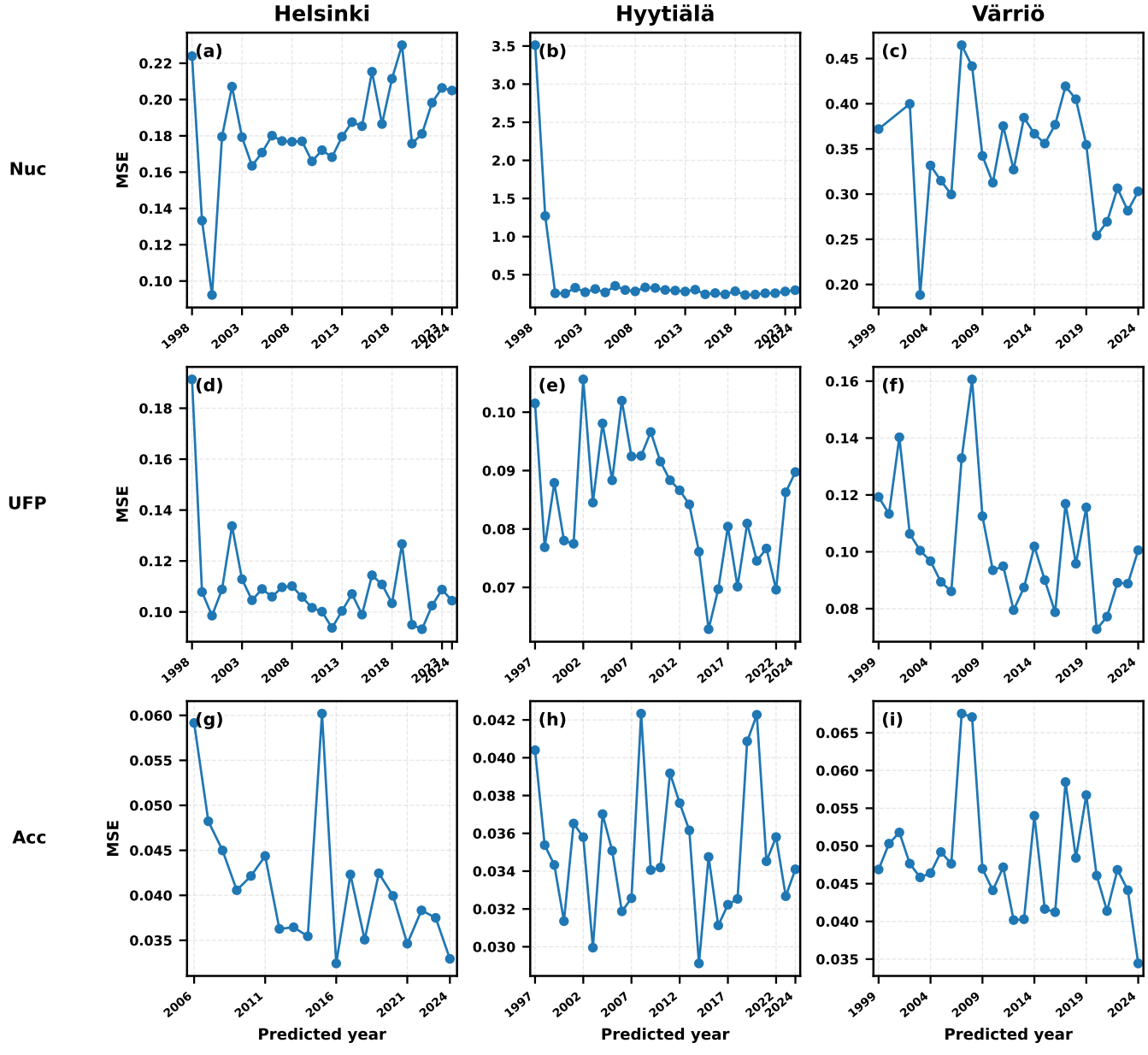


Figure 12. plot showing the Mean Squared Error (MSE) for Scenario I in Helsinki (SMEAR III) ,Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode ($10 \leq D_p \leq 25$ nm), (d-f) UFP ($10 \leq D_p \leq 100$ nm), and (g-i) accumulation ($100 \leq D_p \leq 300$ nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

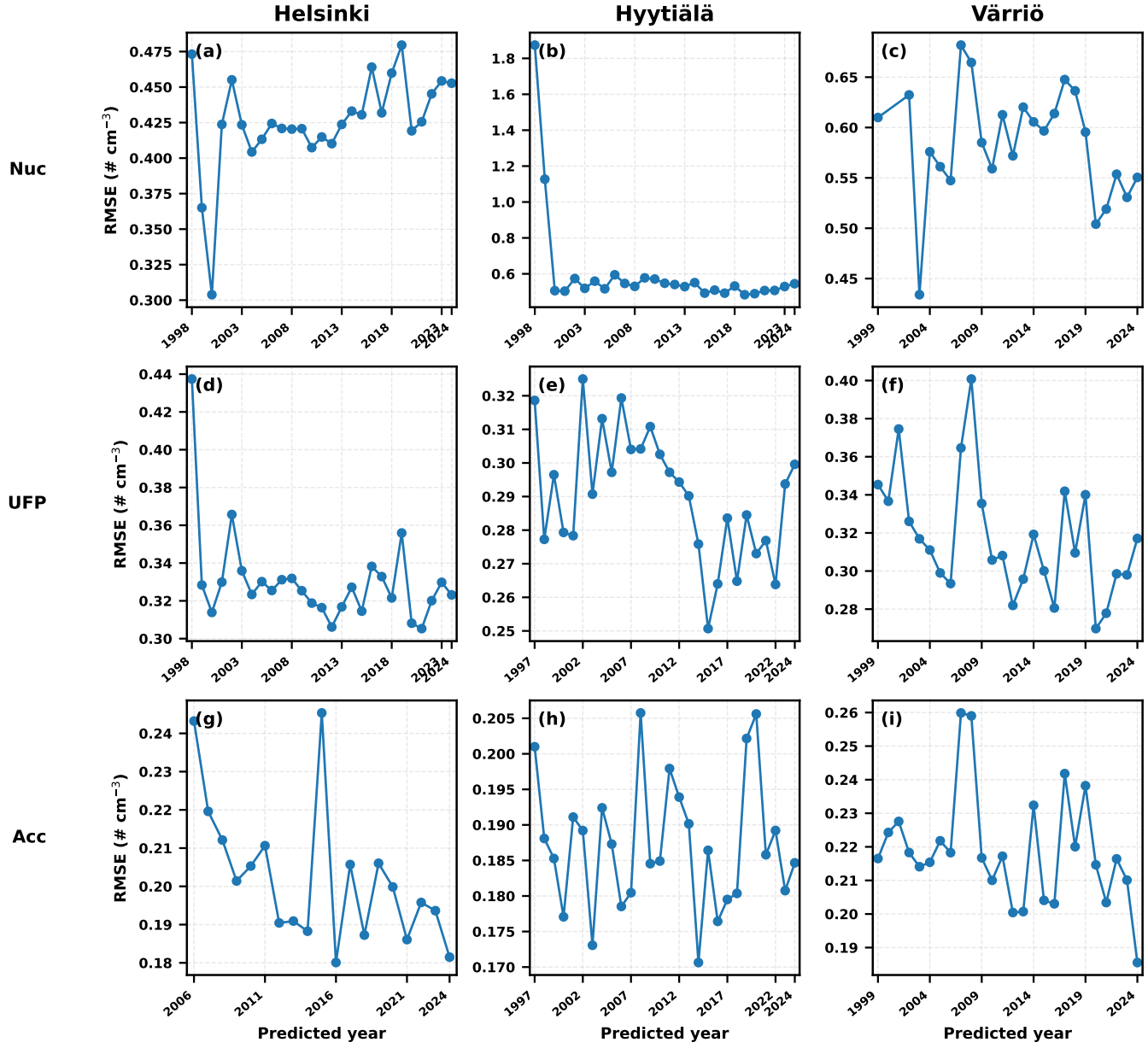


Figure 13. plot showing the Root Mean Squared Error (RMSE) for Scenario I in Helsinki (SMEAR III) ,Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode ($10 \leq D_p \leq 25$ nm), (d-f) UFP ($10 \leq D_p \leq 100$ nm), and (g-i) accumulation ($100 \leq D_p \leq 300$ nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

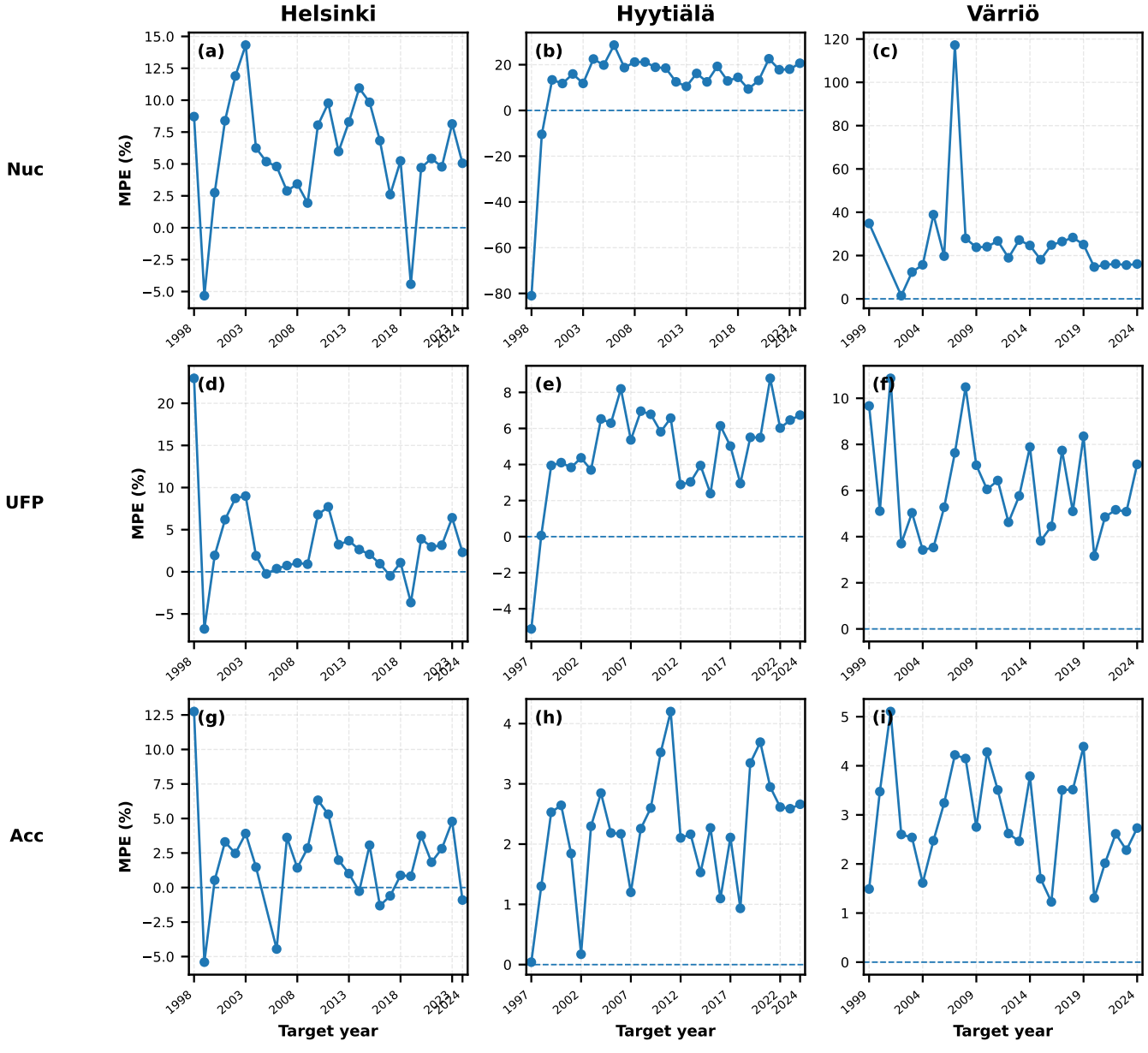


Figure 14. plot showing the Mean Percentage Error (MPE) for Scenario I in Helsinki (SMEAR III) ,Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p \leq$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



3.2.2 Scenario II:

Scenario II is a single-year training in which the model is trained separately on each year and then used to predict every year in the dataset. The heatmaps show clear differences in predictive performance among particle-size fractions and measurement sites.

Across Helsinki (SMEAR III), Hyytiälä (SMEAR II), and Värriö (SMEAR I), the R^2 values are generally moderate to high for N_{nuc}, UFP, and N_{acc}, with N_{acc} showing the highest and most consistent performance. Several vertical bands of lower R^2 are observed for specific training years, indicating that some individual years are less transferable to other target years. In Helsinki, this is particularly evident for 2000 and 2024, which are affected by substantial missing data for the particle-size fractions. The MAE, RMSE, MSE, and MPE heatmaps show a similar overall pattern, with lower prediction errors for well-represented training years and higher errors for years with limited data availability.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

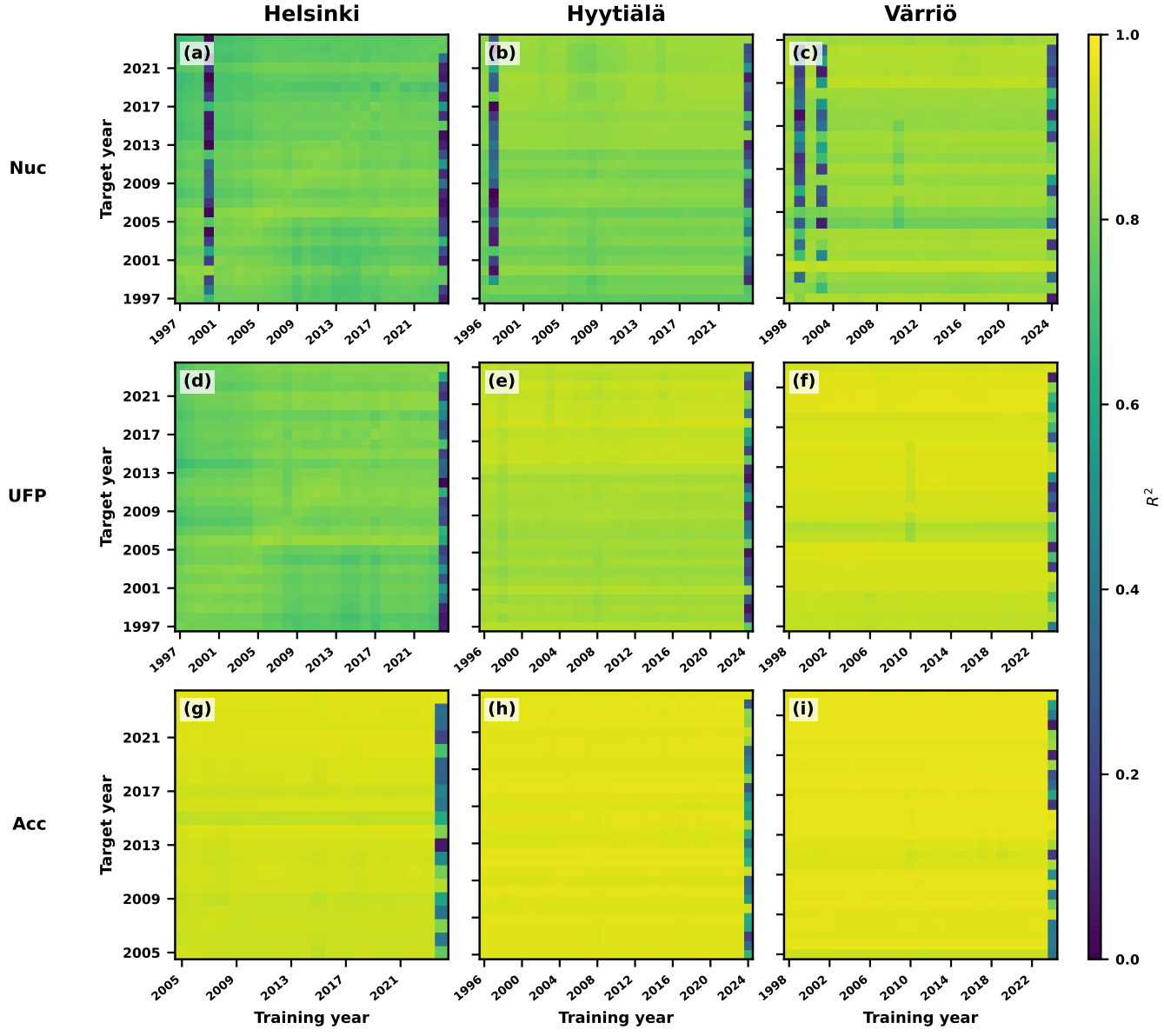


Figure 15. plot showing the R^2 for Scenario II in Helsinki (SMEAR III) ,Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

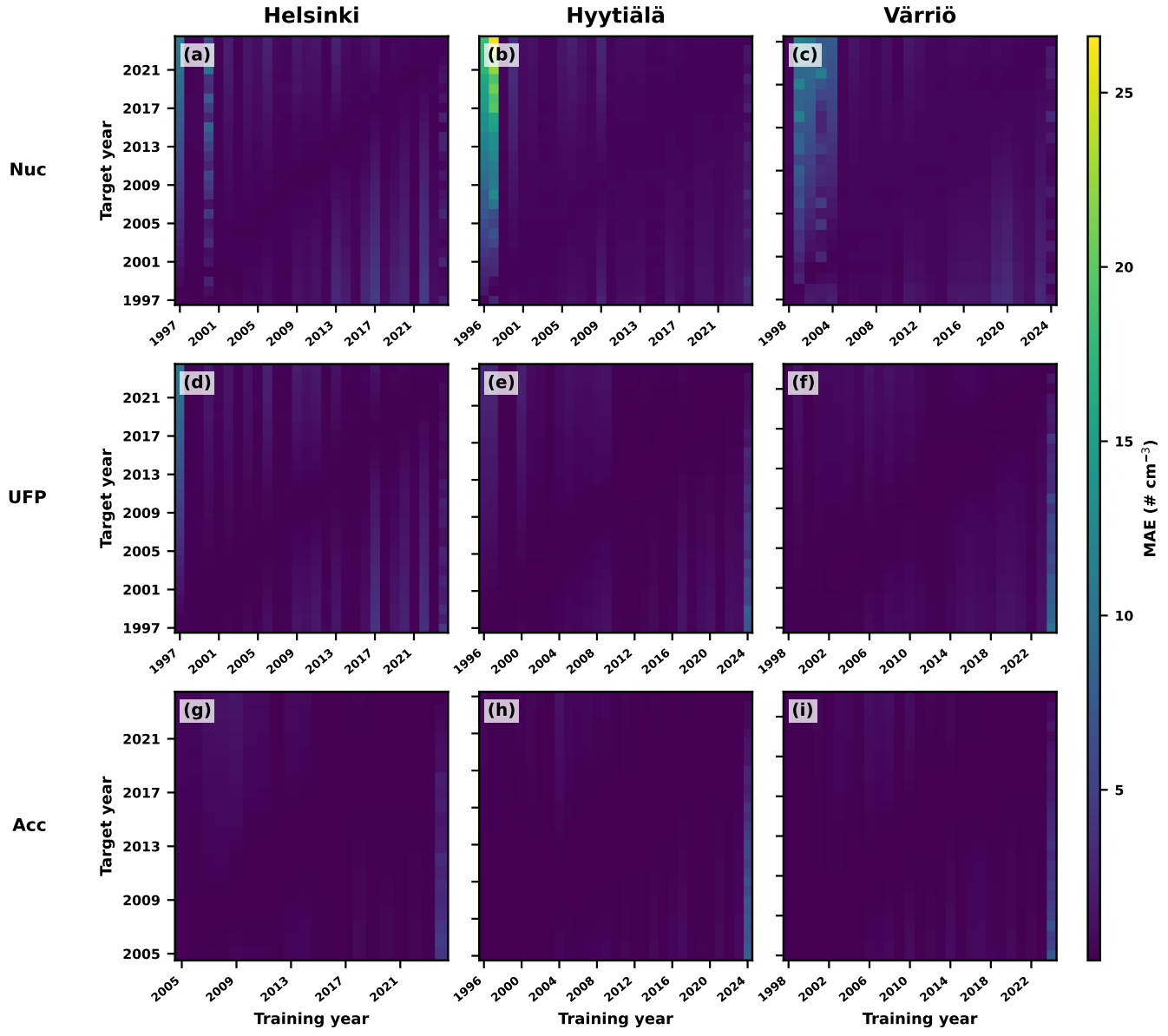


Figure 16. plot showing the Mean Absolute Error (MAE) for Scenario II in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

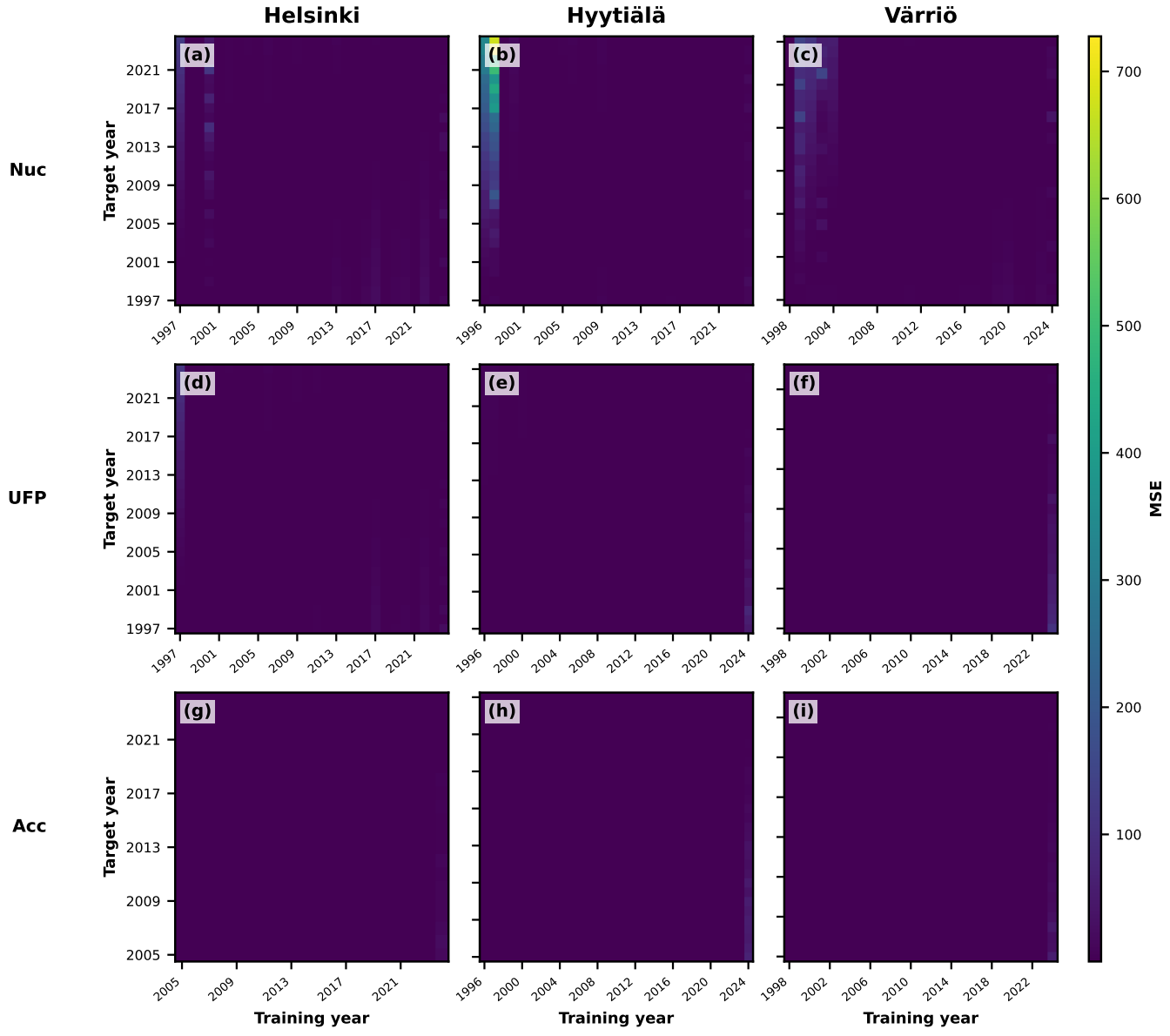


Figure 17. plot showing the Mean Squared Error (MSE) for Scenario II Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode ($10 \leq D_p \leq 25$ nm), (d-f) UFP ($10 \leq D_p < 100$ nm), and (g-i) accumulation ($100 \leq D_p \leq 300$ nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

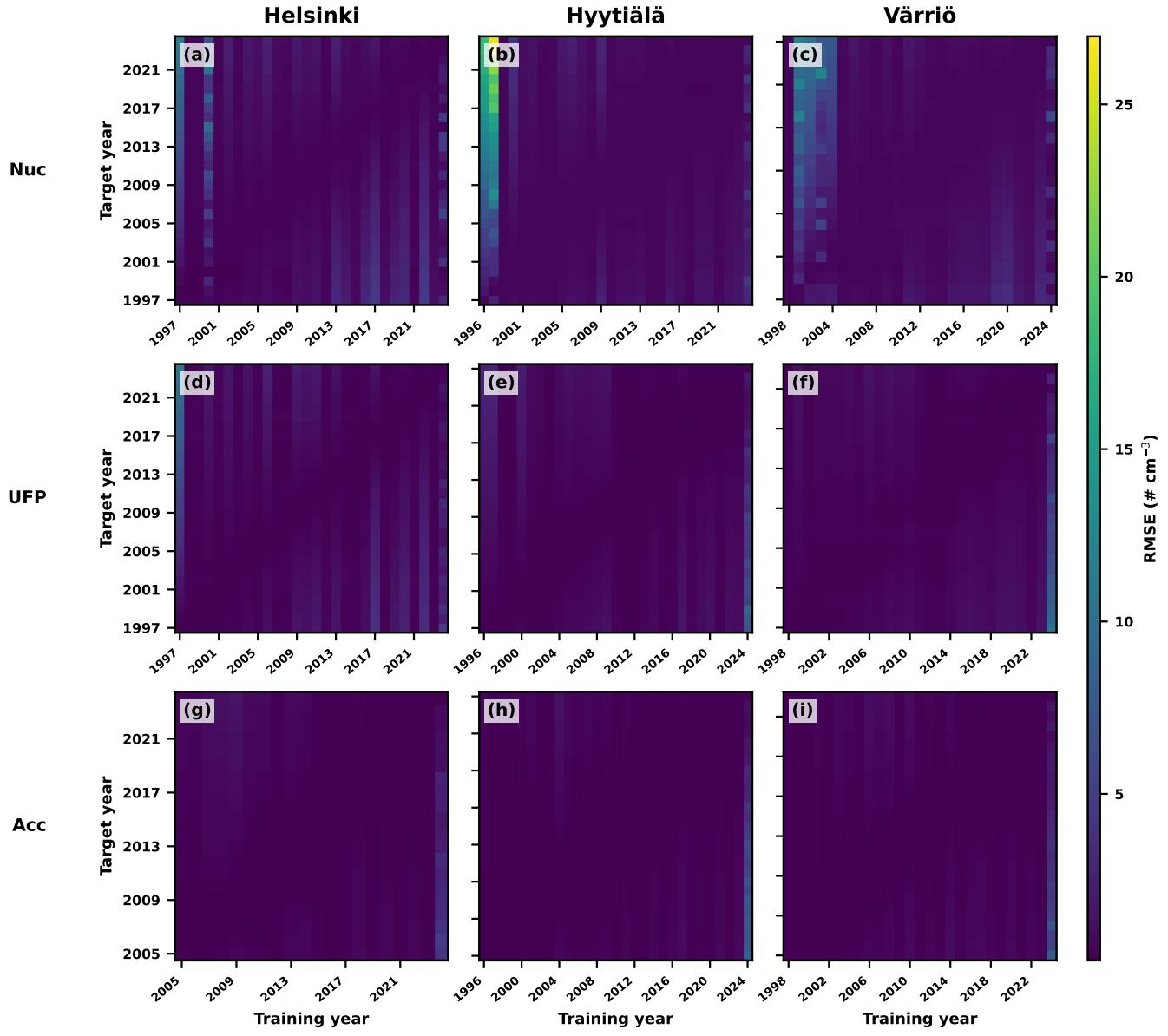


Figure 18. plot showing the Root Mean Squared Error (RMSE) for Scenario II in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

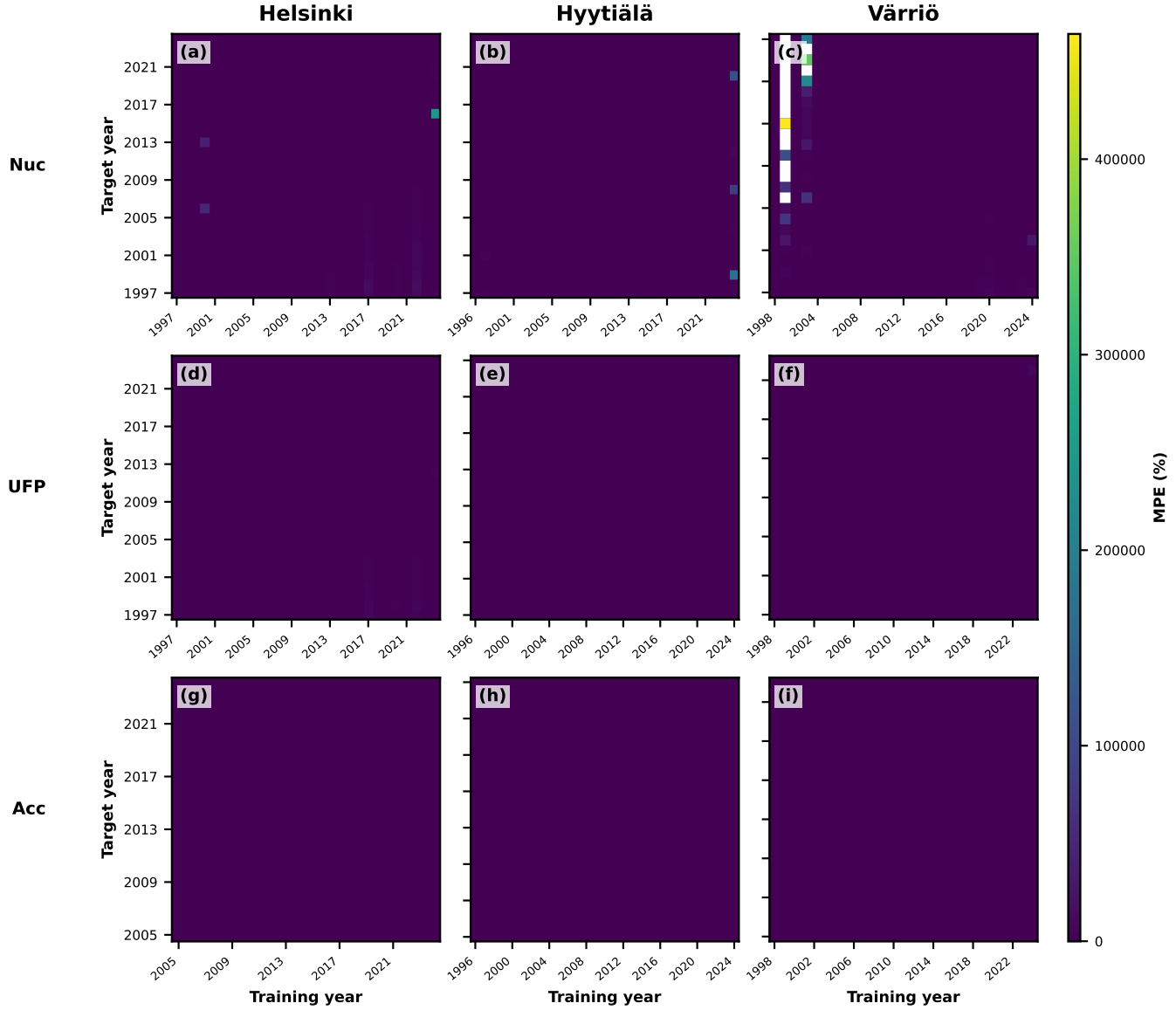


Figure 19. plot showing the Mean Percentage Error (MPE) for Scenario II in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode ($10 \leq D_p \leq 25$ nm), (d-f) UFP ($10 \leq D_p < 100$ nm), and (g-i) accumulation ($100 \leq D_p \leq 300$ nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

3.2.3 Scenario III

The main aim of Scenario III is to address whether adding older observations improves prediction once the recent years before the target are already included. In Helsinki (SMEAR III), for N_{nuc} and UFP, theR^2 gradually decreases as older years are added to the training period. The MAE, RMSE, and MSE generally remain stable for many of these target years rather than decreasing. N_{acc} is less sensitive to the length of the training period, where R^2 remains high across all training years. MAE, MSE, and RMSE are the same as older years are added to the training set. additionally MPE remains very small several few percentage from the 0.

In Hyytiälä (SMEAR II), for N_{nuc},theR^2 is constant across the training years, meaning that increasing old historical data does not influence the model prediction; additionally, the difference in R^2 is not significant.

For UFP, the different target years have clearly different levels of R^2, with later target years generally showing higher predictive performance , but individual curves remain mostly stable as older historical years are added. MAE and RMSE likewise show relatively small changes; MPE shows an increase in percentage as training years increase.

for N_{acc}, the R^2 shows a very high values across the training years. MAE, RMSE, and MSE change only slightly, and MPE is approximately close to 0.

In Värriö (SMEAR I), results show particularly clear stability across training-window lengths. For N_{nuc}, individual target years have different R^2, but within each target year the curves are mostly flat as additional training years are included. MAE, RMSE, and MSE also change relatively little with training-window length. This means that differences in predictive performance are associated with the target year itself rather than with the number of historical years used for training.

For UFP. R^2 values differ among target years but remain relatively constant within each line as the training period becomes longer. The error metrics similarly show only minor changes.

N_{acc} particles show the greatest stability. R^2 remains consistently very high across almost all training lengths, generally around 0.95–0.97, while MAE, RMSE, and MSE remain low and nearly constant.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

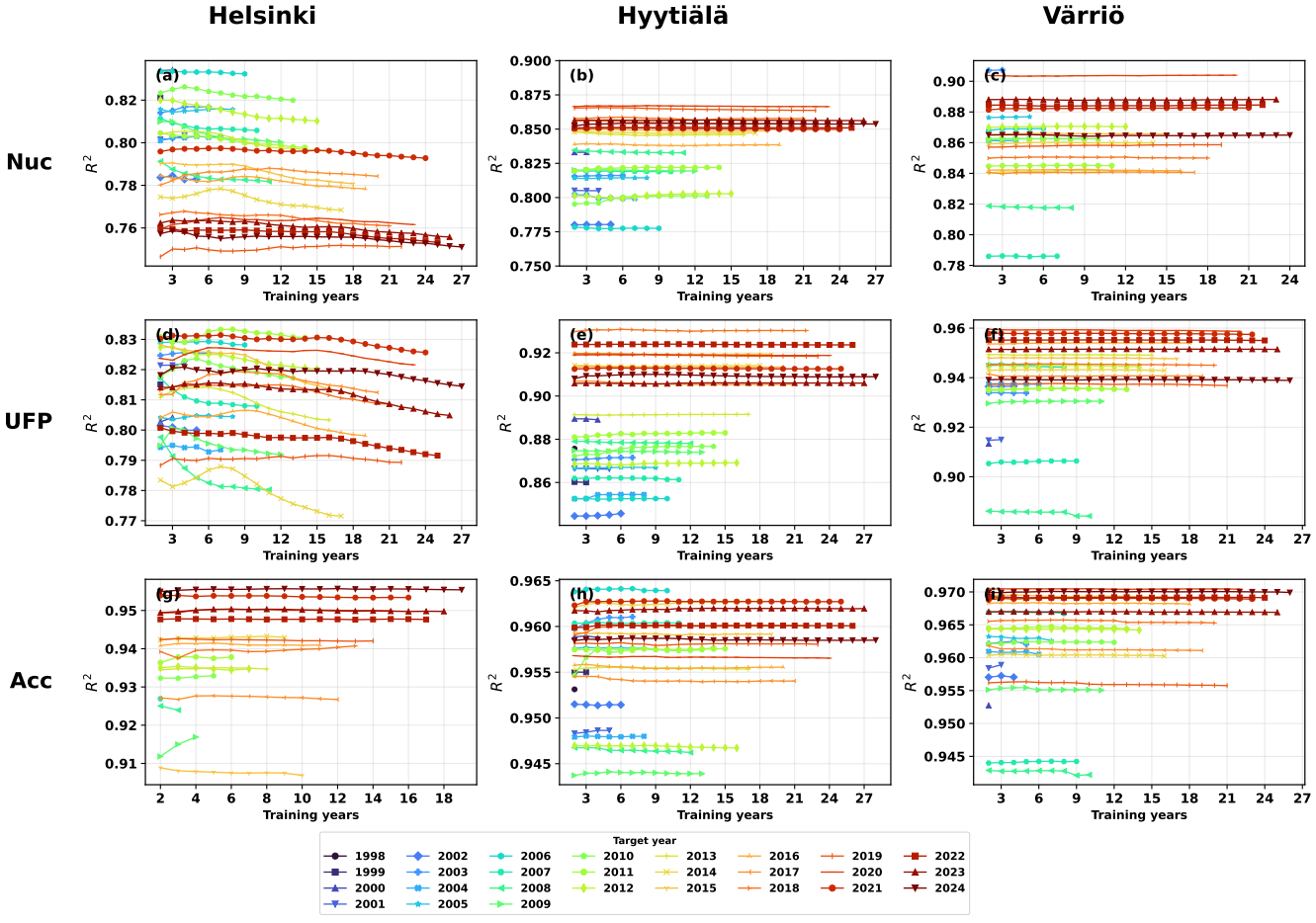


Figure 20. plot showing the R^2 for Scenario III Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode ($10 \leq$D$_p <\leq 25$ nm), (d-f) UFP ($10 \leq D_p < 100$ nm), and (g-i) accumulation ($100 \leq D_p \leq 300$ nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

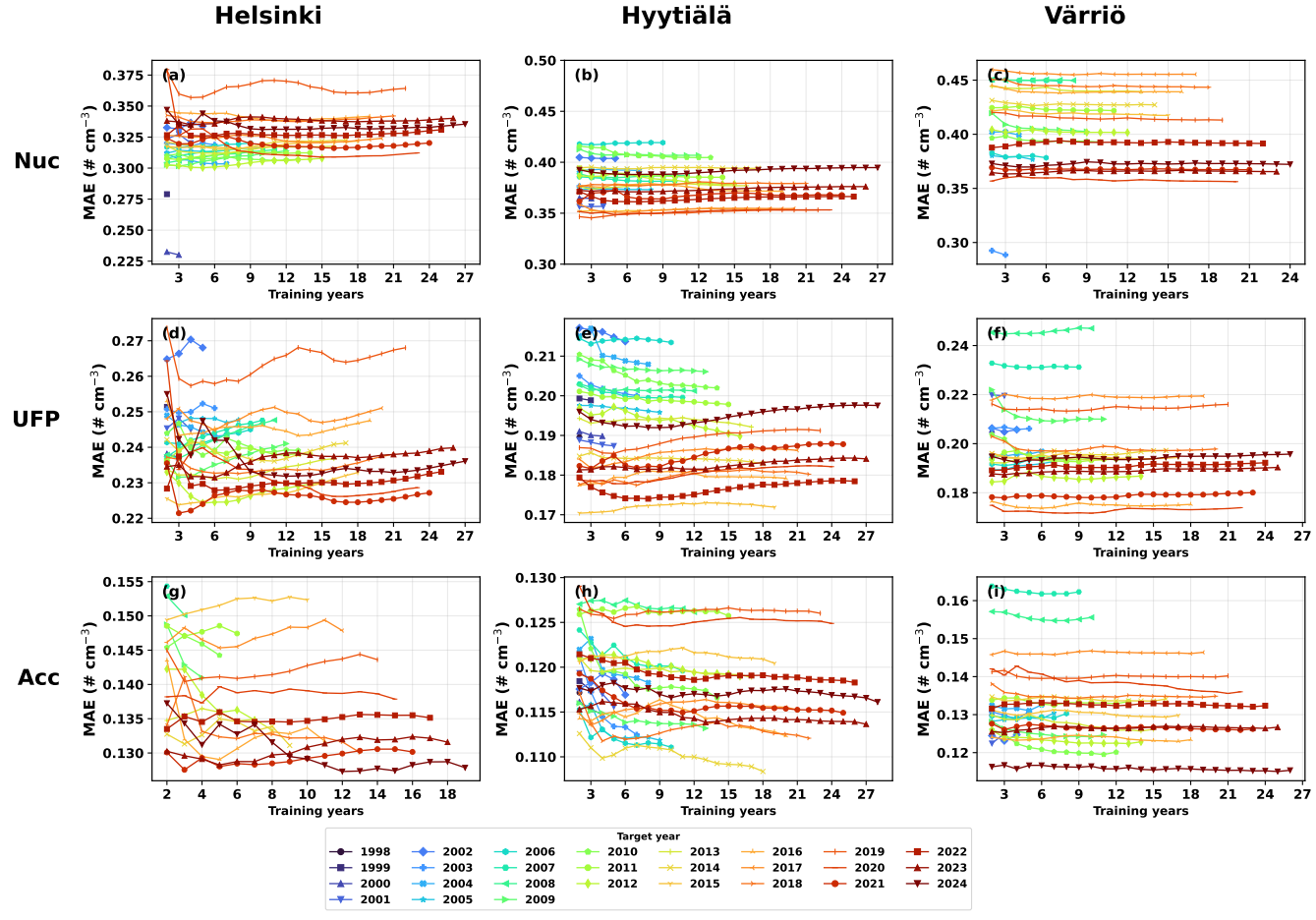


Figure 21. plot showing the Mean Absolute Error (MAE) for Scenario III in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

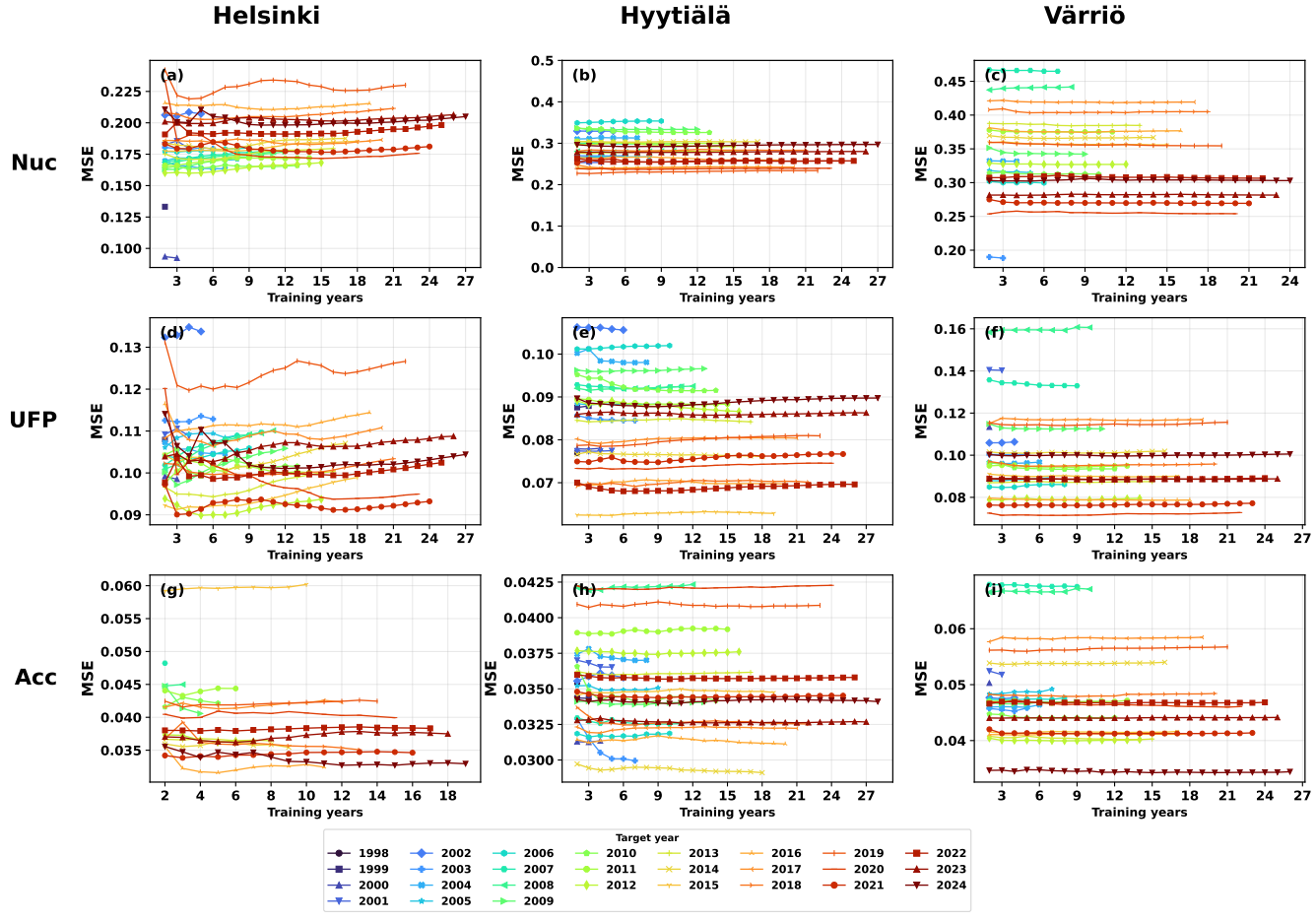


Figure 22. plot showing the Mean Squared Error (MSE) for Scenario III in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

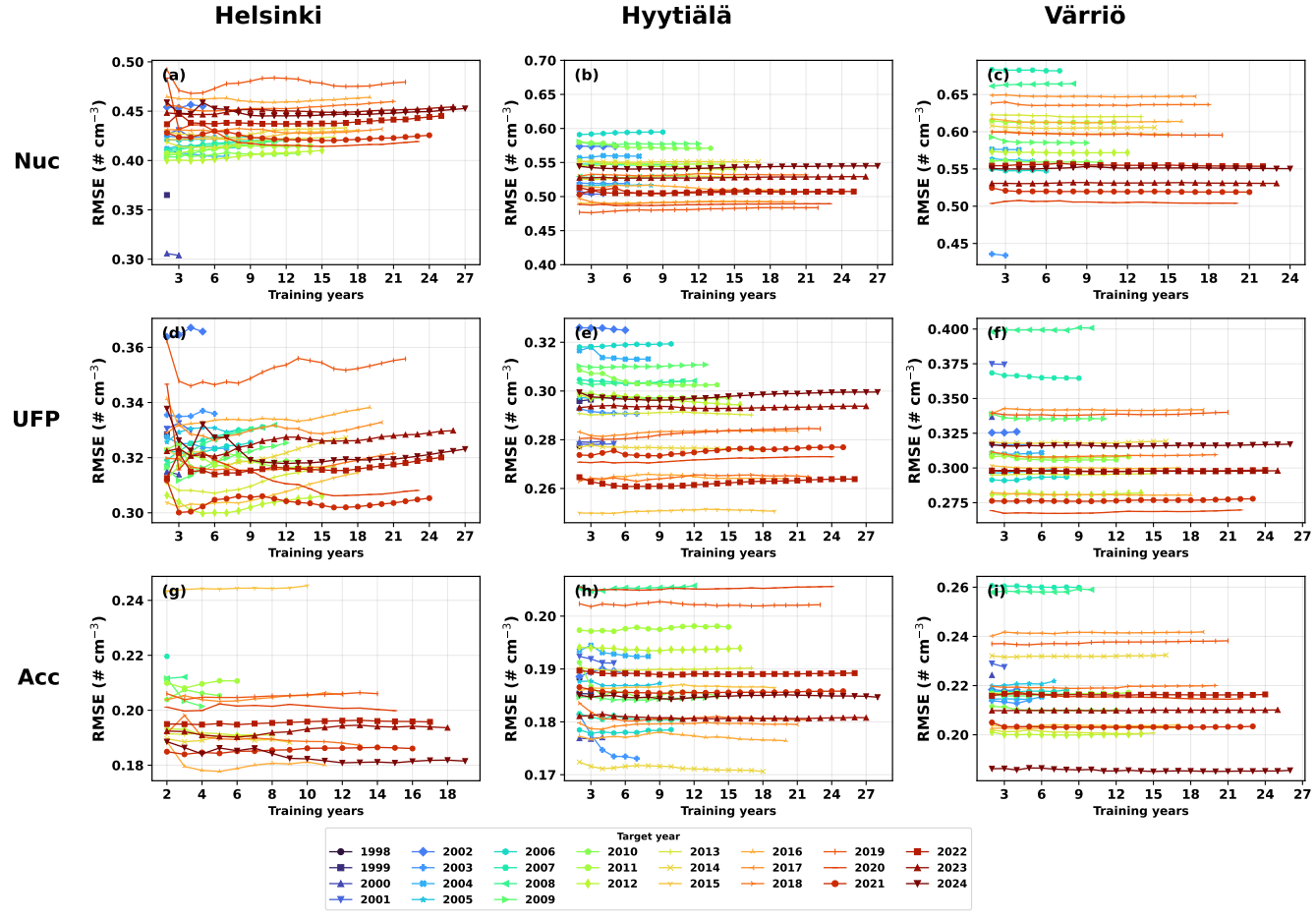


Figure 23. plot showing the Root Mean Squared Error (RMSE) for Scenario III Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

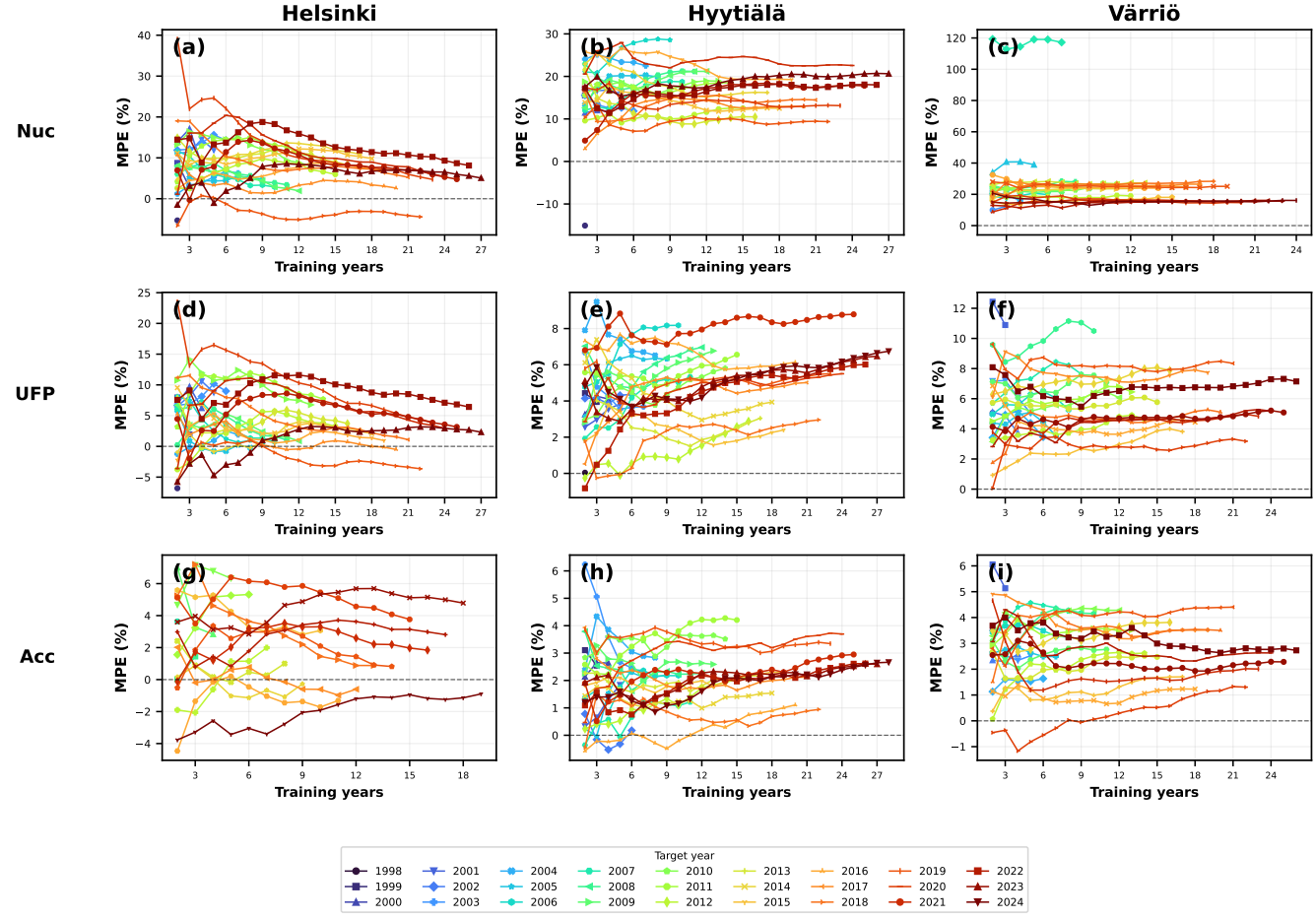


Figure 24. plot showing the Mean Percentage Error (MPE) for Scenario III in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode ($10 \leq D_p <\leq 25$ nm), (d-f) UFP ($10 \leq D_p < 100$ nm), and (g-i) accumulation ($100 \leq D_p \leq 300$ nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

3.2.4 Scenario IV

In Scenario IV, the starting year is fixed at the first available year, while the end of the training period is progressively extended toward the target year. The main goal of Scenario IV is to address how prediction improves when a model initially trained only on early historical data is progressively updated with more recent observations approaching the target year.

In Helsinki (SMEAR III), for N_{nuc}, R^2 generally increases as the training period expands, with the greatest improvement occurring during the shorter training windows before performance begins to stabilize. UFP shows an even clearer increasing tendency: R^2 rises substantially as more years are added and then approaches a relatively stable level for longer training periods.

Accumulation-mode particles also benefit from the expanding training period. R^2 increases rapidly during the first few training years and subsequently changes only slightly, while MAE and RMSE decrease markedly before reaching relatively stable values.

MPE results show that very short training periods also produce large systematic biases, particularly for nucleation-mode particles and UFP; as the training period expands, these large biases are substantially reduced, and the MPE generally moves closer to zero.

In Hyytiälä (SMEAR II),N_{nuc}, UFP, the earlier short training windows show lower R^2, particularly for several earlier target years, whereas adding more observations generally increases R. Later target years remain around or above approximately 0.90 after the model contains sufficient historical information. For N_{acc} accumulation-mode particles, performance is already comparatively high, but there is still a clear initial improvement as the training window expands. R^2 rises during the first few years and subsequently becomes relatively stable at approximately 0.95–0.96.

In Hyytiälä (SMEAR II), the R^2 changes among target years even when the same two-year training window is used. This is because each target year has different meteorological conditions, gas concentrations, seasonal patterns, and particle behavior. The autoregressive terms are also important, since the relationship between the current and previous-hour N_{nuc} concentrations changes from year to year. Years with stronger temporal persistence generally have higher R^2, while years with weaker persistence are harder to predict.

In Värriö (SMEAR I), For nucleation-mode particles and for UFP Accumulation-mode particles s, R^2 generally increases slightly or remains stable as the training period expands. The error metrics show larger changes during the shortest training periods, followed by substantially more stable values once several years have been added.

In Värriö (SMEAR I), for nucleation-mode particles and for UFP Accumulation-mode particles s, R^2 generally increases slightly or remains stable as the training period expands. The error metrics show larger changes during the shortest training periods, followed by substantially more stable values once several years have been added.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

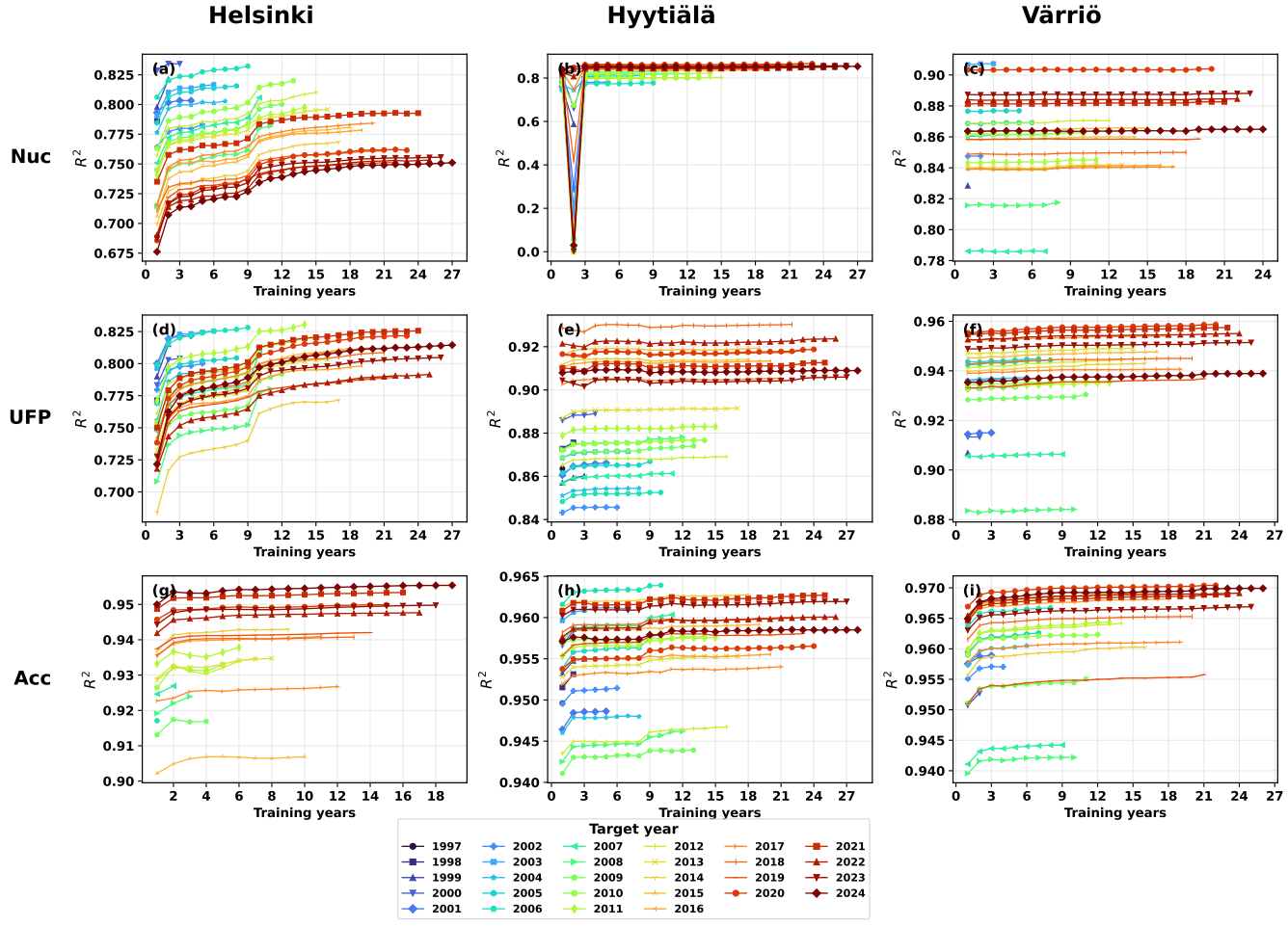


Figure 25. plot showing the R^2 for Scenario IV in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

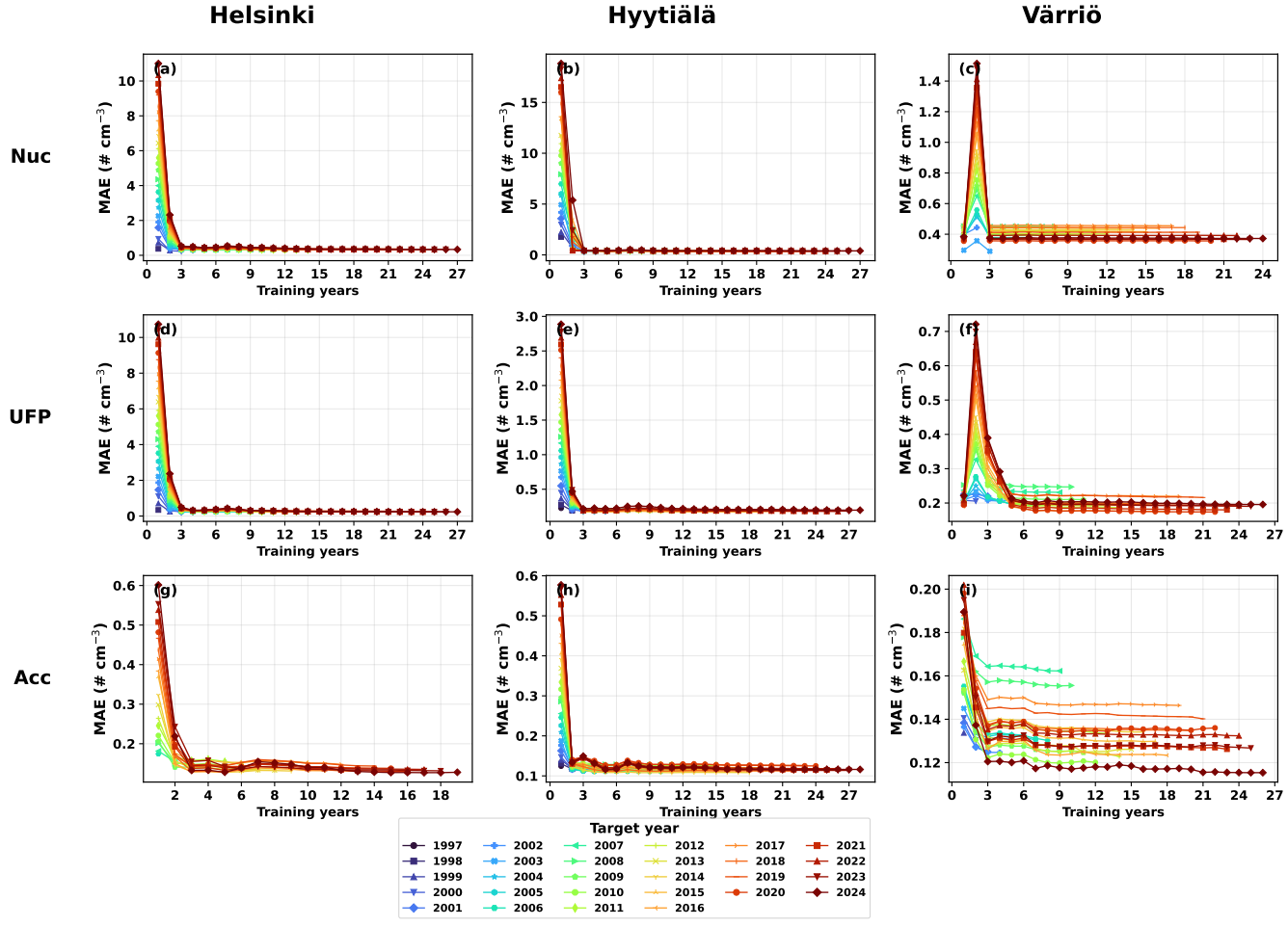


Figure 26. plot showing the Mean Absolute Error (MAE) for Scenario IV in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

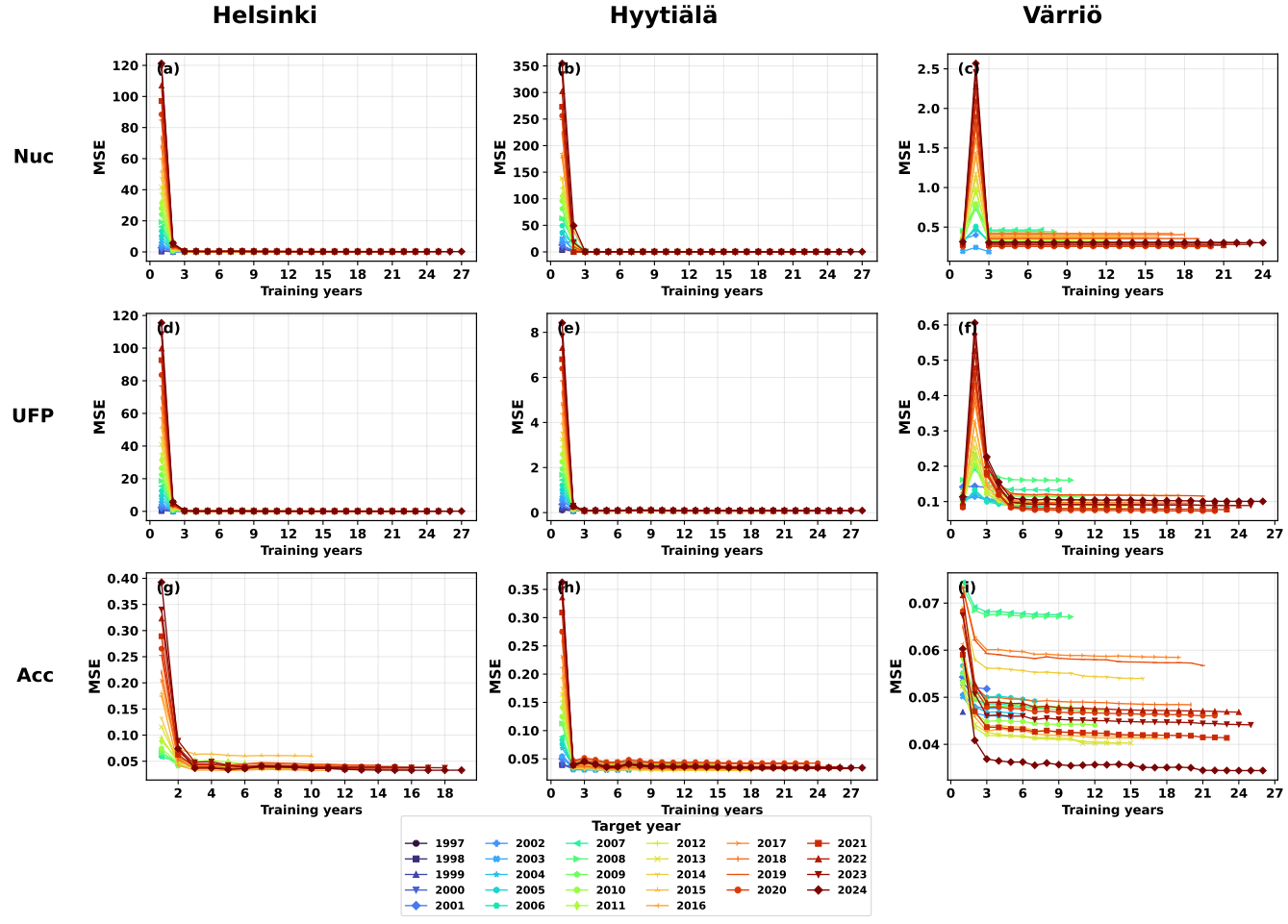


Figure 27. plot showing the Mean Squared Error (MSE) for Scenario IV in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode ($10 \leq D_p <\leq 25$ nm), (d-f) UFP ($10 \leq D_p < 100$ nm), and (g-i) accumulation ($100 \leq D_p \leq 300$ nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

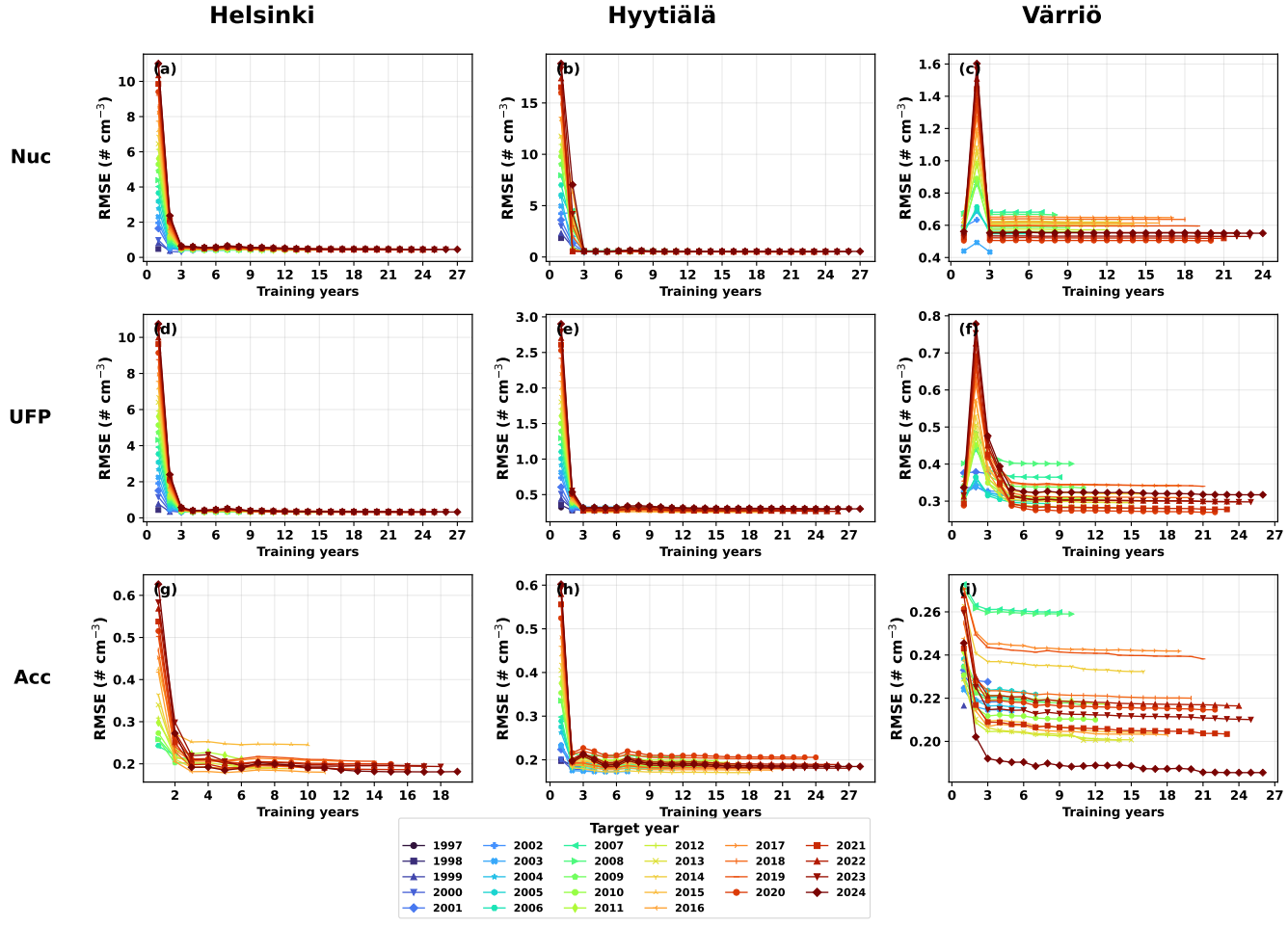


Figure 28. plot showing the Root Mean Squared Error (RMSE) for Scenario IV Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

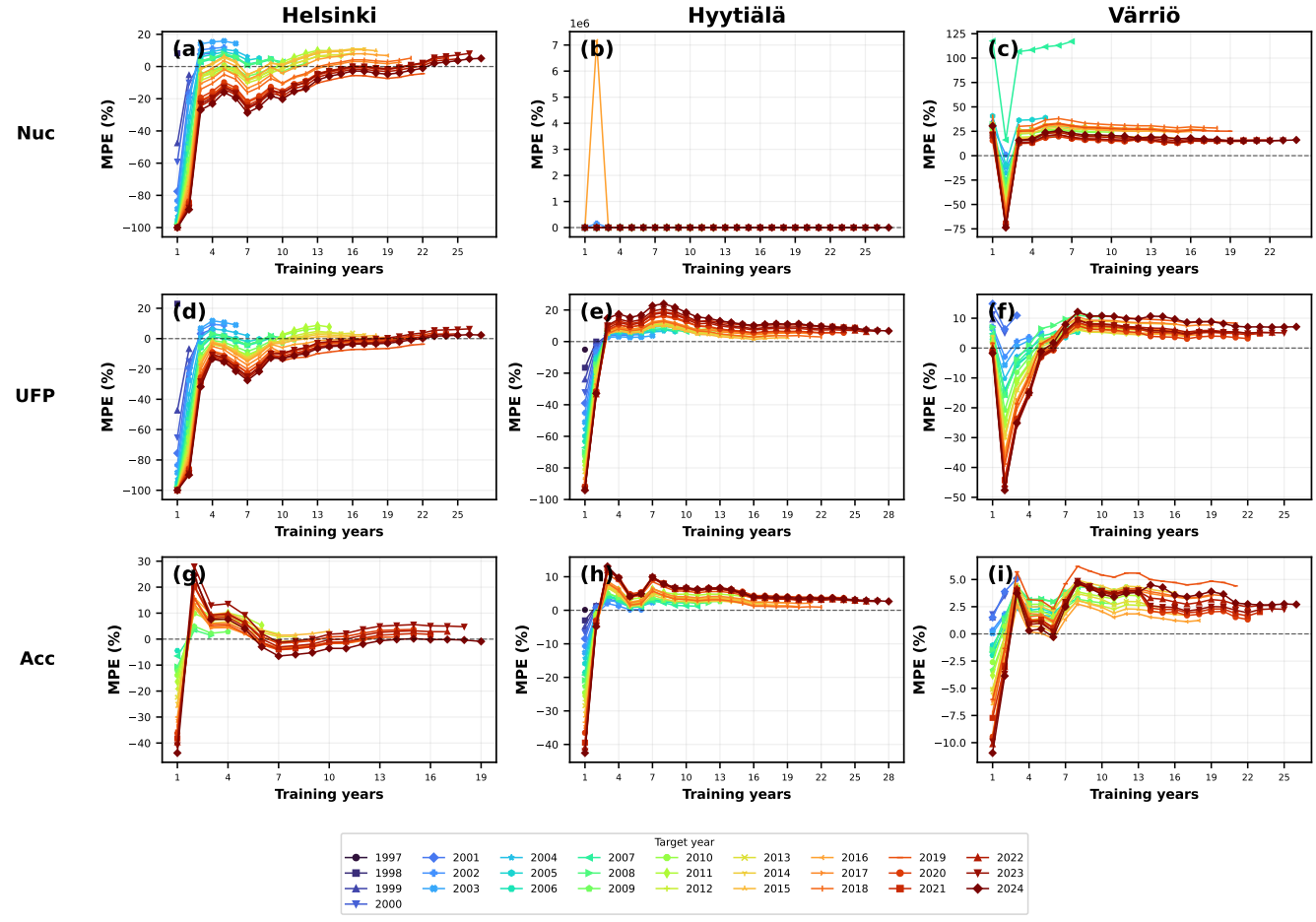


Figure 29. plot showing the Mean Percentage Error (MPE) for Scenario IV Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode ($10 \leq D_p <\leq 25$ nm), (d-f) UFP ($10 \leq D_p < 100$ nm), and (g-i) accumulation ($100 \leq D_p \leq 300$ nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



3.2.5 Scenario V

The aim of scenario v is to show how predictive performance changes as the length of a consecutive training window increases, when the training window can occur anywhere in the historical record before the target year.

In Helsinki (SMEAR III), for N_{nuc}, UFP, and N_{acc}, the R^2 results generally show the complementary pattern: for many target years, R^2 increases as the window length increases, although the magnitude of improvement differs substantially among target years. MAE, RMSE, MSE, and MPE decrease markedly as the training period increases from the shortest windows, with the largest reductions occurring within approximately the first few training years.

In Hyytiälä (SMEAR II),N_{nuc}, UFP, and N_{acc} show a particularly pronounced effect for some early target years. Very short training windows can produce exceptionally large prediction errors and low R^2, whereas performance improves sharply when the training window is extended. For UFP, the effect is more gradual. MAE, RMSE, and MSE generally decrease with increasing training-window length across most target years. R^2, however, changes relatively little for many of the target years compared with the substantial reductions observed in the error metrics. This is important: a longer window can improve the absolute accuracy of the predictions even when the proportion of explained variability changes only slightly.

In Värriö (SMEAR I), the results are somewhat different because R^2 is already comparatively high and stable for many particle-size fractions. For nucleation-mode particles, the absolute level of predictive performance varies considerably among target years, but within individual target years R^2 generally changes only moderately as training length increases. In contrast, MAE, RMSE, and MSE show a clearer decrease during the first few increases in training-window length before approaching relatively stable values.

For UFP, R^2 remains very high for many target years—frequently around 0.93–0.96—and changes relatively little once several training years are available. Nevertheless, the error metrics decline substantially over the shorter training windows before becoming almost flat. This suggests that the benefit of additional training data at Värriö is more evident in reducing prediction error than in substantially increasing explained variance.

N_{acc} exhibit the highest and most stable performance. R^2 is generally around 0.95–0.97 and shows only modest improvement with increasing training-window length.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

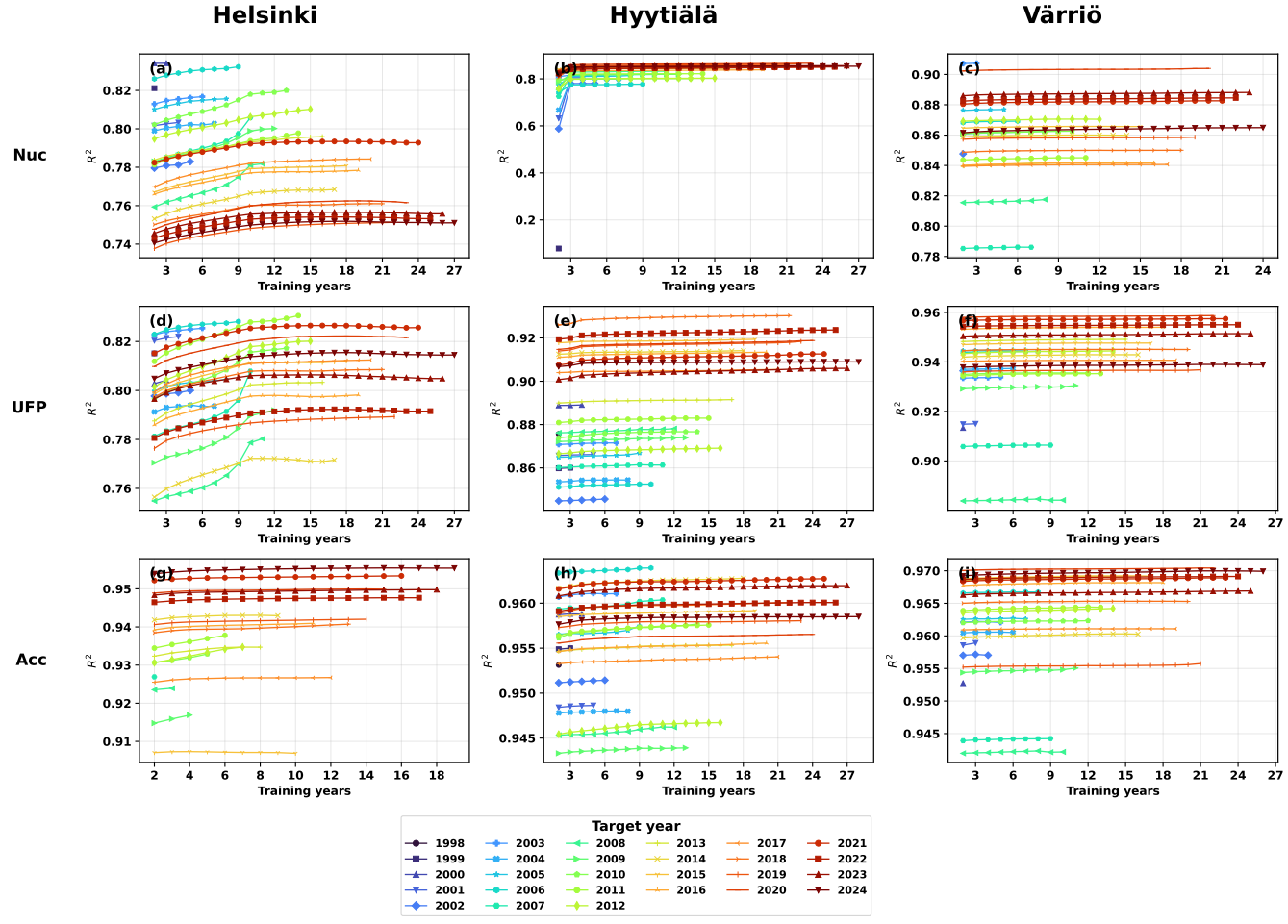


Figure 30. plot showing the R^2 for Scenario V in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

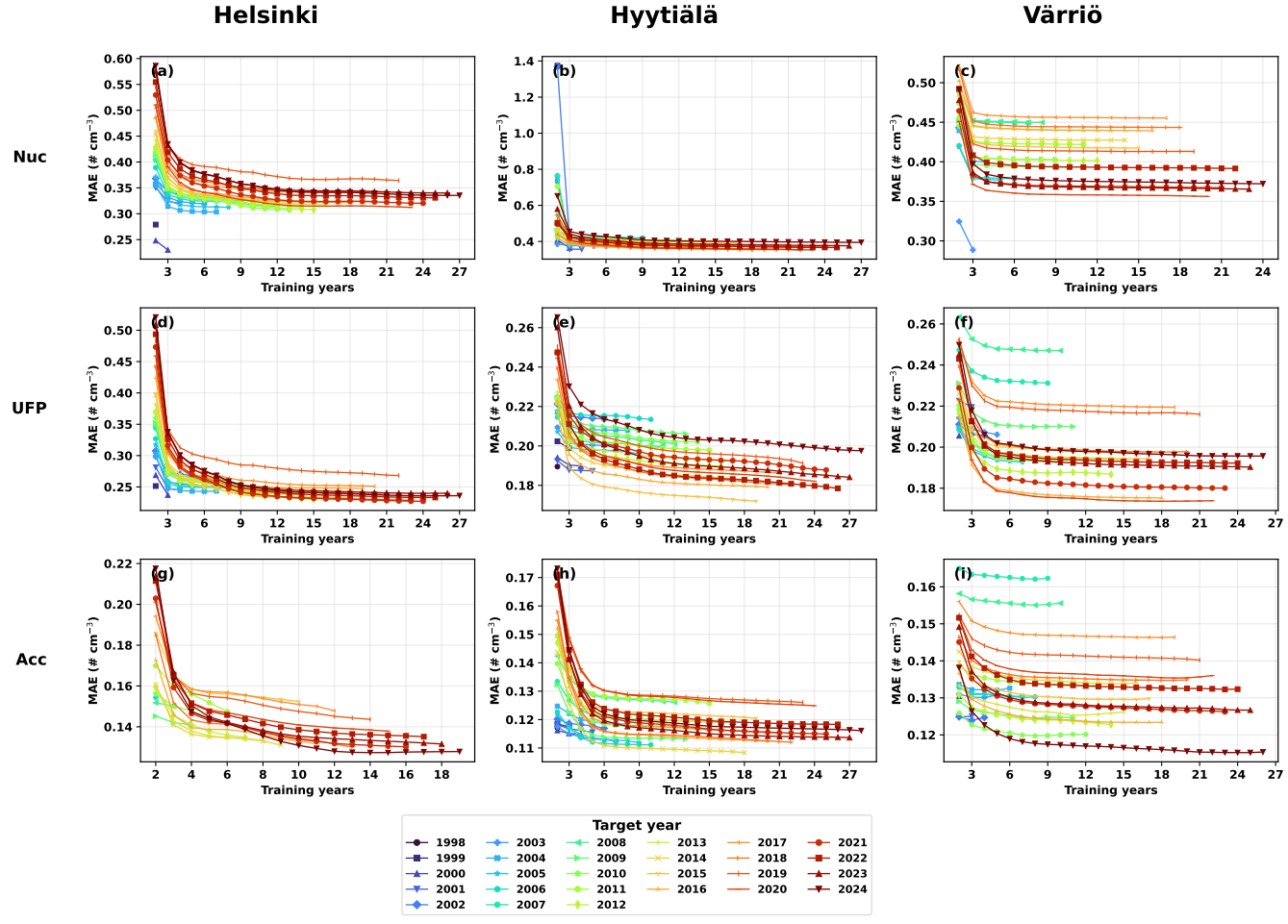


Figure 31. plot showing the Mean Absolute Error (MAE) for Scenario V in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

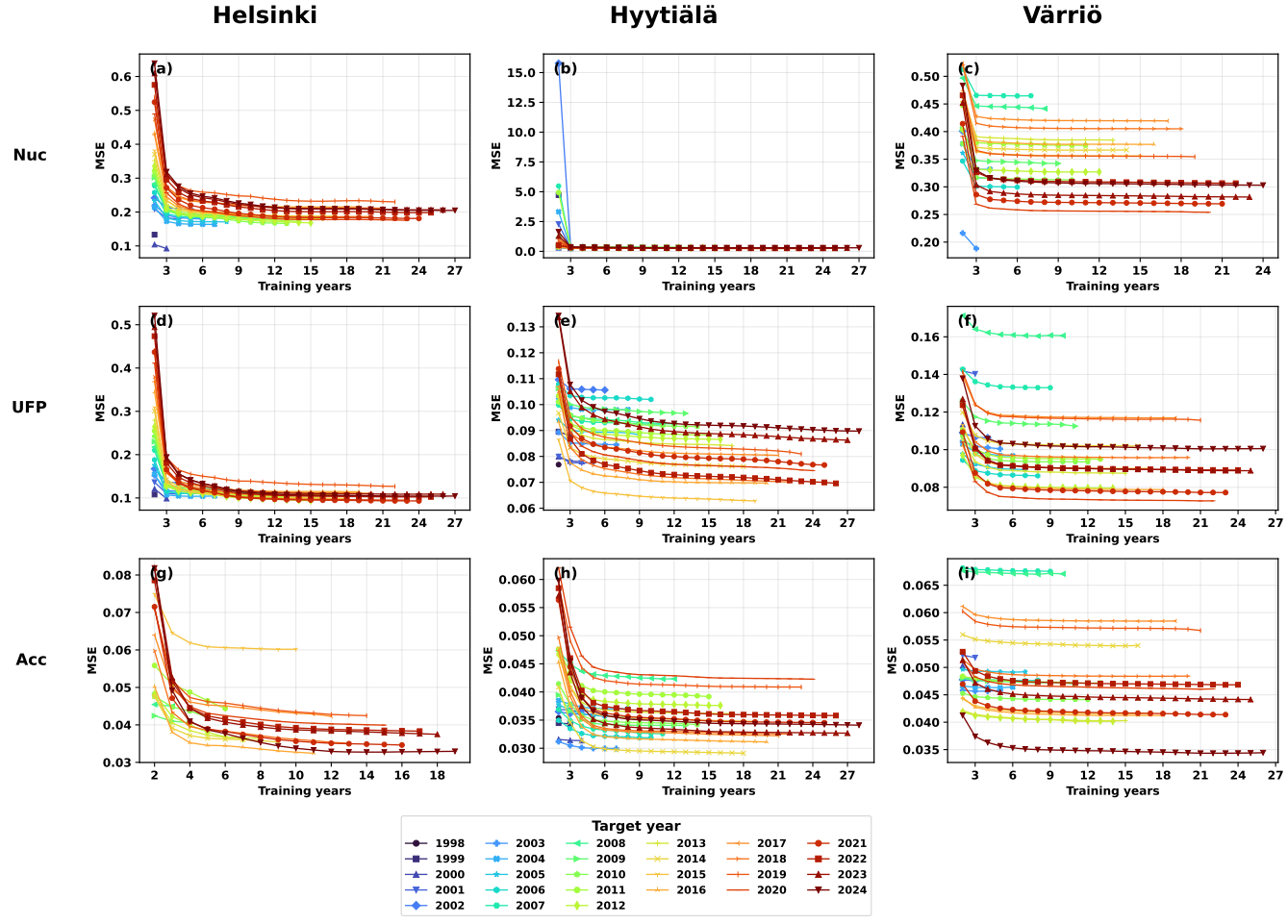


Figure 32. plot showing the Mean Squared Error (MSE) for Scenario V in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

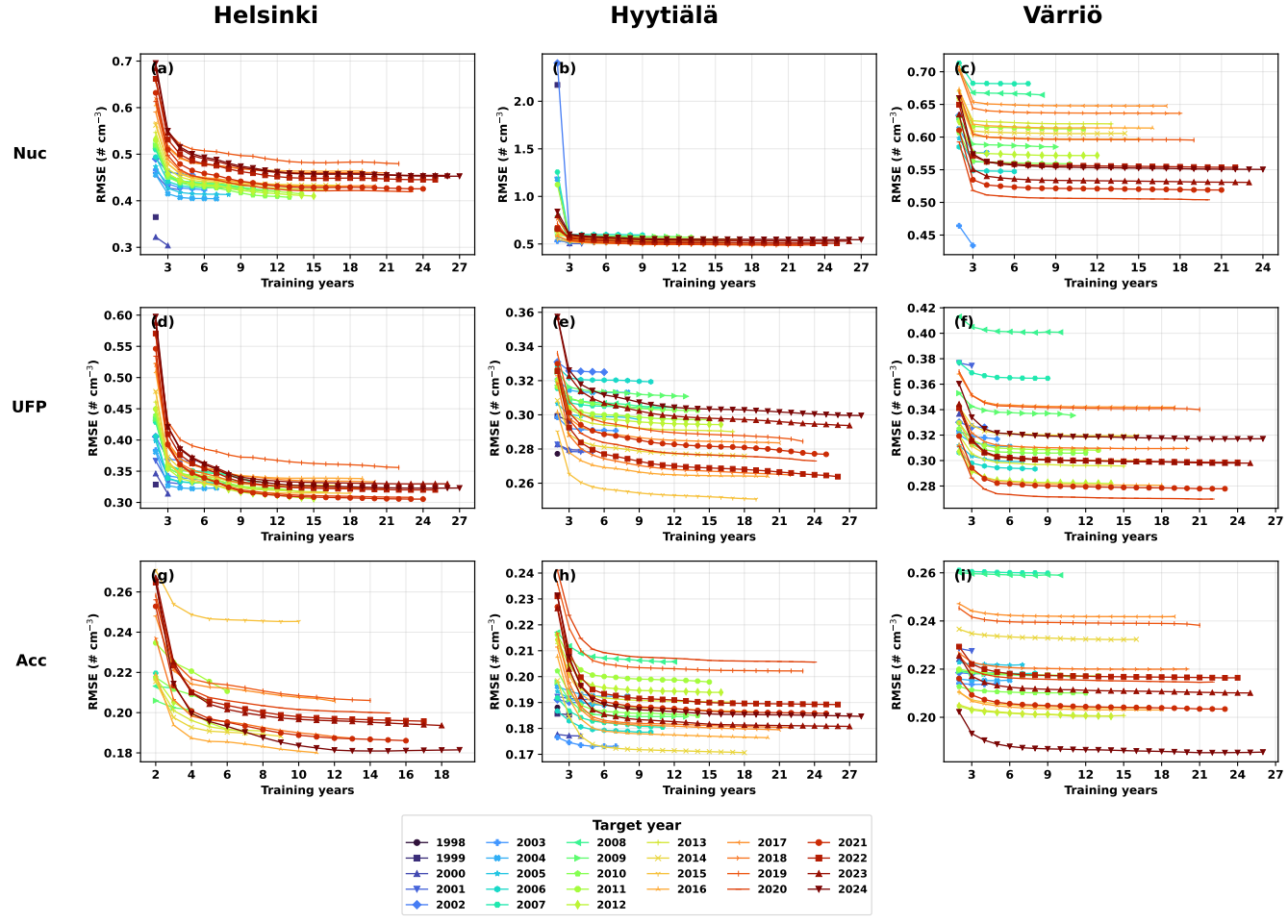


Figure 33. plot showing the Root Mean Squared Error (RMSE) for Scenario V in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode ($10 \leq D_p <\leq 25$ nm), (d-f) UFP ($10 \leq D_p < 100$ nm), and (g-i) accumulation ($100 \leq D_p \leq 300$ nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

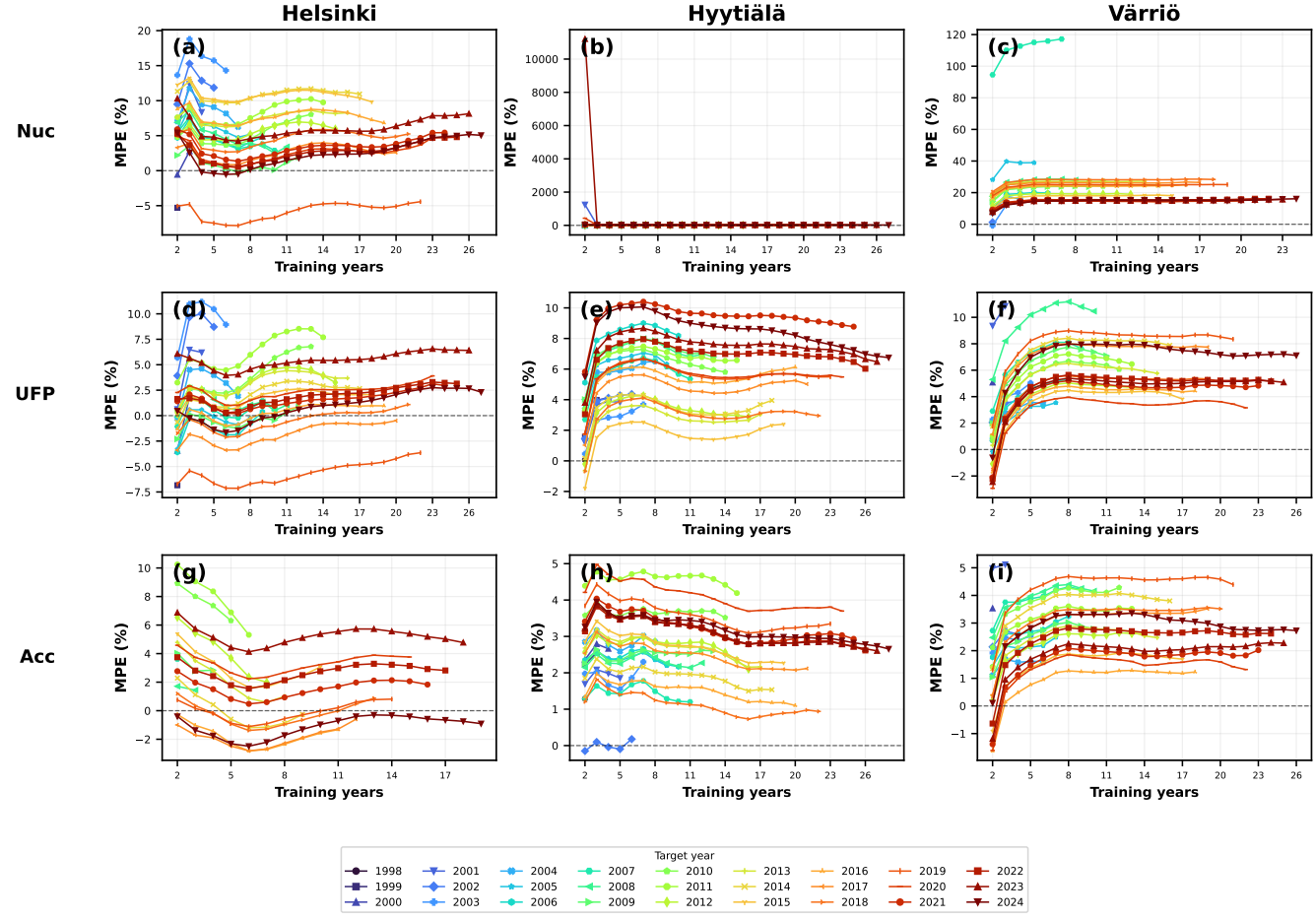


Figure 34. plot showing the Mean Percentage Error (MPE) for Scenario V in Helsinki (SMEAR III),Hyytiälä (SMEAR II), and Värriö (SMEAR I):(a-c) nucleation-mode (10 $\leq$D$_p$ $<\leq$ 25 nm), (d-f) UFP (10 $\leq D_p <$ 100 nm), and (g-i) accumulation (100 $\leq D_p \leq$ 300 nm)

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

3.2.6 limitations of the model

The limitations of the model for the performance parameters (R^2, MSE, MAE, RMSE, and MPE) show some anomalies in the values because of the lack of data in the training data, which is expected.

3.2.7 variation in predictions

Our model demonstrates high prediction skill for the nucleation-mode, UFP, and accumulation modes at 1 h resolution, as measured by the fraction of variance explained, R^2. Much of this performance stems from the inclusion of NO, NO_x, and SO_2 among the predictors. NO and NO_x are emitted locally from traffic and are strongly correlated with UFP, while SO_2 is an important precursor for particle growth and contributes to the accumulation mode, including through long-range transport.

In addition, the model incorporates temperature (T), wind speed (U), and wind direction (WD) to represent boundary-layer mixing; NO_2 as a vehicle-emission tracer; and O_3 together with solar radiation (SR) to capture the photochemical oxidant capacity. An explicit time structure $f(t)$—diurnal, weekly, and seasonal harmonics combined with autoregressive errors—encodes recurring cycles and short-memory dynamics, further improving predictive skill.

In Helsinki (SMEAR III), the model predicts the number concentration in all size fractions. For nucleation-mode particles, the high coefficient of determination shows that the model explains most of the variability. The Ultrafine particle fraction is also predicted well, likely because its variability is strongly linked to traffic- and combustion-related emissions. Predictors such as NO, NO_x, wind speed, temperature, and wind direction help explain these changes. In contrast, accumulation-mode particles are less affected by short-term local fluctuations and are more strongly influenced by regional transport and background conditions.

In Hyytiälä (SMEAR II), the high R^2 for the nucleation mode indicates that the model captures the main variability in aerosol dynamics at this boreal forest site. The predictors explain the effects of biogenic precursor emissions, solar radiation, condensation sink, and boundary-layer development. The R^2 for ultrafine particles indicates that their variability is controlled by atmospheric chemistry, particle growth, and meteorological conditions that are well represented in the model. Accumulation-mode particles show the most stable agreement because they are mainly influenced by regional background conditions and long-range transport.

In Värriö (SMEAR I), aerosol concentrations are less affected by local anthropogenic emissions and more influenced by clean air masses, regional transport, atmospheric chemistry, and meteorological conditions. For nucleation-mode particles, the model captures a substantial part of the variability, indicating that the predictors describe the main controlling processes.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

3.2.8 Acceptability of the model formulation

The model is developed under the assumption that the u error terms are independently and normally distributed. Based on this assumption, the autocorrelation is zero for all the non-zero lags is expected to be zero. This assumption is supported by our analysis, which shows the autocorrelation of the u error terms remains statistically insignificant, indicating a lack of temporal dependence. In our model formulation, we assumed that u_t error term u_t is independent and normally distributed; under this assumption, the autocorrelation is 0 for all non-zero lags (Fig. 35). shows that the autocorrelation of generated u values remains close to 0; additionally, we have plotted the autocorrelation of ε. Clearly, the term ε is autocorrelated; thus the autoregressive structure was mandatory in our model formulation. We have run our model with additional nonzero autoregressive coefficients, which reduced the autocorrelation of the error term u, but the main disadvantage was that the computational time increased significantly without improving the prediction accuracy.

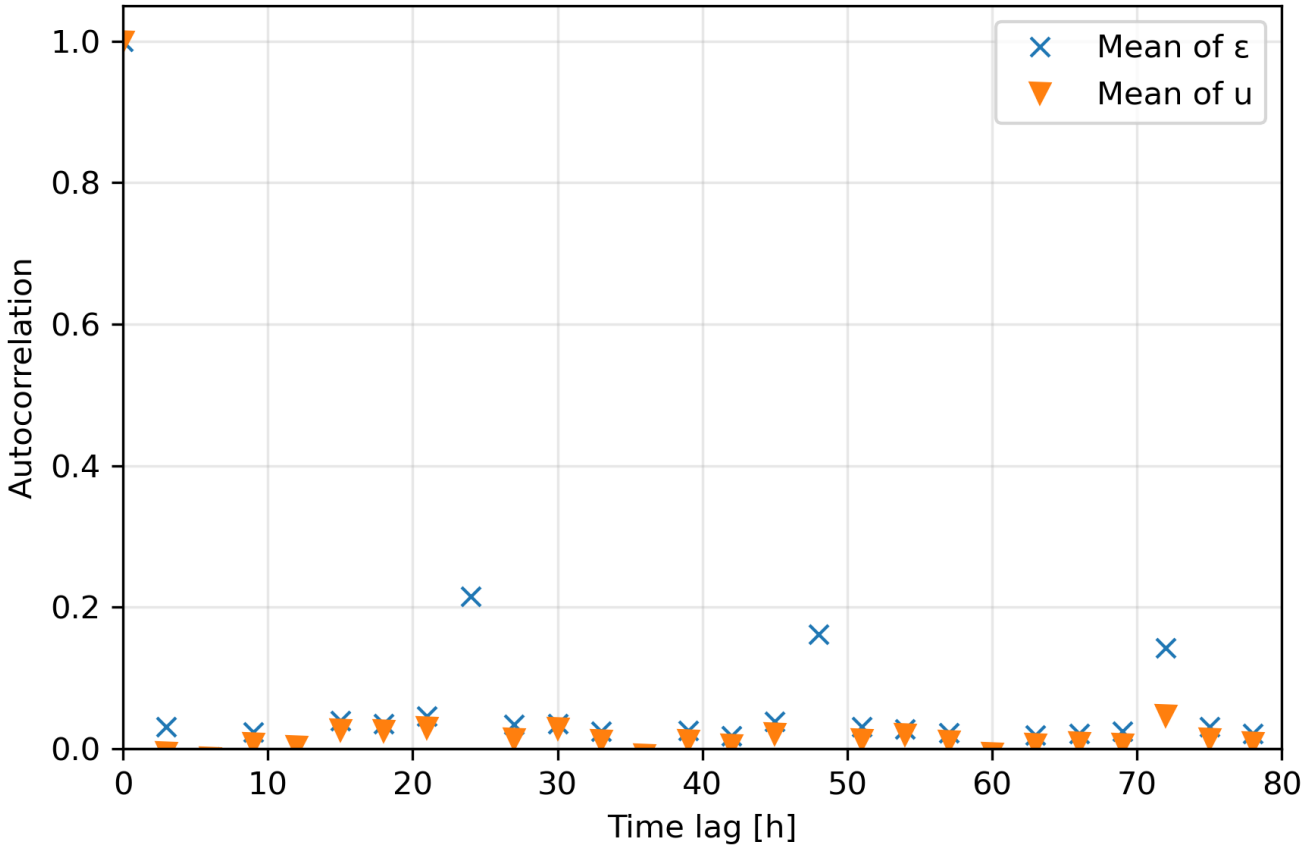


Figure 35. Autocorrelation functions for the AR error u_t (blue) and the model residuals ε_t (orange) for UFP ($10 \leq D_p < 100$ nm) with a 3-hour time step.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



AEROSOL Research Discussions

4 Conclusions

This work demonstrates that a Bayesian regression with autoregressive (AR) errors—driven only by routinely measured particle size distributions, meteorology, and trace gases—can reliably predict size-fractionated particle number concentrations across three contrasting environments: SMEAR I (Värriö; remote Arctic background), SMEAR II (Hyytiälä; boreal forest), and SMEAR III (Helsinki; urban background). By explicitly capturing daily, weekly, and seasonal cycles together with short- to synoptic-scale autocorrelation, the framework reproduces observed temporal cycles and, at 1-hour resolution, explains a large fraction of the variability (high R^2).

We set out to test whether common weather and gas measurements can predict nucleation-mode and other size ranges; whether repeating patterns and day-to-day carryover can be embedded; and whether skill persists across years and sites using long records.

The results support all three aims: the model captures site-specific seasonality (spring–summer nucleation-mode in Hyytiälä/Värriö; winter UFP in Helsinki), preserves diurnal structure (workday/weekend contrasts), and maintains robust predictive skill under sequential forecasting. Solar radiation (SR), O_3, and SO_2 emerge as steady predictors of nucleation-mode, while NO_2 and CO mark combustion-related enhancements in UFP; meteorology then controls dispersion, condensational growth, and wet removal. Across sites, Helsinki has the highest N but shows a long-term decline; Hyytiälä exhibits biogenic spring peaks with no strong trend; and Värriö remains lowest, with variability driven by air-mass origin and precipitation.

Our model was applied on 5 different scenarios. Scenario I aim to predict the next year taken into consideration all previous years into the training data.Scenario II is a single year prediction, where each year in the dataset is being predicted from different year. Scenario III goal was to check how adding older observation into the training dataset would improve the predictions of the recent years. Scenario IV addresses how the prediction of the model will perform when a historical starting point is progressively updated with more recent observations before the target year. The aim of Scenario V is to evaluate how predictive performance changes as the length of a consecutive training window increases, while allowing the training window to occur at different periods within the historical record before the target year.

Our results show that, in Scenario I, expanding the historical training period improves or stabilizes prediction performance depending on the size fractions. In Scenario II the predictive performance of the model depends on the year used for the training. In Scenario III indicates that including increasingly older observations does not substantially improve prediction once recent data are already available. In Scenario IV indicates that progressively updating the model with more recent observations improves prediction performance. In Scenario V, increasing the length of a consecutive training window generally improves prediction performance.

Our statistical prediction model demonstrates high accuracy in predicting nucleation-mode, UFP and accumulation mode particle number concentrations with 1h resolution. The model performance can be attributed to the consideration of different meteorological and gas pollutants that directly or indirectly influence particle formation, growth, and removal. Windspeed, wind-direction, temperature, are consistently among the strongest predictors. Across training years and prediction-window lengths, R^2 stayed consistently high,where MAE, MSE, RMSE and MPE remained almost closed to 0.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



Limitations arise when relationships learned under one mix of emissions and weather shift as background pollution changes (non-stationarity). In addition, missing measurements and instrumentation differences can introduce bias and extra uncertainty.

For operators lacking full PNSD instrumentation, this framework provides a low-cost, interpretable, and deployable pathway to hourly particle-number forecasts—supporting exposure assessment, health alerts, and network design. In sum, our Bayesian AR framework offers a practical, transparent route to operational forecasting across contrasting environments. Future work should couple the model to forecasted meteorology and gases, adapt to regime shifts, and quantify tail risk—turning high skill into robust, real-world guidance.

Acknowledgements

I would like to express my sincere gratitude for my supervisors for their guidance, support and valuable insights through this work.

We acknowledge the SMART-SMEAR data infrastructure and the University of Helsinki for maintaining the SMEAR stations and providing open-access data. This work is supported by Eastern Mediterranean and Middle East Climate and Atmosphere Research (EMME-CARE) project, which received funding from the European Union's Horizon 2020 Research and Innovation Programme (Grant Agreement Number 856612) and the Government of Cyprus.

This project is supported by the Research Council of Finland through the Academy Research Fellowship project (Grant No. 355330) and the University Profiling funding initiative InterEarth (Grant No. 353218).Funded by the European Union ERC-2022-STGERC-BAE-Project: 101076311. Also, supported by the National Foreign Experts Program in China (Granted No. S20260336). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Code availability.

The code used for regression modeling, scenario analysis, and figure generation is available from the corresponding author upon reasonable request and it will be released in a public repository at a later stage.

Data availability.

The data used in this study, DMPS measurements, meteorological variables, and trace gas concentrations, were obtained from the SMEAR database (https://smear.avaa.csc.fi/). Additionally, data is available from the corresponding author upon request.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



References

Aldrin, M. and Haff, I. H.: Generalised additive modelling of air pollution, traffic volume and meteorology, Atmospheric Environment, 39, 2145–2155, https://doi.org/10.1016/j.atmosenv.2004.12.020, 2005.

Chelani, A. B. and Devotta, S.: Air quality forecasting using a hybrid autoregressive and nonlinear model, Atmospheric Environment, 40, 1774–1780, https://doi.org/10.1016/j.atmosenv.2005.11.019, 2006.

Cheon, S. P., Kim, S., Lee, S. Y., and Lee, C. B.: Bayesian networks based rare event prediction with sensor data, Knowledge-Based Systems, 22, 336–343, https://doi.org/10.1016/j.knosys.2009.02.004, 2009.

Chib, S.: Bayes regression with autoregressive errors: A Gibbs sampling approach, Journal of Econometrics, 58, 275–294, https://doi.org/10.1016/0304-4076(93)90046-8, 1993.

Dada, L., Paasonen, P., Nieminen, T., Buenrostro Mazon, S., Kontkanen, J., Peräkylä, O., Lehtipalo, K., Hussein, T., Petäjä, T., Kerminen, V.-M., Bäck, J., and Kulmala, M.: Long-term analysis of clear-sky new particle formation events and nonevents in Hyytiälä, Atmospheric Chemistry and Physics, 17, 6227–6241, https://doi.org/10.5194/acp-17-6227-2017, 2017.

Dada, L., Chellapermal, R., Buenrostro Mazon, S., Paasonen, P., Lampilahti, J., Manninen, H. E., Junninen, H., Petäjä, T., Kerminen, V.-M., and Kulmala, M.: Refined classification and characterization of atmospheric new-particle formation events using air ions, Atmospheric Chemistry and Physics, 18, 17 883–17 893, https://doi.org/10.5194/acp-18-17883-2018, 2018.

Dal Maso, M., Kulmala, M., Riipinen, I., Wagner, R., Hussein, T., Aalto, P. P., and Lehtinen, K. E. J.: Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from SMEAR II, Hyytiälä, Finland, Boreal Environment Research, 10, 323–336, https://researchportal.helsinki.fi/en/publications/formation-and-growth-of-fresh-atmospheric-aerosols-eight-years-of/, 2005.

Dawid, A. P.: Statistical theory: The prequential approach, Journal of the Royal Statistical Society: Series A, 147, 278–292, https://doi.org/10.2307/2981683, 1984.

Ehn, M., Thornton, J. A., Kleist, E., Sipilä, M., Junninen, H., Pullinen, I., et al.: A large source of low-volatility secondary organic aerosol, Nature, 506, 476–479, https://doi.org/10.1038/nature13032, 2014.

Hari, P. and Kulmala, M.: Station for Measuring Ecosystem–Atmosphere Relations (SMEAR II), Boreal Environment Research, 10, 315–322, url={https://researchportal.helsinki.fi/en/publications/station-for-measuring-ecosystem-atmosphere-relations-smear-ii/}, 2005.

Hinds, W. C.: Aerosol Technology: Properties, Behavior, and Measurement of Airborne Particles, Wiley-Interscience, New York, 2 edn., https://hero.epa.gov/reference/13210/, 1999.

Hrušt, L., Klaić, Z. B., Križan, J., Antonić, O., and Hercog, P.: Neural network forecasting of air pollutants' hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations, Atmospheric Environment, 43, 5588–5596, https://doi.org/10.1016/j.atmosenv.2009.07.048, 2009.

Hussein, T., Karppinen, A., Kukkonen, J., Härkönen, J., Aalto, P. P., Hämeri, K., Kerminen, V.-M., and Kulmala, M.: Meteorological dependence of size-fractionated number concentrations of urban aerosol particles, Atmospheric Environment, 40, 1427–1440, https://doi.org/10.1016/j.atmosenv.2005.10.061, 2006.

Hussein, T., Dal Maso, M., Petäjä, T., Koponen, I. K., Paatero, J., Aalto, P. P., Hämeri, K., and Kulmala, M.: Time span and spatial scale of regional new particle formation events over Finland and southern Scandinavia, Atmospheric Chemistry and Physics, 9, 4699–4716, https://doi.org/10.5194/acp-9-4699-2009, 2009.

Hussein, T., Wierzbicka, A., Löndahl, J., Lazaridis, M., and Hänninen, O.: Indoor aerosol modeling for assessment of exposure and respiratory tract deposited dose, Atmospheric Environment, 106, 402–411, https://doi.org/10.1016/j.atmosenv.2014.07.034, 2015.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Jacob, D. J.: Introduction to Atmospheric Chemistry, Princeton University Press, Princeton, New Jersey, ISBN 978-0-691-00185-2, https://doi.org/10.2307/j.ctt7t8hg, 1999.

Jaiswal, A., Samuel, C., and Kadabgaon, V. M.: Statistical trend analysis and forecast modeling of air pollutants, Global Journal of Environmental Science and Management, 4, 427–438, https://doi.org/10.22034/gjesm.2018.04.004, 2018.

Järvi, L., Hannuniemi, H., Hussein, T., Junninen, H., Aalto, P. P., Hillamo, R., Mäkelä, T., Keronen, P., Siivola, E., Vesala, T., and Kulmala, M.: The urban measurement station SMEAR III: Continuous monitoring of air pollution and surface–atmosphere interactions in Helsinki, Finland, Boreal Environment Research, 14, 86–109, https://researchportal.helsinki.fi/en/publications/the-urban-measurement-station-smear-iii-continuous-monitoring-of-/, 2009.

Junninen, H., Lauri, A., Keronen, P., Aalto, P., Hiltunen, V., Hari, P., and Kulmala, M.: SmartSMEAR, Online database, https://smear.avaa.csc.fi/, data accessed via SmartSMEAR portal on 17 November 2025, 2025.

Kerminen, V.-M., Chen, X., Vakkari, V., Petäjä, T., Kulmala, M., and Bianchi, F.: Atmospheric new particle formation and growth: Review of field observations, Environmental Research Letters, 13, 103 003, https://doi.org/10.1088/1748-9326/aadf3c, 2018.

Kerminen, V.-M. et al.: Cloud condensation nuclei production associated with atmospheric nucleation, Atmospheric Chemistry and Physics, https://doi.org/10.5194/acp-12-12037-2012, 2012.

Koo, J. W., Wong, S. W., Selvachandran, G., Long, H. V., and Son, L. H.: Prediction of Air Pollution Index in Kuala Lumpur using fuzzy time series and statistical models, Air Quality, Atmosphere & Health, 13, 77–88, https://doi.org/10.1007/s11869-019-00772-y, 2020.

Kulmala, M., Laaksonen, A., Pirjola, L., Korhonen, H., and Kerminen, V.-M.: Formation and growth rates of ultrafine atmospheric particles: a review of observations, Journal of Aerosol Science, 35, 143–176, https://doi.org/10.1016/j.jaerosci.2003.10.003, 2004.

Kulmala, M., Petäjä, T., Mönkkönen, P., Koponen, I. K., Dal Maso, M., Aalto, P. P., Lehtinen, K. E. J., and Kerminen, V.-M.: On the growth of nucleation mode particles: source rates of condensable vapor in polluted and clean environments, Atmospheric Chemistry and Physics, 5, 409–416, https://doi.org/10.5194/acp-5-409-2005, 2005.

Kulmala, M., Kontkanen, J., Junninen, H., et al.: Direct observations of atmospheric aerosol nucleation, Science, 339, 943–946, https://doi.org/10.1126/science.1227385, 2013.

Kulmala, M., Petäjä, T., Ehn, M., Sipilä, M., Worsnop, D. R., and Vehkamäki, H.: Chemistry of atmospheric nucleation: On the recent advances, Annual Review of Physical Chemistry, 65, 21–37, https://doi.org/10.1146/annurev-physchem-040412-110014, 2014.

Kulmala, M., Junninen, H., Dada, L., Salma, I., Weidinger, T., Thén, W., Vörösmarty, M., Komsaare, K., Stolzenburg, D., Cai, R., Yan, C., Li, X., Deng, C., Jiang, J., Petäjä, T., Nieminen, T., and Kerminen, V.-M.: Quiet new particle formation in the atmosphere, Frontiers in Environmental Science, 10, 912 385, https://doi.org/10.3389/fenvs.2022.912385, 2022.

Kumar, U. and Jain, V. K.: ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO), Stochastic Environmental Research and Risk Assessment, 24, 751–760, https://doi.org/10.1007/s00477-009-0361-8, 2010.

Laakso, L., Hussein, T., Aalto, P. P., et al.: Diurnal and annual characteristics of particle mass and number concentrations in urban, rural and Arctic environments in Finland, Boreal Environment Research, 8, 463–474, https://doi.org/10.1016/S1352-2310(03)00206-1, 2003.

Lee, S.-H., Gordon, H., Yu, H., Lehtipalo, K., Halonen, R., et al.: New particle formation in the atmosphere: From molecular clusters to global climate, Journal of Geophysical Research: Atmospheres, 124, 7098–7146, https://doi.org/10.1029/2018JD029356, 2019.

Lehtipalo, K., Stolzenburg, D., Kangasluoma, J., Sipilä, M., and Kulmala, M.: How the understanding of atmospheric new particle formation has evolved along with the development of measurement and analysis methods, Journal of Aerosol Science, 186, 106 562, https://doi.org/10.1016/j.jaerosci.2024.106494, 2025.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.



Liu, Y., Guo, H., Mao, G., and Yang, P.: A Bayesian hierarchical model for urban air quality prediction under uncertainty, Atmospheric Environment, 42, 8464–8469, https://doi.org/10.1016/j.atmosenv.2008.08.018, 2008.

Ma, L., Zhang, Y., Lin, Z., et al.: Deposition potential of 0.003–10 μm ambient particles in the humidified human respiratory tract: Contribution of new particle formation events in Beijing, Ecotoxicology and Environmental Safety, 243, 114 023, https://doi.org/10.1016/j.ecoenv.2022.114023, 2022.

McMurry, P. H.: A review of atmospheric aerosol measurements, Atmospheric Environment, 34, 1959–1999, https://doi.org/10.1016/S1352-2310(99)00455-0, 2000.

Middya, A. I. and Roy, S.: Pollutant specific optimal deep learning and statistical model building for air quality forecasting, Environmental Pollution, 301, 118 972, https://doi.org/10.1016/j.envpol.2022.118972, 2022.

Mølgaard, B., Hussein, T., Corander, J., and Hämeri, K.: Forecasting size-fractionated particle number concentrations in the urban atmosphere, Atmospheric Environment, 46, 155–163, https://doi.org/10.1016/j.atmosenv.2011.10.004, 2012.

Moustris, K. P., Ziomas, I. C., and Paliatsos, A. G.: 3-day-ahead forecasting of regional pollution index for the pollutants NO2, CO, SO2, and O3 using artificial neural networks in Athens, Greece, Water, Air, and Soil Pollution, 209, 29–43, https://doi.org/10.1007/s11270-009-0179-5, 2010.

Okuljar, M., Kuuluvainen, H., Kontkanen, J., Garmash, O., Olin, M., Niemi, J. V., Timonen, H., Kangasluoma, J., Tham, Y. J., Baalbaki, R., Sipilä, M., Salo, L., Lintusaari, H., Portin, H., Teinilä, K., Aurela, M., Dal Maso, M., Rönkkö, T., Petäjä, T., and Paasonen, P.: Measurement report: The influence of traffic and new particle formation on the size distribution of 1–800 nm particles in Helsinki – a street canyon and an urban background station comparison, Atmospheric Chemistry and Physics, 21, 9931–9953, https://doi.org/10.5194/acp-21-9931-2021, 2021.

Paasonen, P., Asmi, A., Petäjä, T., Kerminen, V.-M., and Kulmala, M.: Warming-induced increase in aerosol number concentration likely to moderate climate change, Nature Geoscience, 6, 438–442, https://doi.org/10.1038/ngeo1800, 2013.

Peltola, M., Thakur, R. C., Spence, K., Lampilahti, J., Mäkelä, R., Karttunen, S., Ezhova, E., Haapanala, S., Vähä, A., Kangasluoma, J., Chan, T., Paasonen, P., Norkko, J., Norkko, A., Kulmala, M., and Ehn, M.: Measurement report: New particle formation and aerosol properties at a newly founded atmospheric observatory at the Finnish Baltic Sea coast, Atmospheric Chemistry and Physics, 26, 489–513, https://doi.org/10.5194/acp-26-489-2026, 2026.

Ruuskanen, T. M., Reissell, A., Keronen, P., Aalto, P. P., Laakso, L., Grönholm, T., Hari, P., and Kulmala, M.: Atmospheric trace gas and aerosol particle concentration measurements in Eastern Lapland, Finland 1992–2001, Boreal Environment Research, 8, 335–349, https://www.borenv.net/BER/archive/ber84.htm#335, 2003.

Schlink, U., Herbarth, O., Richter, M., Dorling, S., Nunnari, G., Cawley, G., and Pelikan, E.: Statistical models to assess the health effects and to forecast ground-level ozone, Environmental Modelling & Software, 21, 547–558, https://doi.org/10.1016/j.envsoft.2004.12.002, 2006.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric Chemistry and Physics: From Air Pollution to Climate Change, John Wiley & Sons, Hoboken, New Jersey, 3 edn., ISBN 978-1-118-94740-1, https://www.wiley.com/en-us/atmospheric-chemistry-and-physics-from-air-pollution-to-climate-change-3rd-edition-p-9781118947401, 2016.

Thakur, R. C., Kangasluoma, J., Kontkanen, J., Junninen, H., Lehtipalo, K., Nieminen, T., Ehn, M., and Kulmala, M.: Characteristics of new particle formation events in urban Helsinki, Atmospheric Chemistry and Physics, 22, 6365–6387, https://doi.org/10.5194/acp-22-6365-2022, 2022.

https://doi.org/10.5194/ar-2026-42
Preprint. Discussion started: 9 November 2026
© Author(s) 2026. CC BY 4.0 License.

Tunved, P., Hansson, H.-C., Kulmala, M., Aalto, P., Viisanen, Y., Karlsson, H., Kristensson, A., Swietlicki, E., Dal Maso, M., Ström, J., and Komppula, M.: One year boundary layer aerosol size distribution data from five Nordic background stations, Atmospheric Chemistry and Physics, 3, 2183–2205, https://doi.org/10.5194/acp-3-2183-2003, 2003.

Vehkamäki, H., Dal Maso, M., Hussein, T., Flanagan, R., Hyvärinen, A., Lauros, J., Merikanto, P., Mönkkönen, M., Pihlatie, K., Salminen, K., Sogacheva, L., Thum, T., Ruuskanen, T. M., Keronen, P., Aalto, P. P., Hari, P., Lehtinen, K. E. J., Rannik, Ü., and Kulmala, M.: Atmospheric particle formation events at Värriö measurement station in Finnish Lapland 1998–2002, Atmospheric Chemistry and Physics, 4, 2015–2023, https://doi.org/10.5194/acp-4-2015-2004, 2004.

Wei, W., Ramalho, O., Malingre, L., Sivanantham, S., Little, J. C., and Mandin, C.: Machine learning and statistical models for predicting indoor air quality, Indoor Air, 29, 704–726, https://doi.org/10.1111/ina.12580, 2019.

Westmoreland, E. J., Carslaw, N., Carslaw, D. C., Gillah, A., and Bates, E.: Analysis of air quality within a street canyon using statistical and dispersion modelling techniques, Atmospheric Environment, 41, 9195–9205, https://doi.org/10.1016/j.atmosenv.2007.07.057, 2007.

Wu, Y., Zaidan, M. A., Cai, R., Duplissy, J., Okuljar, M., Lehtipalo, K., Petäjä, T., and Kangasluoma, J.: Estimating the atmospheric aerosol number size distribution using deep learning, Environmental Science: Atmospheres, 5, 367–377, https://doi.org/10.1039/D4EA00127C, 2025.

Zaidan, M. A., Surakhi, O., Fung, P. L., and Hussein, T.: Sensitivity Analysis for Predicting Sub-Micron Aerosol Concentrations Based on Meteorological Parameters, Sensors, 20, 2876, https://doi.org/10.3390/s20102876, 2020.