



AeroMiG: Multi-modal Inverse-Gamma modeling for aerosol particle number size distributions

Abdur Rahman^{1,2,*}, Juha Kangasluoma², Santtu Mikkonen³, Tareq Hussein², Tuukka Petäjä², Sasu Tarkoma¹, and Martha Arbayani Zaidan^{1,2,*}

¹Department of Computer Science, University of Helsinki, Helsinki, Finland

²Institute for Atmospheric and Earth System Research, University of Helsinki, Helsinki, Finland

³Department of Environmental and Biological Sciences, Faculty of Science, Forestry and Technology, University of Eastern Finland, Kuopio, Finland

Correspondence: Abdur Rahman (abdur.rahman@helsinki.fi) and Martha Arbayani Zaidan (martha.zaidan@helsinki.fi)

Abstract. Aerosol particles play a critical role in air quality, human health, and climate processes, making their accurate characterization essential. One common way to represent aerosol populations is through particle number size distributions (PNSDs), which describe the concentration of particles across different mobility diameters. However, these distributions are often highly skewed, heavy-tailed, and multi-modal, posing significant challenges for conventional fitting approaches. In this study, we propose a robust parametric framework (AeroMiG) for modeling PNSDs using Inverse-Gamma mixture distributions. The method represents observed size distributions as a superposition of multiple Inverse-Gamma components, enabling flexible modeling of asymmetric and heavy-tailed structures that are not well captured by widely-used log-Gaussian based approaches. The performance of the proposed method is evaluated using real-world data from three measurement stations. Results are assessed using multiple criteria, including predictive accuracies, computational efficiency, and a composite score summarizing overall model quality. Comparative analysis against a widely used Aerosol Multi-mode log-Gaussian (AeroMG) demonstrates that the AeroMiG framework consistently achieves improved fitting accuracy, lower error, and faster computation. The proposed approach provides a reliable and computationally efficient solution for modeling complex aerosol size distributions, with potential applications in real-time air quality monitoring and data-driven environmental analysis.

1 Introduction

Atmospheric science is a fundamental discipline for understanding the physical, chemical, and dynamical processes that govern the Earth's atmosphere. It plays a central role in climate research, air quality assessment, and environmental monitoring by providing the scientific basis for interpreting interactions between natural systems and anthropogenic activities (Seinfeld and Pandis, 2016; Intergovernmental Panel On Climate Change (Ipcc), 2023; Masson-Delmotte et al., 2021).

One of the most critical challenges in atmospheric science is the study of climate change, where the atmosphere acts as a key medium for energy balance and radiative forcing (Hinds and Zhu, 2022; Li et al., 2022). Greenhouse gases and aerosol particles influence the Earth's radiation budget through absorption, scattering, and cloud interactions, thereby affecting temperature, precipitation patterns, and extreme weather events (Akinyoola et al., 2024). In particular, aerosols remain one of the largest



sources of uncertainty in climate projections due to their complex spatial variability and interactions with clouds and radiation (Intergovernmental Panel On Climate Change (Ipcc), 2023; Stocker, 2014). Advances in atmospheric measurement and data analysis are critical for reducing uncertainties in climate models, improving air quality forecasts, and enhancing our ability to respond to environmental challenges in a changing climate (Karpatne et al., 2019).

Fine particulate matter, particularly particles with diameters below $2.5 \mu\text{m}$ ($\text{PM}_{2.5}$), has been strongly associated with adverse health outcomes, including respiratory and cardiovascular diseases, and premature mortality (Du et al., 2021), while atmospheric aerosols also play a critical role in the climate system by directly scattering and absorbing solar radiation and indirectly modifying cloud formation, lifetime, and albedo (Seinfeld and Pandis, 2016; Masson-Delmotte et al., 2021; Pope III, 2002).

A key representation of aerosol populations is the PNSD, which describes the concentration of particles as a function of their mobility diameter. PNSDs provide detailed insight into aerosol sources, transformation processes such as nucleation, coagulation, and condensation, and the evolution of atmospheric particle populations (Hinds and Zhu, 2022; Whitby, 1978).

Measured PNSDs are often noisy, highly skewed, and influenced by multiple overlapping physical processes such as nucleation, condensation, and coagulation, which make direct interpretation difficult and necessitate the use of parametric fitting approaches to extract meaningful information about underlying aerosol populations (Seinfeld and Pandis, 2016; Hinds and Zhu, 2022).

Fitting algorithms provide a systematic way to approximate observed distributions using a combination of analytical functions, allowing the decomposition of complex PNSDs into interpretable modes associated with distinct sources or transformation mechanisms (Whitby, 1978). However, the fitting process is inherently challenging due to the nonlinear and multi-parameter nature of the problem, which often leads to multiple local minima and parameter identifiability issues. Traditional optimization methods may struggle to provide stable and consistent solutions, particularly when the data exhibit strong asymmetry and heavy-tailed behavior.

To address these challenges, advanced optimization strategies are required to ensure reliable parameter estimation and global convergence. Evolutionary algorithms, such as Differential Evolution, offer a robust alternative by performing population-based global search, reducing sensitivity to initialization, and improving the likelihood of identifying optimal or near-optimal solutions in complex search spaces (Storn and Price, 1997). These properties make such algorithms particularly suitable for fitting aerosol size distributions, where both model flexibility and numerical stability are essential. Some different models have been used for modeling aerosol size particles. The prevailing approach relies on lognormal modal decomposition, where aerosol populations are represented as a superposition of modes associated with nucleation, Aitken, accumulation, and coarse particles (Whitby and McMurry, 1997; Hinds and Zhu, 2022; Hussein et al., 2005). This framework is widely used because of its physical interpretability and consistency with classical aerosol theory.

The fitting of multicomponent aerosol size distributions using least-squares-based approaches has a long history (Raabe, 1978). Building on this tradition, the AeroMG (Aerosol Multi-mode log-Gaussian) algorithm provides a framework for estimating modal parameters of particle number size distributions using multi-lognormal representations (Hussein et al., 2005; Xiong et al., 2026; Christiansen et al., 2019). While effective in many applications, AeroMG and related approaches depend



on assumptions of lognormality and often require careful initialization and predefined mode structures (Zhu and Wang, 2024; Hussein et al., 2005). As a result, their performance can degrade when applied to highly variable, asymmetric, or heavy-tailed distributions frequently observed in real atmospheric measurements. To improve flexibility, alternative statistical approaches such as Gaussian mixture models and nonparametric density estimation have been explored (Seinfeld and Pandis, 2016). Although these methods can better capture multi-modal behavior, they often lack clear physical interpretability and may suffer from instability or overfitting in the presence of noisy data. More recently, global optimization techniques, including evolutionary algorithms, have been introduced to address the non-convex nature of the fitting problem and reduce sensitivity to initialization (Storn and Price, 1997). However, their integration with physically meaningful distribution models for aerosol data remains limited. Despite these developments, a key limitation persists such that existing approaches struggle to simultaneously capture heavy-tailed behavior, maintain physical interpretability, and ensure stable parameter estimation. This gap motivates the development of alternative modeling frameworks that combine flexible distributional forms with robust global optimization strategies.

To address these limitations, this paper introduces AeroMiG (Aerosol Multi-mode Inverse-Gamma), a framework explicitly engineered to capture the highly skewed, heavy-tailed, and asymmetric structures of ultrafine particles in aerosol PNSDs. This work delivers several key contributions. First, by incorporating a stability-driven objective function, the framework resolves common convergence issues while maintaining strict physical interpretability, enabling the reliable tracking of nucleation representation and multi-mode dynamics. Second, the model is validated using extensive real-world datasets across multiple stations, encompassing a 3-mode configuration spanning 15 years and a 5-mode configuration covering full seasonal cycles. Third, comprehensive evaluations demonstrate that AeroMiG consistently outperforms the state-of-the-art AeroMG framework in both fitting accuracy and computational efficiency. Finally, the model's performance is uniquely benchmarked across distinct NPF event groups to evaluate its robustness under varying atmospheric conditions.

2 Measurement sites and data

This section describes the primary datasets utilized in this study, specifically PNSD data, followed by an overview of the SMEAR (Station for Measuring Ecosystem-Atmosphere Relations) stations where these measurements and relevant variables were collected.

2.1 Aerosol measurement data

PNSD were measured using a differential mobility particle sizer (DMPS), which determines size-resolved aerosol number concentrations based on electrical mobility diameter. The system combines bipolar charge neutralization, size classification using a differential mobility analyser, and detection via a condensation particle counter, enabling continuous measurements across the nucleation, Aitken, and accumulation modes.

A critical atmospheric phenomenon observed via PNSD data is New Particle Formation (NPF). In this study, NPF events are classified following the widely adopted methodology of Dal Maso et al. (Dal Maso et al., 2005), which categorizes days based

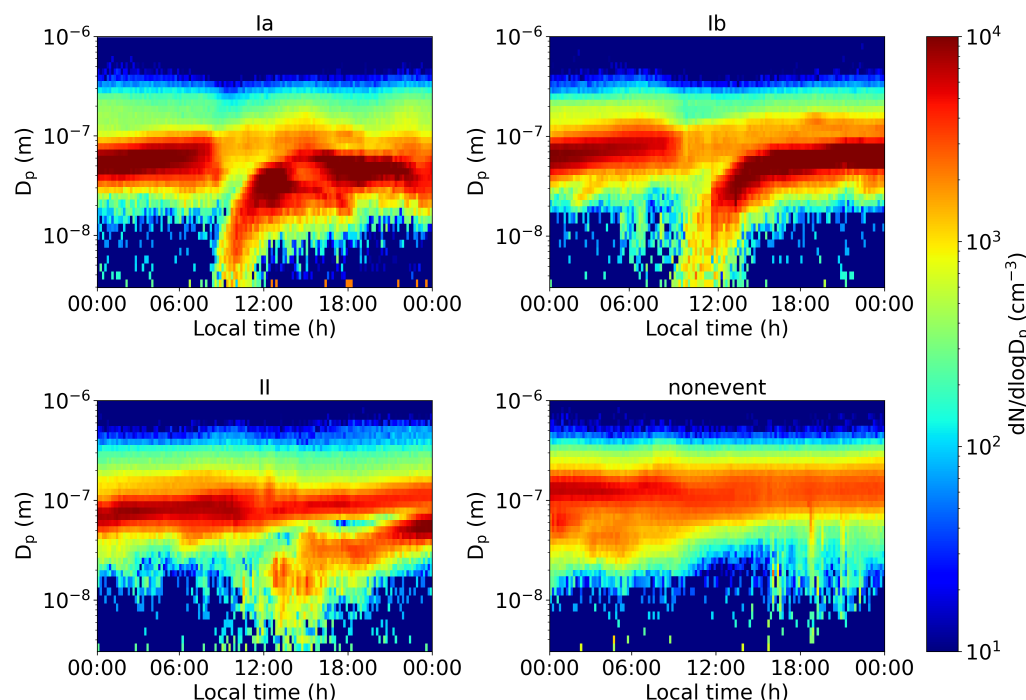


Figure 1. Example PNSD plots from DMPS measurements. Axes show time (x) and particle size in nm (y); color represents particle concentration in cm^{-3} .

on the clarity of the nucleation mode appearance and the persistence of subsequent particle growth. As illustrated in Figure 1, these events are structured into four distinct classes (Zaidan et al., 2018): Class I events (subdivided into Ia and Ib) exhibit robust formation and distinct growth tracks of new particles; Class II events indicate weaker or less well-defined growth trajectories; and nonevent days correspond to the absolute absence of detectable nucleation. In this work, the performance of AeroMiG and AeroMG is evaluated and compared across these distinct NPF event groups to uncover category-specific behavioral differences.

2.2 SMEAR stations

This subsection describes the three SMEAR stations where the PNSD data were collected as shown in Figure 2.

2.2.1 SMEAR I

The Värriö Subarctic Research Station (SMEAR I; 67.75° N , 29.61° E) is located $\sim 120 \text{ km}$ north of the Arctic Circle in the northern boreal zone (Figure 2), near the timberline within a strictly protected nature reserve (125 km^2) (Hari et al., 2013). Due to its pristine conditions and minimal anthropogenic influence, the primary purpose of SMEAR I is the continuous monitoring and study of subarctic aerosol formation, ion dynamics, and biosphere–atmosphere exchange (Ahti et al., 1968; Kyrö et al., 2014). General meteorological variables are collected via a 16 m meteorological mast, alongside outdoor tree chamber mea-

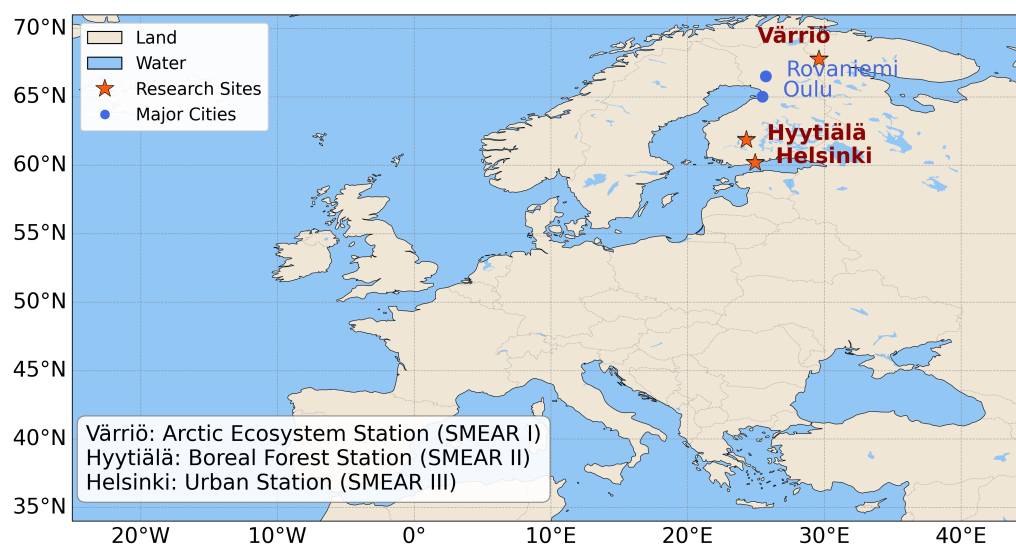


Figure 2. Geographic distribution of the SMEAR stations providing data for this study.

measurements for photosynthesis. Aerosol properties, including PNSDs, and trace gases are continuously sampled and measured using specialized analyzers housed inside a dedicated, climate-controlled measurement hut on site (Kyrö et al., 2014).

105 2.2.2 SMEAR II

The Hyytiälä measurement station (SMEAR II; 61.95° N, 24.28° E) is located 181 m above sea level in a non-urbanized boreal forest environment (Figure 2), approximately 190 km northwest of Helsinki (Hari and Kulmala, 2005). As one of the most comprehensive in-situ observation sites globally, its primary purpose is the long-term, continuous monitoring of ecosystem–atmosphere interactions, atmospheric chemistry, and NPF studies (Hari and Kulmala, 2005; Kulmala et al., 2004).

110 Since 1996, the station’s extensive instrumentation has continuously measured general variables including micrometeorology, weather conditions, and cloud remote sensing. PNSDs and trace gas concentrations are continuously sampled and tracked using specialized online analyzers deployed across the station’s ground-based facilities and dedicated measurement towers (Hari and Kulmala, 2005; Uusitalo et al., 2021).

2.2.3 SMEAR III

115 The SMEAR III station (60.20° N, 24.96° E; 26 m above sea level) is located in Kumpula (Figure 2), approximately 3 km northeast of Helsinki’s center, situated within a heterogeneous urban background environment characterized by a mixture of infrastructure, traffic corridors, and vegetated surfaces (Hussein et al., 2004, 2006). The primary purpose of SMEAR III is to characterize urban aerosol processes, meteorological conditions, and biosphere–atmosphere interactions under diverse anthropogenic and urban source influences (Hussein et al., 2006). General environmental variables are measured from a 31 m



120 tower situated on a rocky hill. PNSDs are continuously sampled using a DMPS housed in a nearby measurement container (Hussein et al., 2006). While measurements originally began at a nearby site in Siltavuori (1997–2001) before relocating to Kumpula, an instrumental upgrade on 9 March 2003 successfully extended the measurable PNSD range from 7–500 nm to 3–1000 nm without introducing discontinuities in the long-term data characterization (Aalto et al., 2001; Hussein et al., 2004).

3 Methods

125 This section provides detailed descriptions of the AeroMG and AeroMiG models, alongside the performance metrics employed to compare them.

3.1 AeroMG-Aerosol Multi-mode log-Gaussian

Atmospheric PNSDs are commonly represented as a superposition of multiple log-Gaussian (lognormal) modes, reflecting contributions from distinct aerosol processes. For a given particle diameter D_p , the modeled PNSD is

$$130 \quad \hat{n}(D_p) = \sum_{k=1}^K \frac{N_k}{\sqrt{2\pi} D_p \ln \sigma_k} \exp \left[-\frac{(\ln D_p - \ln D_{pg,k})^2}{2(\ln \sigma_k)^2} \right], \quad (1)$$

where K is the number of modes, $D_{pg,k}$ is the geometric mean diameter of mode k , σ_k is the geometric standard deviation of mode k , N_k is the total number concentration of mode k . The full parameter vector is therefore

$$\boldsymbol{\theta} = \{D_{pg,1}, \dots, D_{pg,K}, \sigma_1, \dots, \sigma_K, N_1, \dots, N_K\}.$$

Given a measured PNSDs $\{(D_{p,i}, n_i)\}_{i=1}^M$ consisting of M measurement bins, where n_i represents the measured concentration
 135 value at diameter $D_{p,i}$, the fitting procedure minimizes the mean squared error (MSE) between the measured and modeled distributions:

$$\text{MSE} = \frac{1}{M} \sum_{i=1}^M [n_i - \hat{n}(D_{p,i}; \boldsymbol{\theta})]^2. \quad (2)$$

The fitting parameters $\boldsymbol{\theta}$ are obtained by nonlinear optimization subject to physical constraints:

$$D_{pg,k} > 0, \quad \sigma_k > 1, \quad N_k > 0.$$

140 After the initial fit with a maximum number of modes $K_{\max}$, the algorithm automatically reduces the number of modes through two steps:

3.1.1 Elimination of invalid modes

Modes with negligible number concentration or physically invalid parameters are removed:

$$N_k \leq \varepsilon \quad \Rightarrow \quad \text{mode } k \text{ removed}, \quad (3)$$

145 where ε is a small numerical threshold.



3.1.2 Mode overlap reduction

Adjacent modes are tested for overlap using a tolerance-based criterion. If the reduction in fitting error after removing a mode is below a predefined tolerance (TOL_K), the simpler model is preferred:

$$\Delta MSE = MSE_{K-1} - MSE_K < TOL_K \Rightarrow K \leftarrow K - 1. \quad (4)$$

150 This reduction is performed iteratively until removing further modes exceeds the error tolerance.

3.2 AeroMiG-Aerosol Multi-mode Inverse-Gamma

DMPS instruments quantify aerosol particle mobility diameters $D_p > 0$, producing right-skewed, heavy-tailed size distributions (Colbeck and Lazaridis, 2014; Hinds and Zhu, 2022; Seinfeld and Pandis, 2016). In many atmospheric and combustion environments, the probability of observing extremely small particles declines sharply, whereas large particles occur with low
 155 but non-negligible probability (Whitby and McMurry, 1997; Curtius, 2006). Such behavior is characteristic of inverse-scale processes, motivating the use of the Inverse-Gamma distribution as a parametric model for particle diameters.

Let $Y \sim \text{Gamma}(\alpha, \beta)$ with shape $\alpha > 0$ and scale $\beta > 0$. Now consider the random variable $Y = \frac{1}{D_p}$, then the distribution of D_p is exactly the inverse of the gamma variable, which is Inverse-Gamma distribution. If $D_p = g(Y)$ where $g(y) = 1/y$, then $y = 1/D_p$ and the Jacobian $\left| \frac{dy}{dD_p} \right| = \frac{1}{D_p^2}$. Applying this transformation, we get

$$160 \quad f_{D_p}(D_p | \alpha, \beta) = f_Y(1/D_p) \left| \frac{d}{dD_p}(1/D_p) \right| \quad (5)$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\frac{1}{D_p} \right)^{\alpha-1} \exp\left(-\frac{\beta}{D_p}\right) \cdot \frac{1}{D_p^2} \quad (6)$$

$$= \frac{\beta^\alpha}{\Gamma(\alpha)} D_p^{-(\alpha+1)} \exp\left(-\frac{\beta}{D_p}\right), \quad D_p > 0. \quad (7)$$

This distribution is strictly supported on $(0, \infty)$ and exhibits a heavy right tail. Given a measured PNSDs $\{(D_{p,i}, n_i)\}_{i=1}^M$, the log-likelihood function is

$$165 \quad \ln L(\alpha, \beta) = \sum_{i=1}^M n_i \ln f(D_{p,i} | \alpha, \beta) \quad (8)$$

$$= \left(\sum_{i=1}^M n_i \right) (\alpha \ln \beta - \ln \Gamma(\alpha)) - (\alpha + 1) \sum_{i=1}^M n_i \ln D_{p,i} - \beta \sum_{i=1}^M \frac{n_i}{D_{p,i}} \quad (9)$$

Let $N' = \sum_{i=1}^M n_i$, then maximizing $\ln L$ with respect to β yields a closed-form estimator conditional on α such that

$$\hat{\beta} = \frac{N' \alpha}{\sum_{i=1}^M n_i D_{p,i}^{-1}} \quad (10)$$

The nonlinear equation for α becomes:

$$170 \quad \ln \hat{\beta} - \psi(\alpha) - \frac{1}{N'} \sum_{i=1}^M n_i \ln D_{p,i} = 0 \quad (11)$$



where $\psi(\cdot)$ is the digamma function. This equation requires numerical solution. The mean and variance are:

$$\mathbb{E}[D_p] = \frac{\beta}{\alpha - 1}, \quad \alpha > 1, \quad (12)$$

$$\text{Var}(D_p) = \frac{\beta^2}{(\alpha - 1)^2(\alpha - 2)}, \quad \alpha > 2. \quad (13)$$

The Probability Density Function (PDF) (Eq. 7) is commonly computed in log-space such that

$$\ln f(D_p | \alpha, \beta) = \alpha \ln \beta - \ln \Gamma(\alpha) - (\alpha + 1) \ln D_p - \frac{\beta}{D_p}, \quad (14)$$

which preserves floating-point accuracy in high-performance numerical kernels.

Direct evaluation of $\beta^\alpha D_p^{-\alpha-1} e^{-\beta/D_p}$ (Eq. 7) can be unstable when α is large, when D_p is very small (leading to overflow in $D_p^{-\alpha-1}$), or when β/D_p is large (leading to underflow). Therefore, the density is commonly computed using log-space decompositions, especially when high-performance numerical kernels are required. The normalization constant can be written

$$\frac{\beta^\alpha}{\Gamma(\alpha)} = \exp(\alpha \ln \beta - \ln \Gamma(\alpha)) \quad (15)$$

which avoids overflow in β^α and leverages the numerically stable computation of the logarithmic gamma function. Likewise, the power term is evaluated via

$$D_p^{-\alpha-1} = \exp((- \alpha - 1) \ln D_p) \quad (16)$$

to avoid under/overflow. Combining these terms, the entire probability density is evaluated through exponentiating the log-space formulation to preserve floating-point accuracy:

$$f(D_p | \alpha, \beta) = \exp(\alpha \ln \beta - \ln \Gamma(\alpha) - (\alpha + 1) \ln D_p - \frac{\beta}{D_p}) \quad (17)$$

Because α controls the decay rate of large-diameter occurrence and β sets a characteristic inverse scale, the distribution provides interpretable aerosol microphysical parameters.

190 Interpretability of AeroMiG

For the Inverse-Gamma distribution applied to DMPS mobility diameters D_p , the parameters α and β have physically interpretable aerosol micro-physical roles. The shape parameter α primarily controls the heaviness of the right (upper) tail of the distribution. Small values of α ($\alpha \lesssim 2$) correspond to broad size variability and an increased probability of large, aged, or coagulated particles, whereas larger α values indicate a more size-homogeneous aerosol population dominated by ultra-fine particles with limited coarse-mode contribution. Thus, α provides a statistical indicator of aerosol aging, coagulation intensity, and source mixing. The scale parameter β determines the characteristic diameter of the distribution. Because the mode of the Inverse-Gamma distribution is



$$D_{p,mode} = \frac{\beta}{\alpha + 1}, \quad \alpha > 1, \quad (18)$$

larger β values shift the particle-size distribution toward larger modal diameters, indicating an overall redistribution of particle mass toward larger sizes. While condensational growth affects particles across the entire size spectrum, such a shift in the mode may reflect the combined influence of growth processes, secondary formation, and source characteristics. Conversely, smaller β values are associated with freshly nucleated or combustion-derived ultra-fine particles. When $\alpha > 1$, the mean particle diameter is finite and equals

$$\mathbb{E}[D_p] = \frac{\beta}{\alpha - 1}, \quad (19)$$

indicating that the ratio $\frac{\beta}{\alpha - 1}$ can be interpreted as the expected mobility diameter. For $\alpha \leq 1$, the mean diverges, reflecting environments with highly intermittent coarse particles. Taken together, α describes distribution sharpness and tail behavior, while β establishes the dominant particle-size scale. Their combination provides a compact, interpretable parameterization of DMPS aerosol size structure, enabling comparison across sources, air masses, and atmospheric processes.

3.3 Performance metrics

The fitting problem is inherently ill-posed and multi-modal, where multiple parameter combinations may yield similar residual errors. Relying solely on the Mean Squared Error (MSE) to assess model performance can therefore favor over-parameterized solutions that reproduce the data numerically but lack physical interpretability (Bishop, 2016; Huber and Ronchetti, 2009; Murphy, 2012b; Zhao et al., 2019). To achieve a balanced evaluation that accounts for both accuracy and model parsimony, a composite performance metric integrating five complementary statistical measures was adopted: MSE, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), the coefficient of determination R^2 , and adjusted R^2 .

3.3.1 Mean Squared Error (MSE)

The MSE quantifies the average squared deviation between the measured and modeled PNSD across all particle diameter bins:

$$\text{MSE} = \frac{1}{N} \sum_{k=1}^N (y_k - \hat{y}_k)^2, \quad (20)$$

where y_k and $\hat{y}_k$ denote the measured and modeled particle concentrations at the k -th diameter bin, respectively. Thus, MSE provides a single scalar measure summarizing the overall discrepancy between the modeled and observed size distributions. Lower MSE values indicate closer agreement between modeled and observed particle number concentrations. However, because MSE typically decreases with increasing model complexity, it may not reliably indicate the optimal number of modes without additional model selection criteria (Nyaku et al., 2020; Ren et al., 2020; Hastie et al., 2009; Burnham and Anderson, 2010).



3.3.2 Akaike Information Criterion (AIC)

The AIC introduces an information-theoretic penalty for additional model parameters:

$$\text{AIC} = 2v - 2\ln(L) \quad (21)$$

where v is the number of fitted parameters and L is the maximum likelihood. When applying MSE as a fitting proxy, the negative log-likelihood is approximated as proportional to MSE under the assumption of approximately Gaussian residuals. This ensures that improvements in fit quality are justified by corresponding increases in model evidence rather than overfitting (Sugiura, 1978; Akaike, 1974; Burnham and Anderson, 2010).

3.3.3 Bayesian Information Criterion (BIC)

Similar to AIC but with a stronger penalty for model complexity, the BIC is expressed as:

$$\text{BIC} = v \ln(n) - 2\ln(L) \quad (22)$$

BIC tends to favor simpler models when data points (n) are limited, providing a safeguard against overfitting in the presence of noise or measurement uncertainty common in atmospheric datasets (Schwarz, 1978; Kass and Raftery, 1995; Burnham and Anderson, 2010).

3.3.4 Coefficient of Determination (R^2)

The coefficient of determination (R^2) quantifies how well the fitted model reproduces the observed PNSD across all diameter bins. Specifically, R^2 is computed by comparing the measured and modeled concentration values at each particle diameter $D_{p,k}$:

$$R^2 = 1 - \frac{\sum_{k=1}^N (y_k - \hat{y}_k)^2}{\sum_{k=1}^N (y_k - \bar{y})^2}, \quad (23)$$

where y_k and $\hat{y}_k$ denote the measured and modeled particle concentrations at the k -th diameter bin, respectively, and $\bar{y}$ is the mean of the measured distribution. This formulation treats the PNSD as a discrete function sampled over N diameter bins. Thus, R^2 provides a single scalar measure summarizing the goodness-of-fit over the entire size distribution, with higher values indicating that the model effectively captures the structural variability of the PNSD (Gao et al., 2023; Yang et al., 2012).

3.3.5 Adjusted R^2

Adjusted R^2 accounts for model complexity by correcting the coefficient of determination for the number of fitted parameters:

$$R_{\text{adj}}^2 = 1 - (1 - R^2) \frac{N - 1}{N - v - 1}, \quad (24)$$



where N is the number of diameter bins and v is the number of fitted parameters. This metric prevents artificial inflation of R^2 as additional modes are introduced and favors parsimonious models that achieve high explanatory power with fewer components (Wherry, 1931; Kutner et al., 2005; Yang et al., 2012).

255 3.3.6 Composite score

Due to the different scaling and boundedness of the performance metrics, direct comparison or aggregation is not meaningful. Therefore, each metric f_m is transformed into a bounded score using an independent monotonic transformation

$$S_m(f_m) \in (0, 1),$$

such that the resulting score preserves the ordering of model performance while mapping all metrics onto a common scale. For metrics where smaller values indicate better performance (MSE, AIC, BIC), a decreasing logistic transformation is applied:

$$S_m(f_m) = \frac{1}{1 + \exp(\delta g(f_m))} \quad (25)$$

where $\delta > 0$ controls the steepness of the transformation and $g(f_m)$ is a compression function defined as

$$g(f_m) = \log(1 + |f_m|).$$

260 This logarithmic compression reduces the dynamic range of large values while preserving the ordering of the metric. Thus, for the error-based metrics,

$$S_v = \frac{1}{1 + \exp(\delta \log(1 + \text{MSE}))} \quad (26)$$

$$S_a = \frac{1}{1 + \exp(\delta \text{sign}(\text{AIC}) \log(1 + |\text{AIC}|))} \quad (27)$$

$$S_b = \frac{1}{1 + \exp(\delta \text{sign}(\text{BIC}) \log(1 + |\text{BIC}|))} \quad (28)$$

The coefficients of determination R^2 and adjusted R^2 are naturally bounded above by unity and therefore require no additional transformation. Their normalized scores are simply defined as

$$S_r = R^2$$

$$S_{ar} = R_{adj}^2.$$



The final composite score integrates these five normalized metrics using empirically determined weights:

$$S_c = \frac{w_v S_v + w_a S_a + w_b S_b + w_r S_r + w_{ar} S_{ar}}{w_v + w_a + w_b + w_r + w_{ar}} \quad (29)$$

or equivalently

$$S_c = \frac{1}{\sum_j w_j} \sum_j w_j S_j, \quad (30)$$

where w_j denotes the weight assigned to the j – th metric and S_j its normalized score. The weights

$$(w_v, w_a, w_b, w_r, w_{ar}) = (0.2, 0.15, 0.15, 0.25, 0.25)$$

assign equal total importance to error minimization (MSE, AIC, BIC) (50%) and statistical robustness and explanatory power (R^2 , adjusted R^2) (50%). This balanced weighting strategy prevents over-parameterization while ensuring high fidelity to the underlying physical data structure (Burnham and Anderson, 2004). The composite score provides a bounded scalar measure of overall model performance:

$$0 < S_c < 1.$$

The model with the highest composite score is selected as the best-fit representation of the measured PNSD. This multi-criteria approach balances goodness-of-fit, model complexity, and explanatory power, ensuring that the selected solution achieves both high reconstruction fidelity and statistical parsimony. Such balanced evaluation is particularly important for atmospheric nucleation studies, where small numerical differences in fit quality can correspond to substantial differences in inferred particle formation and growth mechanisms.

4 Results and discussion

This section presents a comparative analysis between the proposed AeroMiG model and AeroMG (Hussein et al., 2005; Zhu and Wang, 2024). Furthermore, it discusses the practical advantages of AeroMiG within various real-world contexts.

4.1 Results

This subsection evaluates the performance of the models using multiple metrics, including predictive accuracies which are quantified by Mean Square Error (MSE), R^2 , and the composite score metric as well as computational efficiency. These evaluations were conducted using PNSD data collected across all SMEAR stations.

4.1.1 Predictive accuracies

Predictive accuracies serve as performance metrics to quantify the fitting accuracy of the models (AeroMiG vs. AeroMG) against measured data using MSE and R^2 . However, as the objective is not to overfit the data points, as discussed in the previous section, additional metrics such as adjusted R^2 , AIC, and BIC are incorporated into the composite score metric. This

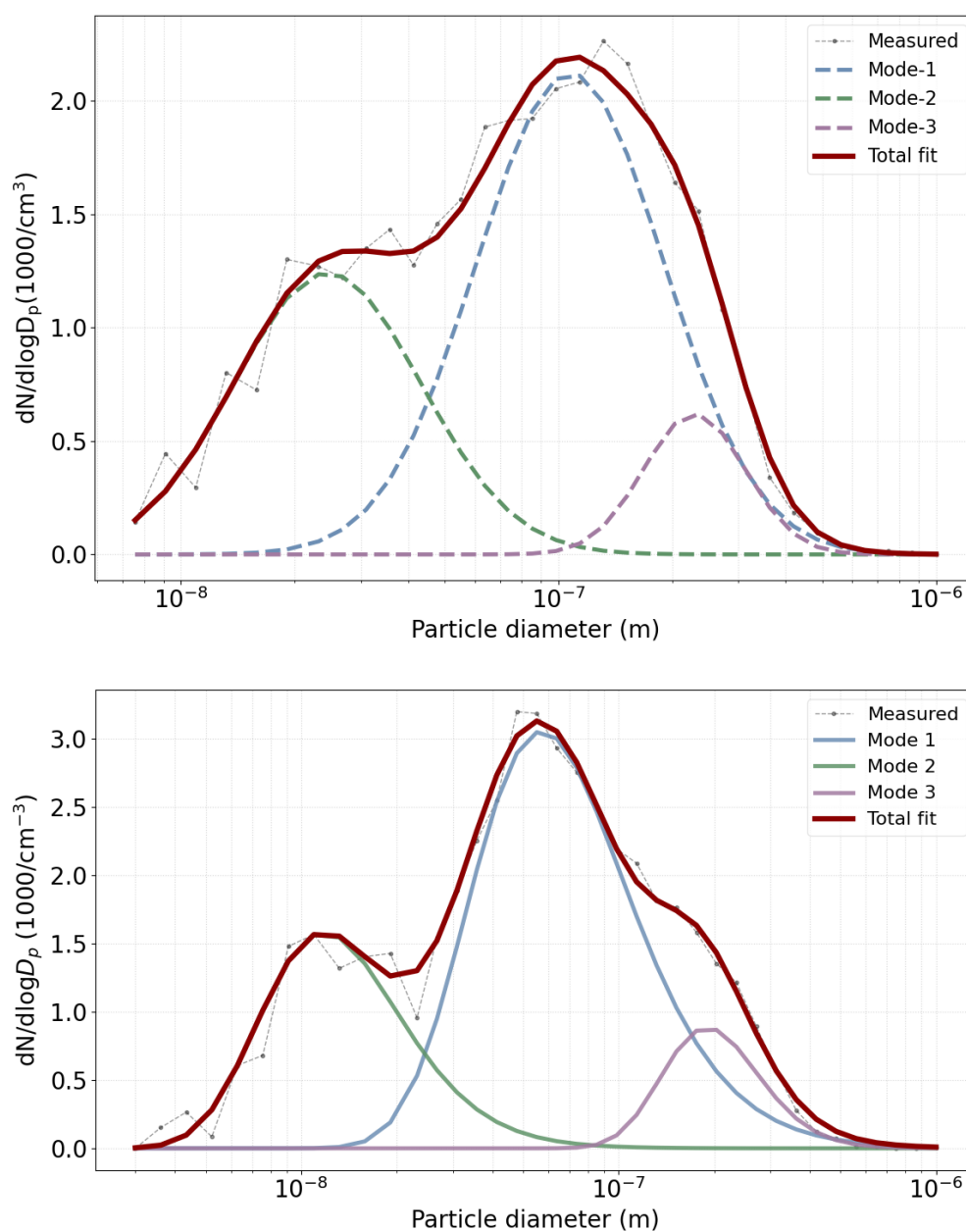


Figure 3. Multi-modal fitting profiles for a single continuous PNSD spectrum using the AeroMG (Figure top) and AeroMiG (Figure bottom).



ensures that the proposed AeroMiG method maintains strong generalization. The results for MSE, R^2 , and the composite score
 285 metrics are presented in Figs. 4, 5, and 6, respectively.

Before evaluating the MSE, R^2 , and the composite score metrics, we first investigate the underlying modal resolution
 capabilities of each algorithm by examining representative multi-modal fitting profiles for a single sample continuous spectrum
 using the AeroMG (Figure 3, top) and AeroMiG (Figure 3, bottom) frameworks. Both models decompose the empirical PNSD
 ($dN/d\log D_p$) into three discrete log-normal sub-modes (Mode 1, Mode 2, and Mode 3) across a particle diameter range
 290 from approximately 5×10^{-9} m to 10^{-6} m. The AeroMG framework achieves an approximate representation of the global
 distribution shape, tracking the overall envelope of the measured data points. However, it exhibits localized systemic misfits,
 notably overestimating the trough near 5×10^{-8} m and underestimating the maximum peak intensity near 1.5×10^{-7} m.
 Conversely, the AeroMiG framework demonstrates high structural flexibility, matching both the local fluctuations and the
 absolute peak amplitudes of the measured data points. The resulting total fit curve closely tracks the experimental data across
 295 the entire sub-micron spectrum. This improved continuity is supported by a distinct spatial allocation of the individual sub-
 modes, where the overlapping boundaries of Modes 1, 2, and 3 closely align with the structural variations present in the raw
 data points.

We assessed model reliability by plotting the MSE for every individual data point across three different stations (SMEAR I,
 SMEAR II, and SMEAR III). The fidelity of both methods (AeroMG and AeroMiG) was quantified by calculating the MSE
 300 for each fitted PNSD profile. As illustrated in the raincloud plot (Figure 4), AeroMiG demonstrates a lower median MSE and
 a highly localized error distribution compared to AeroMG. The distributional analysis shows that AeroMG exhibits a broad
 kernel density estimate with a substantial tail of high MSE values. This error tail for AeroMG is most pronounced at the
 SMEAR III station compared to SMEAR I and SMEAR II. Conversely, the high density of individual data points for AeroMiG
 remains concentrated at the lower error bound across all three stations.

To evaluate the overall goodness-of-fit across the monitoring stations, we analyzed the coefficient of determination (R^2)
 305 distribution for both methods (Figure 5). At all three sites (SMEAR I, SMEAR II, and SMEAR III), AeroMiG consistently
 demonstrates superior fitting performance compared to AeroMG, characterized by a higher median R^2 value and a narrower,
 more compressed interquartile range (IQR). The raincloud plots reveal a distinct left-skewed distribution for both models,
 with a high density of individual data points (rain) clustered near $R^2 = 1.0$. However, AeroMG exhibits a substantially longer
 310 and denser tail stretching down to the lower bound ($R^2 \approx 0.20$). This performance degradation in AeroMG is progressively
 exacerbated from SMEAR I to SMEAR III, as evidenced by the increasing density of low- R^2 outliers in the scatter components.
 In contrast, AeroMiG maintains a highly localized upper-bound density across all stations, restricting the majority of its data
 variance within a tight, high-performing envelope.

To comprehensively evaluate the models across multiple statistical dimensions, we computed a normalized multi-metric
 315 score profile encompassing goodness-of-fit (R^2 , Adjusted R^2), scaled information criteria (sAIC, sBIC), and scaled error vari-
 ance metrics (sMSE) across the three stations (Figure 6). The radar charts demonstrate that AeroMiG consistently outperforms
 AeroMG by enclosing a larger multi-dimensional surface area at all sites. While both models achieve nearly identical peak
 performance along the R^2 and Adjusted R^2 axes, their trajectories diverge substantially across the remaining parameters.

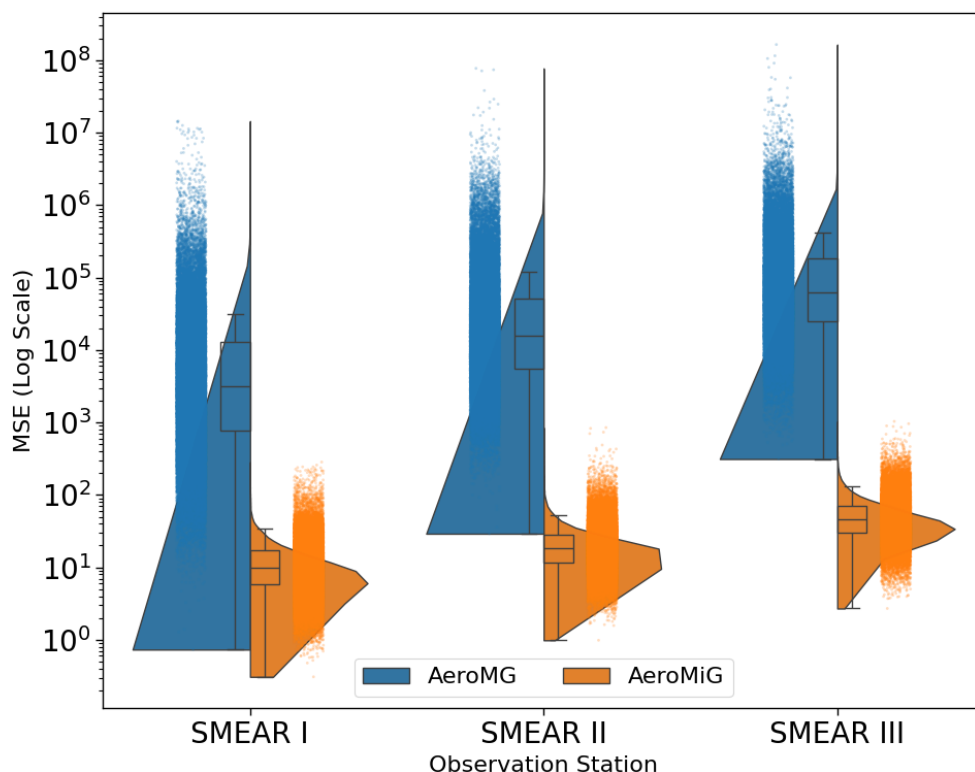


Figure 4. Raincloud plot of Mean Square Error (MSE) in different stations.

AeroMiG displays a marked expansion along the Score, sAIC, and sBIC axes. Notably, the separation between the two models becomes most pronounced in the sMSE and R^2 dimensions. This structural divergence is preserved across the gradient from the clean SMEAR I environment to the highly polluted SMEAR III station, where AeroMG's performance envelope shrinks further inward along the error-centric axes.

4.1.2 Computational efficiency

Computational efficiency evaluates the computational time (in seconds) required for each algorithm (AeroMG and AeroMiG) to process a single day of data (consisting of 144 measured data sets).

To analyze the computational efficiency and operational scalability of both frameworks, we evaluated the execution time distributions (in seconds) across the three tracking stations using letter-value plots (Figure 7). AeroMiG consistently demonstrates significantly lower execution times than AeroMG across all sites, reducing the median computational overhead by approximately 40% to 50%. At SMEAR I, SMEAR II, and SMEAR III, the median execution times for AeroMiG remain well below 200 seconds, whereas AeroMG centers around 330 to 350 seconds.

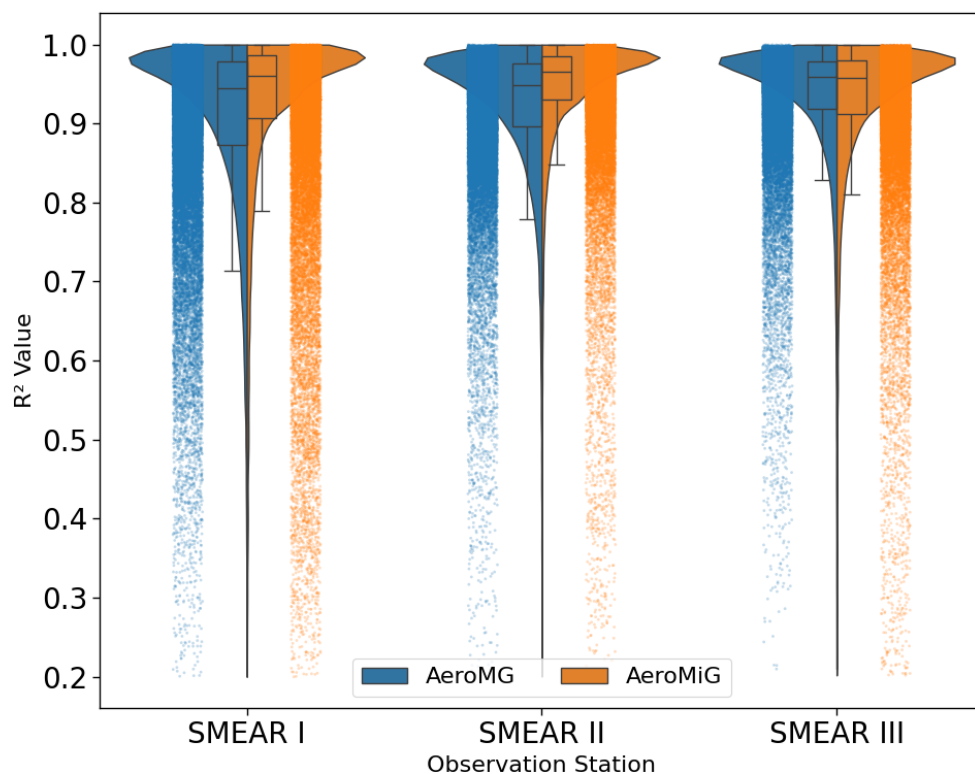


Figure 5. Raincloud plot of R-squared value in different stations.

Furthermore, the dispersion profile highlights that AeroMiG maintains a narrower and more tightly bounded distribution of computational cost. While both methods display symmetric distributions with a low density of extreme outliers stretching toward higher and lower computation limits, AeroMiG's upper confidence tiers scale favorably even under more volatile ambient data conditions. This computational advantage persists systematically across the entire station gradient, confirming that the algorithmic speedup is independent of location-specific data features.

4.2 Discussion

This subsection discusses the various advantages of the AeroMiG algorithm, specifically its consistent performance across different NPF event classifications. It further examines the model's interpretability and assesses its extended capability when extended to handle fitting scenarios involving more than five modes.

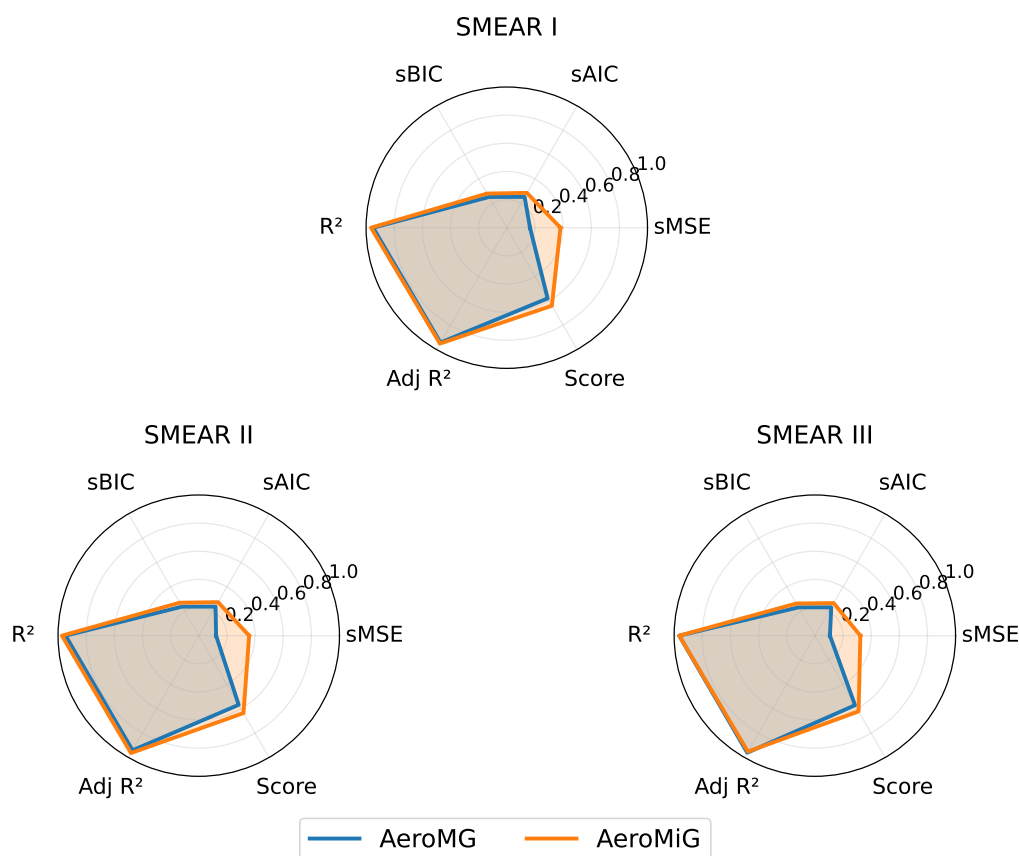


Figure 6. Polar-plot of performance matrix and their composite score in different stations.

340 4.2.1 Predictive accuracies across NPF classes

Predictive accuracies across NPF classes were evaluated to determine if the proposed AeroMiG algorithm maintains superior performance compared to AeroMG when both models were fitted to data categorized by NPF event class. Comparisons using MSE and R^2 metrics are presented via violin plots in Figure 8 and Figure 9, respectively. The comparative modal analysis reveals a fundamental divergence in how each framework resolves complex, overlapping aerosol modes. The systemic discrepancies in the AeroMG fit suggest a rigid optimization landscape that often converges on broad, overly smooth sub-modes. This stiffness limits its capacity to handle fine-grained structural fluctuations, leading to the localized over- and under-estimations observed in the accumulation and Aitken mode transition zones.

In contrast, the high fidelity achieved by AeroMiG highlights its superior mathematical adaptability. By allowing more precise, dynamically constrained parameterization of the sub-modes, AeroMiG effectively captures localized peaks and valleys without introducing unphysical mathematical oscillations. The model's ability to closely track the fine variations of the measured distribution confirms that its underlying algorithm successfully balances individual mode constraints with total fit op-

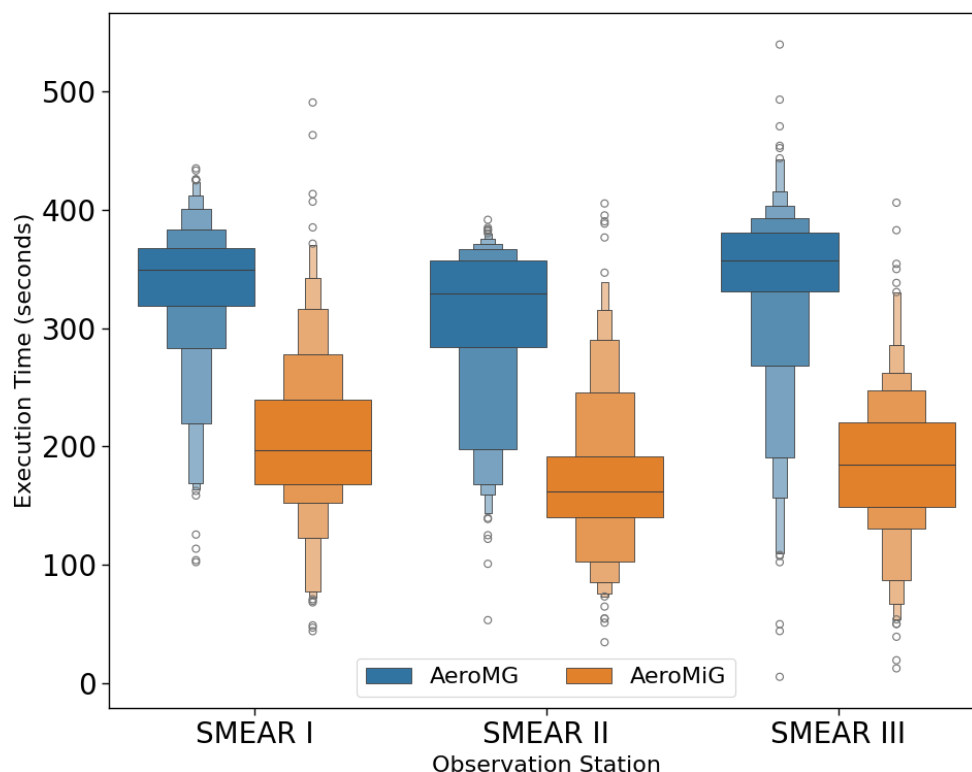


Figure 7. Box-plot of computational time (in seconds) in different stations.

timization. Consequently, AeroMiG provides a more physically representative decomposition of the aerosol spectrum, which is essential for accurately identifying specific atmospheric processes, such as fresh nucleation tracking, coagulation, and regional transport.

355 The broad error distribution of AeroMG (Figure 4) suggests that it frequently fails to capture three-mode fitting. This limitation is particularly evident at the SMEAR III station; because SMEAR III features higher pollution and noise levels, AeroMG struggles to accurately fit the noisy data. In contrast, AeroMiG's consistently low error variance across diverse aerosol concentrations confirms its superior robustness. Notably, AeroMiG handles polluted data (SMEAR II and III) significantly better than AeroMG. This reduction in error variance is critical for ensuring the overall stability of distribution fitting in highly variable
 360 environments.

To verify that the model enhancements are robust against varying atmospheric dynamics, we decoupled the fitting errors by regional aerosol event classification (Figure 8). The stratified distributional analysis reveals that AeroMiG consistently delivers superior, low-error performance ($MSE \approx 10^1$) across all event types (Ia, Ib, II, and nonevent) and at both monitoring stations (SMEAR I and SMEAR II). Conversely, AeroMG shows highly variable performance that degrades depending on the



event class. For class Ia events (characterized by highly distinct and clear multi-mode distributions) AeroMG fails completely to resolve lower-error fits, exhibiting truncated distributions bounded at a minimum MSE of approximately 10^2 to 10^4 . For classes Ib, II, and nonevents, AeroMG exhibits broad, bimodal error distributions with substantial density extending up to an MSE of 10^8 . This widespread variance indicates a frequent, systematic failure to accurately resolve PNSDs, whereas AeroMiG remains tightly bounded and invariant to the specific event conditions.

The comparative R^2 analysis (Figure 5) underscores the enhanced structural robustness of the AeroMiG framework. While both models achieve high target accuracies under ideal, low-noise conditions (reflected by the shared density peaks near unity), their resilience diverges sharply in challenging environments. The protracted low- R^2 tails observed in AeroMG signify recurrent optimization failures or oversimplifications during multi-modal PNSD fitting.

This vulnerability is highly correlated with site environmental characteristics. SMEAR III, typically characterized by elevated background noise and complex multi-modal pollution episodes, induces the highest rate of fitting failures for AeroMG. The fact that AeroMiG suppresses these low-accuracy tails at SMEAR III demonstrates its superior capacity to distinguish true physical aerosol features from complex observational noise. By maintaining structural stability and limiting error variance across diverse ambient concentrations, AeroMiG proves to be a significantly more reliable architecture for automated, long-term atmospheric data inversion.

To complement the error analysis, we evaluated the explanatory power of the models using the coefficient of determination (R^2) stratified by event type (Figure 9). AeroMiG displays exceptional explanatory fidelity across all classified atmospheric conditions, maintaining a tightly bounded distribution with a median $R^2 > 0.95$ and minimal variance at both SMEAR I and SMEAR II. In contrast, AeroMG exhibits structural performance degradation and substantial tailing. While its median remains high, the distribution features a prominent downward tail extending toward $R^2 = 0.0$. This degradation is particularly visible in class Ia events at the SMEAR II station, where the lower interquartile bound drops significantly compared to AeroMiG. The highly concentrated, sharp peaks of the AeroMiG violins close to 1.0 across all event classes demonstrate its superior consistency and statistical robustness in tracking complex, multi-modal aerosol dynamics.

The multi-metric (Figure 6) profile highlights that relying solely on conventional goodness-of-fit parameters like R^2 can mask underlying model inefficiencies. Although AeroMG matches the maximum fit accuracy of AeroMiG under optimal conditions, its reduced coverage across the scaled information criteria (sAIC, sBIC) and sMSE dimensions reveals a substantial penalty in model parsimony and error management.

The expanded penalization profile of AeroMG suggests a tendency toward overparameterization or optimization instability when resolving complex, multi-modal distributions. Conversely, AeroMiG's superior balance across all axes indicates that its algorithmic architecture achieves a more optimal trade-off between statistical complexity and predictive stability. Because this multi-dimensional superiority is robustly maintained even at the noise-heavy SMEAR III site, the framework demonstrates exceptional structural resilience, making it highly suited for autonomous, high-fidelity atmospheric data processing where data quality varies drastically.

The substantial reduction in execution time (Figure 7) achieved by AeroMiG underscores its algorithmic optimization and suitability for high-throughput atmospheric applications. In large-scale climate monitoring and observational networks, pro-

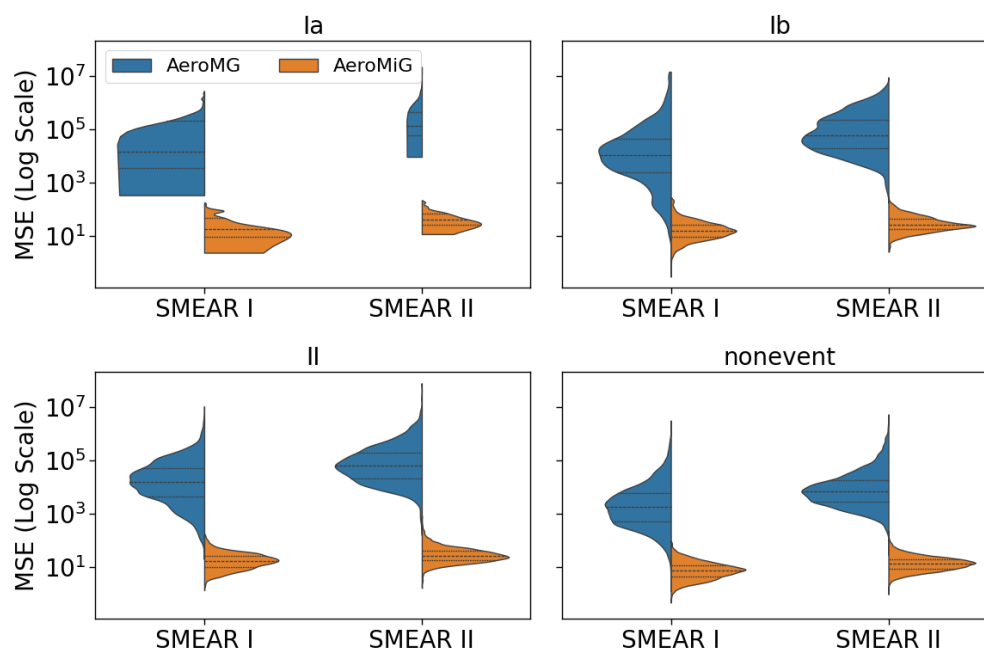


Figure 8. Evaluation of fitting performance across different aerosol event classes. Violin plots showing the distribution of Mean Squared Error (MSE) on a logarithmic scale for AeroMG (blue) and AeroMiG (orange) at the SMEAR I and SMEAR II stations. The distributions are stratified by regional particle formation event classes: Ia, Ib, II, and nonevent. Horizontal dashed lines within the violins indicate the medians and interquartile ranges (IQRs).

cessing multi-decadal PNSD data manually or with computationally heavy inversion techniques creates severe processing bottlenecks. By halving the computational footprint while simultaneously achieving superior fitting accuracy and structural stability (as demonstrated in the preceding R^2 and multi-metric analyses), AeroMiG breaks the traditional trade-off between statistical fidelity and computational cost.

This enhanced speed can likely be attributed to a more efficient convergence pathway within the AeroMiG optimization framework, preventing the algorithm from getting trapped in computationally expensive local minima or redundant iterations. Because the model limits execution times to a predictable and compressed envelope across different data types, it presents a highly scalable architecture. This computational efficiency is vital for enabling real-time or near-real-time automated data processing pipelines across global aerosol monitoring networks.

The close agreement between the data before fitting and after fitting demonstrates the structural fidelity of the numerical framework during dynamic atmospheric phenomena. Resolving NPF event types poses a substantial challenge for automated fitting routines because of the rapid, orders-of-magnitude changes in concentration and the shifting boundaries between nucleation, Aitken, and accumulation modes. The framework's ability to replicate the distinct asymmetry of Class Ia events without over-smoothing the peak concentration indicates an exceptional capability to handle highly localized gradients.

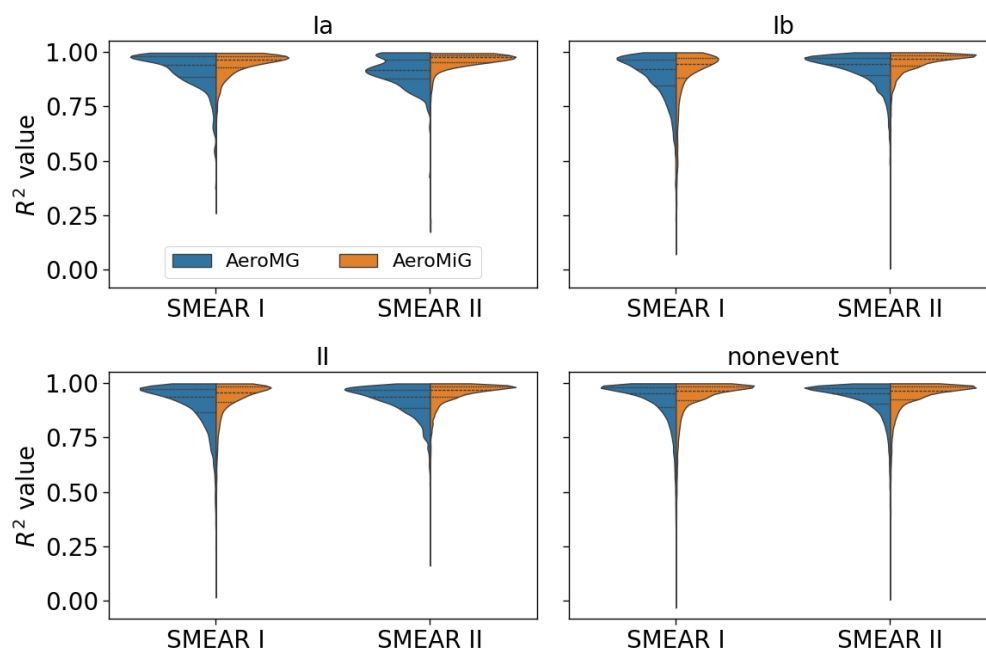


Figure 9. Goodness-of-fit distributions across aerosol event classes. Split-violin plots comparing the coefficient of determination (R^2) distributions for AeroMG (blue) and AeroMiG (orange) at the SMEAR I and SMEAR II stations. The data are stratified by atmospheric particle formation event classes (Ia, Ib, II, and nonevent). Horizontal dashed lines indicate the median and interquartile ranges (IQRs) for each model population.

Furthermore, preserving the precise variance envelopes (1σ) after fitting has major implications for downstream statistical
 415 and climate modeling. The high concentration variance observed in the nucleation mode during Class Ia episodes reflects
 the burst-like, discrete nature of regional nucleation and subsequent growth. By accurately capturing both the mean state and
 the underlying physical variance of the data (while separating them from the stable, accumulation-dominated background of
 nonevents), the model proves it can act as a reliable surrogate for automated event classification pipelines. This high-fidelity
 reproduction of the raw data trends and structural noise boundaries confirms the model's readiness for implementation in
 420 long-term, continuous global climate monitoring arrays.

4.2.2 Interpretability

Interpretability is a key feature that ensures AeroMiG remains transparent and accessible for analysis. While AeroMG is based
 on the log-normal (log-Gaussian) distribution, which is inherently easy to interpret as described in the previous section, the
 parameters of AeroMiG can also be mapped directly to mean and standard deviation values. As a case study, we aggregated
 425 all PNSD data from the SMEAR I station according to NPF classification, as shown in the upper panel of Figure 10. We then
 applied the AeroMiG algorithm to this aggregated data, with the results displayed in the bottom panel of Figure 10. The results

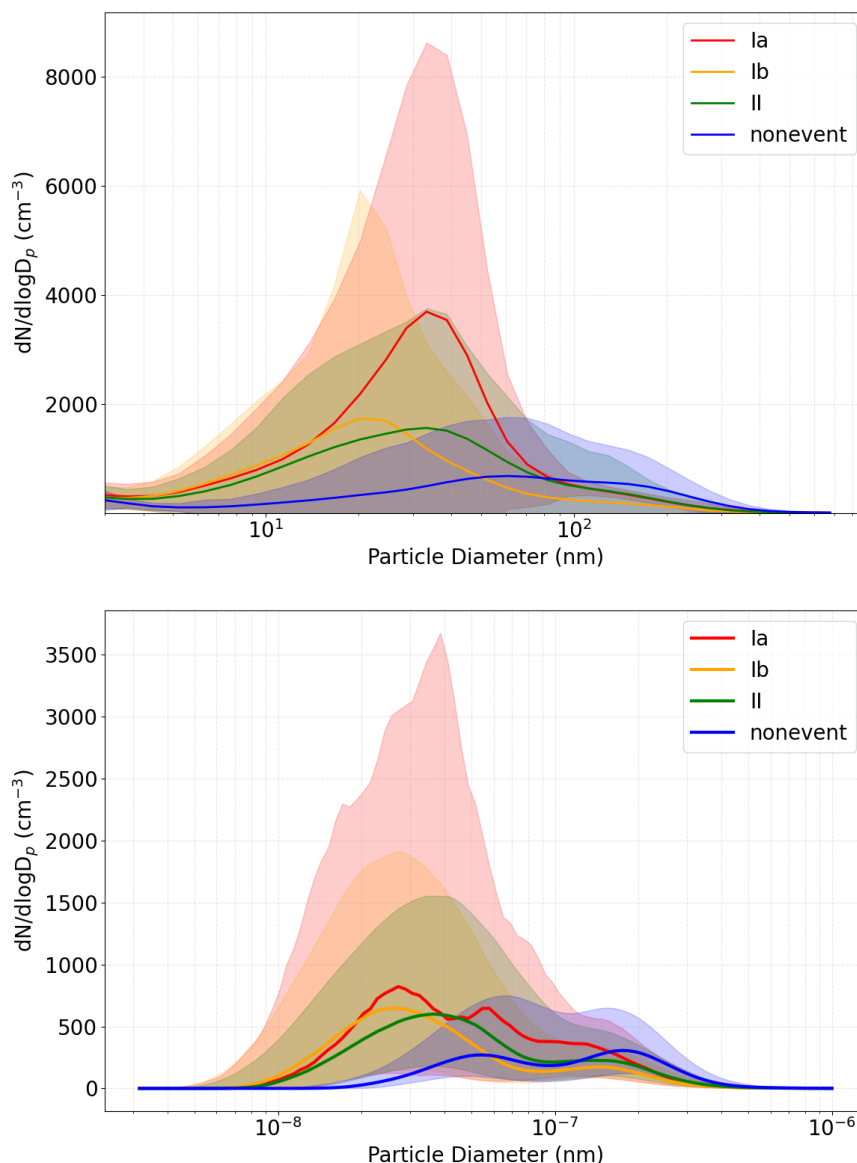


Figure 10. Multi-modal Inverse-Gamma parameterization of aerosol size distributions across NPF classes. Top, Observed sub-micrometer PNSDs ($dN/d\log D_p$) at the SMEAR II station, categorized by NPF event strength (Classes Ia, Ib, II) and nonevent days (solid lines indicate means; shaded regions indicate $\pm 1\sigma$). Down, Reconstructed distributions parameterized via multi-modal Inverse-Gamma functions. The mathematical transition from strong nucleation (Class Ia) to nonevent conditions is driven by a systematic reduction in the shape parameter ($\hat{\alpha}$) and an increase in the scale parameter ($\hat{\beta}$) across the distribution modes (Table 1). High $\hat{\alpha}$ values in Class Ia produce the sharp, tightly bounded peaks characteristic of highly coordinated regional nucleation. Conversely, the suppressed $\hat{\alpha}$ and elevated $\hat{\beta}$ values during nonevents broaden the distribution, mathematically capturing the transition from deterministic particle growth tracks to a highly dispersed, stochastic background aerosol regime.



demonstrate that the fitted curves provide a more structured representation of the physical properties inherent in PNSD data. Furthermore, we extracted the mean and standard deviation variables from the α and β parameters to demonstrate that every mode in AeroMiG is as interpretable as those in AeroMG.

430 To translate the model's latent feature space into physically interpretable physical states, we mapped the learned representations to Inverse-Gamma distributions across distinct NPF event classes (Table 1). While the estimated shape ($\hat{\alpha}$) and scale ($\hat{\beta}$) parameters define the mathematical topology of the distribution, we explicitly converted these parameters into the first two statistical moments (mean and standard deviation (std)) to provide a direct proxy for physical atmospheric dynamics. This conversion uncovers a clear parametric gradient that tracks NPF event strength and clarity: Class Ia and Ib (strong nucle-
 435 ation) are characterized by the lowest mean values (e.g., $\text{mean}_1 = 1.17 \times 10^{-7}$ for Ia) and tightly bounded standard deviations ($\text{std}_1 = 5.50 \times 10^{-8}$). Physically, these low-mean, low-variance profiles represent highly constrained atmospheric conditions. They signify a deterministic and uniform physical trigger, such as a sharp influx of precursor vapors paired with a low pre-existing condensation sink, which is characteristic of ideal "banana-shaped" growth tracks at the SMEAR II station.

As the NPF signatures (class II to nonevents, weak/no nucleation) degrade into nonevent days, the mean and standard
 440 deviation systematically increase across all three distribution modes. For instance, the first mode's mean more than doubles (2.64×10^{-7}), and its standard deviation nearly doubles (9.36×10^{-8}) on nonevent days. Physically, this expansion of the mean and variance captures the transition from highly synchronized aerosol nucleation to a stochastic, unconstrained atmospheric regime. The broader standard deviation on nonevent days directly quantifies the increased environmental noise, background aerosol fluctuations, and regional transport effects that suppress steady nucleation. By converting abstract algebraic parameters
 445 into tangible physical moments (mean and std), the model's internal representations become fully transparent, acting as a quantifiable diagnostic tool for atmospheric stability and particle distribution behavior.

To assess the capacity of the framework to resolve physical atmospheric features across diverse conditions, we analyzed the PNSDs ($dN/d\log D_p$) and associated variances (1σ shaded regions) before fitting (Figure 10 top) and after fitting (Figure 10 bottom), stratified by NPF event types (Class Ia, Ib, II, and nonevent).

450 The profiles before fitting reveal distinct, event-specific multi-modal structures. Class Ia events are characterized by a sharp, dominant nucleation/Aitken mode peaking near 30-40 nm with a maximum concentration exceeding 3500 cm^{-3} , accompanied by a broad, highly volatile envelope extending up to 8500 cm^{-3} . Class Ib and II events present progressively broader, suppressed nucleation peaks shifting toward slightly larger diameters (20-50 nm) and lower mean amplitudes. Conversely, the nonevent profiles exhibit a characteristic accumulation-mode bias, peaking weakly between 100 and 200 nm ($\sim 500 \text{ cm}^{-3}$)
 455 with minimal particle density under 20 nm.

A direct comparison with the profiles after fitting confirms that the framework replicates these highly non-linear, class-specific modal shapes and boundaries with high fidelity. The mathematical inversion accurately captures the steep concentration gradients and maps the shifting peak diameters across all event classes. Crucially, the 1σ variance bounds after fitting closely mirror the structural geometry of the data before fitting, capturing the high variability of active nucleation bursts (Class Ia)
 460 while preserving the tighter, more stable envelopes seen in nonevent regimes.

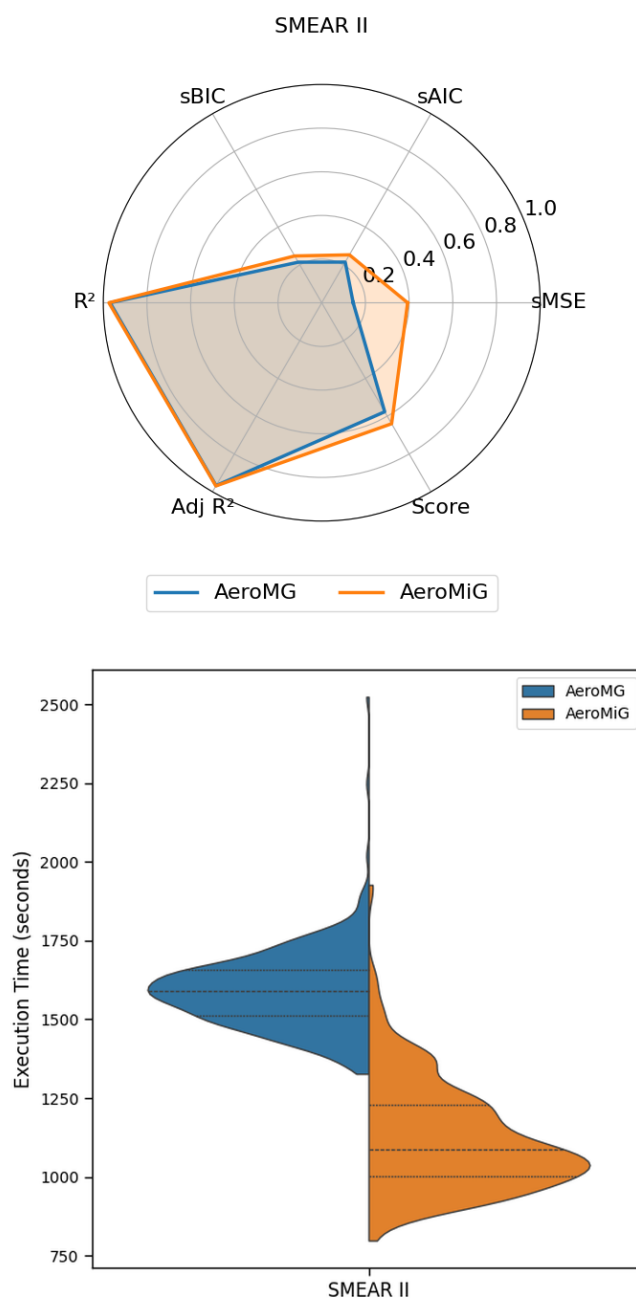


Figure 11. Multi-criteria performance trade-offs and computational efficiency for 5-mode fitting. Top, Radar chart evaluating overall model fit at the SMEAR II station across six standardized statistical metrics: scaled Bayesian Information Criterion (sBIC), scaled Akaike Information Criterion (sAIC), scaled Mean Squared Error (sMSE), holistic composite Score, Adjusted R^2 , and R^2 . Larger shaded areas indicate superior multi-objective performance. Down, Split-violin plot comparing the total execution time (seconds) distributions for AeroMG (blue) and AeroMiG (orange) optimization runs. Horizontal dashed lines denote medians and interquartile ranges (IQRs).



Table 1. Inverse-Gamma Parameters and Distribution Statistics (SMEAR II)

Metric	Ia	Ib	II	nonevent
$\hat{\alpha}_1$	$8.23 \times 10^{+0}$	$7.71 \times 10^{+0}$	$7.60 \times 10^{+0}$	$7.22 \times 10^{+0}$
$\hat{\beta}_1$	8.78×10^{-7}	8.76×10^{-7}	9.10×10^{-7}	1.20×10^{-6}
mean ₁	1.17×10^{-7}	1.85×10^{-7}	2.25×10^{-7}	2.64×10^{-7}
std ₁	5.50×10^{-8}	6.15×10^{-8}	6.45×10^{-8}	9.36×10^{-8}
$\hat{\alpha}_2$	$8.52 \times 10^{+0}$	$7.70 \times 10^{+0}$	$7.68 \times 10^{+0}$	$6.93 \times 10^{+0}$
$\hat{\beta}_2$	8.81×10^{-7}	8.77×10^{-7}	9.31×10^{-7}	1.09×10^{-6}
mean ₂	1.37×10^{-7}	1.74×10^{-7}	2.09×10^{-7}	4.32×10^{-7}
std ₂	5.19×10^{-8}	6.08×10^{-8}	6.61×10^{-8}	9.40×10^{-8}
$\hat{\alpha}_3$	$8.28 \times 10^{+0}$	$7.74 \times 10^{+0}$	$7.67 \times 10^{+0}$	$7.20 \times 10^{+0}$
$\hat{\beta}_3$	9.80×10^{-7}	9.58×10^{-7}	1.00×10^{-6}	1.23×10^{-6}
mean ₃	1.53×10^{-7}	1.72×10^{-7}	2.96×10^{-7}	2.92×10^{-7}
std ₃	6.95×10^{-8}	6.22×10^{-8}	6.47×10^{-8}	9.46×10^{-8}

4.2.3 Extended capability

Extended capability evaluate the AeroMiG algorithm under realistic field conditions, we utilized a multi-instrument field dataset collected at the SMEAR II station spanning a continuous year period from May 2019 to April 2020 ((Sulo et al., 2021)). The preprocessing framework dynamically unifies raw measurements from a Particle Size Magnifier (PSM), a Differential Mobility Particle Sizer (DMPS), and an Aerodynamic Particle Sizer (APS) into a single master time-series matrix. To resolve the sub-3 nm nucleation track, raw counts from the PSMs 1.1-1.7 nm tracking channel are normalized to a differential concentration $dN/d\log D_p$ and assigned to a representative geometric mean diameter (D_g) of 1.366 nm. For the upper coarse boundary, aerodynamic diameters from the APS are iteratively converted to mobility equivalents using the Cunningham slip correction factor assuming an effective particle density of 1.6 g/cm³ and a dynamic shape factor (χ) of 1.0, terminating at a maximum diameter of 15,074 nm (1.5073558×10^{-5} m). The resulting synchronized data array forms the basis for evaluating the five-mode algorithm configuration.

To evaluate the multi-objective trade-offs between fitting accuracy, model parsimony, and computational efficiency, we conducted a comprehensive evaluation at the SMEAR II station (Figure 11). The radar profile (Figure 11 top) demonstrates that AeroMiG outperforms or matches AeroMG across all six standardized statistical criteria. While both frameworks exhibit comparable raw explanatory power (reaching maximum values near 1.0 for R^2 and Adjusted R^2), AeroMiG achieves a markedly larger multi-objective area. This superiority is driven by its substantially higher composite performance Score and optimized information criteria (sAIC, sBIC, and sMSE), proving that its structural design better balances optimization depth with statistical penalization. Crucially, this expanded performance profile does not come at a computational premium. As shown by the



execution time distributions (Figure 11 bottom), AeroMiG provides an approximate 35% reduction in median processing time compared to AeroMG. The execution times for AeroMiG are tightly bounded, with a median of approximately 1000 s and an absence of high-latency tails. Conversely, AeroMG exhibits a higher median execution time (~ 1550 s) alongside a protracted high-latency tail stretching past 2500 s. This combined efficiency and optimization stability establishes AeroMiG as a highly scalable framework for 5-mode large-scale, high-throughput atmospheric data inversion.

4.3 Practical uses of the proposed method

The proposed AeroMiG framework enables rapid, stable characterization of PNSDs with efficient convergence over highly skewed structures, facilitating its seamless integration into real-time air quality systems for the immediate detection of sharp aerosol variations driven by pollution events like traffic emissions or industrial spikes. Because different aerosol sources emit highly distinct particle size profiles (Song et al., 2022), the compact modal descriptors extracted by AeroMiG can be utilized in downstream machine learning pipelines to automate aerosol classification and source attribution. This approach successfully addresses key challenges in urban environmental monitoring, such as differentiating between local, fresh vehicular emissions and regional, long-range-transported backgrounds solely from size distribution dynamics (Harni et al., 2024). Furthermore, by mapping out precise mathematical descriptors across varying locations and timeframes, the framework helps resolve broader global environmental challenges, including the historical difficulty of characterizing highly mismatched particle number concentrations versus mass concentrations across diverse global microenvironments (De Jesus et al., 2019).

5 Conclusions

This study addresses a central challenge in atmospheric data analysis such that the accurate representation of PNSDs, which are inherently skewed, heavy-tailed, and multi-modal. We introduced a flexible parametric framework, AeroMiG, that models these distributions as mixtures of Inverse-Gamma components, providing a more faithful representation of aerosol size structure than conventional lognormal-based approaches. Evaluation using real-world measurements from three independent monitoring stations demonstrates that the proposed method consistently improves fitting accuracy, reduces error, and enhances computational efficiency relative to the existing technique. These improvements are reflected across multiple performance metrics, including R^2 , adjusted R^2 , MSE, AIC, BIC, and a composite performance score, indicating both statistical robustness and practical reliability. Beyond performance gains, the proposed framework offers a conceptually meaningful representation of aerosol populations by capturing asymmetry and tail behavior in a physically interpretable manner. This enables more accurate characterization of aerosol processes and provides a scalable foundation for analyzing large observational datasets.

The broader significance of this work lies in its potential to support real-time air quality monitoring, improve data-driven atmospheric modeling, and enhance the integration of observational data into predictive systems. More generally, the framework provides a transferable approach for modeling complex, heavy-tailed distributions across diverse scientific and engineering domains. For example, capturing the heavy-tailed growth regimes of ultrafine particles improves estimates of NPF rates, lead to more realistic parameterizations of cloud condensation nuclei (CCN) activation in climate models, and refine assessment of



regional health risks driven by fine particulate exposure. Future work will focus on adaptive strategies for selecting the optimal number of mixture components and extending the framework to dynamic and high-dimensional environmental datasets.

Code and data availability. The open accessed data is available in SmartSMEAR (<https://smear.avaa.csc.fi>). Readers interested in accessing the data or code are encouraged to contact the corresponding authors.

515 **Appendix A: Probability distribution functions**

A1 Log-Gaussian distribution

The Gaussian (Normal) distribution is a fundamental continuous probability distribution widely familiar in statistical modeling and machine learning due to the symmetrical bell shape (Casella and Berger, 2024). The PDF for a random variable X following a normal distribution with mean μ and variance σ^2 is given by (Papoulis and Pillai, 1965):

$$520 \quad f(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

Let X be a positive random variable. If the logarithm of X follows a Gaussian distribution,

$$Y = \ln X \sim \mathcal{N}(\mu, \sigma^2), \quad (\text{A1})$$

then X is said to follow a Log-Gaussian (lognormal) distribution, denoted by

$$X \sim \text{LogNormal}(\mu, \sigma^2). \quad (\text{A2})$$

525 Using the change-of-variables theorem, the PDF of the Log-Gaussian distribution is given by

$$f(x | \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right), \quad x > 0. \quad (\text{A3})$$

Assume a sample of n independent and identically distributed (i.i.d.) observations

$$x_1, x_2, \dots, x_n \sim \text{LogNormal}(\mu, \sigma^2). \quad (\text{A4})$$

The likelihood function is then

$$530 \quad L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{x_i\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x_i - \mu)^2}{2\sigma^2}\right). \quad (\text{A5})$$

Taking the natural logarithm of the likelihood function yields

$$\ln L(\mu, \sigma^2) = \sum_{i=1}^n \left[-\ln x_i - \frac{1}{2} \ln(2\pi\sigma^2) - \frac{(\ln x_i - \mu)^2}{2\sigma^2} \right]. \quad (\text{A6})$$



Rearranging terms,

$$\ln L(\mu, \sigma^2) = - \sum_{i=1}^n \ln x_i - \frac{n}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (\ln x_i - \mu)^2. \quad (\text{A7})$$

535 Maximizing the log-likelihood yields the maximum likelihood estimators:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n \ln x_i, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (\ln x_i - \hat{\mu})^2, \quad (\text{A8})$$

which correspond to the sample mean and variance of the logarithmically transformed data.

A2 Inverse-Gamma distribution

540 The Inverse-Gamma distribution is a continuous probability distribution defined on the positive real line and is widely used in Bayesian statistics, particularly as a conjugate prior for variance parameters in Gaussian models (Gelman et al., 2014). Its heavy right tail makes it suitable for modeling positive-valued random variables with high variability whose reciprocal follows a gamma distribution (Gelman et al., 2014; Murphy, 2012a).

Let X be a positive random variable. We say that X follows an Inverse-Gamma distribution with shape parameter $\alpha > 0$ and
 545 scale parameter $\beta > 0$, denoted by $X \sim \text{Inv-Gamma}(\alpha, \beta)$, if its PDF is given by

$$f(x | \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right), \quad x > 0, \quad (\text{A9})$$

where $\Gamma(\cdot)$ denotes the Gamma function. Assume a sample of size n i.i.d. observations $x_1, x_2, \dots, x_n \sim \text{Inv-Gamma}(\alpha, \beta)$. The likelihood function is

$$L(\alpha, \beta) = \prod_{i=1}^n \frac{\beta^\alpha}{\Gamma(\alpha)} x_i^{-(\alpha+1)} \exp\left(-\frac{\beta}{x_i}\right). \quad (\text{A10})$$

550 Taking the natural logarithm of the likelihood function yields:

$$\ln L(\alpha, \beta) = \sum_{i=1}^n \left[\alpha \ln \beta - \ln \Gamma(\alpha) - (\alpha + 1) \ln x_i - \frac{\beta}{x_i} \right]. \quad (\text{A11})$$

Rearranging terms,

$$\ln L(\alpha, \beta) = n\alpha \ln \beta - n \ln \Gamma(\alpha) - (\alpha + 1) \sum_{i=1}^n \ln x_i - \beta \sum_{i=1}^n \frac{1}{x_i}. \quad (\text{A12})$$

Maximization of the log-likelihood with respect to the scale parameter β yields a closed-form estimator conditional on α :

$$555 \quad \hat{\beta} = \frac{n\alpha}{\sum_{i=1}^n x_i^{-1}}. \quad (\text{A13})$$

In contrast, maximization with respect to the shape parameter α leads to a nonlinear estimating equation involving the digamma function. Specifically, $\hat{\alpha}$ is obtained by solving



$$\ln \hat{\beta} - \psi(\alpha) - \frac{1}{n} \sum_{i=1}^n \ln x_i = 0, \quad (\text{A14})$$

where $\psi(\cdot)$ denotes the digamma function. This equation does not admit a closed-form solution and must be solved numerically.

560 The mean of X is defined as:

$$\mathbb{E}[X] = \int_0^{\infty} x f(x | \alpha, \beta) dx \quad (\text{A15})$$

$$= \int_0^{\infty} x \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{-(\alpha+1)} \exp\left(-\frac{\beta}{x}\right) dx. \quad (\text{A16})$$

Combining powers of x :

$$\mathbb{E}[X] = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \int_0^{\infty} x^{-\alpha} \exp\left(-\frac{\beta}{x}\right) dx. \quad (\text{A17})$$

565 Using the change of variables $u = \beta/x$, $dx = -\beta u^{-2} du$, the integral becomes a standard Gamma function:

$$\mathbb{E}[X] = \frac{\beta}{\Gamma(\alpha)} \int_0^{\infty} u^{\alpha-2} e^{-u} du \quad (\text{A18})$$

$$= \frac{\beta \Gamma(\alpha-1)}{\Gamma(\alpha)} = \frac{\beta}{\alpha-1}, \quad \alpha > 1. \quad (\text{A19})$$

The second moment is,

$$\mathbb{E}[X^2] = \int_0^{\infty} x^2 f(x | \alpha, \beta) dx \quad (\text{A20})$$

$$570 \quad = \frac{\beta^2}{\Gamma(\alpha)} \int_0^{\infty} u^{\alpha-3} e^{-u} du \quad (\text{A21})$$

$$= \frac{\beta^2 \Gamma(\alpha-2)}{\Gamma(\alpha)}, \quad \alpha > 2. \quad (\text{A22})$$

Thus, the variance is,

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \quad (\text{A23})$$

$$= \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}, \quad \alpha > 2. \quad (\text{A24})$$



575 **Appendix B: Optimization**

B1 AeroMG optimization

For a fixed number of modes K , parameter estimation is performed by minimizing the sum of squared residuals:

$$\theta_K^* = \arg \min_{\theta_K} \sum_{i=1}^M [n_i - \hat{n}(D_{p,i}; \theta_K)]^2. \quad (\text{B1})$$

580 This formulation corresponds to a nonlinear least-squares problem in linear concentration space. The least-squares objective is equivalent to maximizing the Gaussian likelihood under the assumption

$$n_i = \hat{n}(D_{p,i}; \theta_K) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad (\text{B2})$$

where the residuals ε_i are i.i.d.. After the initial K -mode fit, adjacent modes are tested for redundancy. Two neighboring modes k and $k + 1$ are considered overlapping if

$$|\ln D_{pg,k} - \ln D_{pg,k+1}| < \aleph \max(\ln \sigma_k, \ln \sigma_{k+1}), \quad (\text{B3})$$

585 where $\aleph$ is an empirical overlap threshold. If overlap is detected, the model is reduced to $K - 1$ modes and re-evaluated. The reduced model is accepted if

$$\mathcal{J}_{K-1} \leq \mathcal{J}_K (1 + \text{TOL}_K), \quad (\text{B4})$$

where $\mathcal{J}_K$ is the residual sum of squares for the K -mode model and TOL_K is a mode-dependent tolerance parameter. This reduction process is repeated iteratively until no further mode elimination satisfies the tolerance criterion.

590 **B2 AeroMiG optimization**

To estimate the parameters of the Inverse-Gamma model, the nonlinear, constrained optimization problem was solved using a Differential Evolution (DE) algorithm, which is a stochastic, population-based global optimization algorithm designed for continuous, nonlinear, and non-convex objective functions (Storn and Price, 1997; Price et al., 2005). Unlike gradient-based methods, DE does not require differentiability or convexity of the objective function and is therefore well suited for fitting
 595 multi-modal aerosol size distributions, where the loss surface is typically rugged, high-dimensional, and characterized by multiple local minima (Berkemeier et al., 1020). In this study, DE is employed to estimate the parameters of Inverse-Gamma mixture models describing PNSDs. The algorithm is used to minimize a scalar loss function quantifying the discrepancy between measured and modeled particle number concentrations.

Let $\theta \in \mathbb{R}^d$ denote the vector of model parameters in real numbers, where dimensional space d depends on the number of fitted modes. For each mode k , the parameter set includes

$$\theta_k = (\alpha_k, \beta_k, A_k),$$



representing the Inverse-Gamma shape, scale, and amplitude parameters, respectively. The optimization problem is defined as

$$\theta^* = \arg \min_{\theta \in \Omega} \mathcal{L}(\theta),$$

where $\mathcal{L}$ is the objective (loss) function and $\Omega \subset \mathbb{R}^d$ is a bounded feasible domain defined by physically meaningful constraints on aerosol parameters. DE maintains a population of N_p candidate solutions,

$$P^{(g)} = \{x_1^{(g)}, x_2^{(g)}, \dots, x_{N_p}^{(g)}\},$$

at generation g . Each individual vector is initialized by uniform random sampling within the parameter bounds. DE evolves the population using three fundamental operations: mutation, crossover, and selection.

B2.1 The objective function

Let $\mathbf{D}_p = (D_{p,1}, D_{p,2}, \dots, D_{p,M})$ denote the discrete particle mobility diameter bins measured by the DMPS instrument, and $\mathbf{y}^{\text{meas}} = (y_1, y_2, \dots, y_M)$ the corresponding measured PNSD values, where M is the total number of measurement bins. We assume that the aerosol population can be represented as a superposition of K modes, each parameterized by an Inverse-Gamma distribution with shape parameter α_k , scale parameter β_k , and amplitude A_k . The modeled distribution is given by

$$\mathbf{y}^{\text{model}}(D_p) = \sum_{k=1}^K A_k f_{\text{IG}}(D_p; \alpha_k, \beta_k),$$

where $f_{\text{IG}}(D_p; \alpha, \beta)$ denotes the Inverse-Gamma pdf,

$$f_{\text{IG}}(D_p; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} D_p^{-(\alpha+1)} \exp\left(-\frac{\beta}{D_p}\right), \quad D_p > 0.$$

The residual vector is defined component-wise as

$$r_i = \mathbf{y}^{\text{model}}(D_{p,i}) - y_i^{\text{meas}}, \quad i = 1, \dots, M.$$

The data misfit is quantified using either the mean squared error (MSE) or weighted least squares (WLS):

$$\mathcal{L}_{\text{data}} = \begin{cases} \frac{1}{M} \sum_{i=1}^M r_i^2, & \text{(MSE),} \\ \frac{1}{M} \sum_{i=1}^M \frac{r_i^2}{\max(y_i, \vartheta)}, & \text{(WLS),} \end{cases} \quad (\text{B5})$$

where $\vartheta = 10^{-9}$ is a positive constant introduced to prevent division by zero in low-concentration bins. To discourage redundant or strongly overlapping modes, a smooth separation penalty is introduced:

$$\mathcal{L}_{\text{sep}} = \lambda \sum_{k=1}^K \sum_{j=k+1}^K \exp\left(-\frac{|\beta_k - \beta_j|}{\kappa \bar{D}_p}\right),$$

where $\bar{D}_p = \frac{1}{M} \sum_{i=1}^M D_{p,i}$ is the arithmetic mean particle diameter across all bins, $\lambda = 0.01$ controls the penalty strength, and $\kappa = 0.1$ is a scaling factor determining the sensitivity of mode separation. This penalty softly penalizes modes with similar scale



parameters while preserving differentiability and numerical stability. The total objective function minimized by the optimizer is

$$\mathcal{J}(\theta) = \mathcal{L}_{\text{data}} + \mathcal{L}_{\text{sep}}, \quad (\text{B6})$$

where $\theta = \{(\alpha_k, \beta_k, A_k)\}_{k=1}^K$.

All parameters are constrained to physically meaningful domains:

$$\alpha_k > 0, \quad \beta_k > 0, \quad A_k \geq 0, \quad \forall k \in \{1, \dots, K\}.$$

610 Although maximum likelihood estimation is possible, a loss-based formulation was adopted due to zero inflation, non-Gaussian noise, and the presence of instrumental detection limits in DMPS data. The number of modes K is prescribed a priori, and optimization is performed over continuous distribution parameters only.

Algorithm B1 Objective Function Evaluation for Inverse-Gamma Mode Fitting

Require: Parameter vector $\theta = \{(\alpha_k, \beta_k, A_k)\}_{k=1}^K$, particle diameters $\mathbf{D}_p$, measured PNSD $\mathbf{y}_{\text{meas}}$, loss MAE, weights $w_{\text{data}}, w_{r,2}$

Ensure: Objective value $\mathcal{J}(\theta)$

1: **if** any $\alpha_k \leq 0$, $\beta_k \leq 0$, or $A_k < 0$ **then**

2: **return** large penalty

3: **end if**

4: Compute model prediction:

$$y^{\text{model}}(D_p) = \sum_{k=1}^K A_k f_{\text{IG}}(D_p; \alpha_k, \beta_k)$$

5: Compute residuals:

$$r_i = y^{\text{model}}(D_{p,i}) - y_i^{\text{meas}}$$

6: Compute data error: $E = \frac{1}{M} \sum_{i=1}^M |r_i|$

7: Compute coefficient of determination R^2

8: Combine loss:

$$\mathcal{L}_{\text{data}} = w_{\text{data}} \cdot E + w_{r,2} \cdot (1 - R^2)$$

9: Normalize:

$$\mathcal{L}_{\text{data}} = \frac{\mathcal{L}_{\text{data}}}{w_{\text{data}} + w_{r,2}}$$

10: Compute mode-separation penalty:

$$\mathcal{L}_{\text{sep}} = \lambda \sum_{k=1}^K \sum_{j=k+1}^K \exp\left(-\frac{|\beta_k - \beta_j|}{\kappa \bar{D}_p}\right)$$

11:

12: **return**

$$\mathcal{J}(\theta) = \mathcal{L}_{\text{data}} + \mathcal{L}_{\text{sep}}$$



B2.2 Optimization method

To prevent numerical instability in the Inverse-Gamma fitting procedure and to reduce the occurrence of spuriously low or negative values of the coefficient of determination (R^2), we implemented a multi-level adaptive stabilization strategy. For each measurement row, the model was first fitted using the original data. If this initial fit produced an unacceptable fit quality (negative R^2), the data were re-examined to determine whether zero or near-zero observations were contributing to instability in the likelihood evaluation and associated statistical metrics. To stabilize such cases, the measured signal was first regularized by the addition of a small positive offset ε . Unlike conventional single-shift approaches, the proposed method evaluates a sequence of progressively larger offsets. Let $y = [y_1, y_2, \dots, y_n]$ denote the measured signal, and let y_+ represent the subset of strictly positive values. The largest positive value in the row is defined as

$$y_{\max} = \begin{cases} \max(y_+) & \text{if } y_+ \neq \emptyset \\ 10^{-12} & \text{otherwise} \end{cases} \quad (\text{B7})$$

A discrete set of regularization levels is then introduced: $\varepsilon \in 0.01, y_{\max}; 0.05, y_{\max}; 0.10, y_{\max}$. For each ε , a modified signal

$$y^{(\varepsilon)} = y + \varepsilon \quad (\text{B8})$$

is constructed, and the full fitting procedure is repeated. Among the original and all regularized fits, the model yielding the highest R^2 value is retained. This ensures that additive stabilization is applied only when necessary and at the minimal magnitude required to restore numerical robustness. In highly polluted or coarse-particle regimes, where signal amplitudes are large and mean squared error (MSE) can become extreme, additive regularization alone may be insufficient. In such cases, a secondary stabilization step is applied in which the signal is scale-normalized prior to refitting. Specifically, the data are rescaled by the mean of the positive entries,

$$y^{(s)} = \frac{y}{\max(\bar{y}_+, 10^{-12})}, \quad (\text{B9})$$

and the model is refitted using the normalized signal. If this scale-normalized fit produces a higher R^2 than all previous attempts, it is selected as the final model.

Together, multi-level additive regularization addresses near-zero instability, while scale-normalized refitting resolves high-amplitude regimes problem under conventional Inverse-Gamma optimization. Because these corrective steps are applied adaptively and only to rows exhibiting poor initial fit quality, the procedure preserves the integrity of well-behaved measurements while substantially improving robustness across heterogeneous environmental conditions.

B2.3 Mathematical formulation of the Adaptive Stabilization Strategy

This appendix presents the formal mathematical description of the adaptive stabilization procedure employed to improve robustness of the AeroMiG Inverse-Gamma fitting under conditions of zero inflation and extreme signal amplitudes.



Algorithm B2 Multi-Level Adaptive Stabilization for AeroMiG Fitting

Require: Measurement vector $y \in \mathbb{R}^n$, fitting routine $\text{FitModel}(\cdot)$

Ensure: Best model fit BestFit

```

1: // Step 1: Initial fit
2:  $\text{Fit}_0 \leftarrow \text{FitModel}(y)$ 
3:  $R_0^2 \leftarrow \text{Fit}_0.R^2$ 
4: if  $R_0^2 \geq 0.30$  then
5:     return  $\text{Fit}_0$ 
6: end if
7: // Step 2: Determine positive scale reference
8:  $y_+ \leftarrow \{y_i \mid y_i > 0\}$ 
9: if  $y_+ \neq \emptyset$  then
10:     $y_{\max} \leftarrow \max(y_+)$ 
11: else
12:     $y_{\max} \leftarrow 10^{-12}$ 
13: end if
14: // Step 3: Multi-level epsilon candidates
15:  $\mathcal{E} \leftarrow \{0.01, 0.05, 0.10\} \cdot y_{\max}$ 
16:  $\text{BestFit} \leftarrow \text{Fit}_0$ ,  $\text{BestR2} \leftarrow R_0^2$ 
17: // Step 4: Epsilon-stabilized refits
18: for each  $\varepsilon \in \mathcal{E}$  do
19:     $y^{(\varepsilon)} \leftarrow y + \varepsilon$ 
20:     $\text{Fit}_\varepsilon \leftarrow \text{FitModel}(y^{(\varepsilon)})$ 
21:    if  $\text{Fit}_\varepsilon.R^2 > \text{BestR2}$  then
22:         $\text{BestFit} \leftarrow \text{Fit}_\varepsilon$ ,  $\text{BestR2} \leftarrow \text{Fit}_\varepsilon.R^2$ 
23:    end if
24: end for
25: // Step 5: Scale-normalized fallback for high-MSE cases
26: if  $\text{BestR2} < 0$  then
27:    if  $y_+ \neq \emptyset$  then
28:         $s \leftarrow \max(\text{mean}(y_+), 10^{-12})$ 
29:    else
30:         $s \leftarrow 1$ 
31:    end if
32:     $y^{(s)} \leftarrow y/s$ 
33:     $\text{Fit}_s \leftarrow \text{FitModel}(y^{(s)})$ 
34:    if  $\text{Fit}_s.R^2 > \text{BestR2}$  then
35:         $\text{BestFit} \leftarrow \text{Fit}_s$ 
36:    end if
37: end if
38: return  $\text{BestFit}$ 

```



Let $y = (y_1, y_2, \dots, y_n)$ denote the measured signal corresponding to a single row of the dataset, and let $\widehat{\mathcal{M}}(y)$ denote the fitted AeroMiG model obtained via the numerical optimization procedure. Each fitted model yields a coefficient of determination,

$$R^2(y) = R^2(\widehat{\mathcal{M}}(y)).$$

B2.4 Initial Fit

The fitting procedure begins with the unmodified signal,

$$\widehat{\mathcal{M}}_0 = \widehat{\mathcal{M}}(y), \quad R_0^2 = R^2(y).$$

645 If $R_0^2 \geq 0$, the fit is considered numerically stable and no further correction is applied.

B2.5 Multi-level additive regularization

If $R_0^2 < 0$, instability is assumed to arise from zero or near-zero observations. Define the set of strictly positive measurements,

$$y_+ = \{y_i \mid y_i > 0\}.$$

The characteristic scale of the row is defined as

$$650 \quad y_{\max} = \begin{cases} \max(y_+), & y_+ \neq \emptyset, \\ 10^{-12}, & y_+ = \emptyset. \end{cases}$$

A discrete set of additive regularization levels is introduced,

$$\mathcal{E} = \{0.01 y_{\max}, 0.05 y_{\max}, 0.10 y_{\max}\}.$$

For each $\varepsilon \in \mathcal{E}$, a regularized signal is constructed,

$$y^{(\varepsilon)} = y + \varepsilon,$$

655 and the corresponding fitted model is computed,

$$\widehat{\mathcal{M}}_\varepsilon = \widehat{\mathcal{M}}(y^{(\varepsilon)}), \quad R_\varepsilon^2 = R^2(y^{(\varepsilon)}).$$

Among the original and all regularized fits, the intermediate best model is selected as

$$\widehat{\mathcal{M}}^{(1)} = \arg \max_{\widehat{\mathcal{M}} \in \{\widehat{\mathcal{M}}_0\} \cup \{\widehat{\mathcal{M}}_\varepsilon : \varepsilon \in \mathcal{E}\}} R^2(\widehat{\mathcal{M}}).$$



B2.6 Scale-normalized fallback for high-amplitude regimes

660 In certain regimes characterized by extreme signal amplitudes (e.g., heavy pollution or coarse particle dominance), additive regularization alone may be insufficient. If the best fit obtained from the previous step still satisfies

$$R^2(\widehat{\mathcal{M}}^{(1)}) < 0,$$

a scale-normalized refitting step is applied.

The signal is normalized by the mean of its positive values,

$$665 \quad s = \max\left(\frac{1}{|y_+|} \sum_{y_i \in y_+} y_i, 10^{-12}\right), \quad y^{(s)} = \frac{y}{s}.$$

A final candidate model is then computed as

$$\widehat{\mathcal{M}}_s = \widehat{\mathcal{M}}(y^{(s)}), \quad R_s^2 = R^2(y^{(s)}).$$

The final model retained for the row is given by

$$\widehat{\mathcal{M}}^* = \arg \max_{\widehat{\mathcal{M}} \in \{\widehat{\mathcal{M}}^{(1)}, \widehat{\mathcal{M}}_s\}} R^2(\widehat{\mathcal{M}}).$$

670 This adaptive strategy ensures that: Stabilization is applied only when the initial fit is unstable; Additive regularization addresses near-zero numerical degeneracy; Scale normalization resolves extreme-amplitude regimes with large MSE; No global or arbitrary transformation is imposed on well-behaved data.

Empirically, this two-stage adaptive stabilization substantially reduces the occurrence of high-amplitude regimes problem across heterogeneous measurement regimes while preserving the physical interpretability of the fitted Inverse-Gamma parameters.
 675 eters.

Author contributions. AR and MAZ contributed to the conceptualization and visualization of the study; AR, SM, and MAZ developed the methodology and conducted the investigation; AR was responsible for data curation; AR, JK, SM, and MAZ performed the formal analysis; AR, JK, SM, TH, and MAZ carried out the validation; MAZ, TP, and ST managed the project administration, secured funding, and supervised the work; AR and MAZ prepared the original draft. All authors contributed to reviewing and editing the final manuscript.

680 *Competing interests.* The authors declare that they have no conflict of interest.

Acknowledgements. We acknowledge the use of OpenAI's ChatGPT-4 (<https://chat.openai.com/>) and Google's Gemini-3.5 Flash (<https://gemini.google.com/>) to refine the academic language and coherence of selected sections.



Financial support. This project is supported by the Research Council of Finland through the Academy Research Fellowship projects (Grant No. 355330 and 364223) and the University Profiling funding initiative InterEarth (Grant No. 353218), Center
685 of Excellence in Atmospheric Sciences (ACCS) (Grant No. 356134) and PROFi for the University of Eastern Finland (Grant No. 325022 and 352968).



References

- Aalto, P., Hämeri, K., Becker, E., Weber, R., Salm, J., Mäkelä, J. M., Hoell, C., O’dowd, C. D., Karlsson, H., Hansson, H.-C., Väkevä, M., Koponen, I. K., Buzorius, G., and Kulmala, M.: Physical characterization of aerosol particles during nucleation events, *Tellus B: Chemical and Physical Meteorology*, 53, 344, <https://doi.org/10.3402/tellusb.v53i4.17127>, 2001.
- Ahti, T., Hämet-Ahti, L., and Jalas, J.: Vegetation zones and their sections in northwestern Europe, *Annales Botanici Fennici*, 5, 169–211, <http://www.jstor.org/stable/23724233>, 1968.
- Akaike, H.: A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19, 716–723, <https://doi.org/10.1109/TAC.1974.1100705>, 1974.
- Akinyoola, J. A., Oluleye, A., and Gbode, I. E.: A Review of Atmospheric Aerosol Impacts on Regional Extreme Weather and Climate Events, *Aerosol Science and Engineering*, 8, 249–274, <https://doi.org/10.1007/s41810-024-00223-x>, 2024.
- Berkemeier, T., Takeuchi, M., Eris, G., and Ng, N. L.: Kinetic modeling of formation and evaporation of secondary organic aerosol from NO₃ oxidation of pure and mixed monoterpenes, *Atmospheric Chemistry and Physics*, 20, 15 513–15 535, <https://doi.org/10.5194/acp-20-15513-2020>, 1020.
- Bishop, C. M.: *Pattern Recognition and Machine Learning*, Information science and statistics, Springer New York, ISBN 978-1-4939-3843-8, softcover reprint of the original edition 2006 (corrected at 8th printing 2009), 2016.
- Burnham, K. P. and Anderson, D. R., eds.: *Model Selection and Multimodel Inference*, Springer New York, ISBN 978-0-387-95364-9, <https://doi.org/10.1007/b97636>, 2004.
- Burnham, K. P. and Anderson, D. R.: *Model selection and multimodel inference: a practical information-theoretic approach*, Springer, 2. edn., ISBN 978-0-387-22456-5 978-1-4419-2973-0, 2010.
- Casella, G. and Berger, R.: *Statistical inference*, CRC press, 2024.
- Christiansen, S., Salter, M. E., Gorokhova, E., Nguyen, Q. T., and Bilde, M.: Sea spray aerosol formation: Laboratory results on the role of air entrainment, water temperature, and phytoplankton biomass, *Environmental science & technology*, 53, 13 107–13 116, 2019.
- Colbeck, I. and Lazaridis, M.: *Aerosol science: Technology and applications*, John Wiley & Sons, 2014.
- Curtius, J.: Nucleation of atmospheric aerosol particles, *Comptes Rendus. Physique*, 7, 1027–1045, <https://doi.org/10.1016/j.crhy.2006.10.018>, 2006.
- Dal Maso, M., Kulmala, M., Riipinen, I., Wagner, R., Hussein, T., Aalto, P. P., and Lehtinen, K. E.: Formation and growth of fresh atmospheric aerosols: eight years of aerosol size distribution data from SMEAR II, Hyytiälä, Finland, *Boreal environment research*, 10, 323, 2005.
- De Jesus, A. L., Rahman, M. M., Mazaheri, M., Thompson, H., Knibbs, L. D., Jeong, C., Evans, G., Nei, W., Ding, A., Qiao, L., Li, L., Portin, H., Niemi, J. V., Timonen, H., Luoma, K., Petäjä, T., Kulmala, M., Kowalski, M., Peters, A., Cyrys, J., Ferrero, L., Manigrasso, M., Avino, P., Buonano, G., Reche, C., Querol, X., Beddows, D., Harrison, R. M., Sowlat, M. H., Sioutas, C., and Morawska, L.: Ultrafine particles and PM_{2.5} in the air of cities around the world: Are they representative of each other?, *Environment International*, 129, 118–135, <https://doi.org/10.1016/j.envint.2019.05.021>, 2019.
- Du, S., Li, T., Yang, Y., and Horng, S.-J.: Deep Air Quality Forecasting Using Hybrid Deep Learning Framework, *IEEE Transactions on Knowledge and Data Engineering*, 33, 2412–2424, <https://doi.org/10.1109/tkde.2019.2954510>, 2021.
- Gao, L., Chen, H., Chen, G., and Deng, J.: Particle Size Distributions and Extinction Coefficients of Aerosol Particles in Land Battlefield Environments, *Remote Sensing*, 15, 5038, <https://doi.org/10.3390/rs15205038>, 2023.
- Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.: *Bayesian data analysis*, Chapman and Hall/CRC, 2014.



- 725 Hari, P. and Kulmala, M.: Station for Measuring Ecosystem-Atmosphere Relations (SMEAR II), *Boreal Environment Research*, 10, 345–322, 2005.
- Hari, P., Nikinmaa, E., Pohja, T., Siivola, E., Bäck, J., Vesala, T., and Kulmala, M.: Station for Measuring Ecosystem-Atmosphere Relations: SMEAR, in: *Physical and Physiological Forest Ecology*, edited by Hari, P., Heliövaara, K., and Kulmala, L., pp. 471–487, Springer Netherlands, ISBN 978-94-007-5603-8, https://doi.org/10.1007/978-94-007-5603-8_9, 2013.
- 730 Harni, S. D., Aurela, M., Saarikoski, S., Niemi, J. V., Portin, H., Manninen, H., Leinonen, V., Aalto, P., Hopke, P. K., Petäjä, T., Rönkkö, T., and Timonen, H.: Source apportionment of particle number size distribution at the street canyon and urban background sites, *Atmospheric Chemistry and Physics*, 24, 12 143–12 160, <https://doi.org/10.5194/acp-24-12143-2024>, 2024.
- Hastie, T., Tibshirani, R., and Friedman, J.: *The Elements of Statistical Learning*, Springer Series in Statistics, Springer New York, ISBN 978-0-387-84857-0 978-0-387-84858-7, <https://doi.org/10.1007/978-0-387-84858-7>, 2009.
- 735 Hinds, W. C. and Zhu, Y.: *Aerosol technology: properties, behavior, and measurement of airborne particles*, Wiley, ISBN 978-1-119-49404-1, 2022.
- Huber, P. J. and Ronchetti, E. M.: *Robust Statistics*, Wiley Series in Probability and Statistics, Wiley, 1 edn., ISBN 978-0-470-12990-6 978-0-470-43469-7, <https://doi.org/10.1002/9780470434697>, 2009.
- Hussein, T., Puustinen, A., Aalto, P. P., Mäkelä, J. M., Hämeri, K., and Kulmala, M.: Urban aerosol number size distributions, *Atmospheric Chemistry and Physics*, 4, 391–411, <https://doi.org/10.5194/acp-4-391-2004>, 2004.
- 740 Hussein, T., Maso, M. D., Petäjä, T., Koponen, I. K., Paatero, P., Aalto, P. P., Hämeri, K. J., and Kulmala, M.: Evaluation of an automatic algorithm for fitting the particle number size distributions, *Boreal Environment Research*, 10, 337–355, <https://api.semanticscholar.org/CorpusID:73658875>, 2005.
- 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.
- 745 Intergovernmental Panel On Climate Change (Ipcc): *Climate Change 2021 – The Physical Science Basis: Working Group I Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, 1 edn., ISBN 978-1-009-15789-6, <https://doi.org/10.1017/9781009157896>, 2023.
- Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., and Kumar, V.: *Machine Learning for the Geosciences: Challenges and Opportunities*, *IEEE Transactions on Knowledge and Data Engineering*, 31, 1544–1554, <https://doi.org/10.1109/tkde.2018.2861006>, 2019.
- 750 Kass, R. E. and Raftery, A. E.: Bayes Factors, *Journal of the American Statistical Association*, 90, 773–795, <https://doi.org/10.1080/01621459.1995.10476572>, 1995.
- Kulmala, M., Vehkamäki, H., Petäjä, T., Dal Maso, M., Lauri, A., Kerminen, V.-M., Birmili, W., and McMurry, P.: 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.
- 755 Kutner, M., Nachtsheim, C., Neter, J., and Li, W.: *Applied statistical linear models*, McGraw Hill, 2005.
- Kyrö, E.-M., Väänänen, R., Kerminen, V.-M., Virkkula, A., Petäjä, T., Asmi, A., Dal Maso, M., Nieminen, T., Juhola, S., Shcherbinin, A., Riipinen, I., Lehtipalo, K., Keronen, P., Aalto, P. P., Hari, P., and Kulmala, M.: Trends in new particle formation in eastern Lapland, Finland: effect of decreasing sulfur emissions from Kola Peninsula, *Atmospheric Chemistry and Physics*, 14, 4383–4396, <https://doi.org/10.5194/acp-14-4383-2014>, 2014.
- 760



- Li, J., Carlson, B. E., Yung, Y. L., Lv, D., Hansen, J., Penner, J. E., Liao, H., Ramaswamy, V., Kahn, R. A., Zhang, P., Dubovik, O., Ding, A., Lacis, A. A., Zhang, L., and Dong, Y.: Scattering and absorbing aerosols in the climate system, *Nature Reviews Earth & Environment*, 3, 363–379, <https://doi.org/10.1038/s43017-022-00296-7>, 2022.
- Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., et al.: Climate change 2021: the physical science basis, Contribution of working group I to the sixth assessment report of the intergovernmental panel on climate change, 2, 2391, 2021.
- Murphy, K. P.: *Machine Learning: A Probabilistic Perspective*, MIT Press, 2012a.
- Murphy, K. P.: *Machine learning: a probabilistic perspective*, MIT Press, ISBN 978-0-262-30524-2, OCLC: 810414751, 2012b.
- Nyaku, E., Loughman, R., Bhartia, P. K., Deshler, T., Chen, Z., and Colarco, P. R.: A comparison of lognormal and gamma size distributions for characterizing the stratospheric aerosol phase function from optical particle counter measurements, *Atmospheric Measurement Techniques*, 13, 1071–1087, <https://doi.org/10.5194/amt-13-1071-2020>, 2020.
- Papoulis, A. and Pillai, S. U.: *Random variables and stochastic processes*, McGraw Hill, 1965.
- Pope Iii, C. A.: Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution, *JAMA*, 287, 1132, <https://doi.org/10.1001/jama.287.9.1132>, 2002.
- Price, K. V., Lampinen, J. A., and Storn, R. M.: *Differential Evolution: A Practical Approach to Global Optimization*, Natural Computing Series, Springer-Verlag Berlin Heidelberg, ISBN 978-3-540-20950-8 978-3-540-31306-9, <https://doi.org/10.1007/3-540-31306-0>, 2005.
- Raabe, O. G.: A general method for fitting size distributions to multicomponent aerosol data using weighted least-squares, *Environmental science & technology*, 12, 1162–1167, 1978.
- Ren, Y., Mao, J., Zhao, H., Zhou, C., Gong, X., Rao, Z., Wang, Q., and Zhang, Y.: Prediction of Aerosol Particle Size Distribution Based on Neural Network, *Advances in Meteorology*, 2020, 1–11, <https://doi.org/10.1155/2020/5074192>, 2020.
- Schwarz, G.: Estimating the Dimension of a Model, *The Annals of Statistics*, 6, <https://doi.org/10.1214/aos/1176344136>, 1978.
- Seinfeld, J. H. and Pandis, S. N.: *Atmospheric chemistry and physics from air pollution to climate change*, John Wiley & Sons, Hoboken, New Jersey, third edition. edn., ISBN 9781119221166, <https://ebookcentral.proquest.com/lib/ed/detail.action?docID=4462549>, 2016.
- Song, C., Becagli, S., Beddows, D. C., Brean, J., Browse, J., Dai, Q., Dall'Osto, M., Ferracci, V., Harrison, R. M., Harris, N., et al.: Understanding sources and drivers of size-resolved aerosol in the high Arctic islands of Svalbard using a receptor model coupled with machine learning, *Environmental Science & Technology*, 56, 11 189–11 198, 2022.
- Stocker, T., ed.: *Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, ISBN 978-1-107-05799-9, 2014.
- Storn, R. and Price, K.: Differential Evolution – A Simple and Efficient Heuristic for global Optimization over Continuous Spaces, *Journal of Global Optimization*, 11, 341–359, <https://doi.org/10.1023/A:1008202821328>, 1997.
- Sugiura, N.: Further analysis of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by akaike's, *Communications in Statistics - Theory and Methods*, 7, 13–26, <https://doi.org/10.1080/03610927808827599>, 1978.
- Sulo, J., Sarnela, N., Kontkanen, J., Ahonen, L., Paasonen, P., Laurila, T., Jokinen, T., Kangasluoma, J., Junninen, H., Sipilä, M., Petäjä, T., Kulmala, M., and Lehtipalo, K.: Long-term measurement of sub-3 nm particles and their precursor gases in the boreal forest, *Atmospheric Chemistry and Physics*, 21, 695–715, <https://doi.org/10.5194/acp-21-695-2021>, 2021.
- Uusitalo, H., Kontkanen, J., Ylivinkka, I., Ezhova, E., Demakova, A., Arshinov, M., Belan, B. D., Davydov, D., Ma, N., Petäjä, T., et al.: Occurrence of new particle formation events in Siberian and Finnish boreal forest, *Atmospheric Chemistry and Physics Discussions*, 2021, 1–21, 2021.



- Wherry, R. J.: A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation, *The Annals of Mathematical Statistics*,
 2, 440–457, <https://doi.org/10.1214/aoms/1177732951>, 1931.
- Whitby, E. R. and McMurry, P. H.: Modal Aerosol Dynamics Modeling, *Aerosol Science and Technology*, 27, 673–688,
<https://doi.org/10.1080/02786829708965504>, 1997.
- Whitby, K. T.: The physical characteristics of sulfur aerosols, *Atmospheric Environment* (1967), 12, 135–159, [https://doi.org/10.1016/0004-6981\(78\)90196-8](https://doi.org/10.1016/0004-6981(78)90196-8), 1978.
- 805 Xiong, C., Zheng, G., Zhang, Q., Li, Y., Wang, D., Wang, Z., Jiang, J., and He, K.: Characterizing Long-Term Multimodal Condensational
 Growth of Atmospheric Aerosols with the Synchronous Monitor of Aerosol Size and Hygroscopicity (SMASH), *Environmental Science
 & Technology*, 60, 18 767–18 777, 2026.
- Yang, X., Lee, J., Barker, D. E., Wang, X., and Zhang, Y.: Comparison of six particle size distribution models on the goodness-
 of-fit to particulate matter sampled from animal buildings, *Journal of the Air & Waste Management Association*, 62, 725–735,
 810 <https://doi.org/10.1080/10962247.2012.671148>, 2012.
- Zaidan, M. A., Haapasilta, V., Relan, R., Junninen, H., Aalto, P. P., Kulmala, M., Laurson, L., and Foster, A. S.: Predicting atmospheric
 particle formation days by Bayesian classification of the time series features, *Tellus B: Chemical and Physical Meteorology*, 70, 1530031,
<https://doi.org/10.1080/16000889.2018.1530031>, 2018.
- Zhao, S., Song, J., and Ermon, S.: InfoVAE: Balancing Learning and Inference in Variational Autoencoders, *Proceedings of the AAAI
 Conference on Artificial Intelligence*, 33, 5885–5892, <https://doi.org/10.1609/aaai.v33i01.33015885>, 2019.
- 815 Zhu, K. and Wang, L.: A comprehensive study on the validation and application of multi-lognormal distribution models for atmospheric
 particles, *Atmospheric Environment*, 338, 120 813, <https://doi.org/10.1016/j.atmosenv.2024.120813>, 2024.