Supplements of

Parameterization of particle formation rates in distinct atmospheric environments

Xinyang Li et al.

Correspondence to: xinyang Li (xinyang.li@helsinki.fi) and Lubna Dada (lubna.dada@helsinki.fi)

Contents of this file

About Tracer Model 5 Massively Parallel version (TM5-MP)

The global chemical transport model TM5-MP (Tracer Model 5, Massively Parallel version; using TM5 for short below; Williams et al., 2017; van Noije et al., 2021; Bergman et al., 2022) was applied to simulate the particle formation and growth in a global scale. The model is driven by the input meteorological datasets which are produced from ECMWF (European Centre for Medium-range Weather Forecasts) ERA5 reanalysis (Hersbach et al., 2020). In TM5, a detailed chemistry scheme including gas-phase, aqueous-phase and heterogeneous reactions are accounted for (more details refer to Williams et al., 2017 and van Noije et al., 2021). The aerosol dynamic processes are calculated with the two-moment modal model M7 (Vignati et al., 2004), which has seven log-normally distributed modes, including four water-soluble modes (nucleation, Aitken, accumulation and coarse) and three insoluble modes (Aitken, accumulation and coarse). The dry diameter range is less than 10 nm, 10 nm to 100 nm, 100 nm to 1 μm, and larger than 1 μm for nucleation mode, Aitken mode, accumulation mode and coarse mode, respectively. The aerosol components in the current TM5 version contain SOA, sulfate, ammonium, nitrate, methane sulfonic acid (MSA), primary organic aerosol, black carbon, sea salt, and mineral dust (Bergman et al., 2022). More detailed description of TM5 can be found in van Noije et al. (2021) and Bergman (2022), and below we only present the part which is most related to this study.

The new particle formation occurs in the nucleation mode, and is calculated in four phases (Bergman et al., 2022). First, the binary homogeneous water-sulfuric acid scheme (Vehkamäki et al., 2002) is applied to calculate the formation rate of 1 nm particles. Secondly, the growth scheme proposed by Kerminen and Kulmala (2002) is used to calculate how many particles survive during the growth from 1 nm to 1.7 nm. Thirdly, the formation rate of additional 1.7 nm particles is calculated by the sulfuric acid-ELVOC parametrization method in Riccobono et al. (2014) (Ricco2014 in the following). Finally, the growth of particles from 1.7 nm to 5 nm is calculated with the same growth scheme in Kerminen and Kulmala (2002). The ELVOCs are produced from the oxidation of two precursors, isoprene and monoterpenes, by OH and O_3 in the air. The sources of these two precursors are natural emissions calculated by MEGANv2.1 (Model of Emissions of Gases and Aerosols from Nature version 2.1; Guenther et al., 2012; Sindelarova et al., 2014) and biomass burning emissions based on van Marle et al. (2017).

In our sensitivity experiments, we assess the sensitivity of simulated aerosol size distributions to underlying nucleation scheme by replacing Ricco2014 by the four parameterizations (Eqs. 2-5). In addition, we run TM5 without any nucleation mechanism to quantify impact of nucleation. For each nucleation mechanism we perform global simulation with 2018 meteorological data. Resulting aerosol size distributions are evaluated against data retrieved from ACTRIS.

Figures:

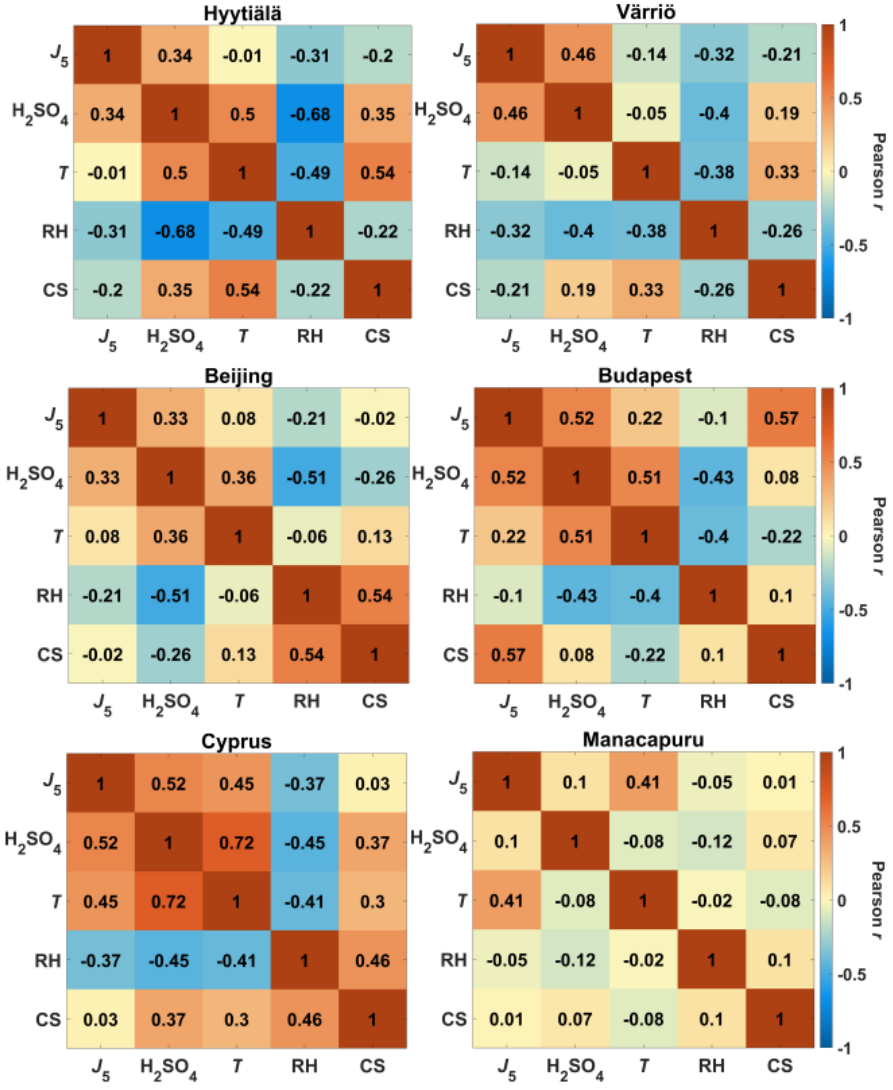

Figure S1. Heatmaps of variables for six measurement sites. The numbers shown in each block were Pearson correlation coefficients, and the color bars on the right side of each row represented the scale of the coefficients. Particle formation rate at 5 nm (J_5) is in cm^{-3} s^{-1}. H_2SO_4 is measured sulfuric acid concentrations in cm^{-3}. T is the ambient temperature in ^oC. Relative humidity (RH) is in %. Condensation sink (CS) is in s^{-1}. CS is not corrected for hygroscopic growth, since the correction factors were only developed for the boreal forest environment Hyytiälä, and they are not applicable for other sites.

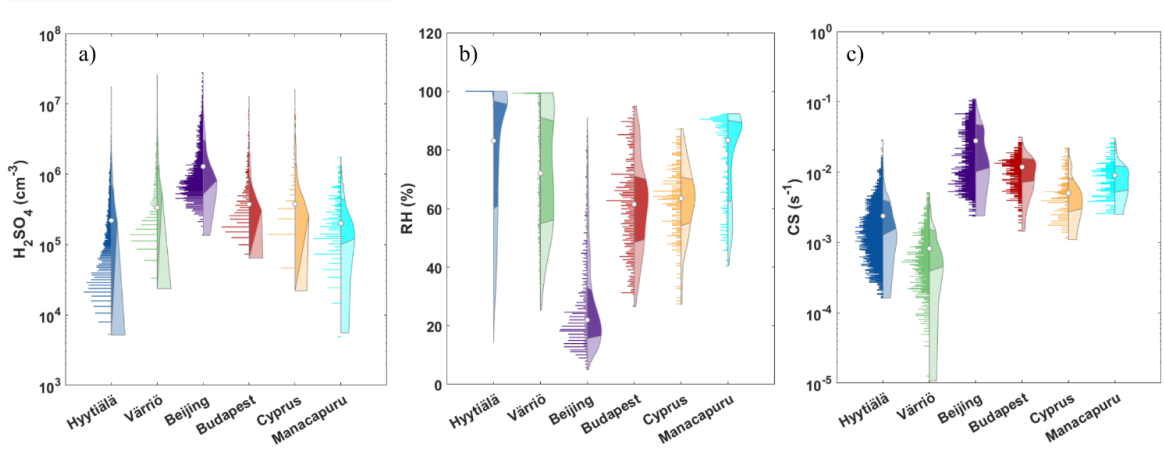

Figure S2. Violin and data distribution plots of input variables a) H_2SO_4 concentrations, b) relative humidity (RH) and c) condensation sink (CS) for each site. The white dots at the center of the violin plots represent the median values of the variable for the corresponding site. The shaded areas on the right side of each violin box cover the entire data distribution, as well as their 25th and 75th percentiles marked by the darker color of the box. The left side of the violin box shows the histograms of the variables.

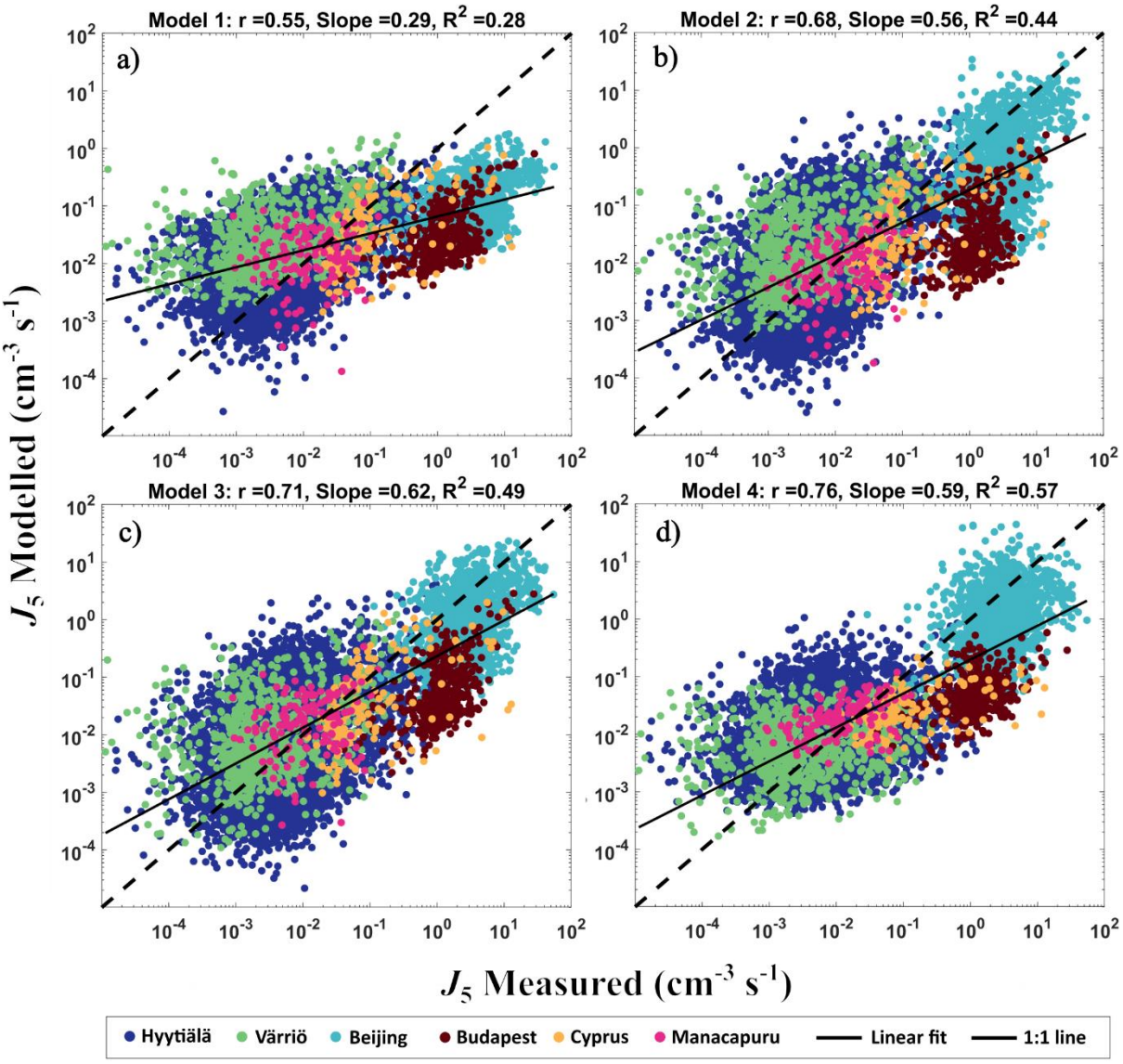

Figure S3. Modelled and measured J_5 scatterplots in logscale from four models using the training dataset containing data from all sites in hourly time resolution. Each color represents the data from one measurement site. The straight line showed the robust linear fit between the logscale modelled and the measure J_5 values, and the dashed line represented the 1:1 line. The correlation coefficient r, slope of linear fit, and the coefficient of determination R^2 are shown in the title of each subplot.

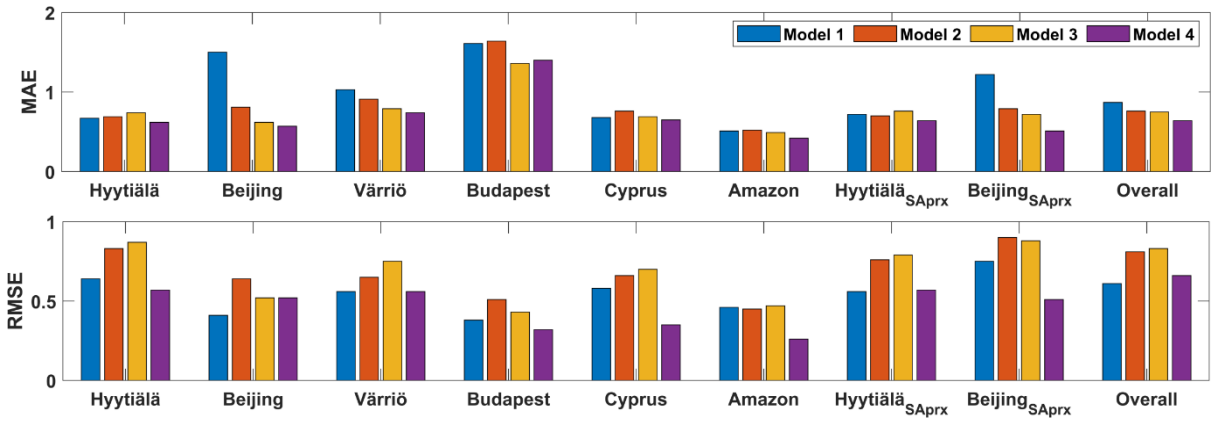

Figure S4. Summary of the mean absolute errors (MAE, upper panel) and the root mean absolute errors (RMSE, lower panel) from four models using testing dataset. The MAE and RMSE values were calculated between the logarithmic values modelled and measured J_5.

Tables:

Table S1. Summary of measurement sites and the instrumentations details for the measured variables. All the data were adjusted into local time for each measurement site.

Location	Type	Measurement Period	Particle Size Distribution	Trace Gases	Temperature (*T*) and relative humidity (RH)
Hyytiälä, Finland	Boreal forest (rural environment, biogenic emissions)	21 March 2016 – 18 August 2019	Twin-DMPS; ground level	H_2SO_4 (CI-APi-ToF); 35m a.g.l.	*T*: Pt100 inside custom shield, 16.8 m; RH: Rotronic MP102H RH sensor, 16.8 m
Beijing, China	Megacity (polluted, anthropogenic emissions)	29 May 2018 – 3 April 2019	3 – 10,000 nm, PSD system: a nano-DMA (3085, TSI), a long DMA (3081, TSI) and an Aerodynamic Particle Sizer (APS, TSI 3321); 12 m a.g.l.	H_2SO_4 (API-LToF-CIMS, model: Aerodyne); ~12 m a.g.l.	*T*, RH: Automatic Weather Station, model: AWS310, Vaisala Ltd.
Värriö, Finland	Boreal forest (cleaner rural environment than Hyytiälä, biogenic emissions)	5 April – 13 August 2019	DMPS with CPC (TSI 3776); ground level	H_2SO_4 (CI-Api-ToF)	*T*: PT-100 Rotronic MP106A, 2 m; RH: capacitive sensor, Rotronic MP106A, 2 m
Budapest, Hungary	City center (anthropogenic emissions)	22 March – 17 April 2018	6-1000 nm using flow-switching type DMPS	H_2SO_4 (CI-APi-ToF)	*T*, RH: BpART Lab
Agia Marina, Cyprus	Rural background (biogenic and anthropogenic emissions)	22 February – 4 March 2018	2-20 nm using Airel NAIS and 20-800 nm using TSI SMPS	H_2SO_4 (CI-APi-ToF)	*T*, RH: meteorological stations

Manacapuru, Brazil	Amazonian basin (biogenic emissions)	22 August – 9 October 2014	IOP1: SMPS (TSI model 3963); IOP2: CPC (TSI model 3772). (ARM, 2014c)	H_2SO_4 (SICIMS)	*T*, RH: Atmospheric Radiation Measurement (ARM) user facility, precision spectral pyranometer

Table S2. Summary of the mean, median, 5th-95th percentiles and standard deviation (SD) of sulfuric acid concentrations (H_2SO_4), relative humidity (RH), and condensation sink (CS) at all locations and time periods included in this study.

Locations		Hyytiälä, Finland	Beijing, China	Värriö, Finland	Budapest, Hungary	Agia Marina, Cyprus	Manacapuru, Brazil
Environment types		Rural boreal forest	Mega-city	Remote boreal forest	Urban background	Rural	Amazon basin
Measurement period		21 March 2016 – 18 August 2019	29 May 2018 – 3 April 2019	5 April – 13 August 2019	22 March – 17 April 2018	22 February – 4 March 2018	22 August 2014 – 9 October 2014
H_2SO_4 *(10^5 cm^{-3})*	mean	5.96	23.52	11.7	8.27	13.16	3.18
	median	2.55	12.43	3.37	3.74	3.76	2.04
	5th percentile	0.27	3.64	0.69	1.19	0.58	0.4
	95th percentile	21.52	68.6	43.97	28.56	62.67	10.13
	SD	9.96	27.88	25.17	13.15	22	3.02
T *(°C)*	mean	7.2	8.8	4.5	11.8	14	26.8
	median	9.1	7.5	4.5	11	13.2	25.6
	5th percentile	-10	-5.5	-5.2	1.9	11.6	23.2
	95th percentile	22	30	13.2	21.7	18	32.3
	SD	10.2	10.8	5.6	5.9	2.3	3.2
RH *(%)*	mean	75	27	72	61	63	78
	median	81	22	72	62	63	85
	5th percentile	35	11	40	35	42	51
	95th percentile	100	65	99	88	82	92
	SD	22	16	19	16	12	15
CS *(10^{-3} s^{-1})*	mean	3.02	32.12	1.07	11.82	6.33	9.7
	median	2.42	28.25	0.82	11.82	5.07	8.94
	5th percentile	0.49	5.27	0.14	4.17	1.77	3.18
	95th percentile	7.52	77.03	2.95	21.12	15.35	22.25
	SD	2.37	23.22	0.88	5.32	4.61	5.47

Table S3. Summary of mean absolute errors (MAE) and root mean squared errors (RMSE) from four models using testing set. The RMSE values were calculated using logarithmic scale from the predicted and measured J_5, and the number in brackets under the name of each site represents the count for data points. The unit of MAE and RMSE is the same as particle formation rate (J), and the magnitude of MAE and RMSE indicates the fitting level between the predicted and measured J_5 values. The higher MAE or RMSE represents that the model is under great influence from large prediction errors, which cause the over-/under-fitting of data and vice versa.

		Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), logscale								
	Models	Hyytiälä (1642)	Beijing (501)	Värriö (248)	Budapest (109)	Cyprus (34)	Manacapuru (40)	$Hyytiälä_{SA}$ (797)	$Beijing_{SA}$ (164)	Overall
MAE	1	0.67	1.50	1.03	1.61	0.68	0.51	0.72	1.22	0.87
	2	0.69	0.81	0.91	1.64	0.76	0.52	0.70	0.79	0.76
	3	0.74	0.62	0.79	1.36	0.69	0.49	0.76	0.72	0.75
	4	0.62	0.57	0.74	1.40	0.65	0.42	0.64	0.51	0.64
RMSE	1	0.64	0.41	0.56	0.38	0.58	0.46	0.56	0.75	0.61
	2	0.83	0.64	0.65	0.51	0.66	0.45	0.76	0.90	0.81
	3	0.87	0.52	0.75	0.43	0.70	0.47	0.79	0.88	0.83
	4	0.57	0.52	0.56	0.32	0.35	0.26	0.57	0.51	0.66

Table S4. Table of statistical parameters utilized to derive the Akaike Information Criterion (AIC) for boreal forests, Hyytiälä and Värriö. Model and equation numbers correspond to those in the main text (see Section 3). N_{train} represents the sample size of training set, and N_{test} for the testing set (Table 1). X denotes the counts for coefficient numbers. The correlation coefficient (r), slope, root mean square error (RMSE) and sum of squared error (SSe) were calculated based on a robust linear fit between the predicted and measured J_5 using the training set in logarithmic scale. AIC is calculated as AIC = $2X + N_{test} \ln(SSe)$. The relative likelihood term L ($L = e^{(AICmin-AICi)/2}$) indicates the likelihood of the ith model minimizing information loss. For example, for Hyytiälä, model 3 (Eq. 4) is approximately 3.26E-303 times more probable than model 4 (Eq. 5) in minimizing the information loss.

Boreal forests				
Total Training dataset number		10294		
Model number	1	2	3	4
X	1	2	3	4
	Hyytiälä			
N_{train}		5003		
N_{test}		1642		
r	0.43	0.47	0.37	0.31
Slope	0.43	0.62	0.48	0.25
RMSE	0.64	0.83	0.87	0.57
SSe	681.59	1125.79	1244.48	532.13
AIC	10715.1	11541.1	11707.7	10314.7

L	1.11E-87	4.84E-267	3.26E-303	1
Hyytiälä$_{SAprx}$				
N_{train}		-		
N_{test}		797		
r	0.41	0.48	0.36	0.33
Slope	0.34	0.57	0.42	0.28
RMSE	0.56	0.76	0.79	0.57
SSe	250.9	463.5	494.0	259.0
AIC	4405.4	4896.7	4949.5	4436.7
L	1	2.05E-107	7.07E-119	1.5E-07
Värriö				
N_{train}		728		
N_{test}		248		
r	0.44	0.47	0.35	0.18
Slope	0.32	0.40	0.32	0.12
RMSE	0.56	0.65	0.75	0.56
SSe	76.1	104.3	137.3	76.8
AIC	1076.2	1156.4	1226.6	1084.6
L	1	3.84E-18	2.17E-33	0.015

Table S5. Table of statistical parameters utilized to derive the Akaike Information Criterion (AIC) for urban environments, Beijing and Budapest. Model and equation numbers correspond to those in the main text (see Section 3). N_{train} represents the sample size of training set, and N_{test} for the testing set (Table 1). *X* denotes the counts for coefficient numbers. The correlation coefficient (*r*), slope, root mean square error (RMSE) and sum of squared error (*SSe*) were calculated based on a robust linear fit between the predicted and measured J_5 using the training set in logarithmic scale. AIC is calculated as AIC = $2X + N_{test} \ln(SSe)$. The relative likelihood term *L* ($L = e^{(AICmin-AICi)/2)}$) indicates the likelihood of the ith model minimizing information loss. For example, for Beijing, model 4 (Eq. 5) is approximately 3.10E-53 times more probable than model 1 (Eq. 2) in minimizing the information loss.

Urban				
Total Training dataset number		10294		
Model number	1	2	3	4
X	1	2	3	4
Bejing				
N_{train}		1342		
N_{test}		501		
r	0.30	0.32	0.30	0.22
Slope	0.33	0.57	0.43	0.28
RMSE	0.41	0.64	0.52	0.52

SSe	83.7	205.7	133.5	134.0
AIC	2220.0	2672.6	2458.0	2461.8
L	1	5.08E-99	2.01E-52	3.10E-53
Beijing$_{SAprx}$				
N_{train}		-		
N_{test}		164		
r	0.004	0.07	0.02	0.09
Slope	0.07	0.23	0.12	0.12
RMSE	0.75	0.90	0.88	0.51
SSe	92.0	130.3	125.1	42.2
AIC	3605.8	3885.1	3854.5	2991.2
L	3.42E-134	7.67E-195	3.36E-188	1
Budapest				
N_{train}		367		
N_{test}		109		
r	0.54	0.46	0.61	0.51
Slope	0.58	0.66	0.85	0.48
RMSE	0.38	0.51	0.43	0.32
SSe	15.2	28.4	20.0	11.0
AIC	298.4	368.6	332.7	268.9
L	4.05E-07	2.33E-22	1.41E-14	1

Table S6. Table of statistical parameters utilized to derive the Akaike Information Criterion (AIC) for Mediterranean rural environment and amazon basin, Cyprus and Manacapuru. Model and equation numbers correspond to those in the main text (see Section 3). N_{train} represents the sample size of training set, and N_{test} for the testing set (Table 1). X denotes the counts for coefficient numbers. The correlation coefficient (*r*), slope, root mean square error (RMSE) and sum of squared error (*SSe*) were calculated based on a robust linear fit between the predicted and measured J_5 using the training set in logarithmic scale. AIC is calculated as AIC = $2X + N_{test} \ln(SSe)$. The relative likelihood term *L* ($L = e^{(AICmin-AICi)/2}$) indicates the likelihood of the ith model minimizing information loss. For example, in Manacapuru, model 1 (Eq. 2) is approximately 4.6E-07 times more probable than model 4 (Eq. 5) in minimizing the information loss.

Rural				
Total Training dataset number		10294		
Model number	1	2	3	4
X	1	2	3	4
Cyprus				
N_{train}		140		
N_{test}		34		

r	0.42	0.49	0.38	0.37
Slope	0.35	0.48	0.37	0.17
RMSE	0.58	0.66	0.70	0.35
SSe	10.9	13.9	15.5	3.9
AIC	83.2	93.4	99.2	54.0
L	4.6E-07	2.78E-09	1.55E-10	1
Manacapuru				
N_{train}		140		
N_{test}		40		
r	0.04	0.19	0.21	0.46
Slope	0.04	0.15	0.18	0.24
RMSE	0.46	0.45	0.47	0.26
SSe	7.9	7.8	8.4	2.5
AIC	84.9	86.2	91.3	45.2
L	2.39E-09	1.26E-09	9.88E-11	1

Table S7. Information of the 14 global sites and the corresponding PNSD measurements, in alphabetic order. All the data are from the year 2018. The main instruments included differential mobility particle sizer (DMPS), scanning mobility particle sizer (SMPS), and tandem scanning mobility particle sizer (T-SMPS).

	Site	**Environment**	**Instrument**	**Size range (nm)**	**Measurement altitude (m)**	**Time resolution**	**Reference**
1	Birkenes	Rural	DMPS	10-800	220	357 sec	Yttri et al., 2021
2	Hyytiälä	Rural, boreal forest	DMPS	3-1000	179	10 min	Hari and Kulmala, 2005
3	K-puszta	Rural regional background	DMPS	6-800	125	10 min	Salma et al., 2016
4	La Reunion	Coastal	DMPS	10-600	2160	6 min	Foucart et al., 2018
5	Melpitz	Rural regional background	T-SMPS	5-800	87	20 min	Sun et al., 2024
6	Montseny	Urban regional background	SMPS	9-855	720	5 min	Minguillón et al., 2015
7	Mount Cimone	High mountain	SMPS	8-840	2165	357 sec	Brattich et al., 2020
8	Pallas	Arctic remote	DMPS	7-500	565	5 min	Aalto et al., 2002

9	Storm Peak	High mountain (limited vegetation coverage)	SMPS	8-333	3220	5 min	Borys and Wetzel, 1997
10	Trollhaugen	Antarctic	DMPS	10-800	1309	6 min	Hansen et al., 2009
11	Utö	Coastal	DMPS	7-500	7	10 min	Engler et al., 2007
12	Värriö	Arctic remote	DMPS	3-700	390	10 min	Aalto et al., 2002
13	Waldhof	Rural regional background	SMPS	10-800	74	5 min	Sun et al., 2024
14	Zeppelin	Polar	DMPS	5-710	473	20 min	Aalto et al., 2002

127

References

Bergman, T., Makkonen, R., Schrödner, R., Swietlicki, E., Phillips, V. T. J., Le Sager, P., and van Noije, T.: Description and evaluation of a secondary organic aerosol and new particle formation scheme within tm5-mp v1.2. Geoscientific Model Development, 15(2), 683–713, 2022.

Guenther, A. B., Jiang, X., Heald, C. L., Sakulyanontvittaya, T., Duhl, T., Emmons, L. K., and Wang, X.: The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions, Geosci. Model Dev., 5, 1471–1492, https://doi.org/10.5194/gmd-5-1471-2012, 2012.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society,146, 1999-2049, 2020.

Kerminen, V.-M. and Kulmala, M.: Analytical formulae connecting the "real" and the "apparent" nucleation rate and the nuclei number concentration for atmospheric nucleation events, J. Aerosol Sci., 33, 609–622, 2002.

Riccobono, F., Schobesberger, S., Scott, C. E., Dommen, J., Ortega, I. K., Rondo, L., Almeida, J., Amorim, A., Bianchi, F., Breitenlechner, M., David, A., Downard, A., Dunne, E. M., Duplissy, J., Ehrhart, S., Flagan, R. C., Franchin, A., Hansel, A., Junninen, H., Kajos, M., Keskinen, H., Kupc, A., Kürten, A., Kvashin, A. N., Laaksonen, A., Lehtipalo, K., Makhmutov, V., Mathot, S., Nieminen, T., Onnela, A., Petäjä, T., Praplan, A. P., Santos, F. D., Schallhart, S., Seinfeld, J. H., Sipilä, M., Spracklen, D. V., Stozhkov, Y., Stratmann, F., Tomé, A., Tsagkogeorgas, G., Vaattovaara, P., Viisanen, Y., Vrtala, A., Wagner, P. E., Weingartner, E., Wex, H., Wimmer, D., Carslaw, K. S., Curtius, J., Donahue, N. M., Kirkby, J., Kulmala, M., Worsnop, D. R., and Baltensperger, U.: Oxidation Products of Biogenic Emissions Contribute to Nucleation of Atmospheric Particles, Science, 344, 717–721, https://doi.org/10.1126/science.1243527, 2014.

Sindelarova, K., Granier, C., Bouarar, I., Guenther, A., Tilmes, S., Stavrakou, T., Müller, J.-F., Kuhn, U., Stefani, P., and Knorr, W.: Global data set of biogenic voc emissions calculated by the megan model over the last 30 years. Atmospheric Chemistry and Physics, 14(17), 9317–9341, 2014.

van Marle, M. J., Kloster, S., Magi, B. I., Marlon, J. R., Daniau, A. L., Field, R. D., Arneth, A., Forrest, M., Hantson, S., Kehrwald, N. M., Knorr, W., Lasslop, G., Li, F., Mangeon, S., Yue, C., Kaiser, J. W., and Van Der Werf, G. R.: Historic global biomass burning emissions for CMIP6 (BB4CMIP) based on merging satellite observations with proxies and fire models (1750-2015). Geoscientific Model Development, 10(9), 3329–3357, 2017.

van Noije, T., Bergman, T., Le Sager, P., O'Donnell, D., Makkonen, R., Gonçalves-Ageitos, M., Döscher, R., Fladrich, U., von Hardenberg, J., Keskinen, J.-P., Korhonen, H., Laakso, A., Myriokefalitakis, S., Ollinaho, P., Pérez García-Pando, C., Reerink, T., Schrödner, R., Wyser, K., and Yang, S.: EC-Earth3-AerChem: a global climate model with interactive aerosols and atmospheric chemistry participating in CMIP6. Geoscientific Model Development, 14(9), 5637–5668, 2021.

Vehkamäki, H., Kulmala, M., Napari, I., Lehtinen, K. E. J., Timmreck, C., Noppel, M., and Laaksonen, A.: An improved parameterization for sulfuric acid/water nucleation rates for tropospheric and stratospheric conditions. J. Geophys. Res., 107, 4622–4631, 2002.

Vignati, E., Wilson, J., and Stier, P.: M7: An efficient size-resolved aerosol microphysics module for large-scale aerosol transport models. Journal of Geophysical Research D: Atmospheres, 109(22), 1–17, 2004.

Williams, J. E., Folkert Boersma, K., Le Sager, P., and Verstraeten, W. W.: The high-resolution version of TM5-MP for optimized satellite retrievals: Description and validation. Geoscientific Model Development, 10(2), 721–750, 2017.