Supplement of Atmos. Meas. Tech., 18, 5763–5782, 2025
https://doi.org/10.5194/amt-18-5763-2025-supplement

Atmospheric Measurement Techniques
Open Access
EGU

Supplement of

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

Antti Vartiainen et al.

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

Table S1. Variables generated by ECHAM-HAMMOZ and considered as inputs to the ML models. Symbol * refers to the SALSA size bins 1a1-1a3, 2a1-2a7, and 2b1-2b7, except for the WAT variable, for which the insoluble type (b) is not defined. Symbol ** refers to five species of aerosol: dust (DU), sea salt (SS), black carbon (BC), organic carbon (OC), and sulfate (SO4). In addition to these five species, the list of variables marked with *** includes sulfur dioxide (SO2). emi_SS and emi_DU were not used, as their only value in all datasets was zero.

Category	**Name**	**Description**	**Units**
Size distributions	num_*	Particle number mixing ratio	1 / kg
	SO4_*	Sulfate mass mixing ratio	kg / kg
	OC_*	Organic carbon mass mixing ratio	kg / kg
	WAT_*	Aerosol water mass mixing ratio	kg / kg
Meteorological variables	ps	Surface air pressure	Pa
	rhoam1	Air density	kg / m^3
	geom1	Geopotential height	m^3 / s^2
	pbl_z	Boundary layer height	m
	aprl	Large scale precipitation	kg / m^2s
	aprc	Convective precipitation	kg / m^2s
	aprs	Snowfall	kg / m^2s
	wind_ew	East-west component of 10 meter wind	-
	wind_ns	North-south component of 10 meter wind	-
	wind_speed	Wind speed	m / s
	precip	Total precipitation	mm
	temp2	2 meter temperature	K
	tsurf	Surface temperature	K
	sradsd	Surface solar radiation, downward	W / m^2
Other variables	pm25 and pm25_simpler	$PM_{2.5}$ mass concentration (computed in two ways)	kg / m^3
	**	Mass concentration of five aerosol species (see caption)	kg / m^3
	emi_***	Emissions of six aerosol species (see caption)	kg / m^2s
	burden_***	Burdens of six aerosol species	kg / m^2
	PM25_**	$PM_{2.5}$ mass concentration of five aerosol species	kg / m^3
	time_ws	Winter-summer component of cyclical time	-
	time_sa	Spring-autumn component of cyclical time	-
	TAU_2D_550nm	Aerosol optical thickness at 550 nm	-

Table S2. Optimized hyperparameters and their search bounds for GLM. Model-specific feature selection was performed during hyperparameter optimization through two additional parameters, rel_thresh and red_thresh. The number of features after selection, controlled by the threshold parameters, is also reported in parentheses. A dash in the threshold columns indicates that no feature selection was done. For hyperparameters that were optimized logarithmically, the bounds are written in the exponent. *power* is a categorical hyperparameter, taking a value of 0 or 2. The selected value of *power* is indicated by having a larger number in the corresponding *power* column (e.g., 1 means selected and 0 not selected). Bold text means that GLM was the best model for the dataset in question.

	alpha	power_0	power_2	red_thresh	rel_thresh	tol
Bounds	$10^{[-2,1]}$	[0, 1]	[0, 1]	[0.45, 0.99]	[0, 0.25]	$10^{[-4,-1]}$
Helsinki Acc	1.19	0.414	0.367	0.894	0.074 (58)	0.0
Helsinki Ait	0.024	0.245	0.556	0.918	0.144 (35)	0.053
Helsinki Nuc	0.171	0.0	1.0	0.773	0.026 (41)	0.078
Leipzig Acc	**0.01**	**1.0**	**0.0**	**0.492**	**0.01 (19)**	**0.001**
Leipzig Ait	1.22	0.0	1.0	0.844	0.03 (49)	0.0
Melpitz Acc	0.125	1.0	0.0	0.817	0.25 (36)	0.001
Melpitz Ait	4.14	1.0	0.0	-	- (100)	0.009
Melpitz Nuc	2.43	0.438	0.164	-	- (100)	0.031

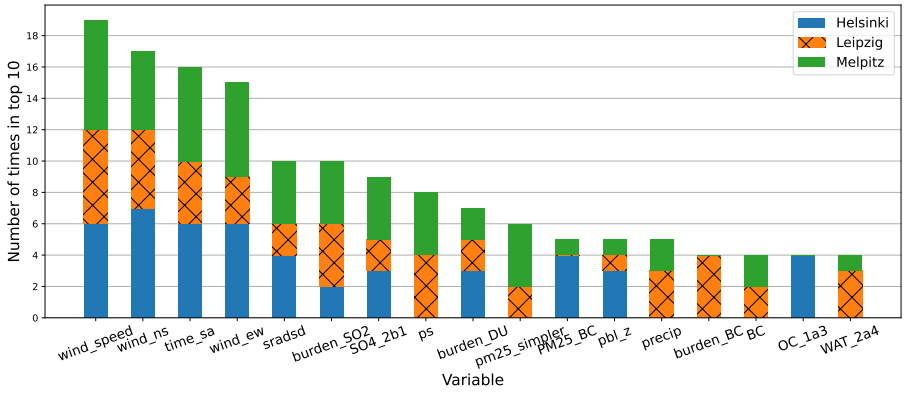

Figure S1. Most important input variables across all three accumulation subrange datasets, measured by mean absolute SHAP values. All seven ML models were analyzed for each dataset. Therefore, the maximum possible height of a bar is 21. Bars with height less than four are not shown.

Table S3. Optimized hyperparameters and their search bounds for RF. Model-specific feature selection was performed during hyperparameter optimization through two additional parameters, rel_thresh and red_thresh. The number of features after selection, controlled by the threshold parameters, is also reported in parentheses. A dash in the threshold columns indicates that no feature selection was done.

	max_depth	max_features	min_impurity_decrease	n_estimators	red_thresh	rel_thresh
Bounds	[1, 50]	[0.01, 1]	[1, 10000.0]	[50, 300]	[0.45, 0.99]	[0, 0.25]
Helsinki Acc	31.3	0.729	3690	60.8	-	- (100)
Helsinki Ait	7.59	0.302	9410	184	0.913	0.208 (16)
Helsinki Nuc	48.0	0.067	162	229	-	- (100)
Leipzig Acc	11.9	0.389	310	176	0.578	0.106 (21)
Leipzig Ait	6.69	0.975	9820	53.9	-	- (100)
Melpitz Acc	45.4	0.314	3530	186	0.698	0.037 (31)
Melpitz Ait	16.0	0.014	4550	116	0.569	0.032 (21)
Melpitz Nuc	45.1	0.036	779	189	0.63	0.028 (26)

Table S4. Optimized hyperparameters and their search bounds for XGBoost. Model-specific feature selection was performed during hyperparameter optimization through two additional parameters, rel_thresh and red_thresh. The number of features after selection, controlled by the threshold parameters, is also reported in parentheses. A dash in the threshold columns indicates that no feature selection was done. For hyperparameters that were optimized logarithmically, the bounds are written in the exponent. Bold text means that XGBoost was the best model for the datasets in question.

	gamma	learning_rate	max_depth	min_child_weight	n_estimators	red_thresh	reg_alpha	reg_lambda	rel_thresh
Bounds	$10^{[-2,4]}$	$10^{[-2,0]}$	[2, 8]	[1, 40]	[20, 200]	[0.45, 0.99]	$10^{[-3,3]}$	$10^{[-3,3]}$	[0, 0.25]
Helsinki Acc	11.1	0.033	3.0	3.0	144	0.987	0.079	11.1	0.094 (86)
Helsinki Ait	**126**	**0.148**	**6.12**	**7.36**	**81.6**	**-**	**11.5**	**350**	**- (100)**
Helsinki Nuc	28.0	0.189	7.0	21.0	168	0.512	225	403	0.216 (5)
Leipzig Acc	1330	0.547	2.19	36.1	151	0.513	708	783	0.004 (20)
Leipzig Ait	0.731	0.07	7.02	27.4	51.3	0.719	1.47	50.7	0.038 (33)
Melpitz Acc	**102**	**0.073**	**2.97**	**28.7**	**128**	**0.695**	**2.82**	**1.64**	**0.015 (33)**
Melpitz Ait	**2230**	**0.129**	**8.0**	**39.0**	**91.0**	**0.823**	**0.001**	**785**	**0.042 (42)**
Melpitz Nuc	**1200**	**0.049**	**4.0**	**18.0**	**26.0**	**0.655**	**0.003**	**0.293**	**0.038 (24)**

Table S5. Optimized hyperparameters and their search bounds for NN1. Model-specific feature selection was performed during hyperparameter optimization through two additional parameters, rel_thresh and red_thresh. The number of features after selection, controlled by the threshold parameters, is also reported in parentheses. A dash in the threshold columns indicates that no feature selection was done. For hyperparameters that were optimized logarithmically, the bounds are written in the exponent. *Activation* is a categorical hyperparameter, taking a value of 0 (logistic) or 1 (tanh). The selected value of *activation* is indicated by having a larger number in the corresponding *activation* column (e.g., 1 means selected and 0 not selected).

	activation0	activation1	alpha	hidden_layer_sizes	learning_rate_init	max_iter	red_thresh	rel_thresh
Bounds	[0, 1]	[0, 1]	$10^{[-6,1]}$	[10, 300]	$10^{[-3,0]}$	[200, 500]	[0.45, 0.99]	[0, 0.25]
Helsinki Acc	1.0	0.0	2.19	249	0.892	249	0.819	0.11 (42)
Helsinki Ait	1.0	0.0	4.01	238	0.774	448	-	- (100)
Helsinki Nuc	1.0	0.0	0.232	244	0.782	351	-	- (100)
Leipzig Acc	1.0	0.0	0.0	73.0	0.963	467	0.769	0.019 (43)
Leipzig Ait	0.123	0.005	0.019	271	0.071	249	0.971	0.021 (83)
Melpitz Acc	1.0	0.0	0.001	132	0.038	416	0.798	0.015 (48)
Melpitz Ait	1.0	0.0	8.17	78.0	0.819	407	-	- (100)
Melpitz Nuc	0.578	0.141	0.089	12.2	0.431	252	-	- (100)

Table S6. Optimized hyperparameters and their search bounds for NN2. Model-specific feature selection was performed during hyperparameter optimization through two additional parameters, rel_thresh and red_thresh. The number of features after selection, controlled by the threshold parameters, is also reported in parentheses. A dash in the threshold columns indicates that no feature selection was done. For hyperparameters that were optimized logarithmically, the bounds are written in the exponent. *Activation* is a categorical hyperparameter, taking a value of 0 (logistic) or 1 (tanh). The selected value of *activation* is indicated by having a larger number in the corresponding *activation* column (e.g., 1 means selected and 0 not selected). The two different columns for hidden layer size represent the number of neurons in the first and second layers of the two-layer network.

	activation0	activation1	alpha	hidden_layer_sizes0	hidden_layer_sizes1	learning_rate_init	max_iter	red_thresh	rel_thresh
Bounds	[0, 1]	[0, 1]	$10^{[-6,1]}$	[10, 300]	[10, 300]	$10^{[-3,0]}$	[200, 500]	[0.45, 0.99]	[0, 0.25]
Helsinki Acc	0.815	0.429	0.618	141	189	0.423	475	0.526	0.097 (18)
Helsinki Ait	1.0	0.0	0.161	103	224	0.473	323	0.799	0.2 (14)
Helsinki Nuc	0.564	0.383	0.552	274	296	0.161	261	-	- (100)
Leipzig Acc	1.0	0.0	1.49	85.0	206	0.251	488	0.879	0.12 (52)
Leipzig Ait	1.0	0.0	0.02	18.0	264	0.728	268	0.748	0.001 (40)
Melpitz Acc	0.894	0.749	0.45	298	164	0.32	327	0.557	0.012 (22)
Melpitz Ait	1.0	0.0	0.059	85.0	268	0.477	496	0.91	0.225 (24)
Melpitz Nuc	0.722	0.413	4.97	285	173	0.356	422	0.862	0.242 (13)

Table S7. Optimized hyperparameters and their search bounds for SVM. Model-specific feature selection was performed during hyperparameter optimization through two additional parameters, rel_thresh and red_thresh. The number of features after selection, controlled by the threshold parameters, is also reported in parentheses. A dash in the threshold columns indicates that no feature selection was done. For hyperparameters that were optimized logarithmically, the bounds are written in the exponent. Bold text means that SVM was the best model for the dataset in question.

	C	epsilon	gamma	red_thresh	rel_thresh
Bounds	$10^{[1,4]}$	$10^{[-1,3]}$	$10^{[-4,-1]}$	[0.45, 0.99]	[0, 0.25]
Helsinki Acc	**2330**	**383**	**0.005**	**-**	**- (100)**
Helsinki Ait	4560	55.1	0.011	0.56	0.16 (11)
Helsinki Nuc	1930	204	0.021	0.99	0.0 (95)
Leipzig Acc	4620	0.1	0.01	0.45	0.0 (16)
Leipzig Ait	2330	383	0.005	-	- (100)
Melpitz Acc	5840	404	0.002	0.814	0.0 (51)
Melpitz Ait	3000	901	0.001	0.942	0.0 (78)
Melpitz Nuc	942	132	0.006	0.99	0.0 (95)

Table S8. Optimized hyperparameters and their search bounds for GP. Model-specific feature selection was performed during hyperparameter optimization through two additional parameters, rel_thresh and red_thresh. The number of features after selection, controlled by the threshold parameters, is also reported in parentheses. A dash in the threshold columns indicates that no feature selection was done. The only other hyperparameter was *alpha*, which was optimized logarithmically. To represent that, its bounds are written in the exponent. Bold text means that GP was the best model for the datasets in question.

	alpha	red_thresh	rel_thresh
Bounds	$10^{[-2,1]}$	[0.45, 0.99]	[0, 0.25]
Helsinki Acc	1.1	-	- (100)
Helsinki Ait	1.1	-	- (100)
Helsinki Nuc	**0.71**	**-**	**- (100)**
Leipzig Acc	0.684	0.6	0.004 (27)
Leipzig Ait	**0.714**	**0.802**	**0.101 (31)**
Melpitz Acc	0.272	0.927	0.0 (74)
Melpitz Ait	1.34	0.811	0.075 (33)
Melpitz Nuc	1.12	0.822	0.018 (46)

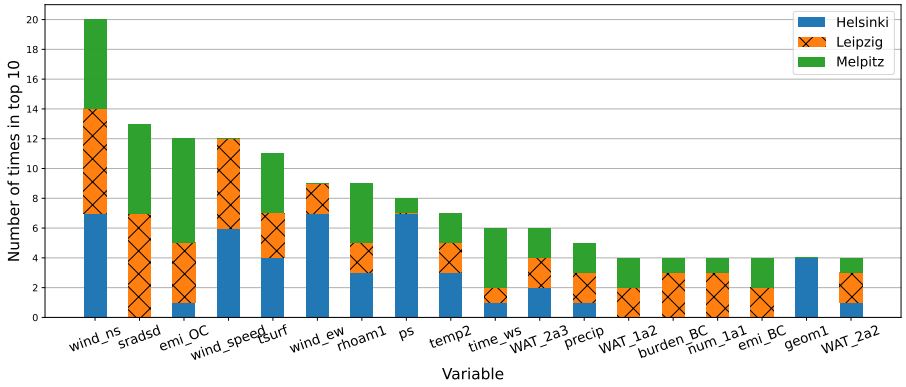

Figure S2. Most important input variables across all three Aitken subrange datasets, measured by mean absolute SHAP values. All seven ML models were analyzed for each dataset. Therefore, the maximum possible height of a bar is 21. Bars with height less than four are not shown.

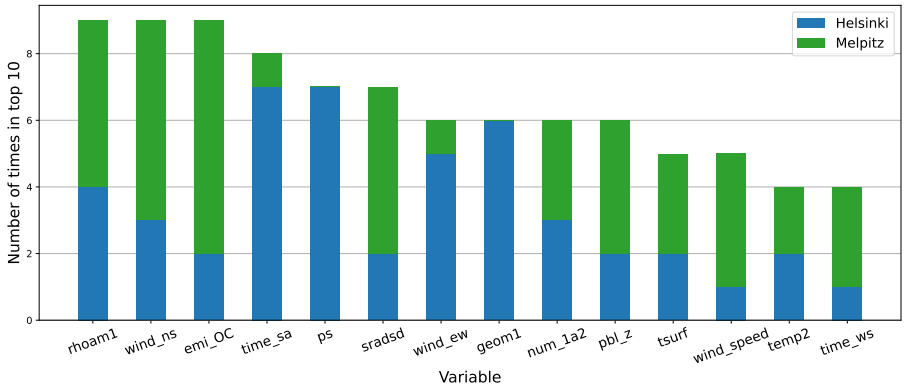

Figure S3. Most important input variables in the two nucleation subrange datasets, measured by mean absolute SHAP values. All seven ML models were analyzed for both datasets. Therefore, the maximum possible height of a bar is 14. Bars with height less than four are not shown.

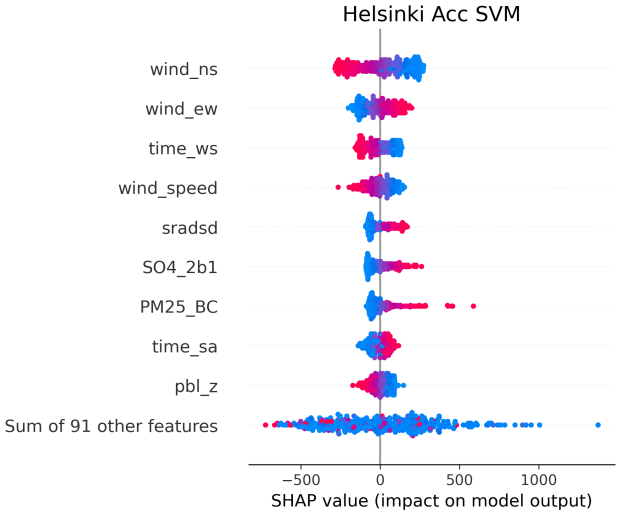

Figure S4. Features of SVM for Helsinki's accumulation subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction.

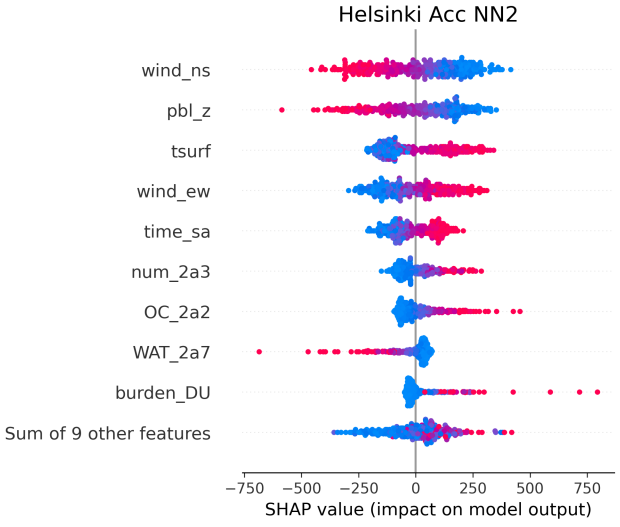

Figure S5. Features of NN2 for Helsinki's accumulation subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction. As all neural networks behave differently depending on their random initializations, one model was arbitrarily selected for this visualization.

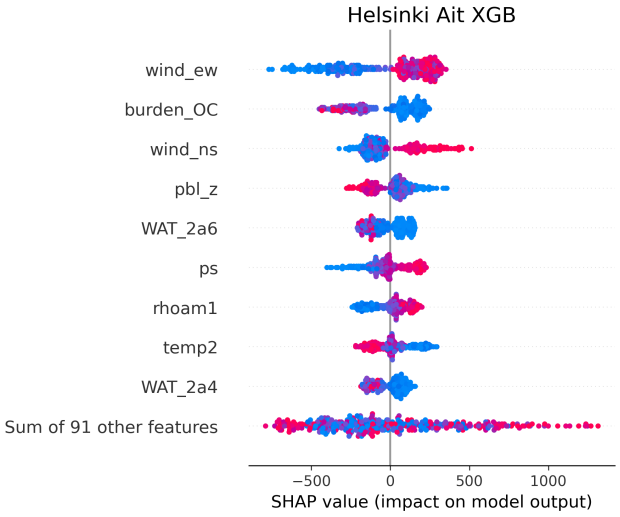

Figure S6. Features of XGBoost for Helsinki's Aitken subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction.

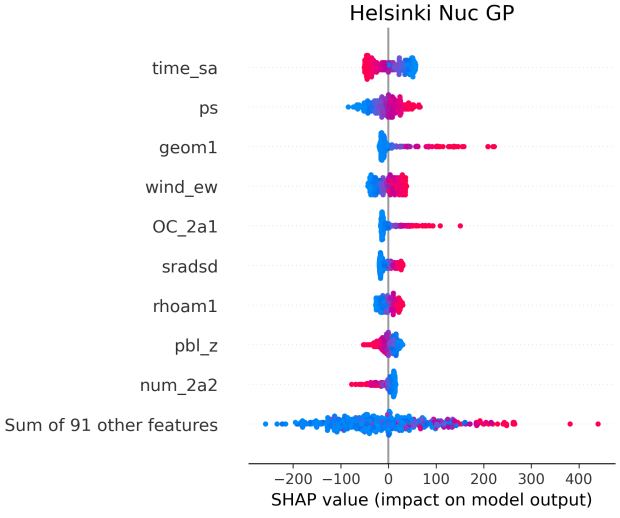

Figure S7. Features of GP for Helsinki's nucleation subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction.

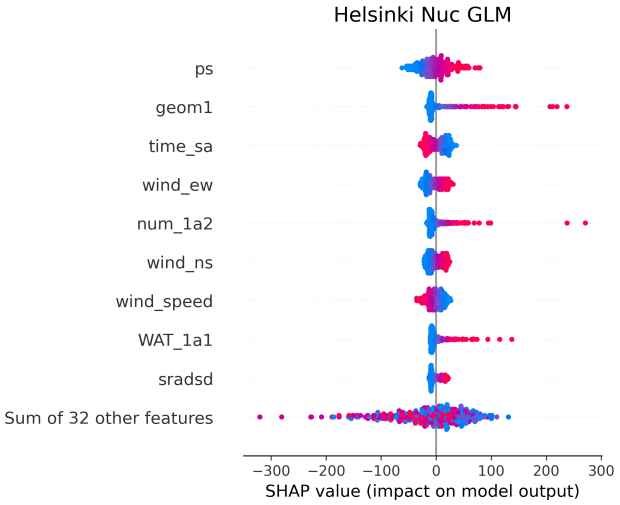

Figure S8. Features of GLM for Helsinki's nucleation subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction.

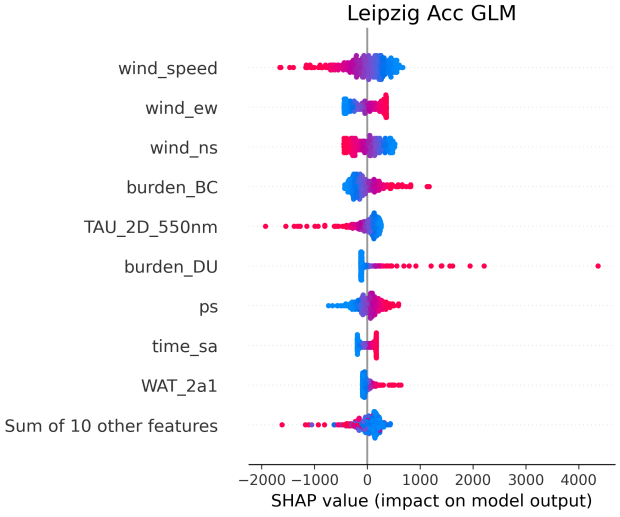

Figure S9. Features of GLM for Leipzig's accumulation subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction.

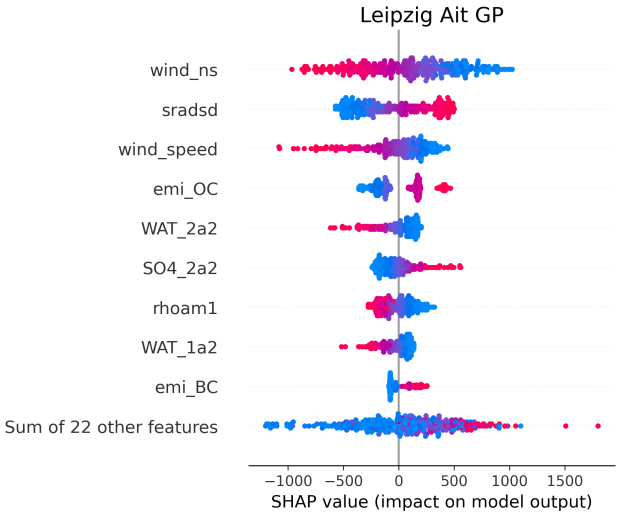

Figure S10. Features of GP for Leipzig's Aitken subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction.

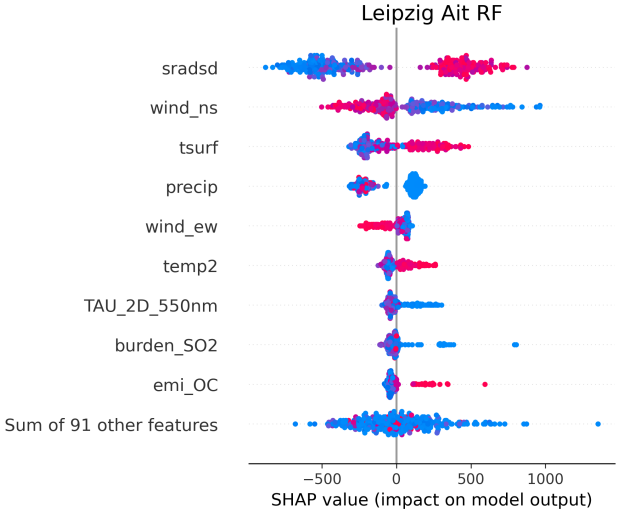

Figure S11. Features of RF for Leipzig's Aitken subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction. As randomness affects the structure of the trees in the ensemble, one model was arbitrarily selected for this visualization.

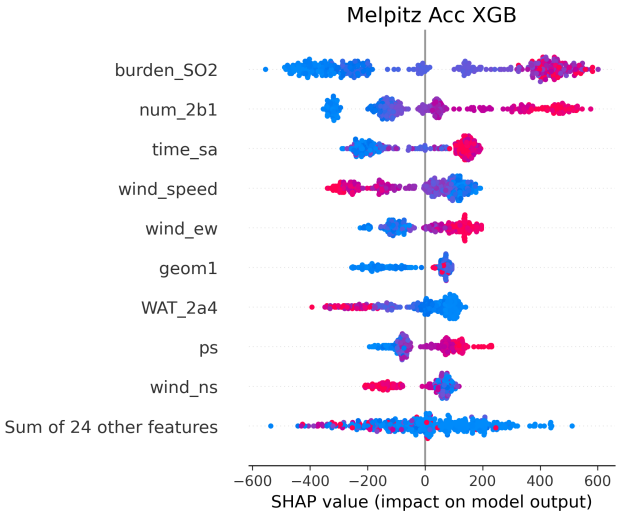

Figure S12. Features of XGBoost for Melpitz's accumulation subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction.

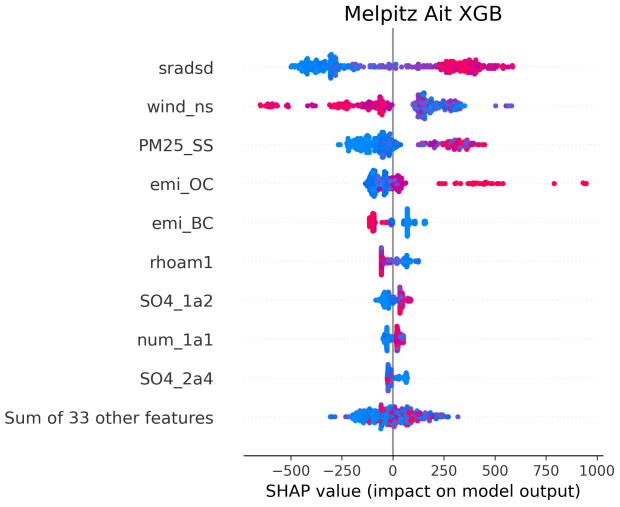

Figure S13. Features of XGBoost for Melpitz's Aitken subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction.

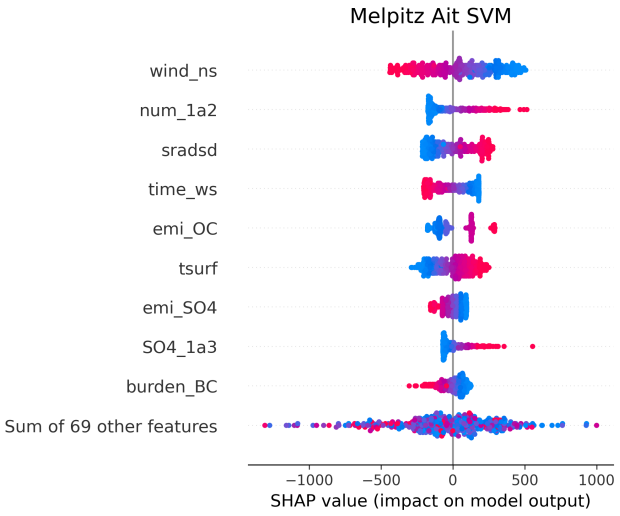

Figure S14. Features of SVM for Melpitz's Aitken subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction.

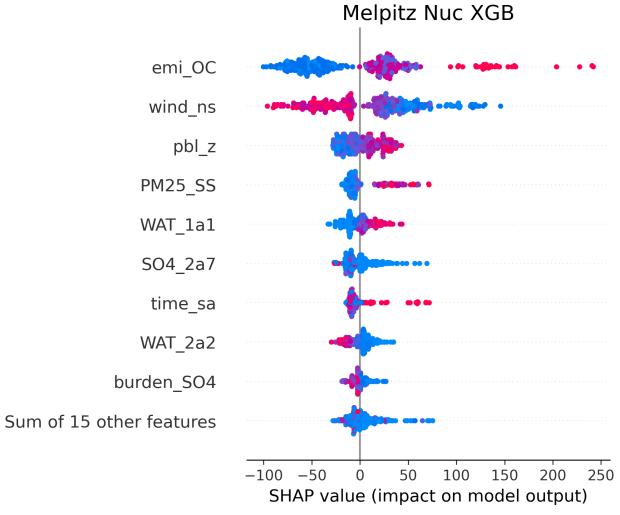

Figure S15. Features of XGBoost for Melpitz's nucleation subrange, ranked by mean absolute SHAP value. Each point corresponds to one prediction. Colors represent feature values; high feature values are shown in red and low values in blue. The SHAP values on the horizontal axis show difference relative to mean prediction.

Table S9. List of variables selected for the NN2 model (shown also in Figure S5) for downscaling the accumulation subrange of Helsinki.

Variable name
tsurf
OC_2a2
OC_2a7
pbl_z
num_2a3
burden_SO2
wind_ns
burden_DU
WAT_2a1
num_2a4
WAT_2a7
wind_ew
num_2b3
aprc
precip
ps
SO4_1a2
time_sa

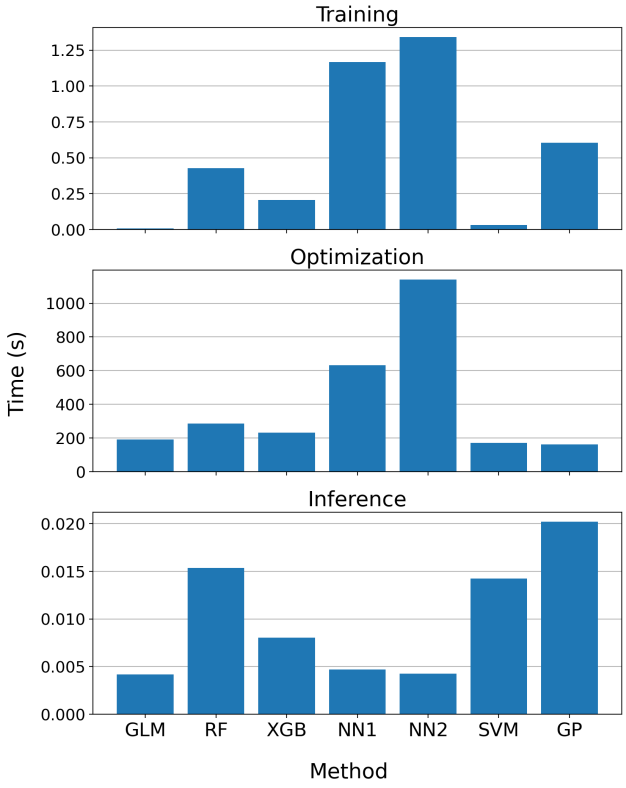

Figure S16. CPU clock times in seconds for different parts of the machine learning procedure, obtained from the Python function *time.perf_counter()*. The results are averaged across all eight datasets for each ML method. Optimization times take into account all 300 iterations, with varying model configurations. Here, training means fitting the model with optimal hyperparameters to the combined training and validation data, and inference refers to producing predictions by inputting the testing data into the trained model. All operations were performed on a laptop with eight Intel® Core™ i7-4810MQ CPUs with 2.80 GHz base frequency.

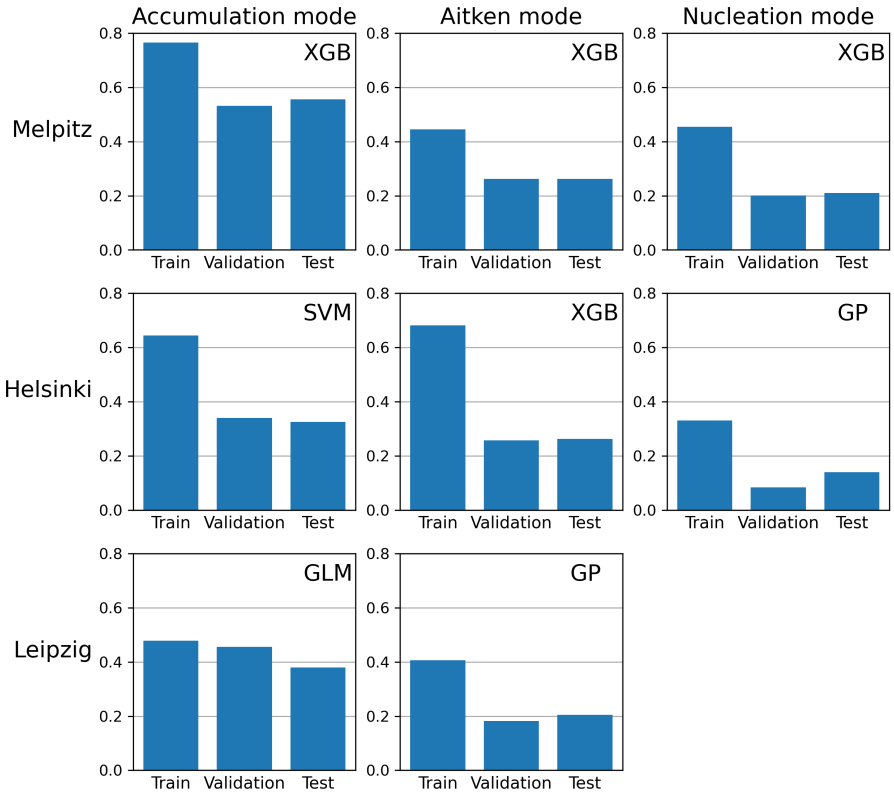

Figure S17. Training, validation and test scores (ρ^2) of the best models for each dataset. Higher test scores compared to validation are likely caused by selecting only the best-performing model, based on the test score, to be further analyzed. This selection was done from among all seven ML models and seven optimization methods for each ML model, for a total of 49 models.