Supplement of Atmos. Meas. Tech., 19, 2225–2244, 2026
https://doi.org/10.5194/amt-19-2225-2026-supplement

Atmospheric Measurement Techniques
Open Access
EGU

Supplement of

A Physics-Constrained Deep-Learning Framework based on Long-Term Remote-Sensing Data for Retrieving Vertical Distribution of $PM_{2.5}$ Chemical Components

Hongyi Li et al.

Correspondence to: Ting Yang (tingyang@mail.iap.ac.cn)

Contents of this file

S1. Lidar instrument specifications

As shown in Table S1, the laser emission at wavelengths of 532 nm and 1064 nm relies on a Nd:YAG laser with a second harmonic generator and is corrected by a beam expander before emission. The emitted laser energies at 532 nm and 1064 nm are 30 mJ and 20 mJ, respectively. The laser pulse repetition frequency can reach up to 20 Hz and is set to 10 Hz in practice. The scattered light is collected by a Schmidt-Cassegrain telescope with a diameter of 20 cm and then is collimated and corrected toward a dichroic mirror to separate the received lidar signals at 532 nm and 1064 nm. The lidar signal at 532 nm is separated into horizontal and vertical polarization components and is measured by a photomultiplier tube. The lidar signal at 1064 nm is directly detected by an avalanche photodiode. Finally, the detected lidar signals are recorded by a digital oscilloscope and then are transferred to a computer for data storage.

S2. Lidar data preprocessing

A comprehensive data quality control procedure was implemented on the original lidar signals to mitigate issues raised by electrical signal errors and signal offsets caused by background radiation. First, background noise was removed by subtracting the average value of signals within the altitudes of 3-9 km from the original lidar signal. Second, the lidar signal was range-corrected by multiplying by the square of the altitude and corrected for the geometric overlap effect using an empirically determined function derived from lidar profiles under well-mixed atmospheric conditions. Third, a cloud-screening algorithm was applied to identify and remove profiles contaminated by clouds. The algorithm operates by first calculating the vertical gradient of the range-corrected signal. It then identifies potential cloud bases as regions where this gradient exceeds a primary threshold of 4×10^{-8} for at least 3 consecutive resolution layers. For each candidate cloud layer, the algorithm determines the cloud top and then validates the layer by checking if the maximum signal within it surpasses a secondary threshold of 5×10^{-6}. Profiles containing such validated cloud layers were entirely excluded from the subsequent aerosol analysis. Finally, the extinction coefficient at a wavelength of 532 nm was retrieved based on Fernald algorithm (Fernald, 1984).

To facilitate data fusion and comply with the input requirements of the machine learning model, all data were vertically re-sampled onto a standardized set of preset

height levels ranging from 50 m to 3 km. The high-resolution lidar data and the low-resolution global reanalysis data were interpolated onto this vertical grid using linear interpolation. The preset height grid with logarithmic intervals can be determined by Eq. S1-S2. Logarithmic interval amplifies vertical resolution within the planetary boundary layer, where fine-mode particles and their chemical components are typically most concentrated (Yang et al., 2024).

$$h_i = 10^{\log_{10}(Z_{min})+(i-1)\times\Delta Z}, i = 1, 2, \dots, n, \quad (S1)$$

$$\Delta Z = \frac{\log_{10}(Z_{max}) - \log_{10}(Z_{min})}{n-1}, \quad (S2)$$

where h_i is the height at i^{th} vertical layer, Z_{min} is the minimum height, ΔZ is the logarithmic interval, Z_{max} is the maximum height, and n is the total number of vertical layers.

S3. Cross validation scheme

We implement a 10-fold time-series cross-validation scheme for the training (and validation) set to preserve its temporal order and prevent future information leakage. As presented in Fig. S2, we repeatedly utilize a forward sliding window to create K (set to 10) validation folds. The training set starts with a subset of the first 80% of the chronological data and is incrementally expanded at each subsequent fold by incorporating an additional block with a length of the forward sliding window, ultimately encompassing the full 80% in the final fold. The validation set immediately follows the training set, comprising 20% of the chronological data.

The length of the forward sliding window is equal to the length of the training set at first fold in practice (Eq. (S3)).

$$l = \frac{r \times N}{K}, \quad (S3)$$

where l is the length of the forward sliding window, r is the proportion of data used for training, N is the total sample size for model construction, and K is the total number of cross-validation folds.

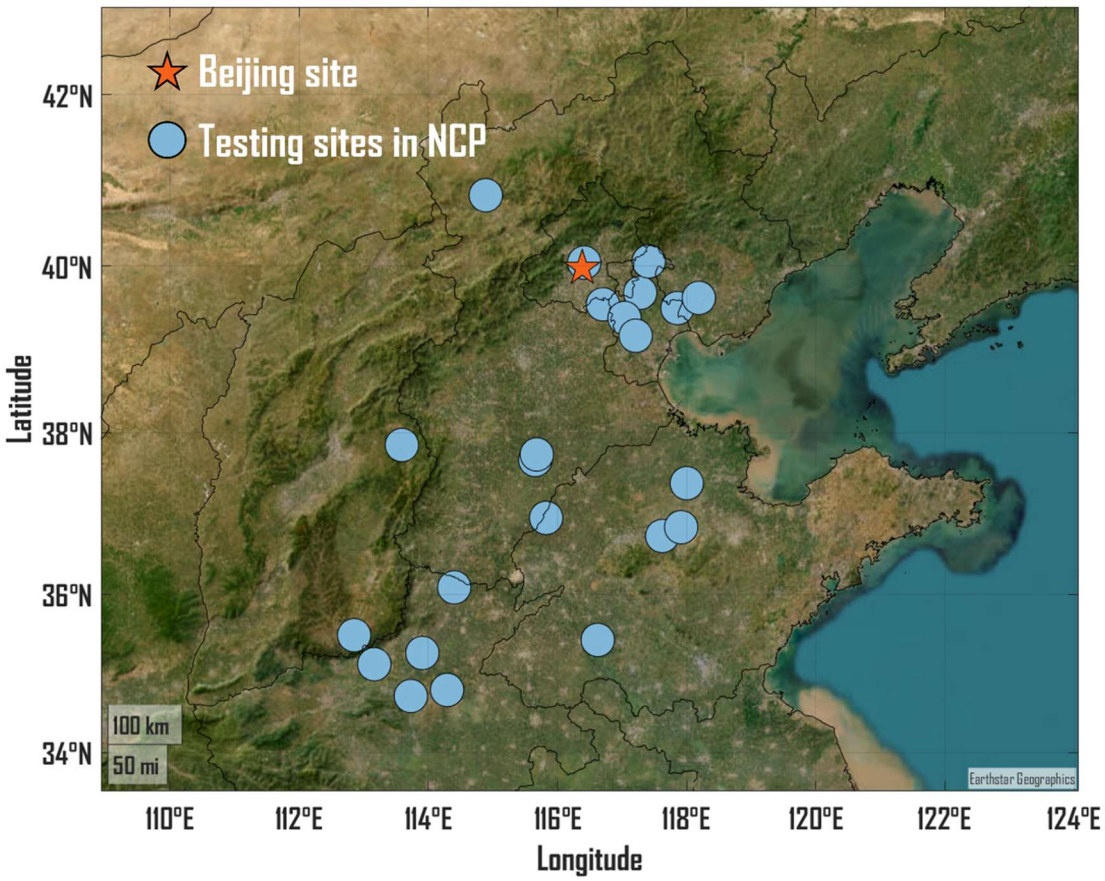

Figure S1: Spatial distribution of Beijing and spatially independent testing sites. Beijing site provides 18-month datasets for the training, validation and temporally independent testing of the deep-learning module. Other 23 sites in North China Plain (NCP) provide 8-day datasets for the spatially independent testing of the final retrieval. The geographic basemap is hosted by Esri | Powered by Esri (https://www.esri.com/en-us/home).

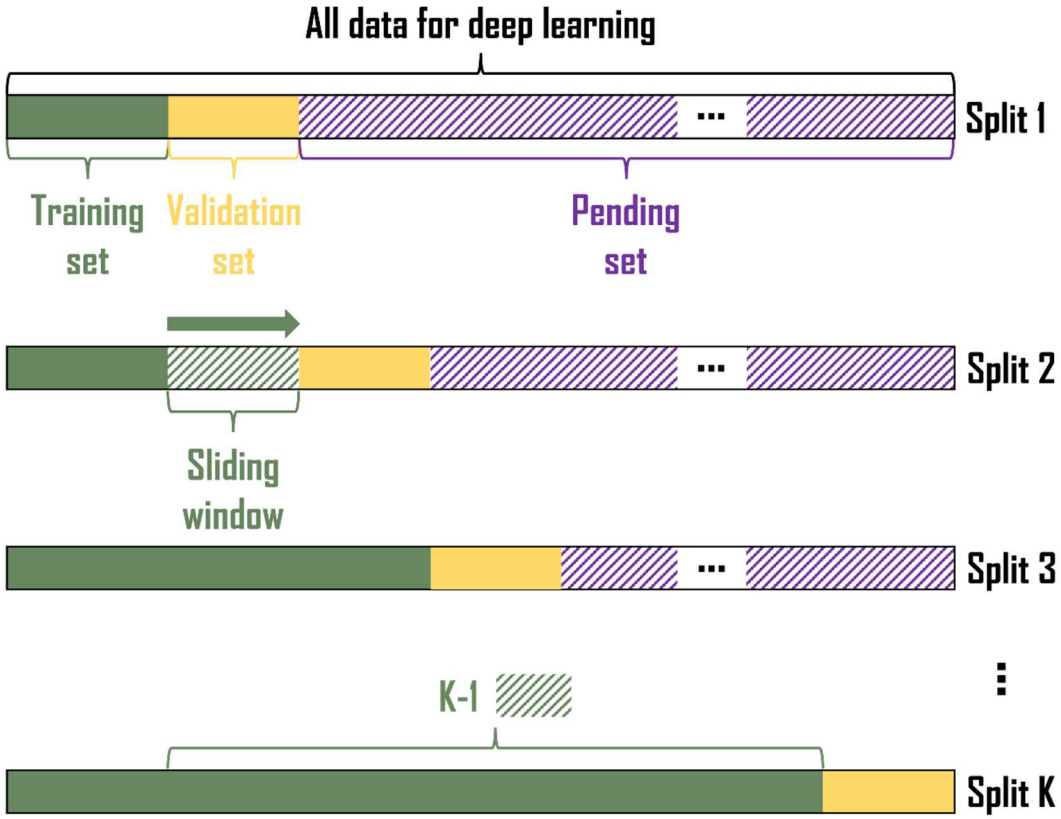

Figure S2: Diagram of the 10-fold cross-validation used in this work.

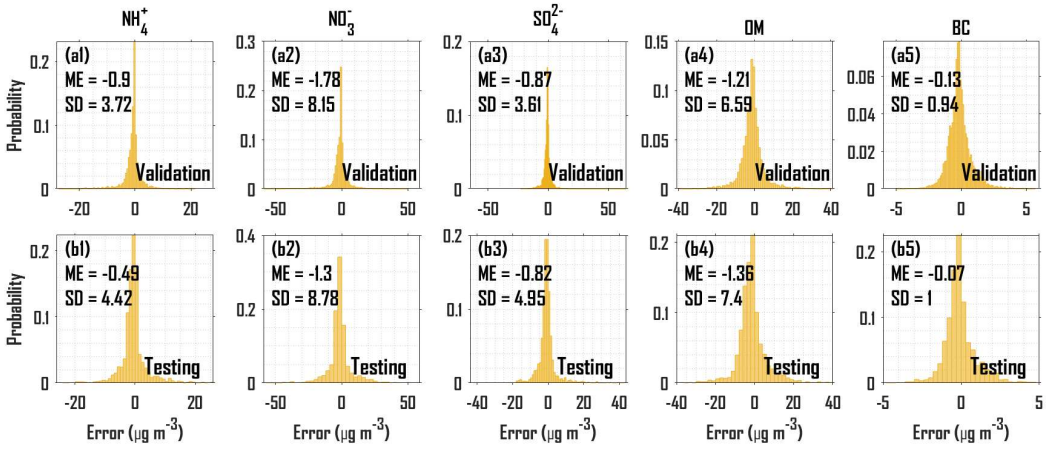

Figure S3: Probability distributions of error (observations minus simulations, μg m^{-3}) for NH_4^+, NO_3^-, SO_4^{2-}, OM and BC during the 10-fold cross-validation phase (a1-a5) and during the temporally independent testing phase (b1-b5). ME: Mean Error; SD: Standard Deviation.

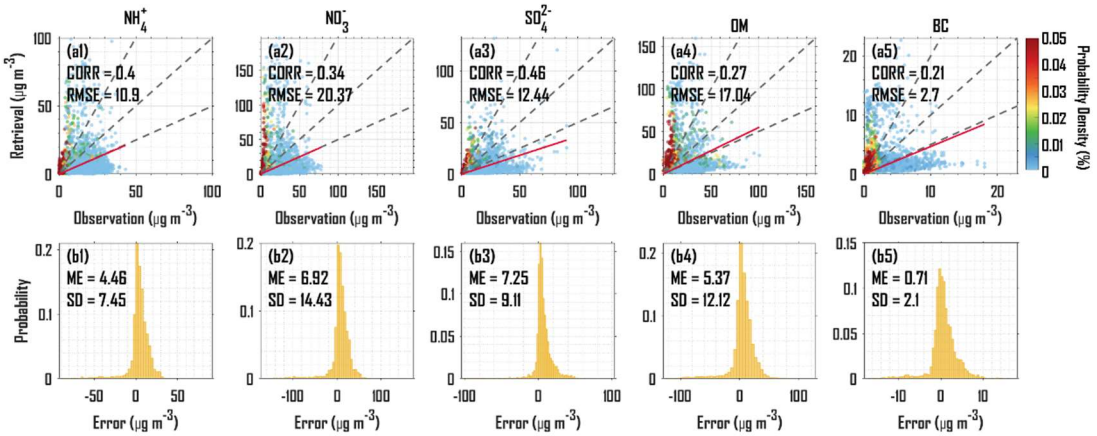

Figure S4: Scatterplots of the retrievals (μg m^{-3}) versus the observations (μg m^{-3}) with probability density (%) for NH_4^+, NO_3^-, SO_4^{2-}, OM and BC across 23 spatially independent testing sites (a1-a5). The dotted grey lines represent the 2:1, 1:1, and 1:2 lines, and the solid red line represents the fitted regression line. CORR represents the correlation coefficient, and RMSE represents root mean square error. Probability distributions of error (observations minus retrievals, μg m^{-3}) for NH_4^+, NO_3^-, SO_4^{2-}, OM and BC across 23 spatially independent testing sites (b1-b5). ME: Mean Error; SD: Standard Deviation.

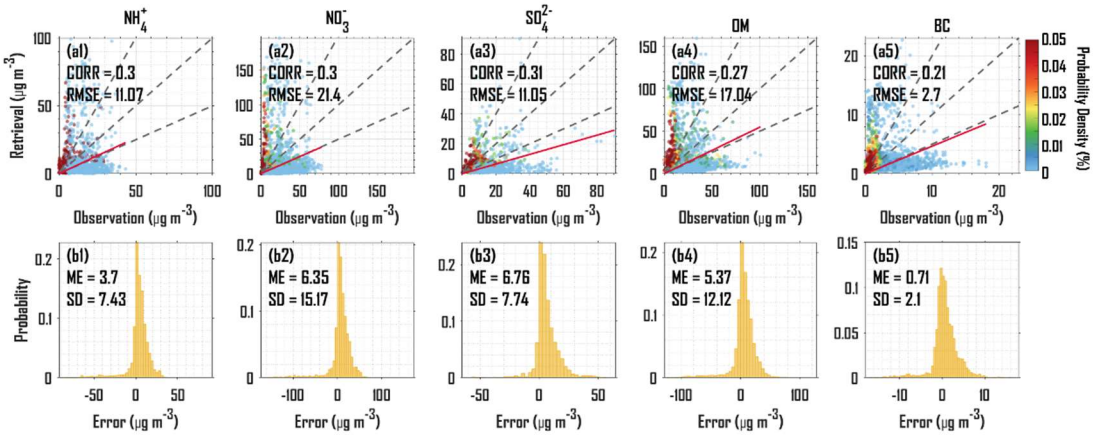

Figure S5: Scatterplots of the retrievals (μg m^{-3}) versus the observations (μg m^{-3}) with probability density (%) for NH_4^+, NO_3^-, SO_4^{2-}, OM and BC across problematic (CORR <0.5) spatially independent testing sites (a1-a5). The dotted grey lines represent the 2:1, 1:1, and 1:2 lines, and the solid red line represents the fitted regression line. CORR represents the correlation coefficient, and RMSE represents root mean square error. Probability distributions of error (observations minus retrievals, μg m^{-3}) for NH_4^+, NO_3^-, SO_4^{2-}, OM and BC across 23 spatially independent testing sites (b1-b5). ME: Mean Error; SD: Standard Deviation.

Table S1. The main specification parameters of dual-wavelength polarization Mie Lidar.

Parameter categories		Description
Laser type		Flashlamp pumped Nd:YAG
Laser pulse energy	532 nm	30 mJ/pulse
	1064 nm	20 mJ/pulse
Pulse Repetition Frequency		≤ 20 Hz, 10 Hz used in this work
Telescope Type		Schmidt Cassegrain
Telescope diameter		20 cm
Field of view		1 mrad
Detector type	532 nm	Photomultiplier tube (PMT)
	1064 nm	Avalanche photodiode (APD)
Data acquisition system		Digital oscilloscope

Table S2. Optimal hyperparameters of the deep learning module.

Hyperparameter	Decision space	Optimal values
Initial learning rate	$[10^{-5}\ 10^{-3}]$	4.71×10^{-4}
Factor for L_2 regularization	$[10^{-10}\ 10^{-2}]$	1.54×10^{-4}
Decay rate of gradient moving average	[0.8 0.98]	0.80
Decay rate of squared gradient moving average	[0.8 0.99]	0.81
Number of filters	1 [8 64]	44
	2 [8 64]	34
Size of filters	1 [3 16]	6
	2 [3 16]	10
Number of layers	[1 4]	3
Number of hidden units	[60 200]	112
Maximum of Epochs	\	100
Size of mini-batch	\	64
Dropout value	\	0.25
Solver	\	adam
Num of cross-validation folds	\	10

Table S3. Statistical metrics quantified by vertical retrievals and tower-based observations during a period from December 30, 2018 to January 2, 2019 for NH_4^+, NO_3^-, SO_4^{2-}, and OM. RMSE: Root Mean Square Error; MAE: Mean Absolute Error; CORR: Pearson correlation coefficient.

	RMSE (μg m^{-3})	MAE (μg m^{-3})	CORR
NH_4^+	4.81	3.14	0.67
NO_3^-	10.48	6.19	0.67
SO_4^{2-}	4.08	2.59	0.66
OM	23.04	15.37	0.67

Reference

Fernald, F. G.: Analysis of atmospheric lidar observations: some comments, Appl. Opt., 23, 652-653, https://doi.org/10.1364/AO.23.000652, 1984.

Yang, T., Li, H., Xu, W., Song, Y., Xu, L., Wang, H., Wang, F., Sun, Y., Wang, Z., and Fu, P.: Strong Impacts of Regional Atmospheric Transport on the Vertical Distribution of Aerosol Ammonium over Beijing, Environ. Sci. Technol. Lett., 11, 29-34, https://doi.org/10.1021/acs.estlett.3c00791, 2024.