Supplement of Atmos. Chem. Phys., 26, 8225–8253, 2026
https://doi.org/10.5194/acp-26-8225-2026-supplement
© Author(s) 2026. CC BY 4.0 License.

Atmospheric Chemistry and Physics
Open Access
EGU

Supplement of

Correcting aerosol extinction coefficient vertical structure biases in GEOS-chem via a physics-informed transformer with physical mechanism diagnosis

Jiajun Xiong et al.

Correspondence to: Yi Wang (wangyi34@cug.edu.cn)

The copyright of individual parts of the supplement might differ from the article licence.

S1. CALIOP Quality Control

To ensure the reliability of the observational target, we implement a rigorous, tiered quality control strategy for the CALIOP Level 2 aerosol profile data. First, to minimize cloud contamination, profiles with a Cloud Layer Fraction (CLF) exceeding 2.0 were excluded. We further refined this dataset by applying strict thresholds to the Cloud Aerosol Discrimination (CAD) score, retaining only retrievals with high confidence (CAD score between -100 and -20). The Extinction QC flag was used to filter for algorithmically stable solutions (unconstrained or constrained), rejecting profiles with divergent retrieval errors. Adjustments to lidar ratios or opaque layer identifications were permitted only when necessary for algorithm convergence. A final multi-step refinement was applied to the vertical profiles: (1) profiles were discarded if the extinction uncertainty exceeded 99.9 km^{-1} at any lower-altitude layer; (2) physically unrealistic extinction coefficients greater than 2 km^{-1} were flagged and removed; and (3) to mitigate surface contamination artifacts, data within 180 m of the terrain were excluded. Observations classified as "clear air" were assigned an extinction coefficient of 0 km^{-1}, with the exception of the lowest 180 m AGL, which was masked to avoid surface return interference. Additional safeguards included the removal of high-altitude data (>4 km) near 0°C ice clouds and the exclusion of spatially isolated aerosol signals (horizontal resolution < 80 km).

S2. AERONET AOD Interpolation

Standard AERONET sun photometers do not measure AOD directly at 532 nm, the operating wavelength of the CALIOP lidar and our GEOS-Chem configuration. To enable direct comparison, we interpolated the AOD to 532 nm using the Ångström exponent (α) derived from measurements at adjacent wavelengths (typically 500 nm and 675 nm, or 440 nm/870 nm if necessary). The interpolation follows the power-law relationship:

$$\tau_{532} = \tau_{\lambda_1} \times \left(\frac{532}{\lambda_1}\right)^{-\alpha}$$

$$\alpha = -\frac{\ln\left(\frac{\tau_{\lambda_1}}{\tau_{\lambda_2}}\right)}{\ln\left(\frac{\lambda_1}{\lambda_2}\right)}$$

where τ represents the AOD at a specific wavelength λ. We prioritized the use of the 500 nm and 675 nm pair for calculating α due to their proximity to 532 nm. In cases where data at 675 nm were unavailable, the 870 nm or 440 nm channels were used as secondary references. This processing ensures spectral consistency between the ground-based validation dataset and the model outputs.

S3. Input Variable

Table S1. Summary of input variables from MERRA-2 reanalysis and GEOS-Chem model simulations used in this study, categorized by data source and physical property.

Category	Variable Name	Physical Description
MERRA-2 (3D Fields)	T	Air Temperature (3D)
	U, V	Eastward & Northward Wind Components
	OMEGA	Vertical Pressure Velocity
	QV	Specific Humidity
	RH	Relative Humidity
	CLOUD	Cloud Fraction
	OPTDEPTH	In-cloud Optical Thickness
	QI, QL	Mass Fraction of Cloud Ice / Liquid Water
	TAUCLI, TAUCLW	Optical Thickness (Ice / Liquid Clouds)
	DTRAIN	Detrainment Mass Flux
	CMFMC	Cumulative Mass Flux
	PFICU, PFILSAN	Ice Precipitation Flux (Conv / LS)
	PFLCU, PFLLSAN	Liquid Precipitation Flux (Conv / LS)
	DQRCU, DQRLSAN	Rainwater Source (Conv / LS)
	REEVAPCN, REEVAPLS	Evaporation of Precip (Conv / LS)
	PV	Ertel's Potential Vorticity
MERRA-2 (Surface/2D)	PS, SLP	Surface Pressure / Sea Level Pressure
	TS, T2M, T10M	Skin Temp / 2m Temp / 10m Temp
	U10M, V10M	10-meter Wind Components
	QV2M / Q850	Specific Humidity (2m / 850hPa)

	PBLH	Planetary Boundary Layer Height
	PRECTOT	Total Precipitation
	PRECCON, PRECLSC	Precipitation (Convective / Large Scale)
	PRECANV, PRECSNO	Precipitation (Anvil / Snow)
	SNODP / SNOMAS	Snow Depth / Snow Mass
	ALBEDO	Surface Albedo
	CLDTOT	Total Cloud Area Fraction
	SWGDN	Incident Shortwave Flux
	LWGNT, LWTUP	Net Downward / Upward Longwave Flux
	EFLUX, HFLUX	Latent / Sensible Heat Flux
	EVAP	Surface Evaporation
	GWETROOT, GWETTOP	Soil Wetness (Root / Top)
	LAI, GRN	Leaf Area Index / Greenness
	FRSNO, FRSEAICE	Fraction of Snow / Sea Ice
	SEAICE00-90	Sea Ice Fraction (Binned 00-90)
	TO3	Total Column Ozone
	TROPPT	Tropopause Pressure
	USTAR, Z0M	Friction Velocity / Roughness Length
	LWI	Land Water Indicator
	PARDF, PARDR	PAR (Diffuse / Direct)
GEOS-Chem (Aerosol Mass)	AerMassSO_4	Sulfate Aerosol Mass
	AerMassBC, AerMassPOA	Black Carbon / Primary Organic Aerosol
	AerMassOPOA	Oxidized Primary Organic Aerosol
	AerMassNIT, AerMassNH_4	Nitrate / Ammonium Aerosol Mass

	AerMassSAL	Sea Salt Aerosol Mass
	AerMassSOA*	Secondary Organic Aerosol
	AerMassHMS	Hydroxymethanesulfonate Aerosol
	AerMassLVOCOA	Low Volatility Oxygenated OA
	TotalOA	Total Organic Aerosol
	PM_{25}, PM_{10}	$PM_{2.5}$ / PM_{10} Mass
	EC_gc	Elemental Carbon Tracer
GEOS-Chem (Optics/Micro)	AODHyg532nm_*	Hygroscopic AOD at 532nm
	AODDust	Dust Optical Depth
	AerHygroscopicGrowth_*	Hygroscopic Growth Factors
	AerSurfAreaHyg_*	Hygroscopic Aerosol Surface Area
	Chem_WetAeroRadi*	Wet Aerosol Radius
	Chem_AeroRadi*	Dry Aerosol Radius
	Chem_WetAeroArea*	Wet Aerosol Surface Area
	Chem_AeroArea*	Dry Aerosol Surface Area
GEOS-Chem (Gas Species)	SpeciesConc_SO_2	Sulfur Dioxide
	SpeciesConc_NO, NO_2	Nitrogen Oxides
	SpeciesConc_O_3	Ozone
	SpeciesConc_NH_3	Ammonia
	SpeciesConc_HNO_3	Nitric Acid
	SpeciesConc_CO	Carbon Monoxide
	SpeciesConc_OH	Hydroxyl Radical
	SpeciesConc_HO_2, H2O2	Hydroperoxyl / Hydrogen Peroxide
	SpeciesConc_CH_2O	Formaldehyde

SpeciesConc_VOCs	VOCs (Isoprene, MVK, MACR, Aromatics...)
SpeciesConc_S-Map	Sulfur Species (DMS, MSA, HMS)

58

S4. Detailed Architecture of the Transformer Components

a. Heterogeneous Input Embedding Strategies

Unlike the standard Transformer which uses a uniform positional encoding, we adopt distinct embedding strategies tailored to the physical nature of each input variable:

① **Hybrid Vertical Height Encoding**: To capture the stratification of the atmosphere, we combine a learnable discrete embedding with a continuous linear projection. For a vertical layer with index l and geopotential height value h_l:

$$e_{height}^{(l)} = Linear\left(Concat\left(E_{idx}[l], Linear(h_l)\right)\right) \quad (S1)$$

where E_{idx} is a learnable lookup table for layer indices.

② **Cyclic Spatiotemporal Encoding**: For variables with intrinsic periodicity (i.e., Month, Latitude, Longitude), we utilize trigonometric cyclic encoding to preserve their continuity (e.g., ensuring December is numerically close to January). For a variable v with a period T (e.g., T=12 for months, T=360 for longitude), the encoding is defined as:

$$e_{cyclic} = Linear\left(\left[\sin\left(\frac{2\pi v}{T}\right), \cos\left(\frac{2\pi v}{T}\right)\right]\right) \quad (S2)$$

③ **Categorical Embedding**: Binary variables (e.g., Day/Night flag $d \in \{0,1\}$) are mapped to dense vectors using standard learnable embeddings:

$$e_{dn} = E_{dn}[d] \quad (S3)$$

b. Variable Identity Embedding

For the global context inputs $X_{global} \in \mathbb{R}^{N_{var}}$, where each scalar x_j represents a distinct physical variable (e.g., PBLH, Surface Pressure), we use a specialized projection:

$$e_{global}^{(j)} = LayerNorm\left(x_j \cdot \omega_j + b_j + id_j\right) \quad (S4)$$

where $\omega_j \in \mathbb{R}^{d_{model}}$ is a variable-specific projection vector (broadcasted multiplication), b_j is a bias term, and id_j is a learnable identity embedding unique to the j-th variable type. This ensures the model distinguishes between variables even if their scalar values are similar.

c. Standard Multi-Head Self-Attention (MSA)

The Transformer encoder utilizes MSA to capture dependencies between different vertical layers. For a given input sequence X, the attention output is calculated as:

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (S5)$$

where $Q = XW^Q$, $K = XW^K$, and $V = XW^V$ are the query, key, and value matrices obtained through linear transformations. In our multi-head setting, the outputs are concatenated and linearly projected:

$$MSA(Q,K,V) = Concat(head_1, \dots, head_h)\, W^O \quad (S6)$$

d. Position-wise Feed-Forward Network (FFN)

Each encoder layer includes a fully connected FFN with a GELU activation function:

$$FFN(x) = Linear_2\left(GELU\left(Linear_1(x)\right)\right) \quad (S7)$$

e. Gated Feature Fusion

$$E_{local}^{(z)} = \sum_{k=1}^{4} \alpha_k^{(z)} \cdot Norm\left(h_k^{(z)}\right)\ s.t.\ \sum \alpha_k = 1 \quad (S8)$$

where $h_k^{(z)}$ represents the embedding of the k-th feature group. The attention weights $\alpha_k^{(z)}$ are generated by a learnable gating network, allowing the model to autonomously identify and prioritize the most physically relevant information source for each specific altitude layer.

f. Output Layer

$$\Delta_{AEC}^{pred} = MLP\left(H_{cross} + H_{phyche}\right) \cdot s + b \quad (S9)$$

Where Δ_{AEC}^{pred} denotes the predicted systematic bias of the AEC. The parameters s and b are learnable scaling and bias terms, respectively, introduced to adaptively map the normalized network outputs to the physical magnitude of extinction biases.

g. Magnitude-Weighted Loss

Aerosol extinction exhibits a high dynamic range and severe spatial heterogeneity, statistically dominated by near-zero background values (e.g., in the clean free

111 troposphere). Standard Mean Squared Error (MSE) treats all data points equally, which
112 often causes the model to over-fit these overwhelming background signals while under-
113 correcting the large systematic biases associated with severe pollution episodes. To
114 resolve this imbalance, we propose a Magnitude-Weighted Loss (L_{MW}).
115

116 $$L_{MW} = \frac{1}{N}\sum_{i=1}^{N}\left(\Delta_{AEC}^{pred} - \Delta_{AEC}^{target}\right)^2 \cdot \omega\left(\Delta_{AEC}^{target}\right) \quad (S10)$$

117 $$\omega\left(\Delta_{AEC}^{target}\right) = \left|\Delta_{AEC}^{target}\right|^p + \lambda \cdot \mathrm{e}^{-\beta\left|\Delta_{AEC}^{target}\right|} + \epsilon \quad (S11)$$

118 Here, the weight function ω dynamically rescales the optimization penalty based on
119 the magnitude of the true bias (Δ_{AEC}^{target}), utilizing three empirically derived components.
120 (1) Large-Error Prioritization ($\left|\Delta_{AEC}^{target}\right|^p$): Governed by the exponent p, this term
121 sharply amplifies the gradient for samples with massive simulation biases. It forces the
122 network to target critical GEOS-Chem failures, such as significantly underestimated
123 dust storms or heavy anthropogenic haze, rather than safely fitting the background
124 average. (2) False Alarm Penalty ($\lambda \cdot \mathrm{e}^{-\beta\left|\Delta_{AEC}^{target}\right|}$): Regulated by constants λ and β,
125 this exponential decay applies a strict penalty when the true GEOS-Chem bias is
126 negligible but the model attempts a non-zero correction. Physically, this prevents the
127 generation of spurious aerosol artifacts in regions where the GEOS-Chem is already
128 accurate, such as the clean free troposphere. (3) Base Stability Term (ϵ): A minor
129 constant added to maintain numerical stability during gradient descent.
130

S5. 5-fold Cross-Validation

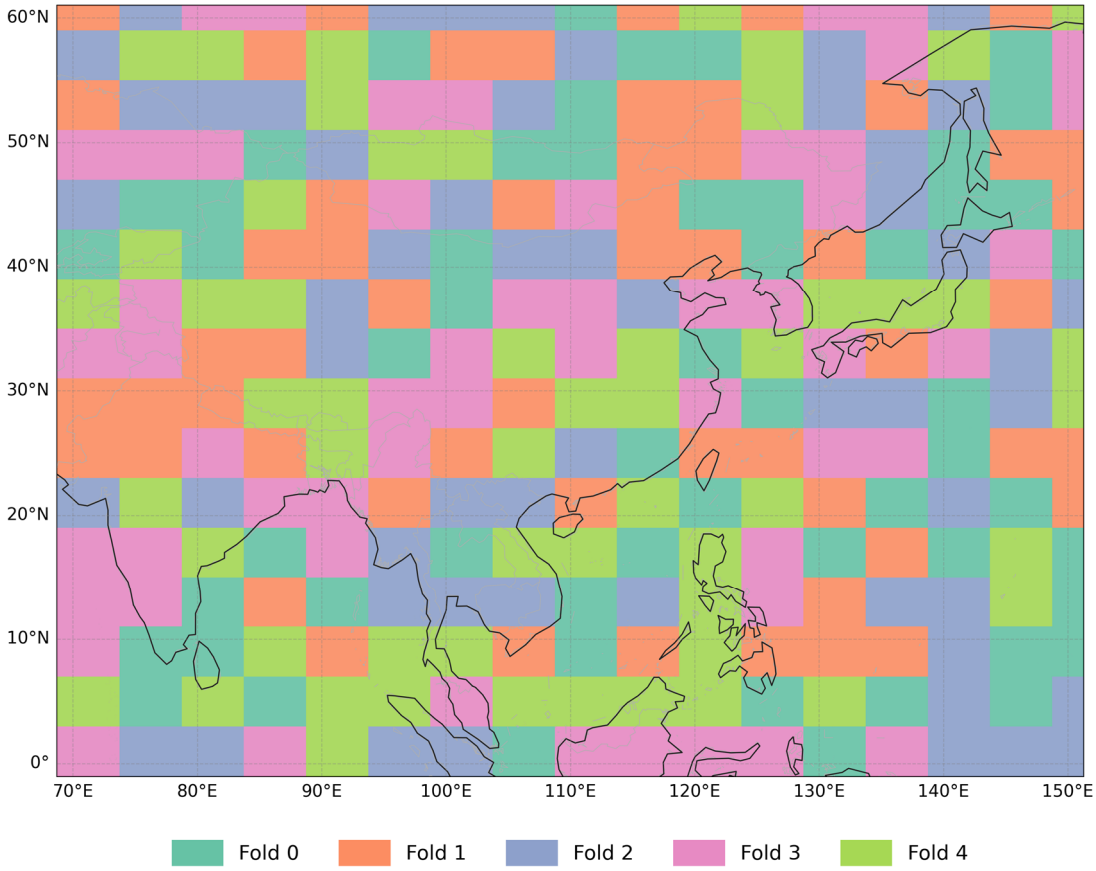


Figure S1. The illustration of the data splitting for the cross-validation. The color-coded checkerboard pattern represents the spatial partitioning of data into five folds used for model evaluation. Specifically, the geographical domain is divided into grid blocks, which are then randomly assigned to one of the five folds.

S6. Statistical Evaluation Metrics

To quantify the model performance in reconstructing AEPs, the following statistical metrics are employed. Let $\Delta_{AEC}^{target}{}_{i}$ and $\Delta_{AEC}^{pred}{}_{i}$ denote the observed (CALIOP) and predicted (Model) AEC for the i-th sample, respectively, and N be the total number of samples.

a. Pearson Correlation Coefficient (R)

Reflects the linear consistency between predictions and observations:

$$R = \frac{\sum_{i=1}^{N}\left(\Delta_{AEC}^{target}{}_{i} - \overline{\Delta_{AEC}^{target}}\right)\left(\Delta_{AEC}^{pred}{}_{i} - \overline{\Delta_{AEC}^{pred}}\right)}{\sqrt{\sum_{i=1}^{N}\left(\Delta_{AEC}^{target}{}_{i} - \overline{\Delta_{AEC}^{target}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(\Delta_{AEC}^{pred}{}_{i} - \overline{\Delta_{AEC}^{pred}}\right)^2}} \quad (S12)$$

where $\overline{\Delta_{AEC}^{target}}$ and $\overline{\Delta_{AEC}^{pred}}$ are the means of observed and predicted values.

b. Error Magnitude Metrics

① Mean Absolute Error (MAE): Represents the average magnitude of errors.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\Delta_{AEC}^{target}{}_{i} - \Delta_{AEC}^{pred}{}_{i}\right| \quad (S13)$$

② Root Mean Square Error (RMSE): Sensitive to large errors and outliers.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\Delta_{AEC}^{target}{}_{i} - \Delta_{AEC}^{pred}{}_{i})^2} \quad (S14)$$

c. Normalized Root Mean Square Error (NRMSE)

To compare performance across vertical layers with varying orders of magnitude in extinction coefficients, we normalize the RMSE by the range of the observed data:

$$NRMSE = \frac{RMSE}{\Delta_{AEC}^{target}{}_{max} - \Delta_{AEC}^{target}{}_{min}} \quad (S15)$$

where $\Delta_{AEC}^{target}{}_{max}$ and $\Delta_{AEC}^{target}{}_{min}$ are the maximum and minimum observed extinction coefficients in the target dataset, respectively.

d. Mean Bias (Bias)

Used to diagnose systematic overestimation ($Bias > 0$) or underestimation ($Bias < 0$):

159 $$Bias = \frac{1}{N}\sum_{i=1}^{N}\left(\Delta_{AEC\ i}^{pred} - \Delta_{AEC\ i}^{target}\right) \quad (S16)$$

S7. Mathematical Formulation of Interpretability Methods

To quantify the contribution of different input features to the bias correction, we employ ensemble-based interpretability methods. The mathematical definitions for the Gradient-based Attribution, Attention Mechanism, and Permutation Feature Importance are detailed below.

a. Gradient-based Feature Attribution (Input × Gradient)

For the vertical profile inputs, we utilize the Input × Gradient method to quantify the sensitivity of the model output to local feature variations. The importance score (I_{feet}) for a specific feature at a specific vertical layer is calculated as the ensemble average of the absolute gradients weighted by the input magnitude:

$$I_{feet} = \frac{1}{M}\sum_{m=1}^{M}\left(\frac{1}{N}\sum_{n=1}^{N}\left|x_{feat}^{(n)} \odot \frac{\partial \mathcal{F}_m\left(x^{(n)}\right)}{\partial x_{feat}^{(n)}}\right|\right) \quad (S17)$$

where: N is the total number of samples in the validation set. M is the number of cross-validation folds (here M = 5). $x_{feat}^{(n)}$ denotes the input value of a specific feature for the n-th sample. $\partial\mathcal{F}_m(\cdot)$ represents the model prediction function for the m-th fold. $\odot$ denotes element-wise multiplication.

b. Global Attention Mechanism

Global Attention Mechanism For global meteorological covariates (i.e., 2D single-layer variables from MERRA-2), we analyze the Cross-Attention weights extracted from the Transformer decoder. The global average attention score ($\bar{A}_{ij}$) represents the interaction strength between the i-th vertical layer of the profile and the j-th global variable (e.g., PBLH, Surface Pressure). It is computed by averaging across all attention heads and all samples:

$$\bar{A}_{ij} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{H}\sum_{h=1}^{H}A_{ij}^{(n,h)} \quad (S18)$$

where: H is the number of attention heads. N is the total number of samples in the dataset. $A_{ij}^{(n,h)}$ is the attention weight matrix from the h-th head for the n-th sample. The inner summation averages the contributions from multi-head attention mechanisms,

while the outer summation computes the dataset-wide average contribution of each global variable, filtering out sample-specific noise.

c. Gated Fusion Weights

The model dynamically fuses information from different sources using learnable scalar weights. The effective contribution weight (W_c) for a specific component c at altitude z is formalized as:

$$W_c(z) = \sigma\left(\frac{1}{N}\sum_{n=1}^{N}\omega_c^{(n)}(z)\right) \quad (S19)$$

where $\omega_c^{(n)}(z)$ is the raw gate value predicted by the model for sample n at altitude z, and $\sigma(\cdot)$ represents the normalization function (e.g., Softmax) ensuring the weights sum to 1.

d. Permutation Feature Importance

To assess the model's reliance on specific feature groups (e.g., Profile Features vs. Global Variables), we calculate the percentage increase in MSE when a feature group is randomly permuted. The importance score (S_k) for feature group k is defined as:

$$S_k = \frac{\mathcal{L}_{perm,k} - \mathcal{L}_{base}}{\mathcal{L}_{base}} \times 100\% \quad (S20)$$

$$\mathcal{L}_{base} = \frac{1}{N}\sum_{n=1}^{N}\left(y^{(n)} - \hat{y}^{(n)}\right)^2 \quad (S21)$$

where: $\mathcal{L}_{base}$ is the baseline MSE on the original validation set. $\mathcal{L}_{perm,k}$ is the MSE calculated after randomly shuffling the feature group k along the batch dimension while keeping other features fixed.

S8. Implementation of SHAP Analysis

To efficiently compute SHAP values for the deep neural network while maintaining statistical representativeness, we adopt a clustering-based sampling strategy for the background dataset:

a. Representative Background Selection

Instead of using a random subset or the zero-mean baseline (which is physically unrealistic for atmospheric variables), we apply K-means clustering on the normalized training dataset. We set the number of clusters k=100 and selected the medoids (samples closest to the cluster centers) to form a background dataset. This ensures that the reference baseline covers the full manifold of atmospheric states, from clean background days to severe pollution episodes.

b. Target Sample Selection

For regional analysis, we select top-50 samples with the highest model corrections (i.e., large initial GEOS-Chem biases that were successfully corrected by the model) within each region of interest. This "Case Study Strategy" ensures that the SHAP explanation focuses on the most significant bias correction events rather than trivial noise.

S9. Supplementary Evaluation of the Overall Model

Table S2. Summary of model performance for AEC simulation biases (Δ_{AEC}) on the validation subsets derived from the 5-fold spatial cross-validation strategy during the training phase.

Experiment ID	R (Val)	MAE (Val)	RMSE (Val)
Exp-2017	0.629	0.015	0.042
Exp-2018	0.652	0.015	0.041
Exp-2019	0.609	0.015	0.042
Average	0.630	0.015	0.042

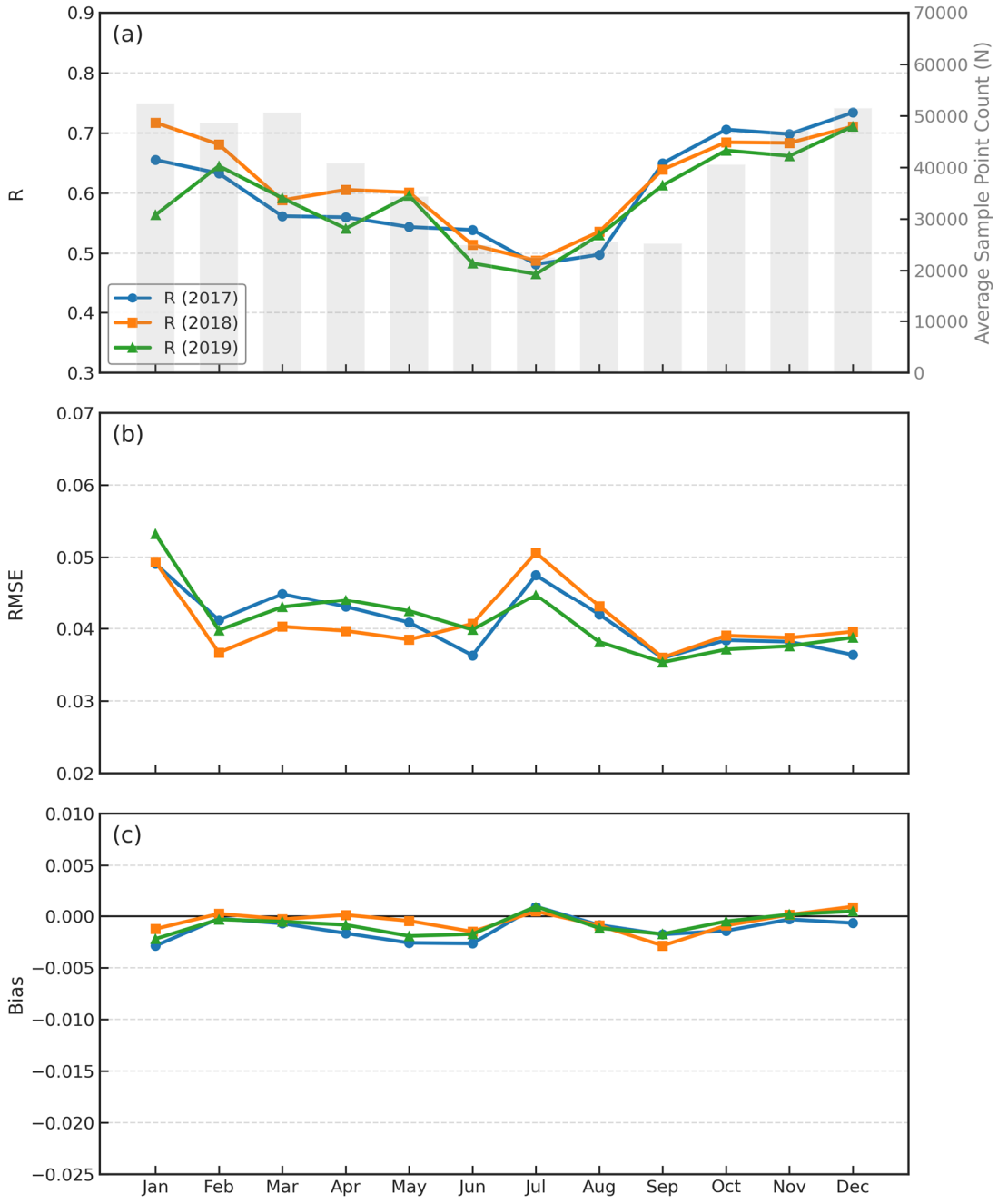


Figure S2. Monthly variations of the statistical evaluation metrics (derived from the 5-fold cross-validation validation subsets) for assessing the physics-informed Transformer model's performance in predicting the AEC simulation bias (Δ_{AEC}). The panels display the time series of R (a) with the multi-year average monthly sample size (N, gray bars), RMSE (b), and mean bias (c).

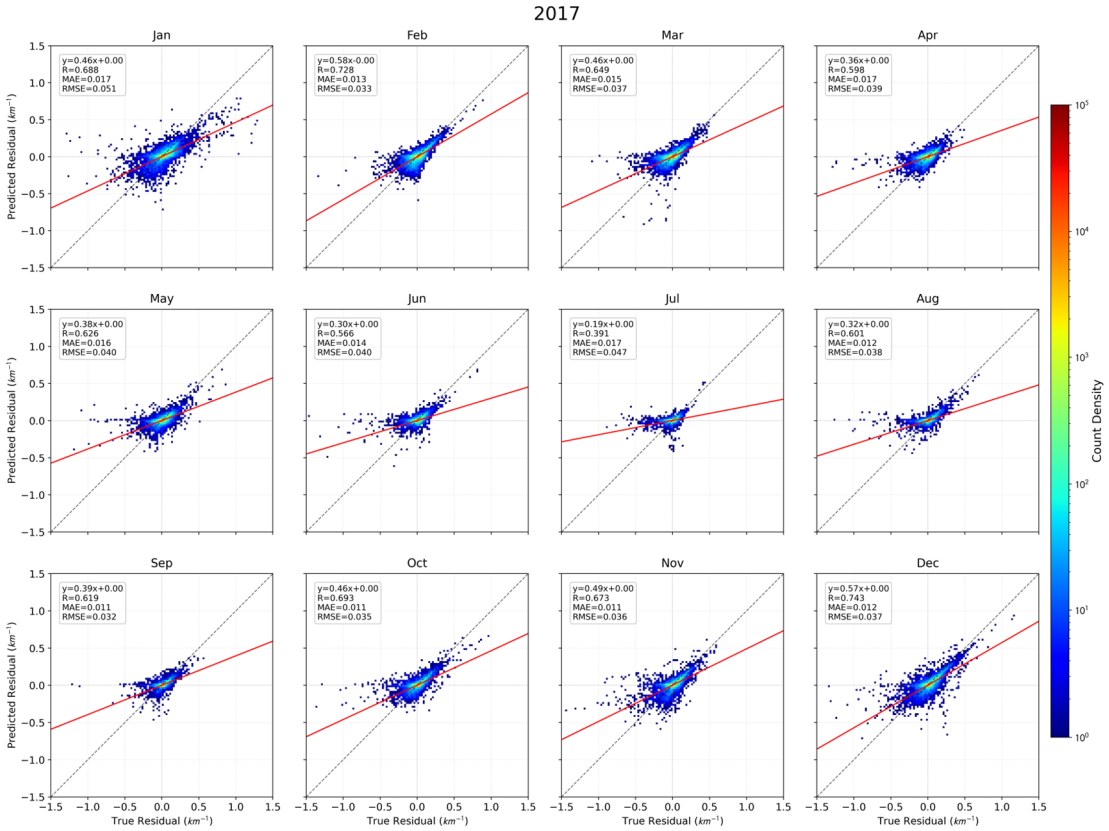


Figure S3. Monthly density scatter plots comparing the predicted AEC simulation bias (y-axis) generated by the physics-informed Transformer against the true residuals (x-axis) for the year 2017. The true residual is defined as the difference between the original GEOS-Chem simulation AEC and CALIOP observation AEC (GEOS-Chem – CALIOP).

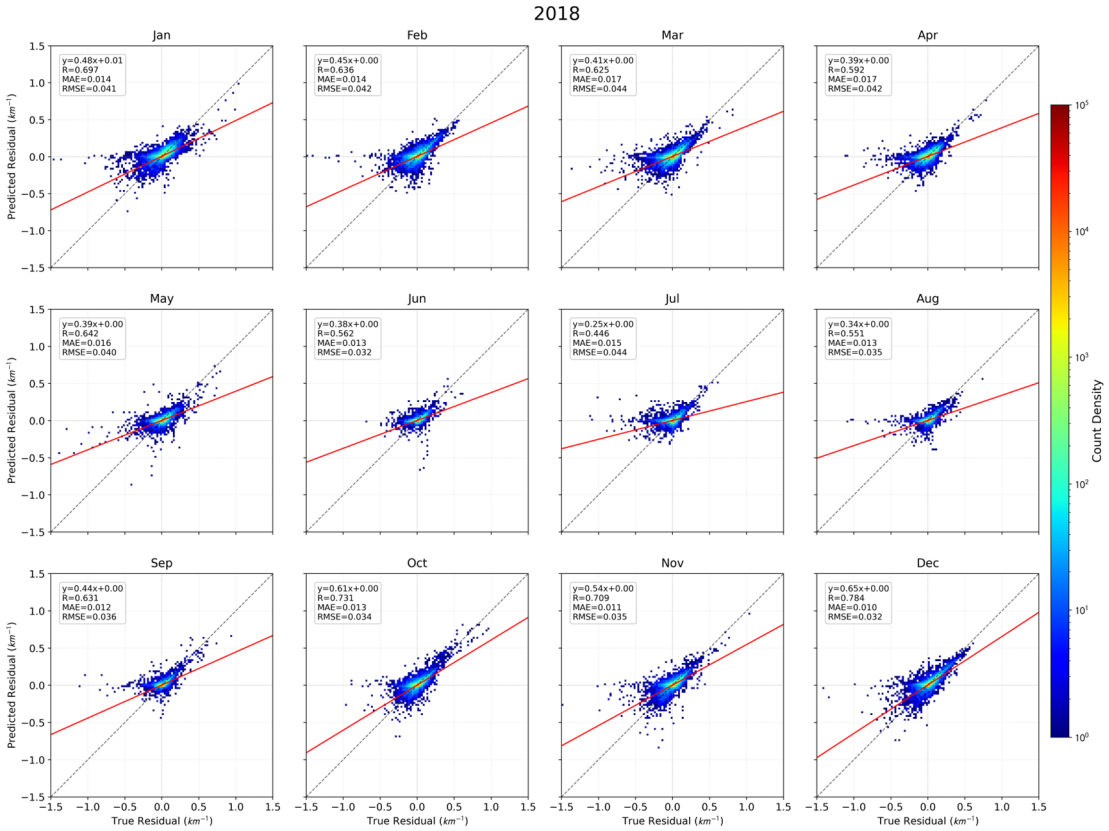


Figure S4. Monthly density scatter plots comparing the predicted AEC simulation bias (y-axis) generated by the physics-informed Transformer against the true residuals (x-axis) for the year 2018.

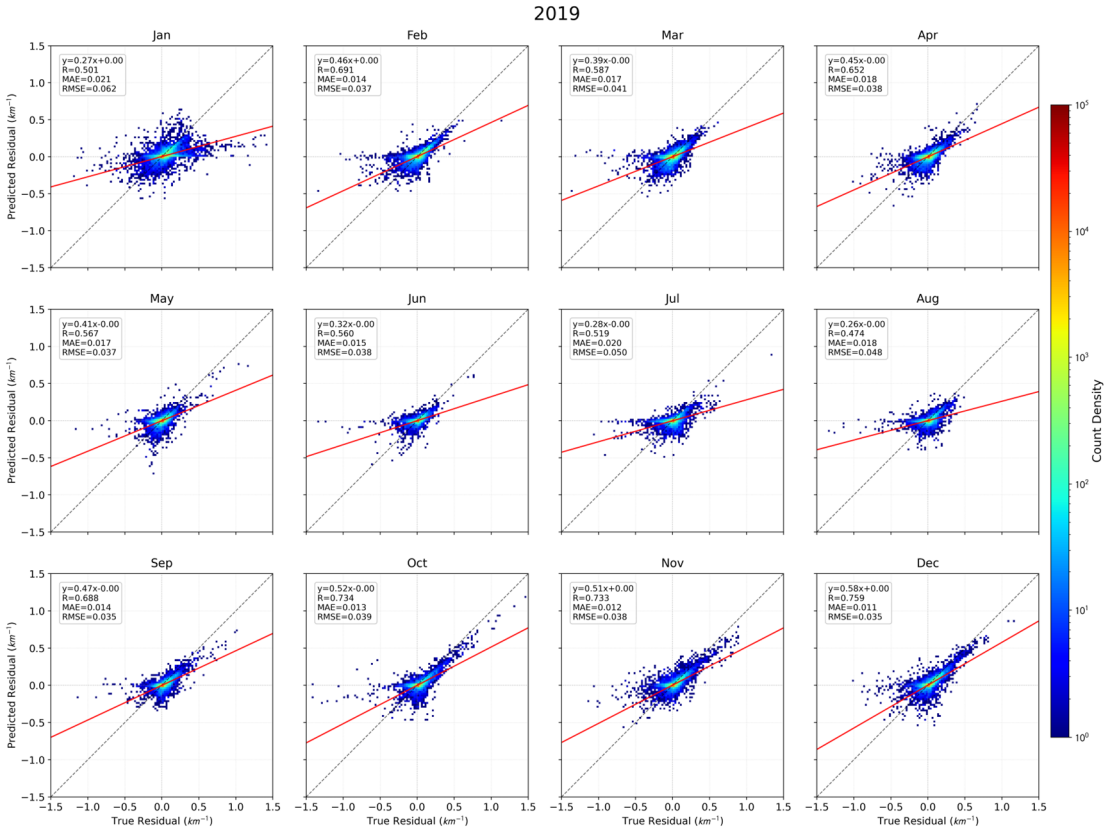
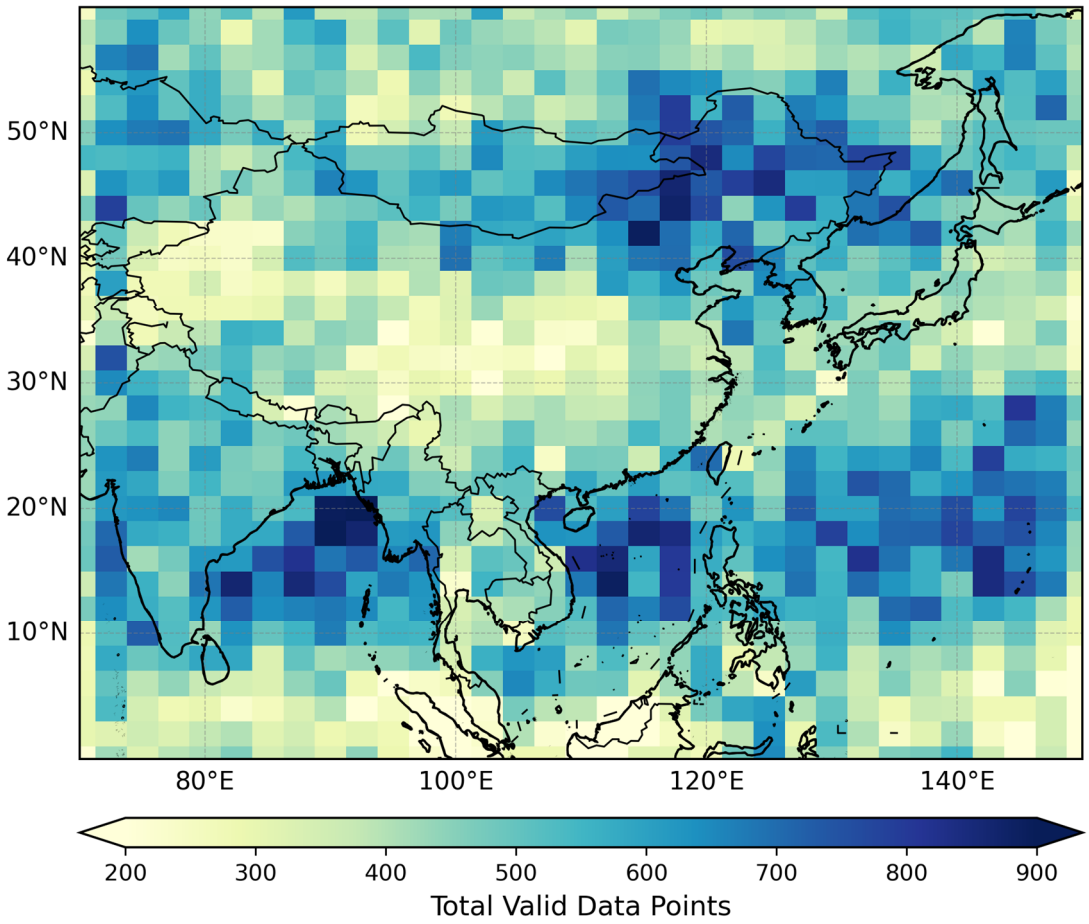


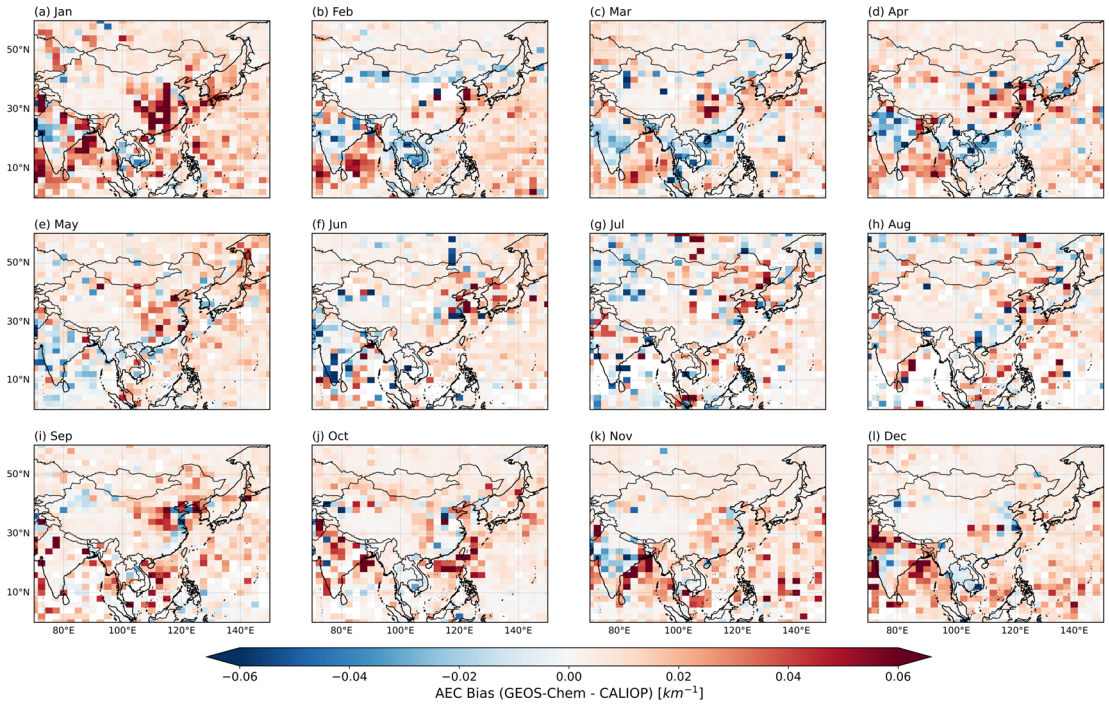
Figure S5. Monthly density scatter plots comparing the predicted AEC simulation bias (y-axis) generated by the physics-informed Transformer against the true residuals (x-axis) for the year 2019.



246

247 Figure S6. Spatial distribution of valid CALIOP data points aggregated for the test year

248 2019.



249

250 Figure S7. Monthly spatial evolution of the Target AEC simulation bias (Δ_{AEC}) for the

251 test year 2019.

S10. Analysis of Residual Outliers

This section analyzes the samples where the absolute residuals between the model-corrected AEC and CALIOP observations exceed 0.15 km^{-1} (i.e., points falling outside the error envelope in Fig. 6). The diagnostic results presented in Figure S8 demonstrate that these outliers are not randomly distributed but exhibit specific spatial and vertical characteristics.

Spatially (Fig. S8a), regions with elevated outlier occurrence are predominantly anchored over major emission source regions, including the IGP, the NCP, and the Indochina Peninsula. Vertically (Fig. S8b), over 80% of these extreme deviations are confined within the PBL (below 1.5 km AGL), an altitude range characterized by the heaviest local aerosol loading and the most intense turbulent mixing.

These distribution patterns confirm that the dispersion observed in the scatter plots fundamentally originates from representativeness errors. In complex source regions, the high-resolution footprint of CALIOP resolves transient, highly concentrated sub-grid plumes. The fine-scale physical structures of these plumes are inherently smoothed out during the spatial averaging onto the coarse 2°×2.5° grid of the GEOS-Chem model. We attribute the remaining 1.20% of outlier samples primarily to the spatial heterogeneity of aerosols that lies below the resolvable scale of the model grid.

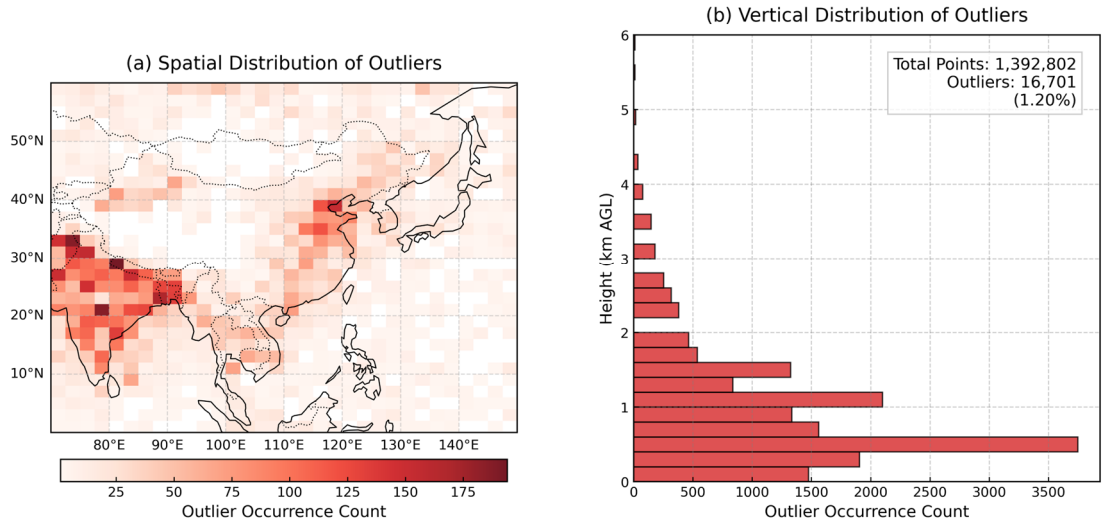


Figure S8. Spatial and vertical distributions of the residual outliers for the combined test years (2017–2019). Outliers are strictly defined as samples where the absolute

residual between the GC-TF prediction and CALIOP observation exceeds ± 0.15 km^{-1}. The panels display (a) the spatial occurrence count of these outliers mapped onto the native 2°×2.5° GEOS-Chem grid, and (b) their vertical distribution as a function of height AGL. The statistical box indicates that these extreme deviations account for merely 1.20% of the total valid dataset.

S11. Supporting Information for Improvement Evaluation of GEOS-Chem Simulations

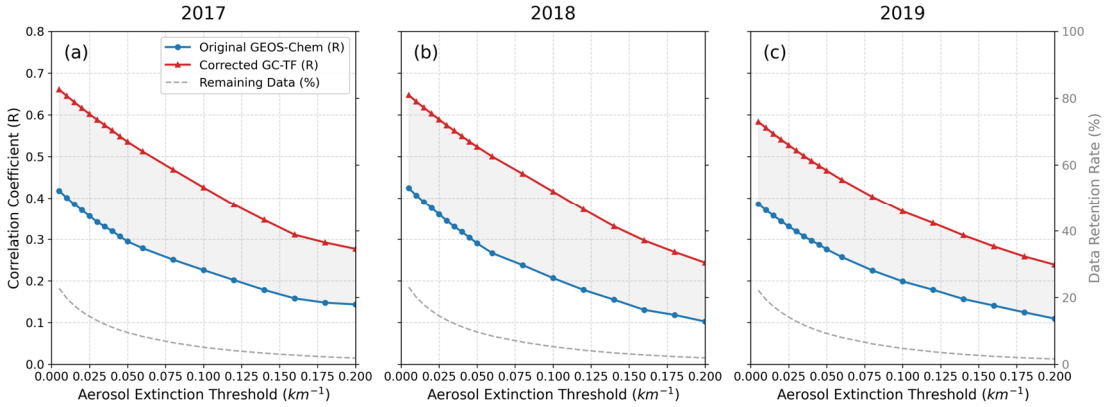


Figure S9. Sensitivity of model predictive performance to AEC thresholds across three independent test years (2017–2019). The gray dashed line (corresponding to the right axis) indicates the data retention rate, representing the percentage of valid samples remaining after filtering out clean background signals below the threshold

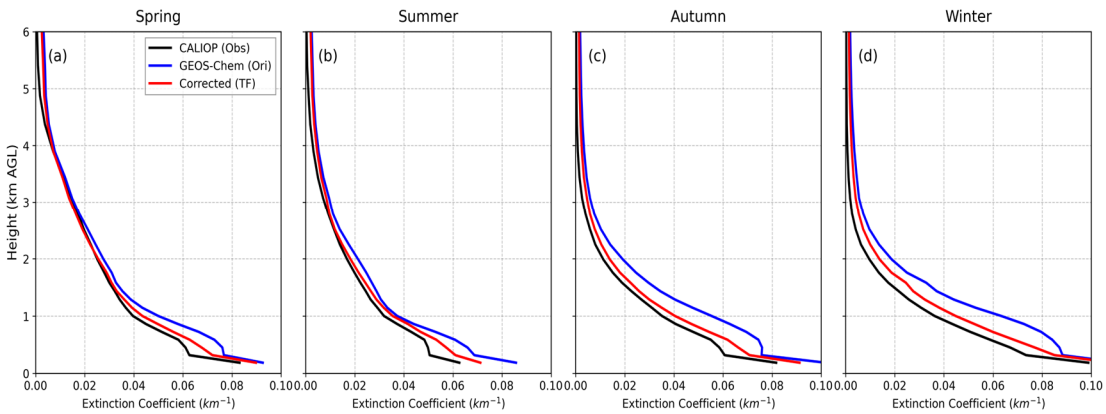


Figure S10. Seasonal mean vertical profiles of AEC (km^{-1}) averaged over the three test years (2017–2019). The panels display the profiles for Spring (MAM, a), Summer (JJA, b), Autumn (SON, c), and Winter (DJF, d). The black lines represent the CALIOP observations, the blue dashed lines denote the original GEOS-Chem simulations, and the red lines show the results corrected by the GC-TF model. The profiles are vertically resolved from the surface to 6 km above ground level (AGL).

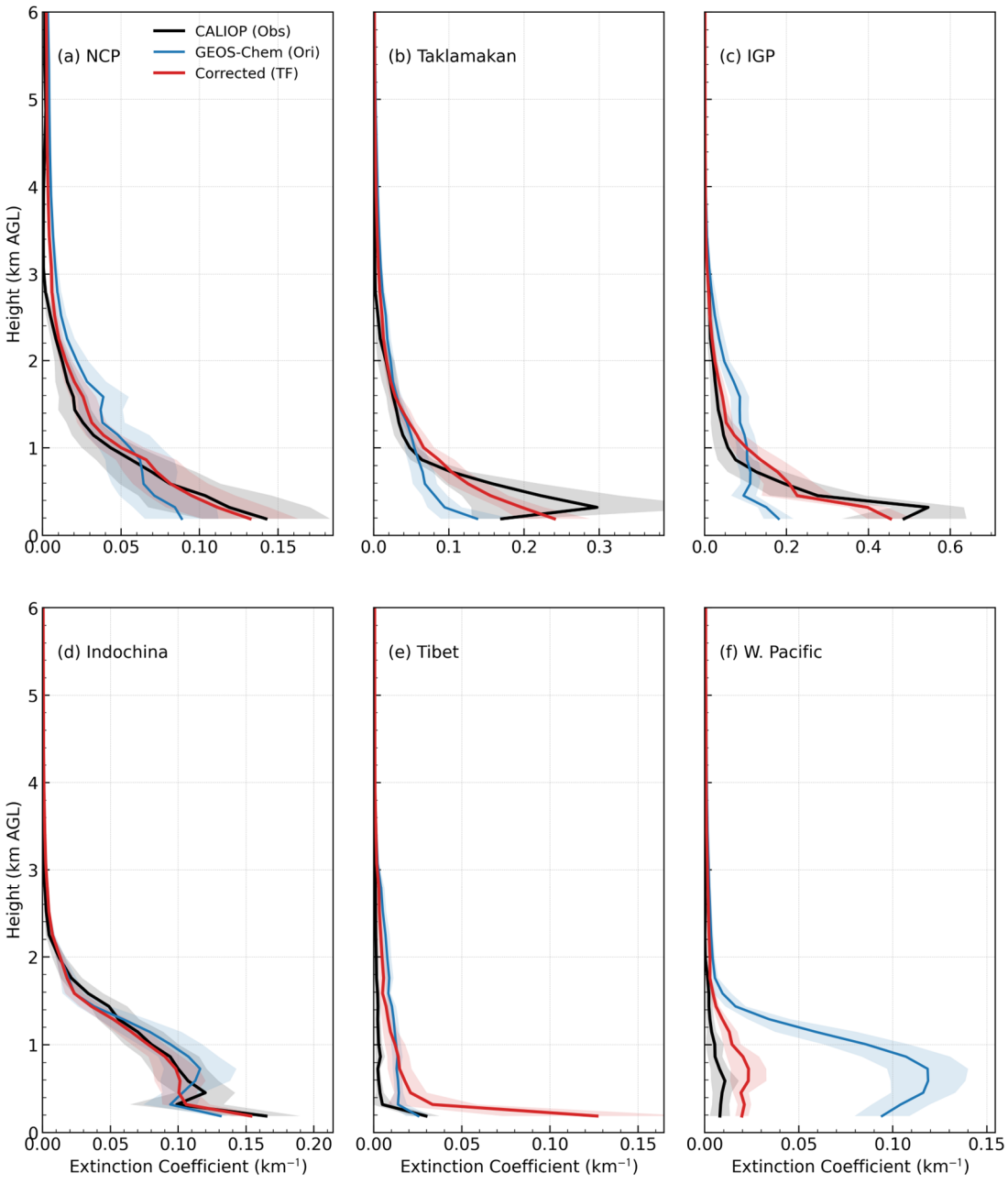


Figure S11. Regional annual mean vertical profiles of AEC over six representative sub-regions for the test year 2017. The panels display the profiles for NCP (a), Taklamakan Desert (b), IGP (c), Indochina (d), Tibetan Plateau (e), and Western Pacific (f). The black lines represent CALIOP observations, the blue dashed lines denote the original GEOS-Chem simulations, and the red lines show the results corrected by the GC-TF model. The shaded areas indicate the regional standard deviation ($\pm 0.5\sigma$), representing the spatial variability within each sub-region.

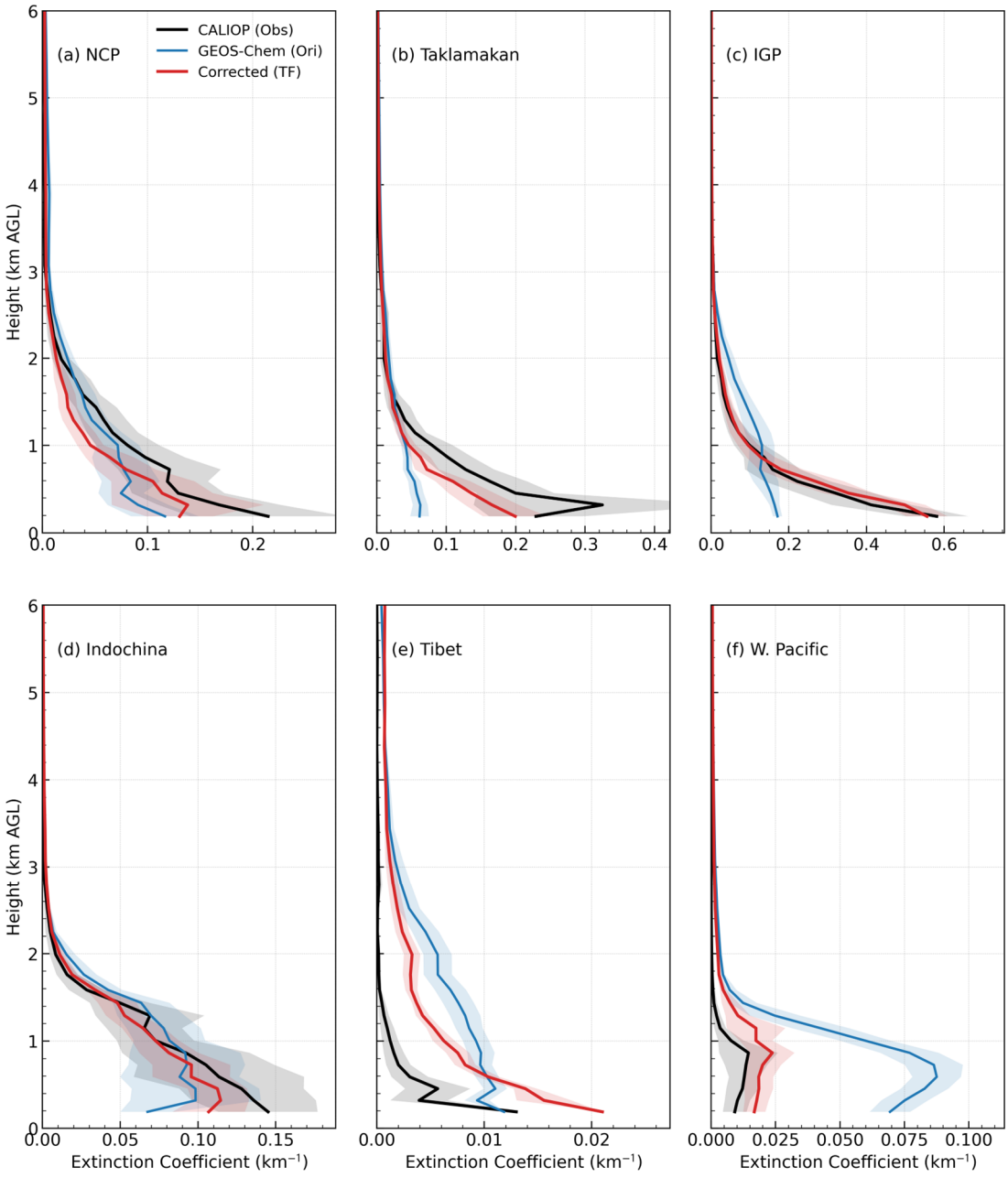


Figure S12. Regional annual mean vertical profiles of AEC over six representative sub-regions for the test year 2018.

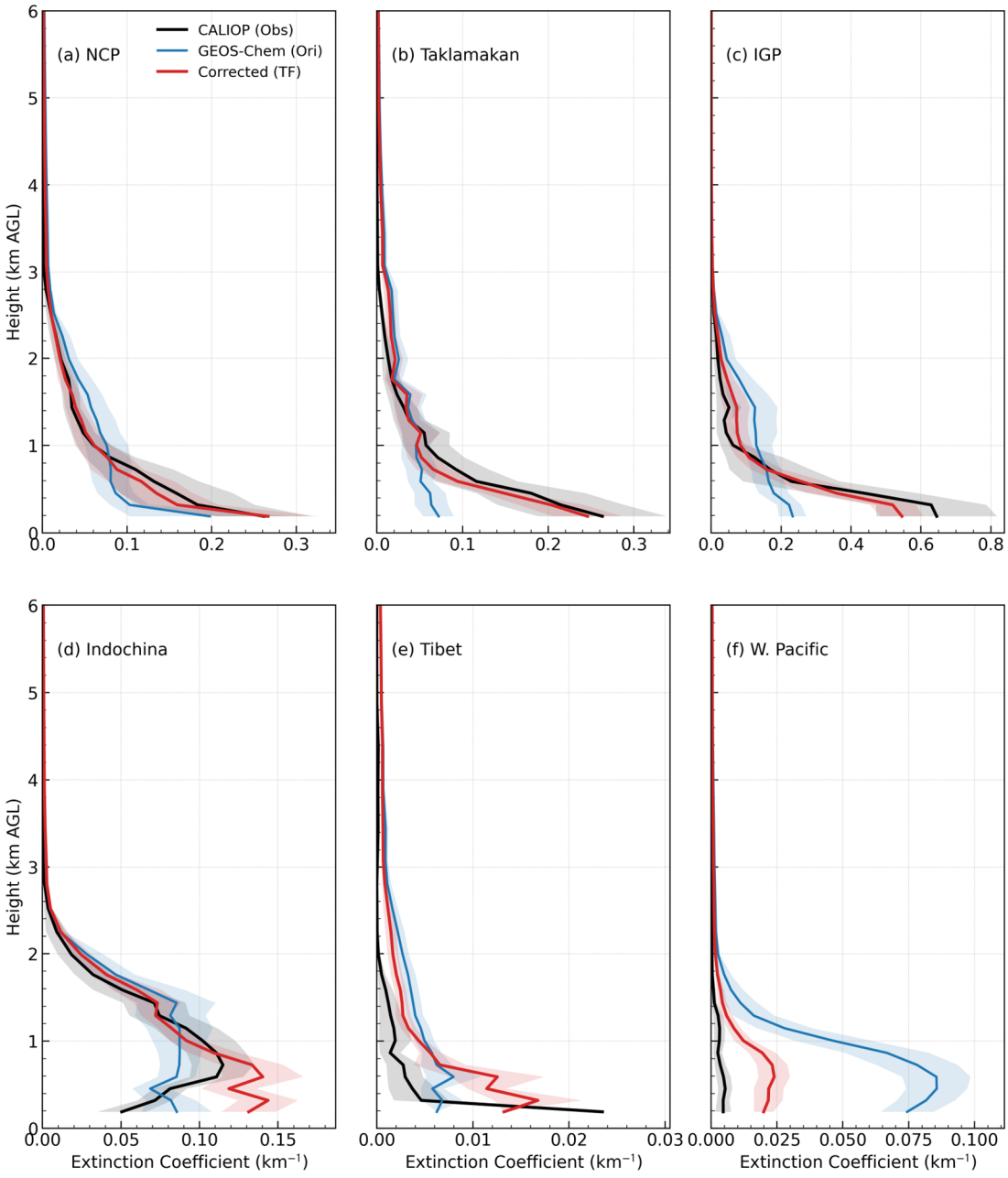


Figure S13. Regional annual mean vertical profiles of AEC over six representative sub-regions for the test year 2019.

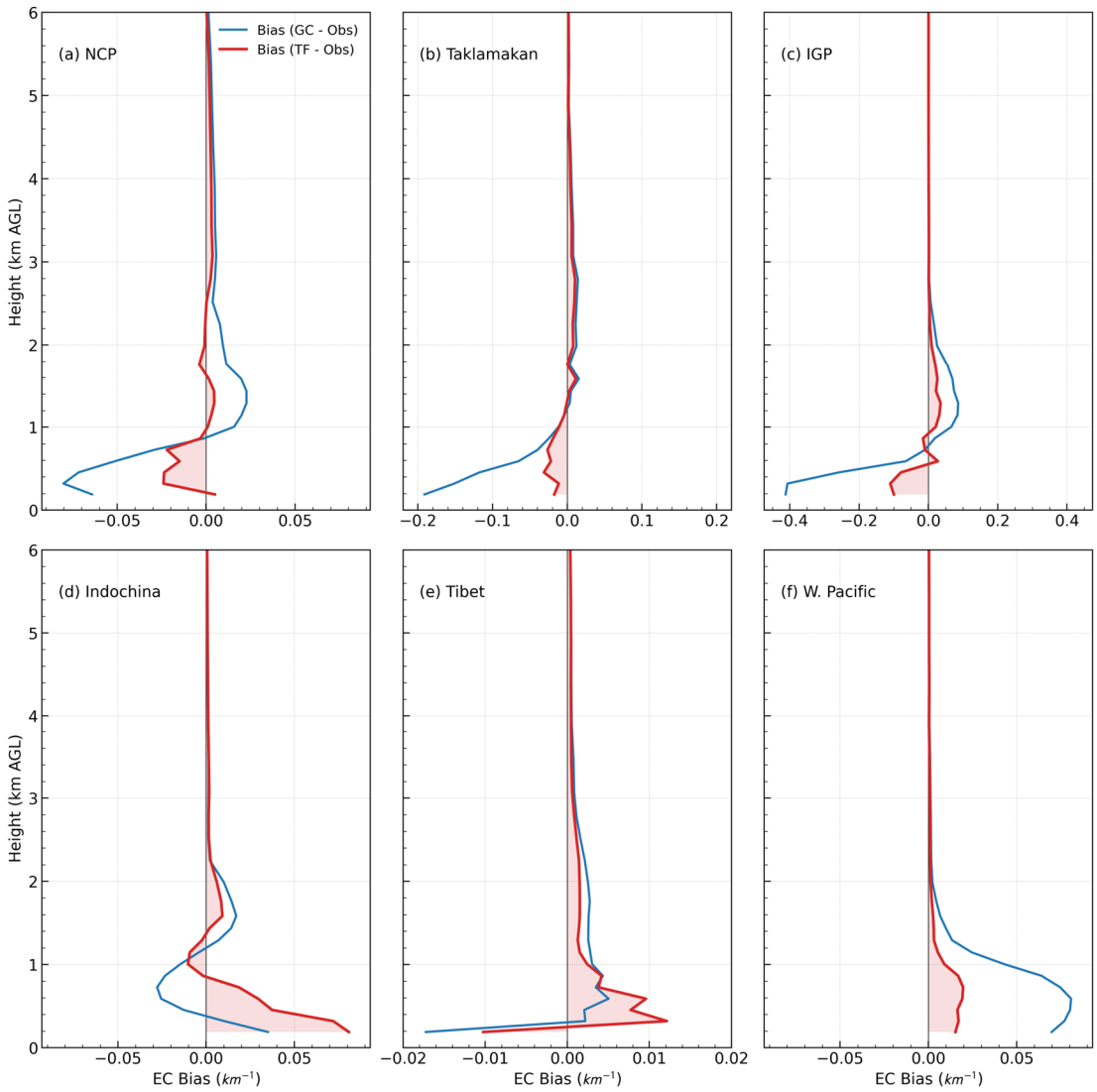


Figure S14. Vertical profiles of the annual mean AEC simulation bias over six representative sub-regions for the test year 2019. The panels display the bias profiles for NCP (a), Taklamakan Desert (b), IGP (c), Indochina (d), Tibetan Plateau (e), and Western Pacific (f). The AEC bias is calculated as Model minus Observation. The blue lines represent the original GEOS-Chem bias (GEOS-Chem−CALIOP), while the red lines denote the remaining bias after GC-TF correction (GC-TF − CALIOP). The vertical black line indicates the zero-bias reference, and the shaded red areas highlight the magnitude of the residual bias after correction.

S12. Supporting Information for Interpretability Analysis

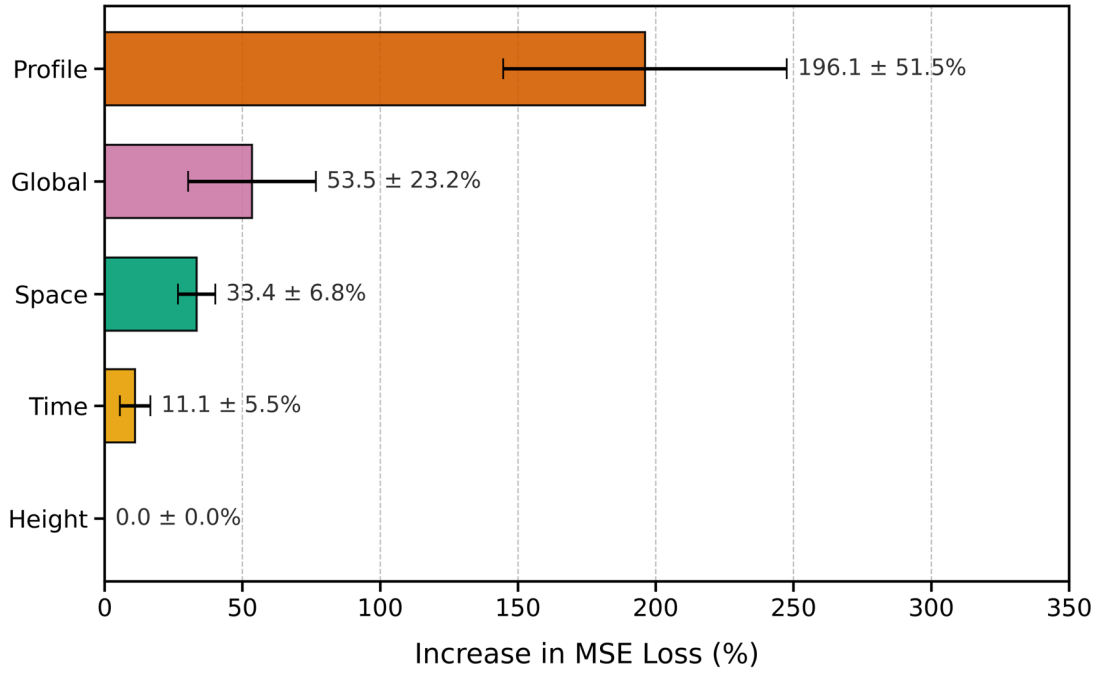


Figure S15. Permutation feature importance analysis evaluated on the test dataset. The importance of each feature group is quantified by the relative increase in MSE loss when the values of that feature group are randomly shuffled, while keeping others unchanged. The error bars represent the standard deviation across the 5-fold cross-validation. Note that Profile Features are identified as the most critical input, causing a ~200% surge in error if disrupted, confirming the model's heavy reliance on physical state variables for magnitude prediction.

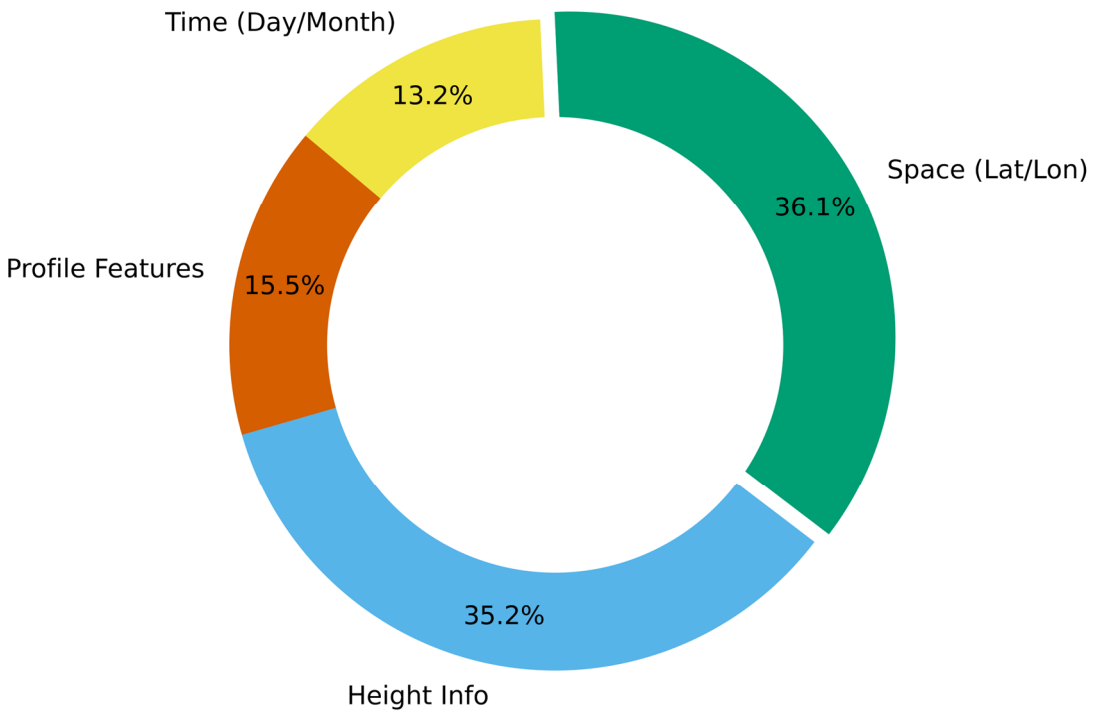


Figure S16. Global average relative importance of the input features learned by the physics-informed model for the test year 2019. The pie chart displays the percentage contribution of four feature groups: Space (36.1%), Height Info (35.2%), Profile Features (15.5%, including GEOS-Chem chemical concentrations and MERRA2 meteorology), and Time (13.2%). The dominance of spatial and height information reflects the strong spatial heterogeneity and vertical stratification of aerosol distributions.

S13. Sensitivity Analysis of CALIOP Observational Uncertainties

To quantitatively evaluate the model's sensitivity to the inherent systematic uncertainties of CALIOP observations, we design a perturbation experiment based on the 2017 independent test set. Considering the reported mean relative bias of CALIOP AOD against AERONET is approximately -5.1% ± 8.5%, we artificially apply a constant ±5% multiplier to the CALIOP AEC targets during the training phase. This creates two extreme scenarios: one simulating a severe systematic underestimation (-5% bias) and another simulating a systematic overestimation (+5% bias). The physics-informed Transformer is then retrained from scratch for both scenarios, and the newly predicted residual profiles are added back to the original GEOS-Chem simulations to obtain the final perturbed corrected AEC profiles.

As presented in Table S3, the GC-TF model exhibits strong resistance to target perturbations. Even with a 5% systematic error injected into the learning target, the model's RMSE and Mean Bias remain highly stable and significantly superior to the original GEOS-Chem baseline. Visually, Figure S17 illustrates that the perturbed predictions (dashed and dotted lines) tightly hug the unperturbed baseline correction (solid red line), forming a narrow perturbation range (red shaded area) that aligns well with the CALIOP observations. This demonstrates that the data-driven correction framework captures robust physical mappings rather than merely overfitting to observational noise.

Table S3. Quantitative evaluation of the GC-TF model's sensitivity to CALIOP observational uncertainties on the 2017 independent test set. The metrics are derived by validating both the original GEOS-Chem simulations and the corrected GC-TF results against the unperturbed CALIOP AEC profiles.

Model/Scenario	RMSE	MAE	Mean Bias	R
Original GEOS-Chem	0.052	0.020	0.006	0.522
Baseline GC-TF (Unperturbed)	0.039	0.014	0.002	0.732
GC-TF (-5% CALIOP Bias)	0.040	0.014	0.001	0.716
GC-TF (+5% CALIOP Bias)	0.040	0.015	0.004	0.718

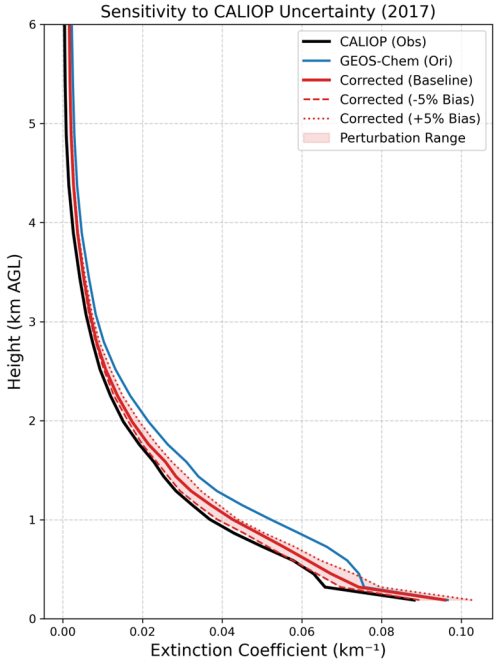


Figure S17. Vertical profiles illustrating the sensitivity of the corrected AEC to target perturbations for the test year 2017. The black solid line represents the CALIOP observations, and the blue solid line represents the original GEOS-Chem simulation. The red solid line indicates the baseline correction of the GC-TF model trained on unperturbed data. The dashed and dotted red lines represent the corrected profiles generated by models trained with -5% and +5% perturbed CALIOP targets, respectively. The light red shaded area denotes the envelope of variation (Perturbation Range) induced by these observational uncertainties.

S14. Stratified Evaluation by Aerosol Subtypes

To evaluate the model performance under distinct aerosol composition regimes, we utilize the Feature Classification Flags embedded in the CALIOP Level 2 Atmospheric Volume Description. The NA evaluation dataset (2018) is partitioned into two representative subsets: a dust-dominated regime (combining 'Dust' and 'Polluted Dust' subtypes) and an SOA-dominated continental regime (combining 'Clean Continental' and 'Polluted Continental/Smoke' subtypes).

Table S4. Statistical evaluation of GC-TF model performance over NA, stratified by dominant CALIOP aerosol subtypes (2018). For each subset, we calculate statistical metrics for both the original GEOS-Chem simulation and the GC-TF corrected predictions against CALIOP AOD observations.

Aerosol Regime	Model	R	RMSE	MAE	Slope
Dust-dominated (Dust and Polluted Dust)	Original	0.41	0.032	0.027	0.21
	Corrected	0.50	0.032	0.027	0.32
SOA-dominated (Clean and Polluted Continental)	Original	0.51	0.034	0.031	0.27
	Corrected	0.50	0.035	0.031	0.30

S15. Methodological Benchmarking and Structural Necessity

To justify the architectural complexity of the proposed framework and isolate the sources of its performance gains, we conduct comprehensive benchmarking and ablation studies using the independent 2017 test dataset. To ensure a strictly fair comparison, all baseline models and ablation variants are trained using the identical set of input predictors—encompassing GEOS-Chem physicochemical states and MERRA-2 meteorological forcings—along with identical hyperparameter configurations and loss functions.

To establish a comprehensive baseline, two representative conventional deep learning architectures were evaluated. The first is a Multilayer Perceptron (MLP), representing point-wise neural networks. By treating vertical layers as independent vectors, the MLP tests whether a simple numerical mapping, devoid of sequential awareness, can resolve AEC biases. The second baseline is a 1-Dimensional Convolutional Neural Network (1D-CNN). This architecture utilizes localized receptive fields to capture vertical gradients between adjacent layers, serving as a benchmark for local structural extraction, contrasting with the global dependency modeling enabled by the Transformer.

Table S5. Performance benchmarking and ablation study of the proposed model against conventional machine learning architectures. Evaluation is conducted on the independent 2017 test dataset. All models are trained utilizing the identical meteorological and chemical state predictors to ensure a rigorous comparison.

Model Configuration	R	MAE (km^{-1})	RMSE (km^{-1})
MLP	0.083	0.019	0.052
1D-CNN	0.540	0.016	0.044
Without Gated Fusion	0.637	0.015	0.040
Without Cross-Attention	0.654	0.014	0.039
Physics-Informed Transformer (Full)	0.666	0.014	0.039