Black color represents the review comments, blue color represents the reply comments, and green color represents the revised contents of the manuscript and supplement.

Reviewer #1:

This manuscript presents a sophisticated physics-informed Transformer framework to correct GEOS-Chem aerosol extinction coefficient profiles using CALIOP observations. The study is ambitious, methodologically advanced, and addresses an important problem in bridging chemical transport models (CTMs) and vertically resolved lidar observations. The reported improvements in correlation and RMSE, along with cross-continental transferability tests, are promising. However, several issues require clarification before the scientific contribution and methodological advantage can be properly evaluated as follows.

First, the scientific objective requires clearer framing. CALIOP observations are used to define simulation bias during training, but they are not included as inputs during inference. Therefore, the framework is not performing data assimilation, but rather learning a state-dependent mapping between atmospheric variables and historical GEOS-Chem biases. If the goal is to generate corrected AEC fields when CALIOP is unavailable, the method should be clearly described as a supervised bias-correction model conditioned on CTM state and meteorology, and its limitations should be acknowledged. For example, if key emissions (e.g., wildfire events) are missing in GEOS-Chem and not represented in the input features, the model cannot reconstruct those missing signals. The correction is inherently constrained by the information content of the CTM and meteorological predictors. The manuscript should therefore distinguish more carefully between correcting systematic state-dependent biases and compensating for missing physical processes. Clarifying this distinction would strengthen the scientific positioning of the study.

Response:

We appreciate the reviewer's insightful assessment regarding the scientific positioning of our framework. We completely agree with the reviewer's assessment: our method operates as a supervised bias-correction model driven by a priori atmospheric

states, rather than a Data Assimilation (DA) system that updates state variables using concurrent observations. Consequently, its corrective capacity is inherently bounded by the information content of the GEOS-Chem and MERRA-2 predictors.

To clarify this critical distinction and acknowledge the model's limitations regarding entirely missing physical processes, we have made systematic revisions throughout the manuscript:

We have revised the Introduction and Method sections to explicitly frame the model as a supervised bias-correction approach and differentiate it from DA.

Added to Section 1:

"…Distinct from traditional DA systems that require concurrent observational inputs to iteratively update state variables, our framework operates as a supervised bias-correction model. It captures the intrinsic state-dependent mapping between CTM structural uncertainties and diverse environmental contexts. By conditioning the correction exclusively on the CTM's a priori state and meteorological drivers, the model effectively mitigates systematic biases without relying on CALIOP data during the inference phase…."

Added to Section 3.1:

"…The model is designed to rectify state-dependent systematic biases rather than to artificially reconstruct aerosol signals from completely unrepresented physical processes that lack corresponding perturbation signatures in the input fields."

Added to Section 5:

"…Functioning as a supervised bias-correction model rather than a DA system, this framework learns a state-dependent mapping to rectify systematic simulation AEC bias…"

We have added a new "Section 4.7 Model Limitations and Scope of Application" to openly discuss the framework's inability to compensate for completely missing physical signals, using the reviewer's excellent example of unrecorded wildfires.

Added Section 4.7:

"**4.7 Model Limitations and Scope of Application**

As a supervised bias-correction framework, the model relies on state-dependent

mapping, meaning its performance is fundamentally constrained by the predictive signals available in the input features. The framework excels at correcting systematic, parameterization-driven bias. For instance, it successfully restores the underestimated dust plumes in the Taklamakan Desert by leveraging wind speed, clear-sky radiation, and vegetation indices as physical proxies for actual dust emission conditions (Section 4.5.4).

However, the model possesses limited capacity to compensate for entirely missing physical processes. If a highly localized or stochastic event is completely absent from the prescribed emission inventory and produces no corresponding anomalies in the input meteorological or chemical precursor fields, the model lacks the necessary physical constraints to reconstruct the resulting aerosol plume. In such scenarios, the correction remains strictly bounded by the prior information provided by the GEOS-Chem and MERRA2."

Second, the model architecture appears to rely on instantaneous vertical profiles and meteorological context, without explicit time-series modeling. It is unclear whether any temporal continuity, lagged predictors, or time-window averaging is incorporated into the inputs. A precise description of the temporal collocation strategy between GEOS-Chem and CALIOP is necessary to assess the robustness of the results. In addition, the manuscript does not discuss how diurnal variability in aerosol vertical structure is handled. Given the strong diurnal cycle of boundary layer evolution, turbulent mixing, hygroscopic growth, and photochemistry, aerosol extinction can vary substantially on hourly timescales. It should be clarified whether simple hour-by-hour matching is sufficient, or whether a temporal window similar to those used in traditional data assimilation frameworks, was considered to reduce representativeness errors. Without such analysis, it remains uncertain whether the reported improvements reflect stable bias correction or sensitivity to sampling timing and diurnal variability.

Response:

We sincerely appreciate the reviewer's rigorous examination of our temporal

collocation strategy and the treatment of diurnal variability. These are indeed critical methodological aspects. We completely agree that a precise description of these processes is necessary to assess the robustness of our data-driven bias correction framework. To address your concerns, we have systematically expanded our methodology and discussion sections. Our responses are detailed below from three perspectives:

1. Temporal Collocation Strategy

To construct the training dataset, we employ a precise nearest-hour collocation approach, generating training pairs exclusively for the specific grid cells and hours where valid CALIOP observations are available. Specifically, the CALIOP Level 2 overpass times are mathematically rounded to the nearest UTC hour, and these observations are strictly paired with the GEOS-Chem 1-hourly instantaneous outputs corresponding to that exact matched hour.

We deliberately refrain from utilizing the time-window averaging commonly applied in traditional Data Assimilation (DA) frameworks. Unlike DA, which typically assimilates sparse observations to adjust an entire regional state over an assimilation window, our deep learning framework operates as a supervised point-to-point mapping. Because the CALIOP instrument records highly localized vertical profiles along its polar orbit at specific instantaneous moments, applying a temporal moving average to the GEOS-Chem outputs inadvertently smooths out transient atmospheric features, such as the sharp peak of the boundary layer depth or narrow smoke plumes. Therefore, aligning the CTM's instantaneous output with the concurrent instantaneous satellite observation directly minimizes representativeness errors and preserves the strict physical consistency required for our modeling task.

2. Handling of Diurnal Variability

Regarding the handling of diurnal variability, we completely concur with the reviewer that aerosol vertical structures are highly sensitive to diurnal cycles driven by boundary layer evolution, turbulent mixing, and photochemistry. While our architecture does not employ lagged predictors or explicit historical time-series modeling, it robustly captures diurnal variability through implicit state-driven modeling. To achieve

this, we precisely match the 3-hourly MERRA-2 meteorological fields with the specific satellite overpass hour using a temporal interpolation strategy, applying nearest-neighbor selection for times within one hour of the reanalysis timestamps and midpoint averaging for the intermediate hours. Consequently, the model is explicitly informed by concurrent meteorological states—such as Planetary Boundary Layer Height (PBLH), Sensible Heat Flux (HFLUX), and Photosynthetically Active Radiation (PAR)—which inherently carry the strong signatures of the diurnal cycle. Furthermore, categorical temporal embeddings, including explicit day/night flags and cyclic month encodings, are incorporated into the network to provide temporal context. By conditioning the bias correction on these physically meaningful, state-dependent meteorological forcings, the model dynamically resolves diurnal atmospheric processes rather than relying on historical sequences.

3. Empirical and Physical Evidence of Robustness

To directly address the concern regarding whether the improvements merely reflect a sensitivity to CALIOP's limited sampling timing (typically ~01:30 and 13:30 local time), we provide both empirical and physical evidence in the manuscript. Empirically, as detailed in Section 4.4, we evaluate the corrected AOD against continuous, high-frequency ground-based AERONET measurements. As shown in the time series for the Kanpur and Nong Khai episodes (Figure 11d-e), the GC-TF model successfully captures the dynamic diurnal evolution and phase fluctuations of aerosols across all daylight hours. This demonstrates that the model generalizes robustly across the entire diurnal cycle and avoids overfitting to CALIOP's specific twice-a-day overpass times. Physically, our interpretability analysis in Section 4.5 confirms that the model explicitly relies on diurnal drivers. For instance, the SHAP analysis (Figure 15) reveals a high reliance on Sensible Heat Flux (HFLUX) in urban regions and diffuse/direct PAR in biomass burning and dust regions. Because HFLUX directly drives turbulent mixing and boundary layer evolution, while PAR governs secondary aerosol photochemistry, their high feature importance proves that the model successfully internalizes the physical mechanisms governing the diurnal cycle. This demonstrates that the model dynamically adjusts its predictions based on concurrent

thermodynamic and photochemical states, rather than employing a static or time-memorized correction based solely on the sensor's overpass hour.

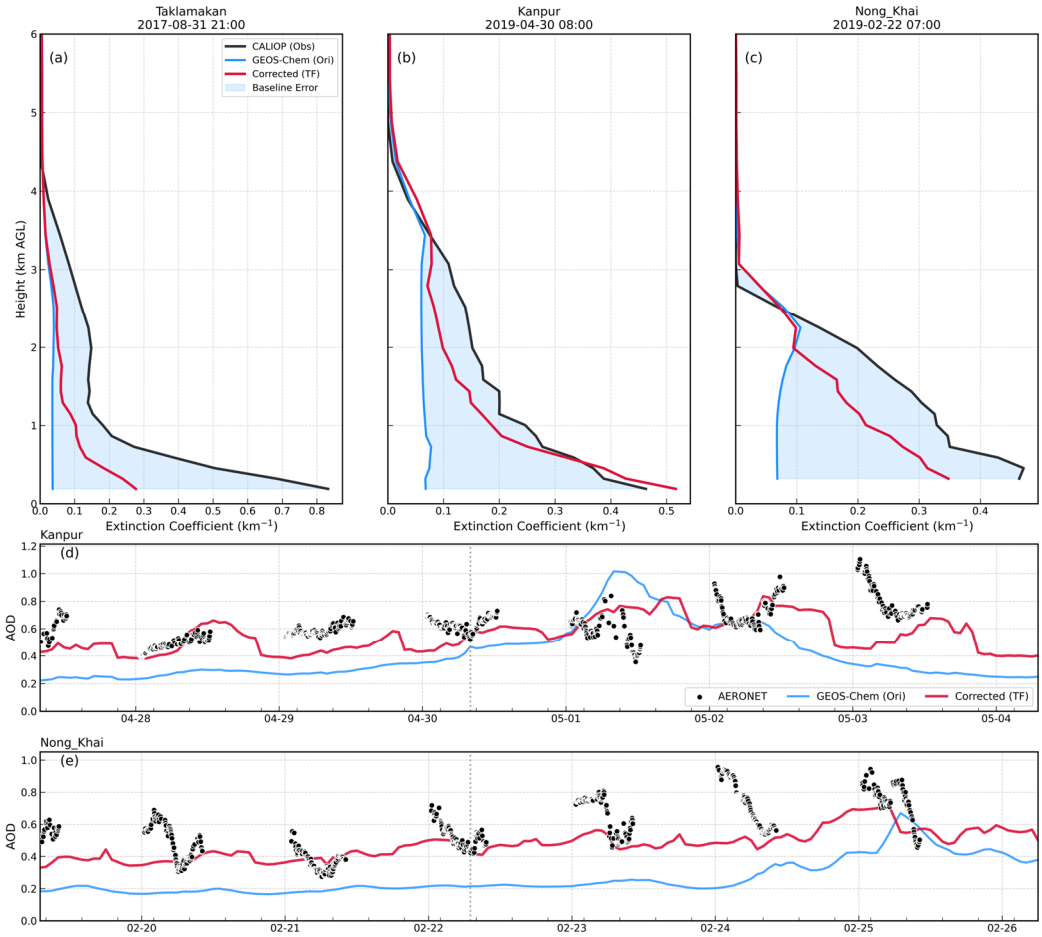


Figure 11. Composite analysis of aerosol vertical structures and temporal evolution during selected pollution episodes. Vertical profiles of AEC at three representative sites: Taklamakan (Dust, a), Kanpur (Anthropogenic Pollution, b), and Nong Khai (Biomass Burning, c). Time series of AOD at the Kanpur (d) and Nong Khai (e) AERONET sites during the corresponding pollution events. The vertical dotted lines mark the CALIOP overpass times (UTC) shown in the top panels.

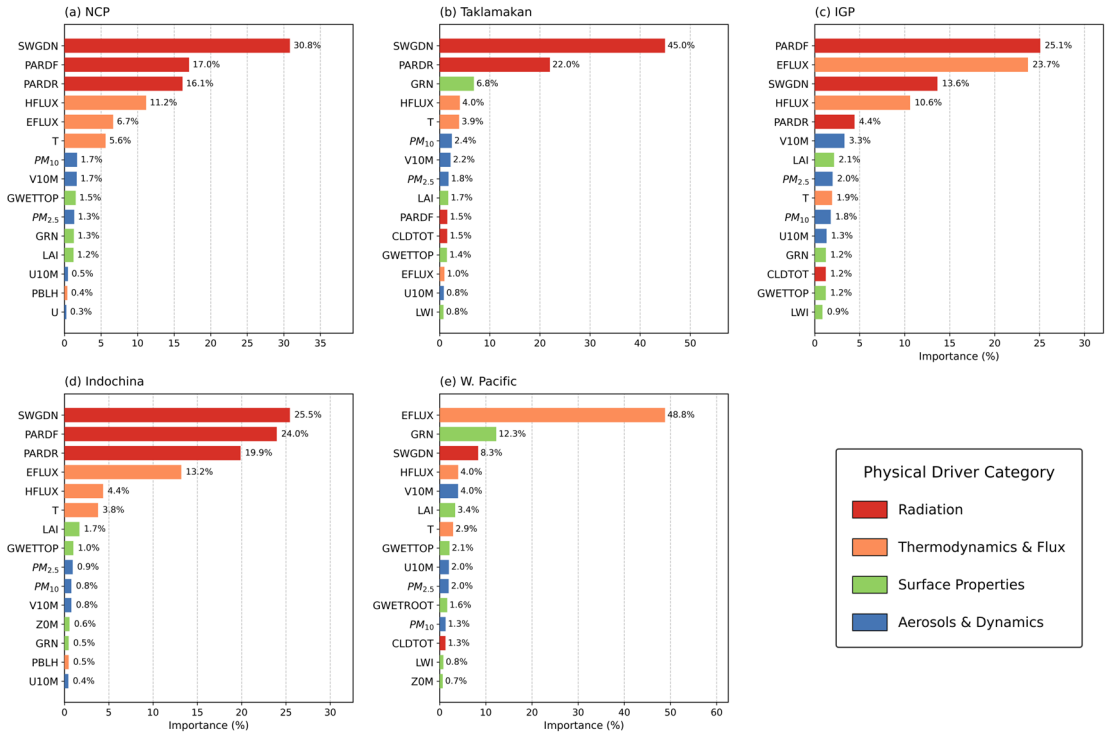


Figure 15. Regional variability in feature importance drivers identified by SHAP analysis for the test year 2019. The panels display the top 15 most influential features for predicting AEC simulation biases in five representative regions: NCP (a), Taklamakan Desert (b), IGP (c), Indochina (d), and Western Pacific (e).

Based on your valuable suggestions, we have made comprehensive revisions to the manuscript to clarify these mechanisms.

Revised Text in Section 2.2:

"To ensure physical consistency between the GEOS-Chem and CALIOP satellite observations, we employ a strict spatiotemporal collocation strategy. Spatially, the high-resolution CALIOP Level 2 profiles are mapped onto the GEOS-Chem grid. All quality-controlled CALIOP profiles falling within a specific grid cell are spatially averaged to represent the observational mean state of that grid box. Temporally, we adopt a precise nearest-hour collocation approach. The CALIOP overpass times are mathematically rounded to the nearest UTC hour and paired strictly with the GEOS-Chem 1-hourly instantaneous outputs. Aligning the instantaneous model output with the concurrent instantaneous observation minimizes temporal representativeness errors (typically constrained within ±30 minutes). However, we acknowledge that this approach

inherently introduces spatial representativeness errors. Averaging the narrow, "curtain-like" nadir view of CALIOP to represent a bulk 2°×2.5° grid box inevitably suffers from sub-grid heterogeneity, particularly in regions with complex terrain or localized intense emissions."

Revised Text in Section 3.1:

"Furthermore, although our architecture does not employ explicit historical time-series modeling, it robustly captures diurnal variability. By rigorously matching the instantaneous MERRA-2 fields with the exact CALIOP overpass time, the model is directly conditioned on the concurrent thermodynamic and dynamic states. Combined with explicit day/night flags, this allows the framework to dynamically resolve meteorology-driven diurnal processes (e.g., boundary layer evolution and photochemistry) without relying on lagged predictors"

Revised Text in Section 4.4:

"In addition to the instantaneous vertical structure, verifying the temporal continuity of the correction results is equally crucial. Given the spatiotemporal sparsity of CALIOP observations, we utilize ground-based high-frequency AERONET AOD data to further evaluate model performance. A critical concern regarding DL models trained on polar-orbiting satellite data is the potential overfitting to the sensor's limited twice-daily sampling timing (e.g., ~01:30 and 13:30 local time for CALIOP). However, the continuous AOD time series output by the GC-TF model demonstrates superior trend consistency with the high-frequency AERONET observations throughout all daylight hours. Figure 11d-e show the corrected model successfully captures the dynamic diurnal evolution and phase fluctuations of aerosols. Specifically, Figure 11d shows that during the study period in Kanpur, the original GEOS-Chem simulation generally underestimates aerosol loading, and the curve exhibits overly smooth characteristics, lacking response to high-frequency fluctuations. In contrast, the GC-TF model closely tracks these dynamic variations. Notably, between May 1 and May 2, the original simulation shows significant overestimation, whereas the GC-TF model successfully adjusts the predicted values back to levels closer to observations. During the active fire period in Nong Khai (Fig. 11e), the original model severely

underestimates the AOD magnitude. The GC-TF model significantly elevates the simulation baseline and captures the phase of diurnal variation trends well (e.g., the fluctuations around 07:00 UTC during February 22–24). This empirical evidence confirms that by conditioning the bias correction on meteorology-driven diurnal processes (as discussed in Sect. 3.1), the framework robustly generalizes across the entire diurnal cycle and avoids overfitting to specific CALIOP overpass times."

Third, the proposed architecture includes multiple advanced components. While the performance improvements are reported relative to the original GEOS-Chem simulation, there is no comparison with simpler machine learning baselines. It is therefore unclear whether the reported gains arise from the Transformer architecture itself, from the inclusion of additional meteorological predictors, or simply from the supervised bias-learning framework. To justify the methodological novelty, the study should include comparisons with at least one conventional model, such as a multilayer perceptron, a CNN-based model, or a tree-based regression approach. Ideally, ablation experiments isolating the contributions of the cross-attention module and gated fusion mechanism would further demonstrate the necessity of the proposed architecture. Without such benchmarks, it is difficult to assess whether the architectural complexity is warranted.

Response:

We fully agree that rigorous benchmarking against conventional algorithms and systematic ablation studies are essential to justify the methodological choice and architectural complexity of our framework. To address this, we have conducted comprehensive benchmarking experiments utilizing the independent 2017 test dataset and added a new Section 4.1.6 and Section S15 to the revised manuscript.

To ensure a strictly fair comparison, all baseline models and ablation variants are trained using the exact same input predictors (GEOS-Chem state variables and MERRA-2 meteorological forcings) and identical hyperparameter configurations. Our findings are summarized as follows:

(1) Benchmarking against Conventional Baselines

To demonstrate the necessity of the Transformer backbone, we evaluate a Multilayer Perceptron (MLP, representing point-wise networks without sequential awareness) and a 1-Dimensional Convolutional Neural Network (1D-CNN, representing localized spatial receptive fields). Table S5 shows the simple MLP struggles to extract vertically resolved spatial correlations (R = 0.083, RMSE = 0.052 km^{-1}). While the 1D-CNN improves predictive capabilities by capturing local vertical gradients (R = 0.540), it still significantly underperforms compared to the Transformer (R = 0.666). This demonstrates that the global sequence modeling enabled by the self-attention mechanism is indispensable for resolving the long-range vertical coupling of atmospheric aerosols (e.g., pollutant exchange between the boundary layer and the free troposphere).

(2) Ablation Study of Proposed Modules

We further isolate the contributions of our specific structural designs by replacing the Gated Feature Fusion and Cross-Attention modules with standard feature concatenation. Removing the Gated Fusion mechanism results in noticeable performance degradation (R drops to 0.637), indicating that dynamically assigning weights to distinct feature groups based on atmospheric stratification is essential. Similarly, removing the Cross-Attention module limits the model's ability to fully utilize macroscopic meteorological constraints (R drops to 0.654). These benchmarks confirm that the reported gains arise directly from the proposed structural designs.

(3) The Diagnostic Necessity of the Architecture

Importantly, beyond statistical improvements, the fundamental objective of introducing Gated Fusion and Cross-Attention is to transition the framework from a "black-box" predictor into a physics-informed diagnostic tool. Without the Gated Feature Fusion, we would be unable to dynamically quantify how the dominance of physical drivers shifts across different altitudes (as analyzed in Section 4.5.1). Likewise, without the Cross-Attention module, the architecture lacks the explicit mechanism required to extract attention weights that map surface meteorological constraints onto vertical aerosol structures (as diagnosed in Section 4.5.3). Therefore, the proposed

complexity is highly warranted for serving as a diagnostic tool to guide physical model improvement.

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

Based on your valuable suggestions, we have made comprehensive revisions to the manuscript to clarify these mechanisms.

The manuscript has added Section 4.1.6:

"**4.1.6 Methodological Benchmarking and Structural Necessity**

To justify the architectural complexity and isolate the sources of performance gains, we conduct comprehensive benchmarking and ablation studies using the independent 2017 test dataset (Table S5). When trained with identical GEOS-Chem and MERRA-2 predictors, the proposed Transformer significantly outperforms conventional machine learning baselines. A standard Multilayer Perceptron (MLP) and a 1-Dimensional Convolutional Neural Network (1D-CNN) yielded substantially lower R (R=0.083 and 0.540, respectively) compared to the Transformer (R=0.666). This performance gap confirms that global sequence modeling via self-attention is critical for capturing the long-range vertical coupling of atmospheric aerosols—such as boundary layer-to-free troposphere exchange—which localized convolutions or point-wise networks fail to resolve.

Furthermore, ablation experiments confirm that the performance enhancements are intrinsically linked to our structural designs. Removing the Gated Feature Fusion or the Cross-Attention module noticeably degrades predictive accuracy (Table S5). More importantly, beyond statistical improvements, these modules are physically indispensable. They transition the framework from a black-box predictor into a diagnostic tool, providing the explicit attention weights necessary to quantify height-dependent physical drivers (Section 4.5.1) and surface environmental modulations (Section 4.5.3)."

The Supporting Information adds Section S15:

"**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

”

In Section 3.4 of the manuscript, the following is added:

“(5) Methodological Benchmarking: We evaluate the proposed Transformer against conventional machine learning baselines and conduct ablation studies to justify the architectural complexity and isolate the sources of performance improvements.”

Reviewer #2:

This study tried to minimize the biases in GEOS-Chem aerosol simulation vertical structure using CALIPSO data. It meets the need for reconstructing aerosols' spatially continuous distributions with high-accuracy vertical profiles. Methodologically, the paper proposes a Physics-Informed Transformer framework, explicitly incorporating physical priors through dual-stream inputs, gated feature fusion, and cross-attention mechanisms, thereby overcoming the limitations of traditional CNNs in capturing vertical dependencies of aerosols. Several issues need to be carefully addressed in the manuscript.

(1) The manuscript is too long to read. Please try to reduce redundant text. In particular, the methodology part contains many technical terms that make it extremely difficult to follow. Figure 2 shows the technical framework. However, I do not understand anything except the input layer when looking at this figure. Please add more details in the figure to show the physical meaning of feature embedding layer, Transformer Encode, and Cross Attention Layer. The methodology part needs to re-write in a way that atmospheric chemists and physicist can understand.

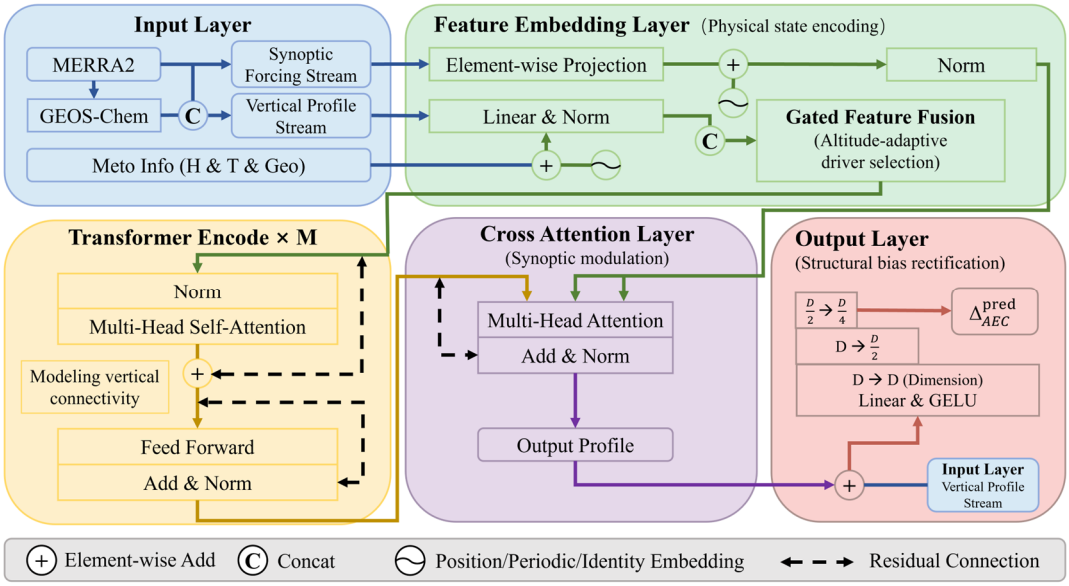
Response:

We highly appreciate the reviewer's valuable feedback. We agree that the original manuscript relied too heavily on deep learning terminology, which created a barrier for readers in the atmospheric sciences.

In response, we have comprehensively rewritten and streamlined Section 3. Redundant mathematical and technical jargon has been moved to the Supplementary Information (Sect. S4), allowing the main text to focus entirely on the physical rationale. The deep learning mechanisms are now explicitly translated into atmospheric processes—for example, framing the "Transformer Encoder" as a model for vertical turbulent mixing, and "Cross-Attention" as synoptic-scale meteorological forcing.

Furthermore, we have redesigned Figure 2 and its caption. Addressing your specific concern, we added detailed annotations to reveal the physical meaning of each network module. The Feature Embedding Layer is now depicted as integrating

chemical and meteorological fields into unified physical states, and the Gated Feature Fusion is explicitly labeled as a mechanism differentiating near-surface emissions from free-tropospheric transport.



"Figure 2. Architecture of the physics-informed Transformer framework. The Feature Embedding Layer integrates the VPS and SFS into unified, high-dimensional physical states. Gated Feature Fusion resolves atmospheric stratification by dynamically prioritizing altitude-dependent drivers. The Transformer Encoder stack models vertical connectivity, simulating exchange processes like turbulent mixing to ensure profile coherence. The Cross-Attention Layer retrieval synoptic constraints to modulate AEC corrections based on the meteorological background. The Output Layer performs residual bias rectification, anchoring the prediction to the initial physicochemical profiles to accurately quantify systematic simulation biases while filtering redundant noise."

Rewritten Section 3:

"**3. Method**

3.1 Input Feature Construction and Target Definition

We design a dual-stream input architecture to decouple local vertical atmospheric states from synoptic meteorological forcing. The detailed inventory of all input variables is provided in Section S3 in the supplement.

The Vertical Profile Stream (VPS) resolves the atmospheric column through three

sub-components. (1) Physicochemical profiles: This includes GEOS-Chem simulated aerosol species and MERRA-2 meteorological profiles. Beyond basic mass concentrations, we explicitly incorporate precursor gases (SO_2, NO_x, NH_3) and microphysical variables (e.g., hygroscopic growth factors and effective radii) to physically constrain secondary aerosol formation and optical extinction. (2) Height information: To maintain vertical stratification within the attention mechanism, we embed explicit altitude information. This allows the model to correctly differentiate near-surface emission interactions from free-tropospheric long-range transport. (3) Spatiotemporal indices: Geographical coordinates (latitude, longitude) and temporal indices (month, day, night) are projected into high-dimensional vectors to capture regional emission patterns and seasonal cycles.

The Synoptic Forcing Stream (SFS) incorporates 2D surface diagnostics to represent synoptic constraints on the atmospheric column. Variables such as Planetary Boundary Layer Height (PBLH) and friction velocity act as indicators for turbulent mixing. Surface fluxes and Leaf Area Index (LAI) parameterize deposition and biogenic emissions, while precipitation rates serve as proxies for wet scavenging.

Furthermore, although our architecture does not employ explicit historical time-series modeling, it robustly captures diurnal variability. By rigorously matching the instantaneous MERRA-2 fields with the exact CALIOP overpass time, the model is directly conditioned on the concurrent thermodynamic and dynamic states. Combined with explicit day/night flags, this allows the framework to dynamically resolve meteorology-driven diurnal processes (e.g., boundary layer evolution and photochemistry) without relying on lagged predictors.

Finally, we define the learning target Δ_{AEC}^{target} as the systematic bias of GEOS-Chem simulated AEC (AEC_{GC}) evaluated against CALIOP observation (AEC_{CAL}):

$$\Delta_{AEC}^{target} = AEC_{GC} - AEC_{CAL} \quad (1)$$

Predicting the simulation bias Δ_{AEC}^{target}, rather than the absolute AEC magnitude, ensures the framework preserves the fundamental physical transport patterns resolved by the CTM, focusing solely on correcting systematic deviations caused by

parameterization or emission uncertainties. It is important to emphasize that while CALIOP observations provide the target during training, they are strictly excluded from the input feature space during inference. Consequently, the framework's corrective capacity is inherently bounded by the information content of the GEOS-Chem and MERRA-2 predictors. The model is designed to rectify state-dependent systematic biases rather than to artificially reconstruct aerosol signals from completely unrepresented physical processes that lack corresponding perturbation signatures in the input fields.

It should be noted that using CALIOP retrievals as the baseline inherently propagates its systematic uncertainties (e.g., a mean relative bias of -5.1%, as discussed in Sect. 2.2) into the learning target. If CALIOP exhibits a systematic negative bias, the model may theoretically learn a tendency to slightly over-compensate the AEC. However, because GEOS-Chem's structural biases are typically an order of magnitude larger than these observational uncertainties, the data-driven correction remains highly beneficial. A detailed quantitative evaluation of the model's sensitivity to these observational uncertainties is presented in Section 4.1.5.

3.2 Physically-Informed Transformer Architecture

The overall architecture of our proposed framework (Figure 2) bridges GEOS-Chem simulations and CALIOP observations. To preserve the distinct structural characteristics of atmospheric profiles and synoptic environmental contexts, the framework processes these two streams through specialized embedding strategies (detailed in Sect. S4a, b). The model comprises an altitude-dependent gated feature fusion mechanism, a Transformer encoder for vertical dependencies, a cross-attention module for synoptic constraints, and an output layer.

3.2.1 Altitude-Dependent Gated Feature Fusion

Physical factors governing AEC vary significantly with altitude. Local emissions and chemical composition dominate near-surface AEC (Xiong et al., 2025; Jiang et al., 2024), whereas long-range transport and regional backgrounds dictate the free troposphere (Uno et al., 2009; Val Martin et al., 2013). To reflect this stratification, we design a gated feature fusion mechanism within the VPS. Instead of statically

concatenating inputs, this module dynamically weights the contributions of physicochemical profiles, height information, and spatiotemporal indices for each altitude layer. This allows the model to autonomously prioritize the most relevant physical drivers at specific heights.

The SFS incorporates diverse meteorological parameters with distinct physical units. To prevent the network from treating these distinct physical quantities merely as dimensionless numbers, we implement a variable identity embedding (Eq. S4). This mechanism assigns a unique physical tag to each 2D variable, ensuring the model accurately distinguishes between different meteorological forcing factors when modulating the AEC simulation bias.

3.2.2 Modeling Vertical Connectivity and Synoptic Modulation

Aerosol layers are inherently coupled through vertical exchange processes such as turbulent mixing, deep convection, and gravitational sedimentation. We employ a Transformer encoder stack to explicitly model this vertical connectivity. Its self-attention mechanism acts as a dynamic vertical covariance operator (detailed in Sect. S4c). It facilitates information flow between near-surface accumulation layers and high-altitude transport layers, ensuring the rectified AEC profile maintains physical continuity.

To constrain this vertical AEC bias correction with synoptic meteorology, we introduce a cross-attention layer. Functionally, this mechanism acts as a dynamic diagnostic process. It allows the aerosol state at each specific altitude to dynamically respond to the prevailing synoptic conditions (e.g., surface wind speed, PBLH), thereby extracting relevant environmental constraints for the local bias adjustment. This design mimics physical reality, where synoptic meteorological backgrounds continuously modulate local microphysical structures.

3.2.3 Output Layer

To predict the final AEC bias, we employ a residual connection that adds the initial baseline state (from the VPS) directly to the output of the cross-attention module (which has already fused the encoded VPS with the SFS). Physically, this residual design serves as a critical prior constraint. It anchors the network to the fundamental atmospheric

state provided by GEOS-Chem, ensuring the model computes a meteorology-driven perturbation rather than generating unphysical aerosol signals. Subsequently, the integrated features undergo a progressive dimension-reduction (represented as $D \rightarrow D/2 \rightarrow D/4$ in Fig. 2). This architecture functions as an information funnel, filtering redundant meteorological noise and distilling the non-linear interactions among diverse drivers to accurately quantify the true magnitude of the AEC biases.

3.3 Magnitude-Weighted Loss Function

To address the statistical imbalance between the predominant clean background signals and the physically critical pollution episodes, we propose a Magnitude-Weighted Loss (L_{MW}, detailed in Sect. S4g). This customized loss function dynamically rescales the correction weighting to enhance the model's sensitivity to large simulation AEC biases while strictly suppressing spurious aerosol artifacts in atmospheric regimes where the CTM already performs satisfactorily.

3.4 Model Evaluation Strategy

To comprehensively assess the robustness and generalization capability of the physics-informed Transformer model, we design a rigorous evaluation framework covering five dimensions.

(1) Spatial block cross-validation: To mitigate information leakage caused by spatial autocorrelation (Geer, 2021), we implement a spatial block K-fold cross-validation strategy (Sect. S5). The study region is divided into non-overlapping 4° × 5° blocks (aggregating 2×2 model grids). In each iteration, the model is trained on four folds and evaluated on the remaining spatially independent fold. This "checkerboard" approach ensures performance metrics reflect the model's ability to extrapolate to unseen geographic locations.

(2) Temporal transferability: Given the interannual variability in emissions and meteorology (Xiong et al., 2025), we adopt a Leave-One-Year-Out (LOYO) cross-validation scheme comprising three independent experiments (Table 1). This tests whether the model learns generalizable physical rules rather than overfitting to specific temporal patterns.

(3) External spatial generalization: To rigorously test the model's transferability

beyond its training distribution, we perform an out-of-domain evaluation on the independent NA defined in Section 2.1. By directly applying the model trained on EA data to this unseen continent—which possesses distinct aerosol sources and meteorological backgrounds—we evaluate whether the learned bias-correction mechanism captures universal physical laws rather than region-specific correlations. Furthermore, to dissect the impact of varying aerosol composition regimes on model transferability, the NA validation results are further stratified using CALIOP aerosol subtype classifications.

(4) Independent ground-based validation: We employ ground-based AERONET observations as an independent physical benchmark. Predicted AEPs are vertically integrated to derive column AODs, which are then compared with AERONET data to assess the reproduction of high-frequency temporal evolution.

(5) Methodological benchmarking: We evaluate the proposed Transformer against conventional machine learning baselines and conduct ablation studies to justify the architectural complexity and isolate the sources of performance improvements.

To quantify the model performance across these dimensions, we employ a comprehensive set of statistical metrics including the Pearson correlation coefficient (R), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Normalized Root Mean Square Error (NRMSE). Detailed mathematical definitions are provided in Sect. S6. NRMSE is specifically used to enable fair comparisons across vertical layers by normalizing biases against the exponentially decaying dynamic range of AEC.

3.5 Model Interpretability Framework

To elucidate the inference logic of the correction framework and ensure physical consistency, we establish a hierarchical diagnostic approach. This framework addresses potential functional overlaps by characterizing model behavior across three scales: micro-scale local sensitivity, domain-wide feature ranking, and regional heterogeneity.

3.5.1 Micro-Scale Local Sensitivity

We employ distinct attribution methods tailored to the hybrid inputs to capture micro-scale responses. For the VPS, we apply gradient-based attribution, utilizing the Input×Gradient method (Shrikumar et al., 2017) to quantify the sensitivity of AEC bias

correction to physicochemical profiles. Simultaneously, Cross-Attention weights are extracted to map the interaction strength between the SFS and the VPS, revealing how synoptic forcing modulates vertical profile rectifications. Furthermore, to understand the model's internal decision-making, we analyze the learnable weights of the gated feature fusion mechanism (detailed in Sect. 3.2.1). This analysis visualizes the altitude-dependent prioritization among the four VPS components: physicochemical profiles, height information, spatial coordinates, and temporal indices.

3.5.2 Domain-wide Feature Ranking

To assess the model's reliance on the overarching input feature groups (the VPS and SFS), we perform permutation feature importance analysis (detailed in Sect. S7d). By measuring the percentage increase in Mean Squared Error (MSE) when specific groups are randomly shuffled, this method provides a domain-wide to identify the fundamental predictors essential for AEC bias correction.

3.5.3 Regional Heterogeneity

Considering the spatial heterogeneity of aerosol sources, SHAP Analysis is used to dissect regional dependencies and feature interactions. We employ a K-means clustering strategy to construct a representative background dataset capturing diverse atmospheric states (detailed in Sect. S8). SHAP values are computed for the designated ROIs to reveal how dominant AEC bias drivers shift under different environmental regimes."

(2) In Section 3.1 (Eq. 1), the learning target is defined as the bias of GEOS-Chem relative to CALIOP. However, Section 2.2 states that CALIOP AOD shows a mean relative bias of −5.1% ± 8.5% against AERONET, and CALIOP backscatter agrees with HSRL within 1.0% ± 3.5%. These results indicate that CALIOP itself contains systematic uncertainties. Consequently, the learned "bias" effectively represents a combination of GEOS-Chem error and CALIOP error. If CALIOP has a negative bias, the model may incorrectly learn a tendency to increase AEC, even in cases where GEOS-Chem is accurate. This issue directly affects the interpretation and reliability of the bias-correction results. Suggestions: (a) Explicitly acknowledge this limitation in

Section 2.2 or 3.1, and discuss the potential impact of CALIOP uncertainty on the training target. (b) Add a sensitivity analysis in the Results section (Section 4): quantify how the bias-correction results change if perturbations are applied to the CALIOP inputs.

Response:

We completely agree that because our framework uses CALIOP as the observational benchmark, its intrinsic systematic uncertainties inevitably propagate into the learning target, potentially causing slight over-compensation in certain layers. We have carefully addressed both of your suggestions as follows:

(a) Acknowledging the limitation:

We have explicitly acknowledged this limitation in the revised Section 3.1, detailing the error propagation mechanism and clearly stating that the learned bias is a combination of the GEOS-Chem structural error and the CALIOP retrieval uncertainty.

Section 3.1 of the manuscript has been modified to correspond to the following content:

"It should be noted that using CALIOP retrievals as the baseline inherently propagates its systematic uncertainties (e.g., a mean relative bias of -5.1%, as discussed in Sect. 2.2) into the learning target. If CALIOP exhibits a systematic negative bias, the model may theoretically learn a tendency to slightly over-compensate the AEC. However, because GEOS-Chem's structural biases are typically an order of magnitude larger than these observational uncertainties, the data-driven correction remains highly beneficial. A detailed quantitative evaluation of the model's sensitivity to these observational uncertainties is presented in Section 4.1.5."

(b) Sensitivity analysis:

To quantitatively evaluate the impact of this uncertainty on our results, we conduct the suggested perturbation experiment using the 2017 independent test set. We artificially inject a ±5% systematic multiplier into the CALIOP AEC learning target to simulate both severe underestimation and overestimation scenarios, and retrained the physics-informed Transformer from scratch. The results have been added to the newly

created Section 4.1.5 (Sensitivity to Observational Uncertainties) and detailed in Section S13, Table S3, and Figure S17 in the Supplementary Material.

As summarized in the newly added Section 4.1.5, the systematic perturbation induces only a narrow envelope of variation in the corrected AEC profiles. For instance, even with a ±5% systematic error injected, the perturbed predictive RMSE (~0.040 km^{-1}) consistently outperforms the original GEOS-Chem simulation (0.052 km^{-1}) by a large margin, and the absolute shift in Mean Bias fluctuates tightly between 0.001 and 0.004 km^{-1}. This quantitative test confirms that the physics-informed Transformer is highly robust against observational noise and does not uncontrollably amplify inherent satellite errors.

The newly added Section 4.1.5 in the manuscript:

"4.1.5 Sensitivity to Observational Uncertainties

As discussed in Section 3.1, using satellite retrievals as the learning target inherently absorbs CALIOP's systematic uncertainties. To quantify how this observational limitation impacts the reliability of our framework, we conduct a perturbation-based sensitivity analysis (detailed in Sect. S13). We retrain the GC-TF model by artificially injecting a ±5% systematic perturbation into the CALIOP AEC learning target.

Table S3 and Figure S17 demonstrate that this systematic perturbation induces only a narrow envelope of variation in the corrected AEC profiles. The absolute shift in the Mean Bias fluctuates tightly between 0.001 and 0.004 km^{-1}, and the perturbed predictive RMSE (0.040 km^{-1}) consistently outperforms the original GEOS-Chem simulation (0.052 km^{-1}) by a large margin. This confirms that while observational uncertainties theoretically bound the absolute precision, the physics-informed Transformer does not uncontrollably amplify these errors, ensuring the robustness of the data-driven correction."

The supplementary Table S3 and Figure S17 are as follows:

"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

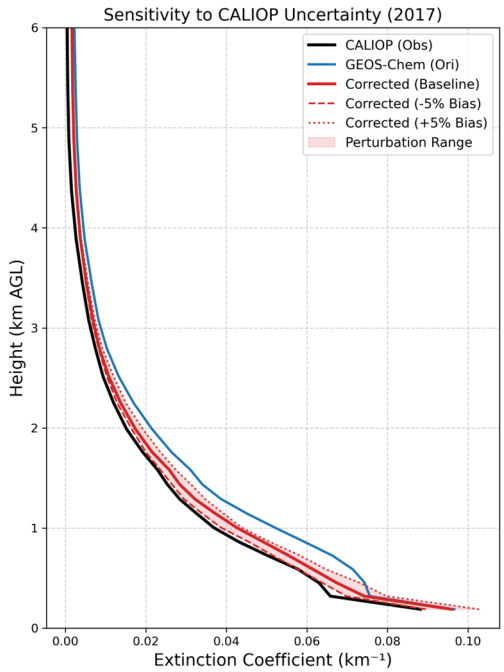


Figure S17. Vertical profiles illustrating the sensitivity of the corrected AEC to target perturbations for the test year 2017. The black solid line represents the CALIOP observations, and the blue solid line represents the original GEOS-Chem simulation. The red solid line indicates the baseline correction 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."

The newly added Section S13 in the Supporting Information:

"**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."

(3) The manuscript states that the interpretability analysis can provide a solid physical basis for improving GEOS-Chem parameterizations and emission inventories, thereby establishing a feedback loop from "data-driven correction" to "physical mechanism improvement." However, in Section 3.5, the interpretability framework is limited to feature-sensitivity approaches such as gradient attribution, permutation importance, and SHAP, without explaining how these results can be translated into concrete parameterization adjustments. For example, if SHAP identifies "sensible heat flux" as a dominant driver of the bias, it is unclear which specific GEOS-Chem

parameters should be modified (e.g., diffusion coefficients in the PBL scheme, surface flux parameterizations, or others), and how such modifications would be implemented. This missing link weakens the claimed feedback loop and makes the statement appear largely conceptual rather than actionable.

Response:

We completely agree that directly translating deep learning feature sensitivities into concrete CTM code modifications remains challenging, and our previous description of establishing a "closed feedback loop" was overly conceptual and lacked actionable physical linkages. To address this and make our findings actionable, we have substantially rewritten and renamed Section 4.6 (now titled "Diagnostic Insights for Refining GEOS-Chem Parameterizations").

Instead of a generic "feedback loop," we now explicitly bridge the interpretability outputs with specific physical schemes within GEOS-Chem. Regarding boundary layer dynamics, we clarified that the high sensitivity to sensible heat flux indicates uncertainties in the vertical eddy diffusion coefficient within GEOS-Chem's non-local boundary layer scheme. This finding underscores the potential necessity of integrating a dedicated urban canopy model to better partition sensible heat in heavily urbanized domains like the North China Plain. Furthermore, in the context of dust emission over the Taklamakan Desert, we connected the model's coupled reliance on surface wind speed and vegetation indices to the required recalibration of the threshold friction velocity and soil erodibility parameters in the emission scheme. Finally, concerning secondary aerosol formation, we explained how the framework leverages diffuse radiation as a physical proxy for altered photolysis rates and accelerated photochemical aging. This specific linkage highlights the need to optimize Secondary Organic Aerosol (SOA) yield parameterizations and refine biomass burning plume injection heights.

To ensure consistency, we have also systematically adjusted the Abstract, Introduction, and Conclusions to align with this "diagnostic insight" framework rather than the previous "closed-loop" terminology. The detailed revisions incorporated into the manuscript are provided below.

1. Section 4.6 in the Revised Manuscript:

"**4.6 Diagnostic Insights for Refining GEOS-Chem Parameterizations**

The interpretability analysis in Section 4.5 demonstrates that the GC-TF model captures physically meaningful relationships rather than merely fitting statistical noise. While directly translating data-driven feature sensitivities into concrete code modifications remains challenging without further sensitivity simulations, this transparency allows the framework to serve as a valuable hypothesis-generation tool. It highlights potential structural uncertainties in CTMs and points toward targeted refinements in physical parameterizations.

4.6.1 Diagnosing Thermodynamic Parameterization Deficiencies

The model heavily relies on temperature and HFLUX to correct AEC profiles (Sect. 4.5.2, 4.5.4), which suggests potential uncertainties in diagnosing PBLH and turbulent mixing intensity within the GEOS-Chem non-local boundary layer scheme. Given that HFLUX drives surface buoyancy and directly modulates the vertical eddy diffusion coefficient, the widespread excessive diffusion biases observed in the lower troposphere indicate that the model may overestimate thermal turbulence under certain stability conditions. In highly urbanized regions like the NCP, the acute sensitivity to HFLUX implies that current surface energy balance calculations struggle to resolve the distinct thermodynamic properties of urban canopies. Future model development could benefit from constraining stability functions within the vertical diffusion module, or alternatively, coupling a dedicated urban canopy model to better represent sensible heat partitioning.

4.6.2 Refining Emission and Formation Schemes via Environmental Proxies

The cross-attention weights, which reveal how synoptic forcing modulates vertical aerosol profiles (Sect. 4.5.3, 4.5.4), highlight potentially inadequately parameterized mechanisms in current emission and chemical modules. Over the Taklamakan Desert, the model explicitly pairs greenness fraction with surface wind speed to capture dust extinction (Fig. 15). This suggests that the GEOS-Chem dust emission scheme might struggle to accurately parameterize threshold friction velocity over complex bare soils, indicating that the non-linear response of wind-blown dust to surface shear stress and

soil erodibility likely requires recalibration. Similarly, high sensitivity to diffuse radiation in the biomass burning region of Indochina points to potentially under-represented SOA formation. Given that high aerosol loading enhances diffuse radiation and alters photolysis rates, the data-driven model likely leverages diffuse radiation as a proxy for accelerated photochemical aging. This highlights a need to optimize SOA yield parameterizations and refine biomass burning plume injection heights to capture rapid aerosol evolution in dense smoke.

4.6.3 Bridging Data-Driven Interpretation with CTM Development

Beyond statistical bias correction, this study highlights the utility of physics-informed DL for model diagnosis. By decoupling the contributions of meteorology and aerosol composition, the framework verifies that CTMs provide a robust physicochemical baseline, yet exhibit uncertainties in representing the complex, non-linear interactions between aerosols and meteorology. The correction strategies derived from the data-driven model offer valuable diagnostic clues. Identifying specific environmental proxies that govern simulation biases bridges the gap between data-driven retrieval and deterministic modeling, ultimately guiding the targeted integration of neglected physical constraints into future parameterization schemes."

2. Corresponding revisions in the Abstract:

"…Furthermore, interpretability analysis serves as a diagnostic tool to guide physical model improvement. The model identifies temperature and sensible heat flux as primary drivers to constrain boundary layer mixing, pointing to potential uncertainties in vertical eddy diffusion. Additionally, it uses environmental proxies (e.g., vegetation indices and diffuse radiation) to diagnose potential deficiencies in dust threshold friction velocity and secondary organic aerosol yields…"

3. Corresponding revisions in the Introduction:

"…This process not only enables an interpretable diagnosis of CTM simulation biases—identifying the dominant drivers of bias within specific altitudes or regions—but also bridges data-driven correction with the targeted refinement of GEOS-Chem's physical parameterizations…"

4. Corresponding revisions in the Conclusions:

"Third, by integrating interpretable DL techniques, this study advances beyond standard bias correction to serve as a diagnostic framework for physical mechanisms. Attribution analysis reveals that the model identifies AEC simulation bias drivers with clear physical significance: (1) In the PBL, the heavy reliance on temperature and HFLUX highlights potential uncertainties in vertical eddy diffusion coefficients within stability-dependent mixing schemes; (2) Over dust source regions, the paired use of vegetation indices and wind speed suggests the need to recalibrate threshold friction velocity and soil erodibility; (3) In biomass burning regions, the sensitivity to diffuse radiation points to under-represented SOA yields and photochemical aging processes; (4) In marine regions, the utilization of latent heat flux and surface wind implies uncertainties in sea-salt generation functions and hygroscopic growth ."

(4) The Introduction appears to be over-cited, which makes it difficult for readers to clearly distinguish foundational studies from more recent developments. It would improve readability and focus to streamline the citations, limiting each statement to approximately three to five representative and/or recent review or key references.

Response:

We fully agree with the reviewer that the Introduction was over-cited, which compromised readability and diluted the focus on key scientific developments. We have systematically reviewed and streamlined the citations throughout the Introduction to adhere to the recommended limit of three to four representative references per statement. Specific revisions can be found in the introduction of the manuscript.

(5) Section 3.5 is divided into 3.5.1 (Dual-Mechanism Attribution), 3.5.2 (Gated Fusion Analysis), and 3.5.3 (Feature Sensitivity and Regional Drivers). However, the Permutation Feature Importance and SHAP analysis in 3.5.3 overlap functionally with the Gradient-based Attribution in 3.5.1—both are essentially feature importance assessments. It is recommended to clearly articulate the complementarity of these three attribution methods: Gradient-based Attribution captures local sensitivity, Permutation

Feature Importance provides global ranking, and SHAP analysis handles feature interactions and regional heterogeneity.

Response:

We thank the reviewer for this constructive recommendation to clarify our interpretability framework. We completely agree that without explicitly defining the physical scope and specific purpose of each attribution method, they may appear functionally redundant to the reader.

Following your specific guidance, we have comprehensively restructured Section 3.5 (Model Interpretability Framework). Rather than presenting these methods as parallel feature importance assessments, we have now explicitly articulated their complementarity across a unified, hierarchical diagnostic framework.

Specifically, our framework first addresses micro-scale local sensitivity (Sect. 3.5.1) by combining gradient-based attribution and cross-attention to diagnose how vertical layers react to variations in the Vertical Profile Stream (VPS, detailed in Sect.3.1) and how they are modulated by synoptic forcing from the Synoptic Forcing Stream (SFS, detailed in Sect.3.1). Within this section, we have thoughtfully integrated the analysis of the gated feature fusion mechanism to explicitly demonstrate how the model internally prioritizes information across different altitudes. Scaling up to the overarching model behavior, Sect. 3.5.2 utilizes permutation feature importance to provide a domain-wide ranking that evaluates the model's fundamental reliance on these primary input streams. Finally, to capture spatial heterogeneity, Sect. 3.5.3 employs SHAP analysis across designated Regions of Interest (ROIs) to parse non-linear feature interactions and shifting bias drivers across distinct environmental regimes (e.g., dust versus anthropogenic pollution).

We believe this hierarchical restructuring eliminates any perceived functional overlap and clearly demonstrates the physical purpose of each diagnostic tool. To ensure clear visibility for the reviewer, the revised text in the manuscript (Section 3.5) reads as follows:

“3.5 Model Interpretability Framework

To elucidate the inference logic of the correction framework and ensure physical consistency, we establish a hierarchical diagnostic approach. This framework addresses potential functional overlaps by characterizing model behavior across three scales: micro-scale local sensitivity, domain-wide feature ranking, and regional heterogeneity.

3.5.1 Micro-Scale Local Sensitivity

We employ distinct attribution methods tailored to the hybrid inputs to capture micro-scale responses. For the VPS, we apply gradient-based attribution, utilizing the Input×Gradient method (Shrikumar et al., 2017) to quantify the sensitivity of AEC bias correction to physicochemical profiles. Simultaneously, Cross-Attention weights are extracted to map the interaction strength between the SFS and the VPS, revealing how synoptic forcing modulates vertical profile rectifications. Furthermore, to understand the model’s internal decision-making, we analyze the learnable weights of the gated feature fusion mechanism (detailed in Sect. 3.2.1). This analysis visualizes the altitude-dependent prioritization among the four VPS components: physicochemical profiles, height information, spatial coordinates, and temporal indices.

3.5.2 Domain-wide Feature Ranking

To assess the model's reliance on the overarching input feature groups (the VPS and SFS), we perform permutation feature importance analysis (detailed in Sect. S7d). By measuring the percentage increase in Mean Squared Error (MSE) when specific groups are randomly shuffled, this method provides a domain-wide to identify the fundamental predictors essential for AEC bias correction.

3.5.3 Regional Heterogeneity

Considering the spatial heterogeneity of aerosol sources, SHAP Analysis is used to dissect regional dependencies and feature interactions. We employ a K-means clustering strategy to construct a representative background dataset capturing diverse atmospheric states (detailed in Sect. S8). SHAP values are computed for the designated ROIs to reveal how dominant AEC bias drivers shift under different environmental regimes.”

(6) Line 198: Explain the specific collocation strategy. How are the two datasets matched in space and time? What level of representativeness error might be introduced by this collocation approach?

Response:

We sincerely thank the reviewer for pointing this out. We completely agree that a transparent description of the collocation strategy and an objective assessment of the associated representativeness errors are essential for evaluating the model. To address this, we have substantially expanded the description in Section 2.2.

Specifically, to align the two datasets, we employ a rigorous spatiotemporal collocation strategy. Spatially, the high-resolution CALIOP Level 2 profiles are averaged within their corresponding GEOS-Chem grid cells. Temporally, we adopt a precise nearest-hour approach, matching the CALIOP overpass times to the closest 1-hourly instantaneous GEOS-Chem outputs. By aligning instantaneous model outputs with concurrent satellite observations, this approach effectively minimizes temporal representativeness errors to within ±30 minutes. However, we explicitly acknowledge that this methodology inherently introduces spatial representativeness errors driven by sub-grid heterogeneity. Because the narrow, "curtain-like" nadir view of the CALIOP instrument is used to represent the bulk atmospheric state of a coarse 2°×2.5° grid box, this spatial scale mismatch constitutes a recognized baseline uncertainty within our data-driven framework.

We have incorporated these clarifications into Section 2.2 of the revised manuscript. The newly added text is provided below:

“To ensure physical consistency between the GEOS-Chem and CALIOP satellite observations, we employ a strict spatiotemporal collocation strategy. Spatially, the high-resolution CALIOP Level 2 profiles are mapped onto the GEOS-Chem grid. All quality-controlled CALIOP profiles falling within a specific grid cell are spatially averaged to represent the observational mean state of that grid box. Temporally, we adopt a precise nearest-hour collocation approach. The CALIOP overpass times are mathematically rounded to the nearest UTC hour and paired strictly with the GEOS-Chem 1-hourly

instantaneous outputs. Aligning the instantaneous model output with the concurrent instantaneous observation minimizes temporal representativeness errors (typically constrained within ±30 minutes) (Ichoku et al., 2002). However, we acknowledge that this approach inherently introduces spatial representativeness errors. Averaging the narrow, "curtain-like" nadir view of CALIOP to represent a bulk 2°×2.5° grid box inevitably suffers from sub-grid heterogeneity, particularly in regions with complex terrain or localized intense emissions."

(7) Line 746-748: The lower transfer performance over North America (R = 0.70) compared to East Asia (R = 0.93) is attributed to a shift in aerosol composition regimes (higher SOA fraction in North America versus sulfate–nitrate–dust dominance in East Asia). While this explanation is reasonable, it remains qualitative and lacks supporting evidence. It would be helpful to further evaluate the performance over North America stratified by CALIOP aerosol types.

Response:

We agree with the reviewer that a quantitative evaluation stratified by aerosol subtypes significantly strengthens our argument.

Following your suggestion, we utilize the Aerosol Subtype classification from the CALIOP Level 2 Atmospheric Volume Description. We partition the NA evaluation dataset 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). We then recalculated the statistical metrics for these subsets.

As shown in the newly added Table S4, the stratified evaluation strongly supports our hypothesis. The model demonstrates clear corrective capability in the dust-dominated regime (R increased from 0.41 to 0.50, and the regression slope improved from 0.21 to 0.32), indicating that the physical constraints governing dust extinction transport learned in EA successfully transfer to NA. Conversely, the correction for the SOA-dominated regime shows negligible improvement. This discrepancy physically

stems from the fact that the thermodynamic-to-optical mapping learned in EA is largely driven by highly hygroscopic inorganic salts, which is less applicable to the weakly hygroscopic biogenic SOA prevalent in NA.

This subtype-based analysis has provided a much more rigorous and physically grounded explanation for the model's performance in unseen domains. We have added Table S4 to the Supplementary Information and substantially revised Section 4.3 in the revised manuscript to include this quantitative evidence and physical discussion.

The newly added Section S14 in the Supporting Information is as follows:

"**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

"

Corresponding revisions in Section 4.3:

"Despite these capabilities, the overall correction performance in NA (R=0.70) remains statistically lower than in the EA training domain (R=0.93). We hypothesize

that this performance gap originates from a fundamental domain shift in aerosol composition. The NA atmosphere features lower background concentrations and a significantly higher fraction of biogenic Secondary Organic Aerosols (SOA) (Goldstein et al., 2009). The optical properties and hygroscopicity of these organic species differ fundamentally from the sulfate-nitrate-dust mixtures dominating EA (Crawford et al., 2021).

To quantitatively verify this impact, we further evaluate the model performance stratified by CALIOP aerosol subtypes over NA (Table S4). The results reveal a clear divergence in the model's corrective capability across distinct aerosol regimes. For dust-dominated regimes, the model effectively mitigates GEOS-Chem's systematic underestimation, enhancing the regression slope from 0.21 to 0.32 and increasing R from 0.41 to 0.50. This confirms that the physical constraints governing dust extinction and vertical transport learned in EA translate reliably to the NA domain. In contrast, the model yields negligible improvements for the SOA-dominated continental regime. Although total mass concentrations are provided as predictors, the specific thermodynamic-to-optical mapping learned in EA—typically dominated by the high hygroscopic growth of inorganic salts—is less applicable to the complex, weakly hygroscopic biogenic SOA prevalent in NA. The stagnant R and regression slope in the SOA group suggest that without locally representative training samples to capture the unique mass-to-extinction relationships of NA-specific organic species, the data-driven framework maintains a conservative correction. This ultimately limits the overall accuracy improvement across the NA background."

(8) Abstract: too technical. Suggest to add several sentences in the beginning to introduce the science context and research gap before jumping into technical details.

Response:

We agree that the original abstract dived into technical details too abruptly. To address this, we have revised the beginning of the abstract to better establish the scientific context and highlight the specific research gap our study aims to fill.

We have added the following sentences to the beginning of the abstract in the revised manuscript:

"Accurately characterizing aerosol vertical distributions is essential for evaluating radiative forcing and air quality. While Chemical Transport Models (CTMs) simulate spatially continuous Aerosol Extinction Coefficient (AEC, km^{-1}), they exhibit systematic AEC biases. Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) observations provide precise AEC profiles but are constrained by sparse spatial sampling. To bridge this gap, we propose a physics-informed Transformer framework to..."

(9) Figure 3. R between model prediction and what data? What are the units for RMSE and Bias?

Response:

We apologize for the ambiguity in the original figure. We have updated both the axis labels and the caption of Figure 3 to ensure clarity. The updated Figure 3 and its caption are as follows.

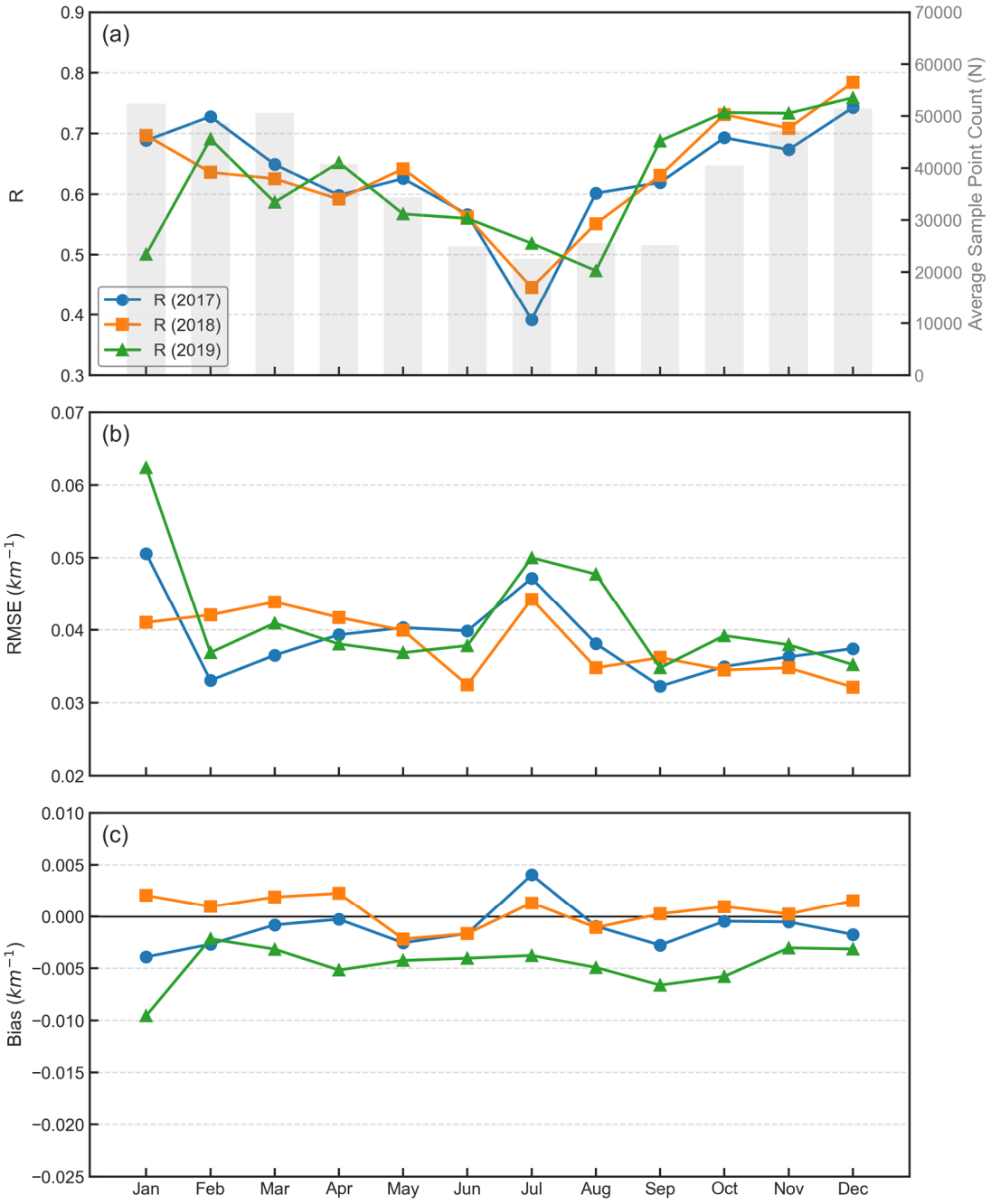


Figure 3. Monthly variations of statistical metrics evaluating the physics-informed Transformer model's predictive performance over EA across the three test years (2017, 2018, and 2019). The panels display the time series of R between the AEC bias predicted by the model and that simulated by GEOS-Chem (a), along with the multi-year average monthly sample size (N, gray bars), RMSE (b), and mean bias (c).

(10) Figure 5. Why India shows much negative results?

Response:

We thank the reviewer for pointing this out. The pronounced negative results over India, particularly the Indo-Gangetic Plain (IGP), represent a systematic underestimation of the AEC by the original GEOS-Chem simulation compared to CALIOP observations. We attribute this regional underestimation to three primary mechanistic factors: (1) the underrepresentation of local residential biofuel and agricultural burning emissions in traditional bottom-up inventories; (2) simplified aerosol external mixing state assumptions in the CTMs, which tend to underestimate the optical lensing effect and hygroscopic enhancement under high ambient humidity; and (3) the excessive numerical diffusion inherent to CTMs, which artificially dilutes the strong near-surface aerosol accumulation trapped by the natural topographic barrier of the Himalayas. To clarify this for readers and strengthen our physical analysis, we have explicitly incorporated these mechanistic drivers into Section 4.1.4 of the revised manuscript. The revised text reads as follows:

"…Specifically, the model accurately captures the systematic underestimation over major anthropogenic and biomass burning source regions, including the NCP, IGP, and Indochina Peninsula. Over regions like the IGP, this negative simulation bias is primarily driven by the underrepresentation of local biofuel and agricultural emissions in traditional inventories (Mcduffie et al., 2020), coupled with simplified aerosol mixing state assumptions that underestimate extinction enhancement under high humidity (Burgos et al., 2020; Zhai et al., 2021). Furthermore, the model's excessive numerical diffusion, a common limitation in CTMs, artificially dilutes the strong near-surface pollutant accumulation bounded by local topography (e.g., the Himalayas) (Eastham and Jacob, 2017). The GC-TF framework effectively identifies and mitigates these state-dependent underestimations…"

(11) Figure 6. I would say that the correlation even after correction is not that good. Can you explain where are those points that are far away from the 1:1 line?

Response:

We agree that despite the substantial improvement in statistical metrics (R increasing from ~0.50 to ~0.70), visual scatter remains in the density plots. We have conducted a thorough diagnostic analysis on these specific points, which reveals that they are primarily a result of representativeness errors (spatial scale mismatch) rather than inherent failures of the bias-correction framework.

Specifically, the visual dispersion observed in the scatter plots (Figure 6) is largely amplified by the logarithmic color scale (Count Density). Statistically, data points exceeding the 0.15 km^{-1} error envelope constitute only 1.20% of the total valid dataset, indicating that these outliers are statistically sparse despite their visual prominence. Further diagnostic analysis reveals a distinct physical origin for these discrepancies, largely attributable to spatial scale mismatch. Spatially, these extreme deviations are highly clustered over major emission hotspots, such as the NCP, the IGP, and the Indochina Peninsula. Vertically, over 80% of them are confined within the PBL (< 1.5 km AGL). In these complex source regions, CALIOP's narrow, high-resolution footprint captures transient, highly concentrated sub-grid aerosol plumes. However, these fine-scale peak values are inherently smoothed out during the spatial averaging process across the coarse 2°×2.5° grid of GEOS-Chem. Consequently, while our model successfully captures the systematic biases of the grid mean, it theoretically cannot resolve stochastic sub-grid extremes that lack corresponding predictors within the coarse meteorological inputs.

To make this clear to the readers, we have explicitly addressed this phenomenon in the revised manuscript.

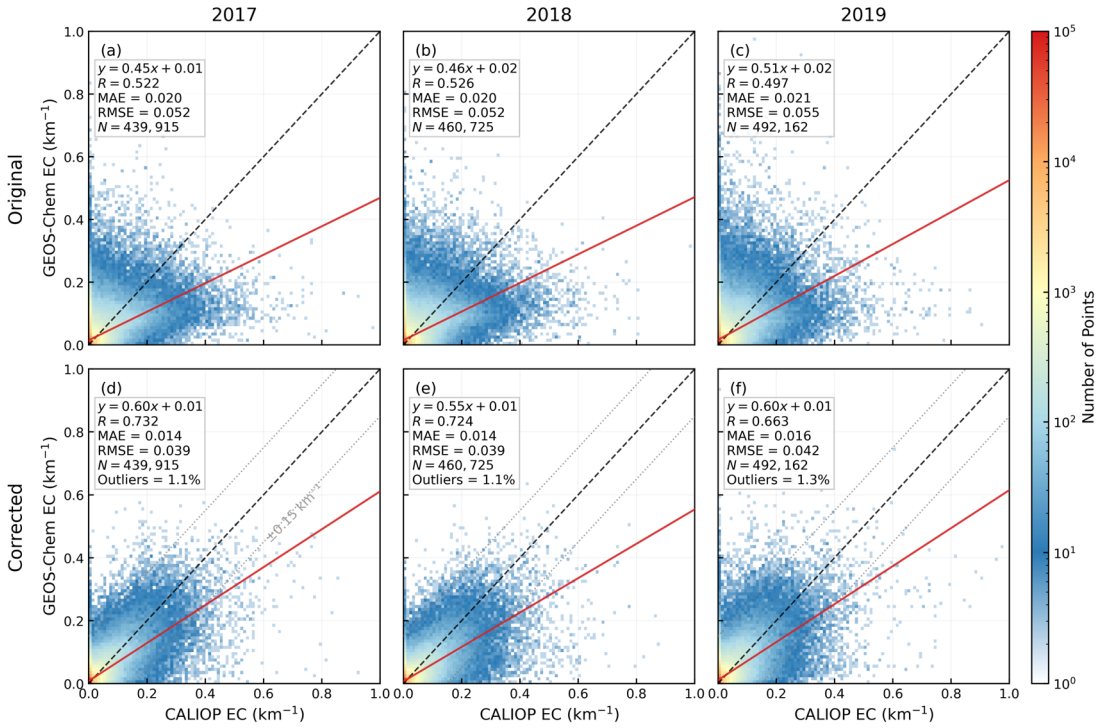


Figure 6. Density scatter plots comparing the simulated AEC against CALIOP observations over EA for the three test years: 2017 (a, d), 2018 (b, e), and 2019 (c, f). The top row (a, b, c) displays the validation results for the original GEOS-Chem simulation, while the bottom row (d, e, f) shows the results after correction by the physics-informed Transformer model. The dashed gray lines in the bottom panels (d–f) delineate the ±0.15 km^{-1} error envelope, with the corresponding percentage of outliers (points falling outside this envelope) indicated in the statistical boxes.

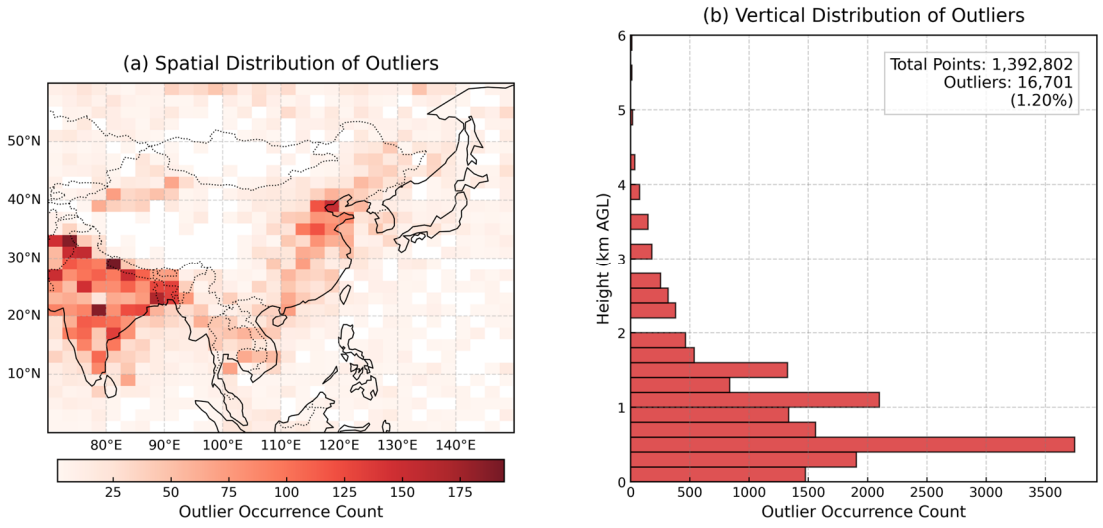


Figure S8. Spatial and vertical distributions of the residual outliers for the combined test years (2017–2019). Outliers are strictly defined as samples where the absolute

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.

We have added the following discussion in Section 4.2.1:

"Despite these substantial statistical improvements, visual scatter remains in the density plots. To rigorously quantify these discrepancies, an error envelope of ±0.15 km^{-1} is introduced in Figure 6(d–f). Statistical analysis indicates that outliers exceeding this threshold account for only 1.20% of the total dataset. Further diagnostic analysis (detailed in Sect. S10) reveals that these extreme deviations are not random noise but exhibit distinct spatial clustering over major emission hotspots (e.g., the IGP, the NCP, and the Indochina Peninsula), and are vertically confined within the PBL (<1.5 km AGL). These residuals are primarily driven by representativeness errors: CALIOP's narrow footprint captures transient, highly concentrated sub-grid aerosol plumes, which are inherently smoothed out during the spatial averaging process across the coarse 2°×2.5° grid of GEOS-Chem. Consequently, the GC-TF model captures the systematic, state-dependent biases of the grid mean, rather than fitting stochastic sub-grid extremes."

Additionally, we have added a dedicated sub-section S10. Analysis of Residual Outliers and Figure S8 in the Supplementary Information to provide the detailed spatial and vertical distribution of these outliers.

"**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."

(12) Figure 10. Font size too small.

Response:

We have enlarged the font in Figure 10. The enlarged Figure 10 is shown below.