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

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.

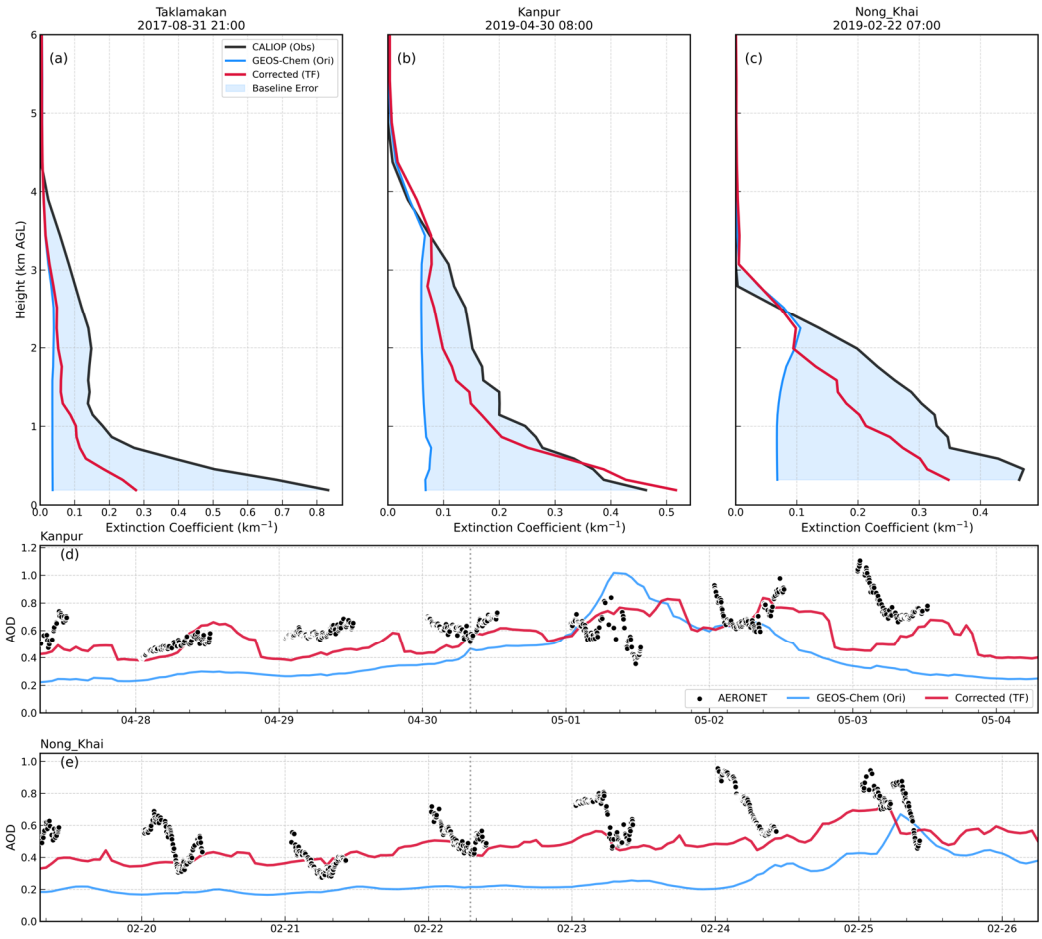

Figure 11. Composite analysis of aerosol vertical structures and temporal evolution during selected pollution episodes. Vertical profiles of AEC at three representative sites: Taklamakan (Dust, a), Kanpur (Anthropogenic Pollution, b), and Nong Khai (Biomass Burning, c). Time series of AOD at the Kanpur (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.

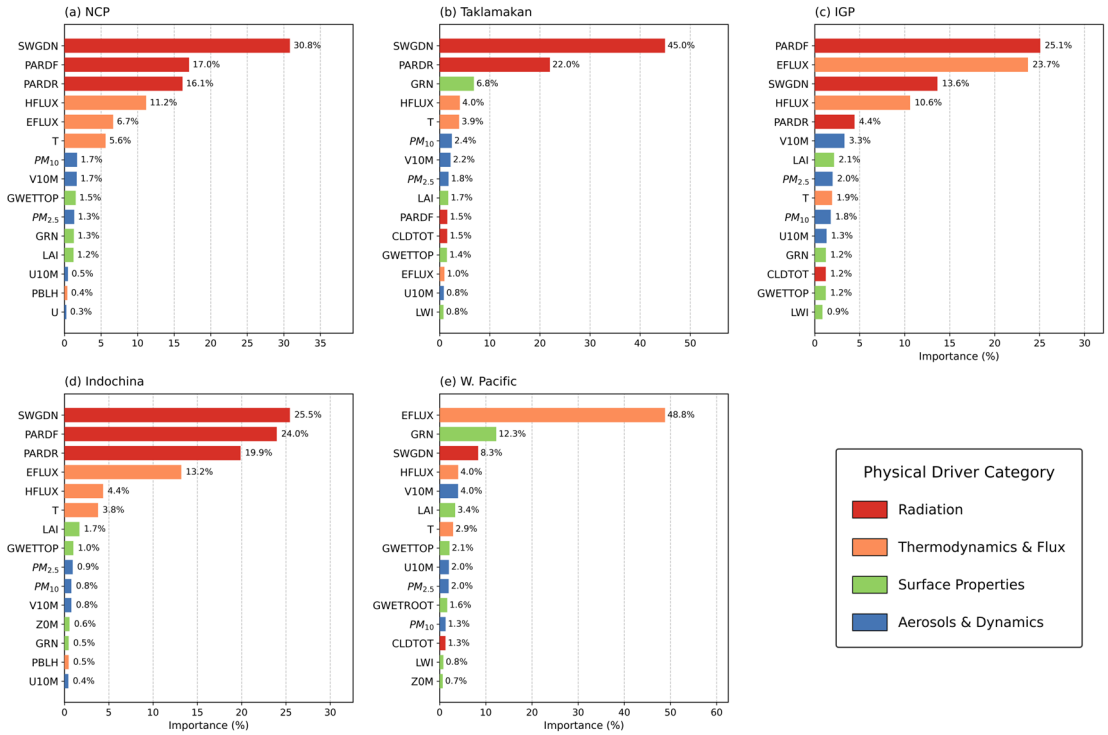

Figure 15. Regional variability in feature importance drivers identified by SHAP analysis for the test year 2019. The panels display the top 15 most influential features for predicting AEC simulation biases in five representative regions: NCP (a), Taklamakan Desert (b), IGP (c), Indochina (d), and Western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