Journal of Environmental Management 405 (2026) 129618

Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Research article

Physics-informed spatiotemporal analysis of methane concentrations in an oil sands region

Yang Xu [a], Hao Wang [b], Jude D. Kong [a,c,d,e,f,*]

[a] Artificial Intelligence and Mathematic Modelling Lab, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Office 662, Toronto, ON, M5T 3M7, Canada
[b] The Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, T6G 2J5, Canada
[c] Institute of Health Policy, Management and Evaluation (IHPME), University of Toronto, Toronto, Canada
[d] Department of Mathematics, University of Toronto, Bahen Centre for Information Technology, Room 6291, 40 St. George Street, Toronto, Ontario, Canada, M5S 2E4
[e] Africa-Canada Artificial Intelligence and Data Innovation Consortium (ACADIC), Canada
[f] Global South Artificial Intelligence for Pandemic and Epidemic Preparedness and Response Network (AI4PEP), Canada

ARTICLE INFO

Keywords:
Methane concentration
Environmental monitoring
Atmospheric transport
Oil sands
Spatiotemporal analysis
Uncertainty quantification
Physics-informed modelling

ABSTRACT

Methane (CH_4) emissions from complex industrial regions, especially oil sands ponds, exhibit strong spatial heterogeneity and episodic extremes, posing persistent challenges for reliable regional-scale concentration estimation. While recent data-driven models have improved predictive accuracy, their physical consistency, robustness to observation sparsity, and behaviour under extreme conditions remain insufficiently examined. Here, we develop a multi-source, physics-informed, and spatially structured prediction framework that integrates ground-based monitoring stations with satellite-derived background information to estimate regional CH_4 concentrations. The proposed framework explicitly enforces transport-consistent structure and spatial regularity while remaining robust to data gaps and sensor failures. Across independent validation experiments, the model achieves coefficients of determination exceeding 0.80 and demonstrates stable performance under simulated station dropout rates of up to 50%. Uncertainty calibration shows near-nominal predictive interval coverage (0.916 for a nominal 95% interval) with a mean interval width of ±215 ppb, indicating reliable probabilistic characterization. Spatial diagnostics further confirm physically realistic smoothness, with a roughness ratio of 0.64 relative to baseline interpolation methods, and strong alignment with wind-resolved transport patterns (Spearman $\rho = 0.817$). Using a minimal post-hoc monotonic correction as a diagnostic tool, the recoverability of extreme concentration amplitudes can be quantitatively assessed without modifying the underlying predictive model. Overall, this study demonstrates that integrating physical, spatial, and observational constraints enables not only accurate but also scientifically interpretable and deployment-ready CH_4 predictions in complex industrial environments.

1. Introduction

In recent years, the trend of global warming has significantly intensified. In January 2025, the World Meteorological Organization (WMO) confirmed 2024 as the warmest year on record at about 1.55 °C above the pre-industrial level (WMO, 2025), and the last decade (2015-2024) has become the warmest since meteorological records began. The primary driver of this phenomenon is caused by greenhouse effect due to human activities. The globally averaged surface concentration of CO_2 reached 420 parts per million (ppm), methane 1934 parts per billion, and nitrous oxide 336.9 parts per billion (ppb) in 2023. These values are 151%, 265%, and 125% of pre-industrial levels (WMO, 2024). Among these greenhouse gases, methane stands out for its exceptionally rapid warming effect over short timescales, with a 20-year global warming potential approximately 80 times that of CO_2, making it the most critical greenhouse gas for near-term climate mitigation despite its lower long-term global warming potential compared to nitrous oxide (Wei et al., 2023). Considering the powerful heating capability of methane, combined with its continuously rising emission rates, makes methane reduction a critical focus for current climate mitigation

* Corresponding author. Artificial Intelligence and Mathematic Modelling Lab, Dalla Lana School of Public Health, University of Toronto, 155 College Street, Office 662, Toronto, ON, M5T 3M7, Canada.
E-mail address: jude.kong@utoronto.ca (J.D. Kong).

https://doi.org/10.1016/j.jenvman.2026.129618
Received 25 January 2026; Received in revised form 19 March 2026; Accepted 6 April 2026

strategies. Industrial regions intensify this concern, particularly where oil and gas operations, coal mining, and other energy-related activities release substantial quantities of methane into the atmosphere (Kong et al., 2019). The accurate detection and quantification of these high-emission sources have become crucial for effective environmental management and policy implementation (Tiemann et al., 2024).

Methane emissions monitoring is quantified through two principal methodologies: a 'bottom-up' approach, which collects data from individual sources, for example the ground-based sensors and manual inspections (Brantley et al., 2014; Fox et al., 2019) and a 'top-down' approach, which infers emissions from atmospheric measurements, including advanced remote sensing technologies, using airborne and satellites with enhanced spatial and spectral resolutions (Jacob et al., 2022).

Current monitoring infrastructure presents its own set of limitations. Ground-based sensor networks provide accurate, high-frequency measurements but suffer from sparse spatial coverage due to deployment costs and maintenance requirements (Jiang et al., 2024; Jacob et al., 2016). Satellite remote sensing offers comprehensive spatial coverage but struggles with temporal resolution, cloud interference, and retrieval uncertainties over certain surface types (Jin et al., 2024). These complementary yet incomplete data sources create an opportunity for innovative multi-modal approaches that can leverage the strengths of each monitoring platform while compensating for its individual weaknesses.

Atmospheric modelling of methane concentrations has increasingly incorporated statistical and machine learning methods, evolving from early linear models such as Autoregressive Integrated Moving Average (ARIMA) to modern deep learning architectures (Sysoeva et al., 2025). While earlier models struggled with complex nonlinear dynamics (Liao et al., 2020), deep learning approaches like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) have proven more effective (Saha et al., 2025). RNNs, especially Long Short-Term Memory (LSTM), excel at modelling time-series data (Meng et al., 2022), while CNNs are powerful for analyzing spatial data, such as satellite imagery (Dumont Le Brazidec et al., 2023). However, improved accuracy alone does not guarantee that the resulting predictions remain physically interpretable or spatially realistic under complex atmospheric conditions.

Recent years have seen the emergence of multi-modal architectures that attempt to fuse heterogeneous data sources (Xu et al., 2026). Hybrid CNN-LSTM models have shown promise in combining temporal ground station data with spatial satellite observations to generate continuous pollution maps (Yazdinejad et al., 2025). However, these approaches face significant challenges in handling data gaps, resolution mismatches, and the irregular spatial distribution of monitoring stations (Zhu et al., 2023). Graph neural networks (GNN) have emerged as a natural solution for modelling the non-Euclidean relationships between sensor locations, while transformer architectures with self-attention mechanisms have shown strong capability in capturing long-range spatio-temporal dependencies. Despite these advances, many models remain difficult to interpret physically and may exhibit unstable behaviour when applied under sparse or dynamically changing observational conditions.

A fundamental limitation of purely data-driven approaches lies in their lack of explicit representation of the physical processes governing atmospheric transport. This "black box" nature can lead to physically implausible predictions and degraded generalization under conditions of data sparsity or regime shifts (Alamu et al., 2025). Moreover, their performance often depends strongly on the availability of large, high-quality training datasets (Wesselkamp et al., 2024). To address the limitations of purely data-driven models, physics-informed machine learning has been increasingly explored. Early approaches incorporated outputs from physical models as auxiliary inputs (Daw et al., 2022). More recent developments embed governing physical relationships, such as partial differential equations (PDEs), into learning frameworks to promote physical consistency (Cuomo et al., 2022). These approaches have demonstrated improved robustness under data sparsity and enhanced generalization in atmospheric applications (Raissi et al., 2019).

Despite these developments, a critical gap remains in the systematic evaluation of whether multi-modal, physics-informed models produce methane concentration estimates that are not only accurate but also physically consistent, spatially coherent, and robust to observational perturbations. In complex industrial regions, where emission patterns are highly heterogeneous and extreme events play a disproportionate role, the lack of such an assessment limits confidence in model-based methane monitoring and decision-making.

In this work, we propose a physics-informed spatiotemporal modelling framework for methane concentration reconstruction in an oil sands region in Canada. The methodology combines ground-based monitoring data with spatial structure and transport-aware constraints to improve the stability and physical consistency of inferred methane fields. Rather than focusing solely on prediction, we emphasize comprehensive evaluation of model behaviour, including uncertainty calibration, robustness to station sparsity, spatial realism, wind-aligned transport consistency, and performance under extreme concentration events. Drawing on real-world monitoring data, this work shows how combining data-driven learning with physically grounded constraints can help to achieve more reliable regional methane monitoring in complex industrial environments.

2. Data and methods

2.1. Study area and station stratification

This study analyzes the Wood Buffalo Environmental Association (WBEA) network in northeastern Alberta, Canada, utilizing data from 28 ground-based monitoring stations from January 1, 2024, to January 1, 2025 (BOE Report, 2025). To address data heterogeneity, stations were stratified into Tier 1 (Core, n = 15) and Tier 3 (Auxiliary, n = 13) based on methane data completeness (see Text S1 in Supplementary Material for detailed classification protocols). Tier 1 stations (≥85% coverage) served as the primary training dataset, while Tier 3 stations (<40% coverage) were treated as a validation set to evaluate the model's spatial generalization capabilities under sparse observation conditions. It should be noted that station classification was based on a composite assessment of multiple data quality dimensions rather than CH_4 coverage alone. Specifically, Tier 1 designation required near-complete records across all monitored variables, including CH_4 concentration, wind speed, wind direction, temperature, and co-pollutant species. Stations were assigned to Tier 3 if they exhibited incomplete records in one or more critical variables, regardless of their CH_4 coverage status. Accordingly, the Tier 3 group comprises three distinct sub-categories: (1) ten stations equipped with no CH_4 monitoring instrument during the study period; (2) three stations with complete CH_4 records (100% coverage) but substantial gaps in key meteorological variables, particularly wind speed and wind direction, which are essential for transport-consistent model training; and (3) portable deployment units with inherently lower measurement reliability. A complete station-level classification with CH_4 data coverage statistics is provided in Table S5 of the Supplementary Material. The spatial distribution of Tier 1 and Tier 3 stations across the study domain is illustrated in Fig. 1.

2.2. Data description

Building upon our previous methodological framework, this study integrates ground-based measurements from the expanded 28-station WBEA area with satellite-derived atmospheric methane observations. The comprehensive dataset provides full spatial coverage for methane emission modelling in the oil sands region.

The ground monitoring dataset consists of continuous time-series data from all 28 WBEA stations. Each station records meteorological

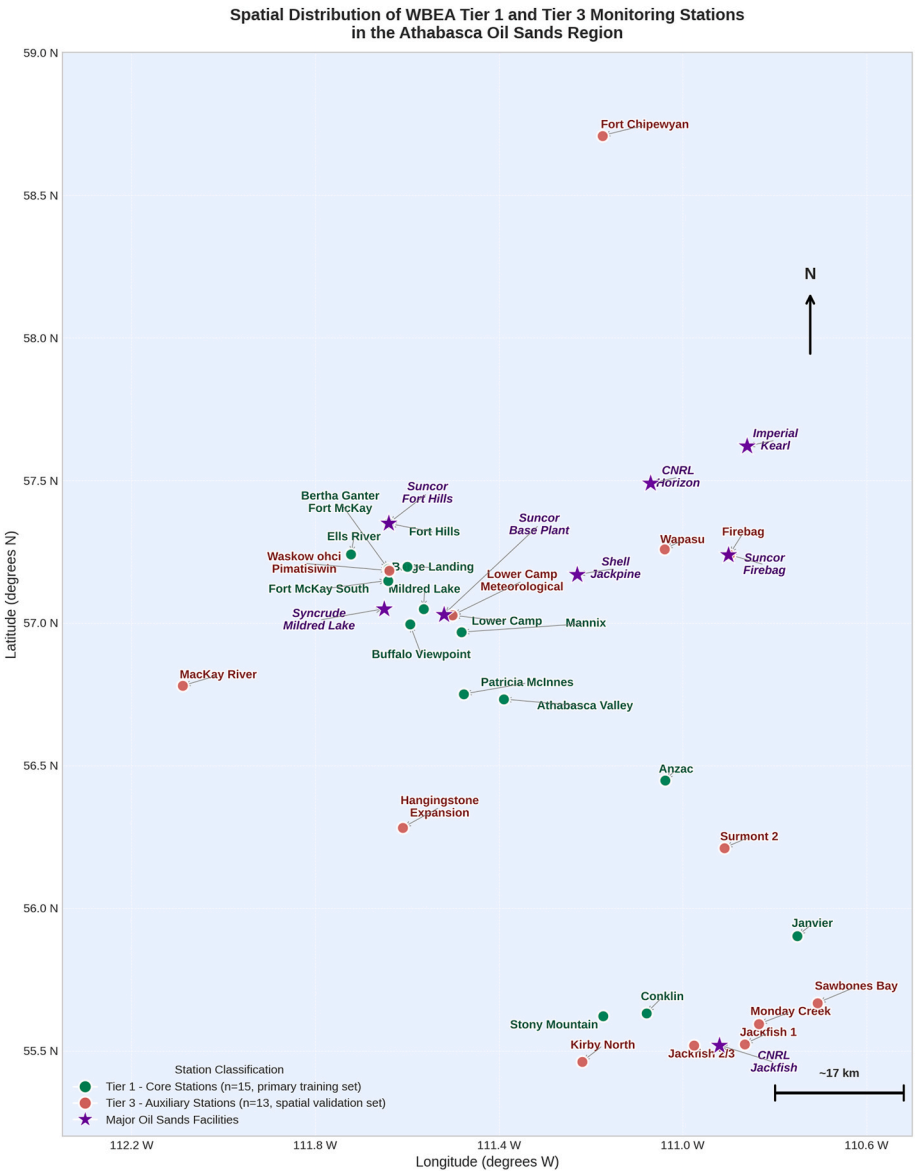

Fig. 1. Spatial Distribution of WBEA Tier 1 and Tier 3 Monitoring Stations in the Athabasca Oil Sands Region. Green circles denote Tier 1 core stations (n = 15) used as the primary model training dataset; red circles denote Tier 3 auxiliary stations (n = 13) reserved for spatial generalization validation. Purple stars indicate major oil sands facilities for geographic reference. Station classification was based on a composite assessment of CH4 measurement completeness, availability of key meteorological variables (wind speed, wind direction), and station deployment type (see Table S5 for full details). Tier 1 and Tier 3 stations are spatially interleaved across the full study domain, with both groups represented across the northern industrial core (57.0°N–57.4°N), the central monitoring zone (56.4°N–57.0°N), and the southern sector (55.5°N–56.3°N), confirming that the Tier 3 validation set provides a geographically representative assessment of the model's spatial generalization capability rather than targeting a specific sub-region.

variables such as temperature, relative humidity, barometric pressure, wind vectors, and concentrations of atmospheric pollutants. The key monitored pollutants include CH_4, total hydrocarbons (THC), non-methane hydrocarbons (NMHC), sulphur compounds (H_2S, SO_2), total reduced sulphur (TRS), nitrogen oxides (NO_x, NO, NO_2), ozone (O_3), carbon monoxide (CO), and particulate matter ($PM_{2.5}$). Among these, THC, CH_4, and NMHC are simultaneously measured using a Thermo Scientific Model 55i gas chromatograph with flame ionisation detection (GC-FID), which provides simultaneous quantification of all three hydrocarbon species at an hourly temporal resolution. This instrumentation has been continuously operated across all Tier-1 stations since January 1999 (WBEA, 2024). A comprehensive summary of all monitored variables, including their instrumentation, measurement principles, temporal resolution, and role in the modelling framework, is provided in Table S2 in the Supplementary Material. This ground-level data provides high temporal resolution insight into local atmospheric dynamics but is limited by its sparse spatial coverage. To complement these ground measurements with a broad spatial context, we incorporated atmospheric methane concentration data from GHGSat's SPECTRA platform (GHGSat, 2024). This satellite product provides area-averaged methane concentration heatmaps at a spatial resolution of 10 km and a weekly temporal resolution. To align with our study's scope, the satellite data were spatially cropped to encompass the entire geographical region containing the 28 WBEA monitoring stations, ensuring complete coverage of the study area. The fusion of these two data sources with the temporally dense ground data and the spatially comprehensive satellite data creates a robust foundation for our model to learn the complex spatio-temporal patterns of methane emissions in the region. It should be noted that the GHGSat SPECTRA Basic product provides area-averaged, cloud-screened weekly methane concentration

composites, in which cloud interference has been addressed at the data product level through the provider's internal retrieval and compositing algorithm. The weekly heatmaps therefore represent cloud-free observations with complete spatial coverage over the study domain, and no additional cloud gap-filling procedure was required in our pre-processing pipeline. This is a particularly important feature for high-latitude regions such as the Athabasca Oil Sands, where persistent cloud cover can otherwise severely compromise satellite data availability and continuity.

2.3. Data pre-processing and feature engineering

2.3.1. Temporal alignment of multi-resolution data streams

A fundamental challenge in fusing the two data modalities used in this study is their substantial difference in temporal resolution: the WBEA ground station network provides continuous measurements at an hourly resolution, while the GHGSat SPECTRA satellite product delivers weekly XCH_4 heatmaps. To enable joint model training within a unified deep learning framework, the weekly satellite observations must be up-sampled to match the hourly ground station timestamps prior to model input.

To determine the most appropriate temporal alignment strategy, three interpolation methods were systematically evaluated: (1) linear interpolation, which assumes a constant rate of change between consecutive weekly observations; (2) nearest-neighbour interpolation, which assigns the value of the temporally closest weekly observation to each hourly timestamp, producing a step-function approximation; and (3) time-weighted interpolation, which computes the value at each hourly timestamp as an inverse-distance-weighted average of the two temporally nearest weekly observations, with weights inversely proportional to the squared temporal distance. This local two-point inverse-distance weighting (IDW) scheme gives greater influence to the nearer weekly anchor while preserving smooth temporal transitions, and is physically motivated by the expectation that the atmospheric state at any given hour is more strongly influenced by the nearest available observation than by distant ones.

Performance of the three methods was quantitatively assessed using a cross-validation experiment, in which each interior weekly observation was withheld in turn and predicted from the remaining weeks, with reconstruction fidelity measured by Mean Squared Error (MSE). As shown in Fig. S8, time-weighted interpolation achieved the lowest reconstruction MSE of 1.903×10^{-2}, outperforming linear interpolation of MSE $= 2.519 \times 10^{-2}$, and nearest-neighbour interpolation of MSE $= 3.618 \times 10^{-2}$. Fig. S9 further illustrates the temporal behaviour of the three methods applied to the four highest-variance XCH_4 spatial context features across the full study period. The high density of weekly satellite observations over the study year results in temporally coherent hourly time series under all three methods, which is itself an indicator of good data coverage; the time-weighted method nonetheless achieves the lowest cross-validation reconstruction error, confirming its superior fidelity at the individual weekly anchor points. Based on this systematic evaluation, time-weighted interpolation was adopted as the standard temporal alignment procedure for all satellite-derived features throughout this study.

We explicitly acknowledge that temporal up-sampling from weekly to hourly resolution is an inherent limitation of the present work. The interpolated satellite features cannot reproduce true sub-weekly atmospheric variability, and should therefore be interpreted as spatially representative background context inputs rather than fine-scale temporal predictors at the hourly scale. This limitation is consistent with the general challenge of temporal resolution mismatch in multi-source environmental data fusion, and overcoming it will require future data from next-generation satellite missions with higher revisit frequencies, such as MethaneSAT or dedicated geostationary methane sensors.

2.3.2. Wind direction transformation

Firstly, all the wind direction values must be handled because wind direction, by nature, is a cyclic variable spanning 0 to 360°, which complicates direct use in machine learning models due to the numerical variances where values like 359° and 1° appear distant despite being physically neighboring. To mitigate this issue, the original wind direction data was transformed into two components, which are *U*, representing the east-west flow vector, and *V*, representing the north-south flow. This conversion translates the circular measurement into two continuous linear variables, making them more suitable for ML modelling and effectively capturing the actual directional movement of air masses. All wind direction values were converted into the *U* and *V* components, which were then used as new features in place of the original wind direction data for model training.

2.3.3. Multi-scale feature extraction via wavelet decomposition

To decouple long-term atmospheric trends from transient emission events, we applied discrete wavelet decomposition to all time-series features. We selected the Daubechies 4 (db4) wavelet as the basis function, which is widely regarded as the standard for non-stationary environmental signal analysis due to its optimal balance of frequency localization and smoothness (Percival and Walden, 2000). We performed a five-level decomposition (A5, D1-D5). To optimize the feature set, we conducted a rigorous signal energy distribution analysis. Results demonstrated that the low-frequency approximation (A5) and the first two high-frequency detail components (D1, D2) cumulatively account for over 99.5% of the total signal energy across the network. Conversely, higher-order components (D3-D5) represented negligible variance (<1%) and were identified primarily as high-frequency noise. Consequently, we retained only the [A5, D1, D2] triplet to represent each variable, ensuring the model captures both the underlying baseline (A5) and critical short-term fluctuations (D2) without overfitting to noise (Sahoo et al., 2024).

2.3.4. Unified regional temporal pattern identification

We implemented a unified regional clustering strategy to provide the model with an interpretable physical context. We aggregated wavelet features from all 15 Tier 1 stations into a single high-dimensional feature space to establish consistent regional environmental semantics. Multiple clustering algorithms have been evaluated, including K-Means clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Gaussian Mixture Models (GMM), combined with dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbour Embedding (t-SNE). The PCA + K-Means combination yielded the most robust cluster separation. The analysis was performed on two time scales to generate dual labels. The A5 represents the baseline patterns that classify the long-term atmospheric state into baseline, moderate, or elevated pollution levels, driven by regional meteorological conditions and persistent emissions. The D2 is also named as dynamic states, which classifies short-term volatility into stable, variable, or volatile conditions, flagging potential acute emission events or turbulent transport. These two dimensions generate nine distinct environmental semantic states. These labels serve as dynamic control signals for the subsequent neural network modules, allowing the model to adaptively weight spatial and temporal features based on the identified physical regime.

2.3.5. Diagnostic feature analysis via surrogate modelling

To rigorously validate that the extracted feature set captures the underlying physical mechanisms governing methane concentrations before training the complex hybrid model, we implemented a diagnostic interpretability step. A tree-based agent model (XGBoost) was trained on the pre-processed dataset to serve as a transparent seeker for feature interactions. We then employed SHapley Additive exPlanations (SHAP) to quantify the contribution of each environmental variable. This diagnostic analysis specifically aims to verify the chemical fingerprints of

industrial emissions and the non-linear response to meteorological dispersion, ensuring the dataset possesses sufficient physical accuracy to support the subsequent PINN architecture.

2.4. Physics-informed regularization

To ensure physical plausibility, the model incorporates regularization terms that guide predictions toward spatially consistent behaviour. The core component is a Transport-Consistency Regularization, which penalizes concentration gradients between neighboring stations that contradict observed wind vectors. Unlike explicit PDE solvers, this acts as a soft constraint, enforcing that spatial gradients align with advection-diffusion principles derived from local meteorology. This is complemented by a Boundary Stabilization constraint to mitigate extrapolation errors at the domain edges (defined via convex hull) and a Domain-Average Stability term to prevent unrealistic network-wide drift. To balance these physical constraints with data-driven objectives, we employed an Adaptive Weighting strategy that dynamically adjusts the contribution of physical loss terms based on training progress. The complete mathematical formulations for these constraints are detailed in Text S2.

2.5. Hybrid deep learning architecture

The proposed framework, as shown in Fig. 2, is built upon a Bidirectional Cross-Attention Mechanism that dynamically fuses temporally dense ground observations with spatially extensive satellite imagery. Unlike static concatenation, this module establishes parallel pathways where ground and satellite features "query" each other to extract context-aware representations. The core innovation lies in the Physics-Informed Neural Network (PINN) Layer, which integrates three advanced constraints into the training objective: (1) Physics-Informed Constraints (*L_physics*) that embed the advection-diffusion principles; (2) GNN-based Spatial Constraints (*L_spatial*) utilizing Graph Attention Networks to encode inter-station topology; and (3) Cluster-Guided Temporal Constraints (*L_temporal*) that align predictions with atmospheric stability regimes. The model is optimized using a composite loss function with adaptive weighting to balance data-driven errors and physical penalties. Following fundamental atmospheric physics, the Physics-Informed Constraints are grounded in the complete advection-diffusion equation governing CH_4 transport:

$$\frac{\partial C}{\partial t} + \nabla\cdot(vC) = \nabla\cdot(K\nabla C) + S(x,t)$$

where v is the wind velocity vector governing advection, K is the turbulent diffusion coefficient, and S (x,t) is the spatiotemporally varying source generation term. In our hybrid framework, each term is addressed at a distinct modelling layer. The advection term $\nabla\cdot(vC)$ and diffusion term $\nabla\cdot(K\nabla C)$ are explicitly enforced by the PINN loss function L_physics using observed wind vectors and estimated diffusion parameters. The source term S (x,t), which governs local CH_4 generation from emission sources through complex thermodynamic and chemical processes, including temperature-driven volatilization and co-emission with THC from oil sands operations, is implicitly approximated by the Baseline Model backbone through its learned representations of Temperature and THC, precisely the variables identified as dominant drivers by the SHAP analysis. This structural design ensures physically consistent closure across both local emission generation and spatial transport dynamics. Mathematical derivations of the attention mechanism and the loss function are detailed in Text S3.

We adopted a diagnostic evaluation protocol that extends beyond standard statistical metrics to examine physical consistency and operational robustness, enabling a comprehensive assessment of the framework's applicability for regional environmental monitoring. Reflecting practical forecasting constraints, the dataset was divided strictly chronologically to prevent temporal data leakage, with the first 70% used for training, the subsequent 15% for validation and hyperparameter tuning, and the final 15% reserved as a held-out test set. All reported metrics are derived exclusively from this test set. Baseline predictive performance was first quantified using the Root Mean Square Error (RMSE) and the coefficient of determination (R^2), ensuring that the model captures the dominant temporal variability of methane concentrations. While these metrics assess point accuracy, they are insufficient to characterize the reliability of a monitoring system under realistic operating conditions.

To evaluate predictive uncertainty, we quantified ensemble-based uncertainty estimates and assessed the reliability of prediction intervals using the Prediction Interval Coverage Probability (PICP) at nominal 90% and 95% confidence levels. To ensure operational usability, a post-hoc scaling calibration was applied using the validation set. Operational robustness was examined through a station dropout

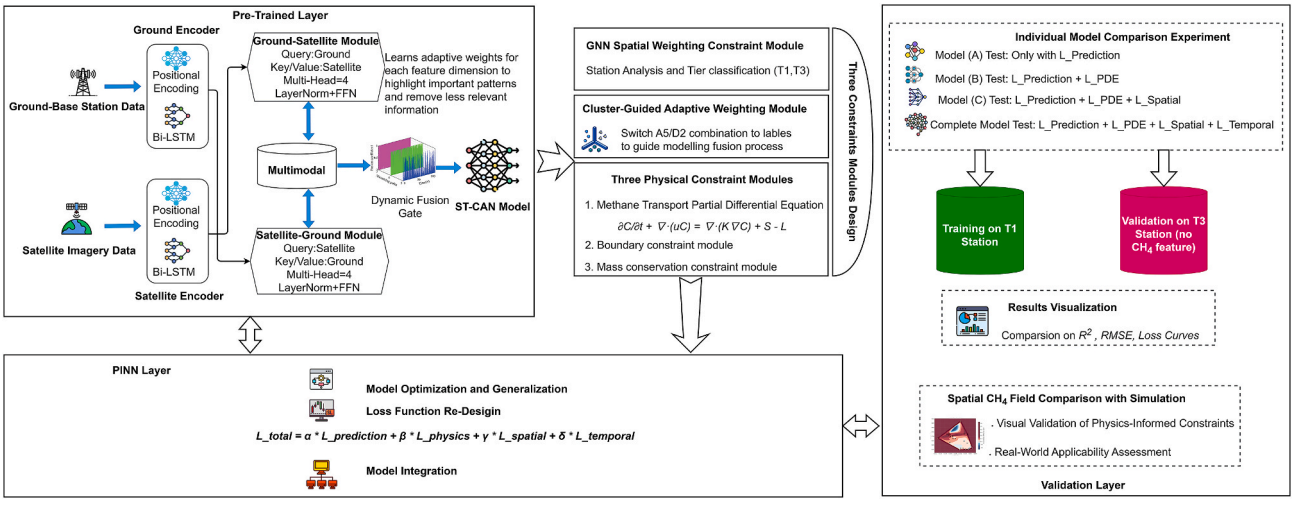

Fig. 2. Schematic Diagram of the Proposed ST-CAN Framework with Physics-Informed and GNN Spatial Constraints. The framework comprises a Pre-Trained Layer with bidirectional cross-attention fusion of ground and satellite encoders, a PINN Layer (composite loss function with three constraint modules), and a Validation Layer (ablation experiments and spatial field comparison). In the Individual Model Comparison Experiment panel (right), Model (A), Model (B), Model (C), and the Complete Model correspond respectively to the Baseline Model, Physics-Informed Model, GNN Model, and Complete Model as defined in Table S1. L_physics denotes the physics-informed PDE-based advection-diffusion transport constraint, L_spatial the GNN-based spatial constraint, and L_temporal the cluster-guided temporal constraint.

stress test, in which 10% to 50% of monitoring stations were randomly removed during inference. The resulting degradation in predictive performance was quantified to determine whether reliable regional estimates can be maintained during periods of sensor maintenance or failure. Beyond accuracy and robustness, additional diagnostics were introduced to verify that the reconstructed methane fields respect atmospheric physical structure rather than merely fitting data points. Spatial realism was assessed by analyzing field roughness on a K-nearest-neighbour (KNN) station graph, which penalizes unrealistic local fragmentation and provides a quantitative measure of spatial smoothness consistent with atmospheric transport.

Transport consistency was further evaluated by conditioning predicted concentration gradients on observed wind vectors. For each time step, monitoring stations were classified as upwind or downwind relative to a regional wind direction derived from observed horizontal wind components. The concentration gradient was defined as the difference between the mean downwind and upwind concentrations. Only time steps satisfying minimum wind speed, angular separation, and station count criteria were retained, and the same procedure was applied to both observations and model predictions to enable direct comparison. Finally, given the disproportionate environmental impact of emission events, we examined model fidelity for extreme concentrations exceeding the 95th percentile. This diagnostic focused on the preservation of event amplitudes and included a post-hoc isotonic regression to quantify the recoverability of peak events that may be attenuated by stability-oriented loss functions.

3. Results and discussion

3.1. Diagnostic analysis of environmental drivers

In the initial stage of analysis, raw observations from all 28 WBEA stations were processed into a quality-controlled and integrated dataset. Due to heterogeneity in instrumentation across the network, an automated feature selection procedure was applied to identify a consistent set of input variables, including common meteorological parameters (wind speed, wind vector components, temperature, and relative humidity) and key co-pollutants, including total hydrocarbons (THC), NO_x, O_3, and $PM_{2.5}$. To examine whether the selected variables exhibit physically meaningful relationships with methane variability, a diagnostic analysis was conducted using a tree-based surrogate model. This analysis was used to characterize dominant associations between input

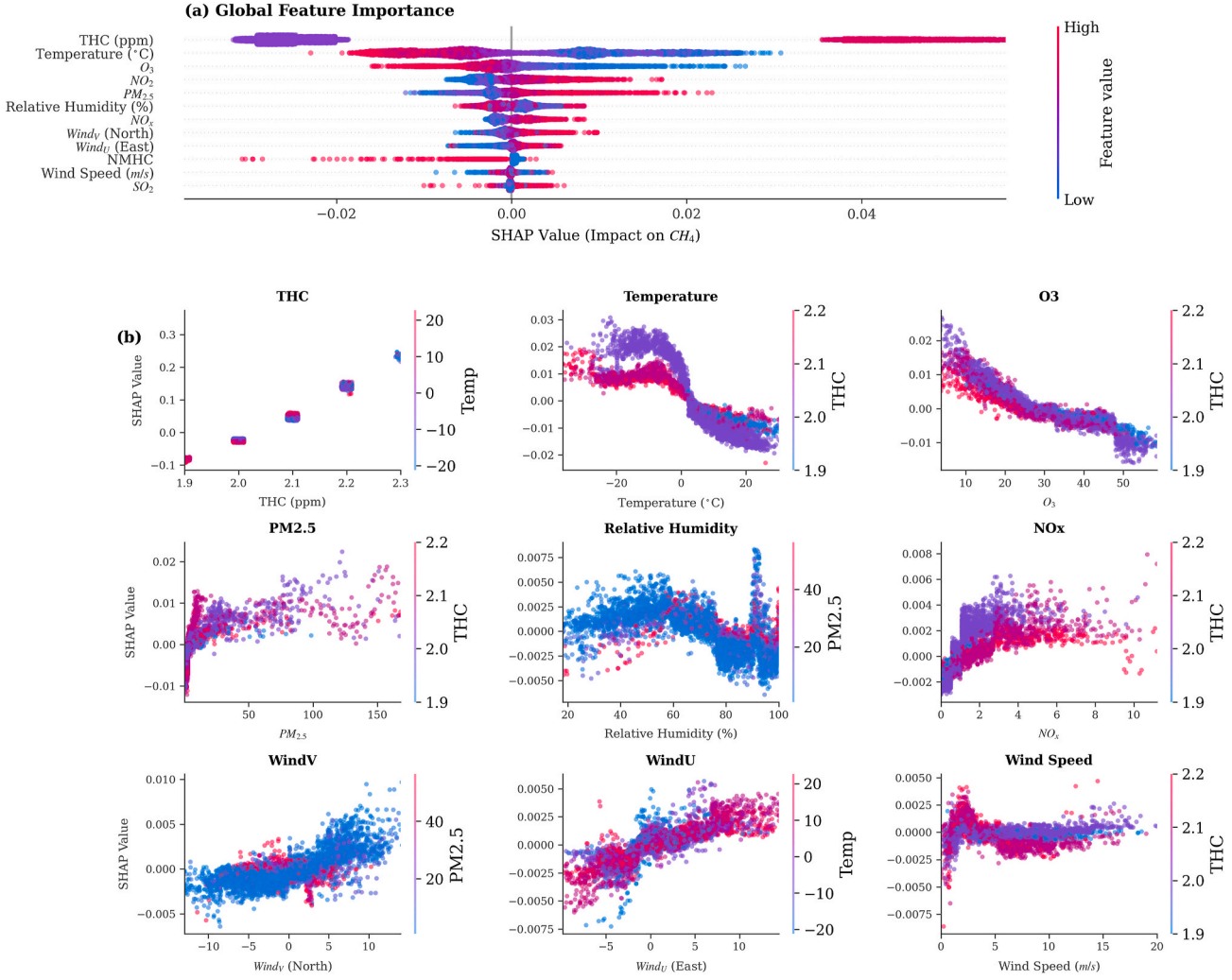

Fig. 3. Diagnostic SHAP Analysis Revealing Relationships Between Methane Concentrations and Key Environmental Variables. (a) Global feature importance ranking, where each dot represents an individual sample. Colour indicates the relative magnitude of each feature value using a per-feature quantile normalisation scale (red/warm: high; blue: low), and horizontal position indicates the SHAP value. (b) Dependence plots for the top nine variables, highlighting nonlinear associations consistent with known emission signatures and atmospheric transport processes, including wind-driven dispersion, temperature-dependent vertical mixing, and co-emission relationships with industrial pollutants. In panel (b), dot colour represents the value of an automatically selected interaction feature (labelled on the right-hand colour axis of each subplot) rather than the focal feature itself, following standard SHAP dependence plot convention.

variables and predicted methane concentrations, providing a data-driven assessment of whether known emission and transport processes are reflected in the observations. SHAP-based feature attribution results from the most comprehensively instrumented monitoring station are shown in Fig. 3. To further validate the robustness of this feature ranking, a comparison with XGBoost native Gain importance is provided in Fig. S7 in the Supplementary Material, confirming consistent rankings across both methods.

The global feature importance ranking, as shown in Fig. 3a, provides a hierarchical view of the driving forces. THC and Temperature emerge as the most dominant predictors, followed closely by O_3 and NO_x. The colour distribution in the plot reveals clear directional impacts. High values (red dots) of industrial pollutants like THC and NO_x are consistently associated with positive SHAP values, driving up predicted methane concentrations, whereas high values of dispersion factors like Wind Speed and Ozone typically exert a negative influence.

Detailed dependence plots illustrated in Fig. 3b further unravel these non-linear physical interactions through three distinct mechanisms. First is the industrial chemical fingerprint. As a primary proxy for fugitive emissions in oil sands extraction, THC exhibits a strongly positive association with model-predicted methane contributions. In the Athabasca Oil Sands region, THC and CH_4 are co-emitted from the same industrial processes, including tailings pond volatilization and equipment fugitive releases, such that elevated THC serves as a chemical fingerprint exclusive to active industrial emission events. Background methane variability driven by biogenic sources or long-range transport, by contrast, occurs in the absence of concurrent THC enhancements. This strong coupling therefore serves as a definitive industrial signature, confirming the data's ability to distinguish active emission plumes from background concentration fluctuations. Similarly, NOx exhibits a rapid positive response, reinforcing the identification of combustion-related sources. Interestingly, Ozone exhibits a notable negative correlation with methane. This inverse relationship captures two complementary mechanisms. It reflects the meteorological dilution effect. High ozone levels typically occur during sunny, convective afternoons when vertical mixing is strongest, naturally diluting surface-level methane. In addition, it aligns with the atmospheric titration effect characteristic of industrial plumes. Fresh emissions contain high levels of NO, which chemically consumes ambient ozone within the plume. Consequently, local dips in ozone concentration serve as a negative marker, pinpointing the precise location of concentrated methane plumes against the cleaner background air.

Second, the data accurately reflect meteorological dispersion dynamics governed by atmospheric transport physics. The dependence plot for Wind Speed exhibits a characteristic unimodal response. In the very low wind speed regime (0–3 m/s), SHAP values rise sharply and reach a positive peak, confirming that near-stagnant conditions are the primary driver of surface-level pollutant accumulation by suppressing vertical mixing and horizontal dispersion. As wind speed increases beyond approximately 3 m/s, SHAP values drop markedly, reflecting the rapid onset of advective dilution that effectively disperses the methane plume. Beyond 5 m/s, SHAP values stabilise near zero, indicating diminishing marginal returns in concentration reduction once the plume is effectively dispersed. This non-monotonic unimodal pattern is physically consistent with the advection-diffusion principles governing atmospheric transport: stagnant air maximises near-source accumulation, while even moderate wind speeds are sufficient to achieve effective plume dispersion.

Finally, the model identifies the effect of vertical air movement. The Temperature curve reveals a distinct physical transition. At low temperatures (<0 °C), SHAP values are consistently positive. This points to the 'trapping effect' of winter inversions, where cold, heavy air prevents pollutants from rising, keeping them near the ground. In contrast, higher temperatures show negative SHAP values. This indicates a 'dilution effect' caused by thermal convection, where warm air rises and mixes with the atmosphere above, effectively reducing the methane concentration at the surface.

3.2. GNN-based adaptive weighting module

To account for spatial heterogeneity in data coverage and site relevance across the monitoring network, an adaptive weighting scheme was implemented using a graph-based model. Rather than directly predicting methane concentrations, this module produces a set of spatial weights that modulate the relative contribution of each monitoring station within the network. The resulting weights reflect differences in spatial context and data characteristics among the 28 WBEA stations.

The spatial distribution of the learned weights is shown in Fig. S1 in the supplementary material. This map displays the geographic location of each station, with the size and colour of the markers corresponding to the magnitude of the learned spatial weight. The results reveal a clear pattern of non-uniform importance across the network. The GNN assigned the highest weights to stations located in the central, high-activity industrial zone, particularly those in proximity to major oil sands facilities. For instance, stations like Mildred Lake (AMS 2) and Fort Hills (AMS 23) were identified as highly influential. In contrast, stations located in the southeastern portion of the domain, including several portable units, exhibit lower weights.

A comparison between the geometric centroid of the monitoring network and the centroid of the weighted station distribution provides further insight into the spatial structure of the weighting scheme. The weighted centroid is displaced toward the central industrial region relative to the geometric centroid. This spatial shift indicates that stations closer to areas of concentrated industrial activity exert a greater influence on the aggregated network signal than would be expected from geometry alone, consistent with known patterns of emission intensity in the study region.

The statistical distribution of the station weights is summarized in Fig. S2 in the supplementary material. The weights are concentrated near the central value (mean = 0.500, median = 0.495), with a modest positive skew. This distribution indicates that most stations contribute comparably, while a smaller subset exhibits elevated influence.

3.3. Semantic clustering-guided temporal constraint module

We systematically evaluated 12 combinations of clustering algorithms (including K-Means, DBSCAN, and GMM) and dimensionality reduction techniques to ensure robust cluster identification. While density-based methods, for example, DBSCAN, were assessed, PCA combined with K-Means was ultimately selected as the optimal strategy based on three decisive criteria critical for environmental monitoring. First, K-Means ensures 100% temporal data coverage, whereas density-based algorithms classify outliers as noise, creating gaps in the time-series that would disrupt the continuous gating mechanism. Second, K-Means offers superior physical interpretability, producing explicit cluster centroids in the original feature space that map directly to characteristic pollutant profiles. Finally, its linear complexity ensures computational efficiency suitable for operational deployment across the network.

The optimal number of clusters ($k = 3$) was rigorously determined through the convergence of three independent methods. As illustrated in Fig. S3 in the supplementary material, the Elbow Method exhibited a clear inflection at $k = 3$, and the Silhouette Score achieved its maximum for the D2 component at this point. Furthermore, the Bayesian Information Criterion (BIC) for GMM showed a continuous decrease beyond $k = 3$, indicating overfitting. This statistical convergence aligns with the established atmospheric pollution classification scheme (Baseline, Moderate, Elevated).

Fig. S6 in the supplementary material visualizes the clustering results of three representative algorithms (K-Means, GMM, Hierarchical) in the PCA-reduced feature space. All three methods successfully identify distinct long-term atmospheric regimes, with K-Means producing the

most compact and well-separated clusters. The cluster centroids in the original 48-dimensional feature space reveal physically meaningful differences in CH_4 concentrations and meteorological conditions, validating the clustering approach.

The unified clustering analysis successfully extracted distinct and physically interpretable environmental patterns from the aggregated T1 station network. Unlike station-specific models, this approach establishes a consistent regional semantic baseline across the heterogeneous monitoring network. As visualized in the targeted radar charts in Fig. 4a, the clustering of long-term trend components (A5) identified three distinct baseline regimes driven primarily by pollution intensity. The dominant background condition, labelled as the Baseline State (Cluster 1), accounts for 66.9% of the dataset and is characterized by the lowest mean methane levels (~2.03 ppm) alongside minimal industrial pollutants, representing typical clean atmospheric days. In contrast, the Elevated State (Cluster 2), which comprises 26.5% of the period, captures high-pollution events. This cluster exhibits a drastic surge in NO_x (mean ~16.19 ppb) and SO_2 with elevated methane (~2.16 ppm), a multi-pollutant signature that strongly suggests industrial plume transport rather than biogenic sources. A third Moderate State (Cluster 0, 6.6%) represents transition periods characterized by intermediate pollution levels.

Meanwhile, the clustering of D2 components effectively segregated periods based on signal volatility, as shown in Fig. 4b. The most frequent Stable Conditions (Cluster 0, 53.1%) are characterized by minimal variance across all pollutants, indicating calm atmospheric conditions. Conversely, the Volatile Conditions (Cluster 2, 20.0%) flag critical dynamic events. This state exhibits a methane volatility of 0.093, nearly double that of the stable state, effectively highlighting periods of active plume passages or acute emission spikes. The robustness of these clusters is quantitatively confirmed by the separation analysis, as shown in Fig. S10 in Supplemental Material, where the Euclidean distance between the Baseline and Elevated centroids in the A5 space reaches 9.93, indicating excellent separation. Furthermore, the identified semantic states persist for an average of 6.36 h, confirming that the model captures sustained atmospheric regimes rather than random noise.

To validate the physical realism of these semantic states, we analyzed their temporal distributions via fingerprint heatmaps as shown in Fig. 5. The results reveal distinct temporal fingerprints that align with atmospheric boundary layer physics. The Elevated states, such as State 6 (Elevated-Stable), show a strong winter dominance from December to February and peak during night and early morning hours. This pattern is consistent with the formation of nocturnal inversion layers that trap pollutants near the surface during colder months. In contrast, State 8 (Elevated-Volatile) displays a distinct diurnal dynamic with a sharp frequency peak at 15:00 during February. This timing coincides with maximum boundary layer height and peak industrial operational hours, confirming its interpretation as a "Critical Operational Event" driven by active emissions and turbulent mixing. By combining these baseline and dynamic dimensions into nine distinct environmental semantic states detailed in Table 1, the proposed framework provides the Physics-Informed Model with a robust, physics-aware contextual input for determining spatial fusion weights.

3.4. Individual model comparison results

The quantitative results of the deterministic ablation study are summarized in Table S1 in the supplementary material, which presents the final validation performance of each architecture configuration. The analysis reveals a compelling, non-linear trend as model complexity increases, highlighting the distinct contributions of data-driven, spatial, and physical components. Our Baseline architecture established a robust performance floor with an R^2 of 0.768 and RMSE of 0.057. Interestingly, the addition of the Cross-Modal Attention module alone resulted in a slight performance regression of −1%. This suggests that without semantic guidance or physical constraints, the direct fusion of interpolated satellite features introduces high-frequency noise or spatial ambiguities that can distract the model from reliable ground-based temporal patterns. This finding underscores the necessity of the advanced guidance mechanisms proposed in this study.

Explicitly modelling the non-Euclidean spatial relationships via the GNN Spatial Weights module recovered performance and provided a solid gain, reaching an R^2 of 0.79. This confirms that capturing topological dependencies, including inter-station connectivity, is critical for regional monitoring. However, the most significant breakthrough was achieved by the Temporal + Cluster guided configuration. This model attained the highest predictive accuracy with an R^2 of 0.82 and the lowest RMSE of 0.0501, representing a substantial 7% improvement over the baseline. This result provides decisive evidence that the Unified Regional Clustering strategy is essential. By using interpretable environmental semantics to dynamically weight the fusion process, the model effectively filters out the noise observed in the pure Cross-Modal variant, adaptively prioritizing satellite data only when the environmental context.

The Physics-Informed model achieved an R^2 of 0.78, yielding a modest but positive improvement (+1%) over the baseline. While its statistical gain is smaller than the Cluster-guided model, its value lies in regularization. Unlike pure data-driven approaches, this configuration ensures that predictions remain consistent with mass balance and advection-diffusion laws, effectively preventing anti-physical prediction

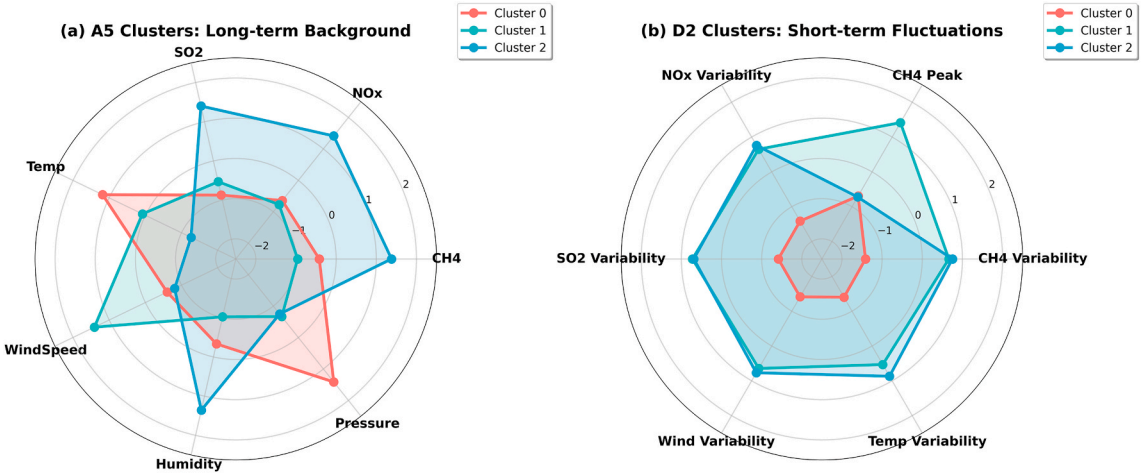

Fig. 4. The Radar Charts Characterizing the Unified Environmental Semantic Clusters. The charts visualize the distinct physical fingerprints of the identified clusters based on standardized feature values.

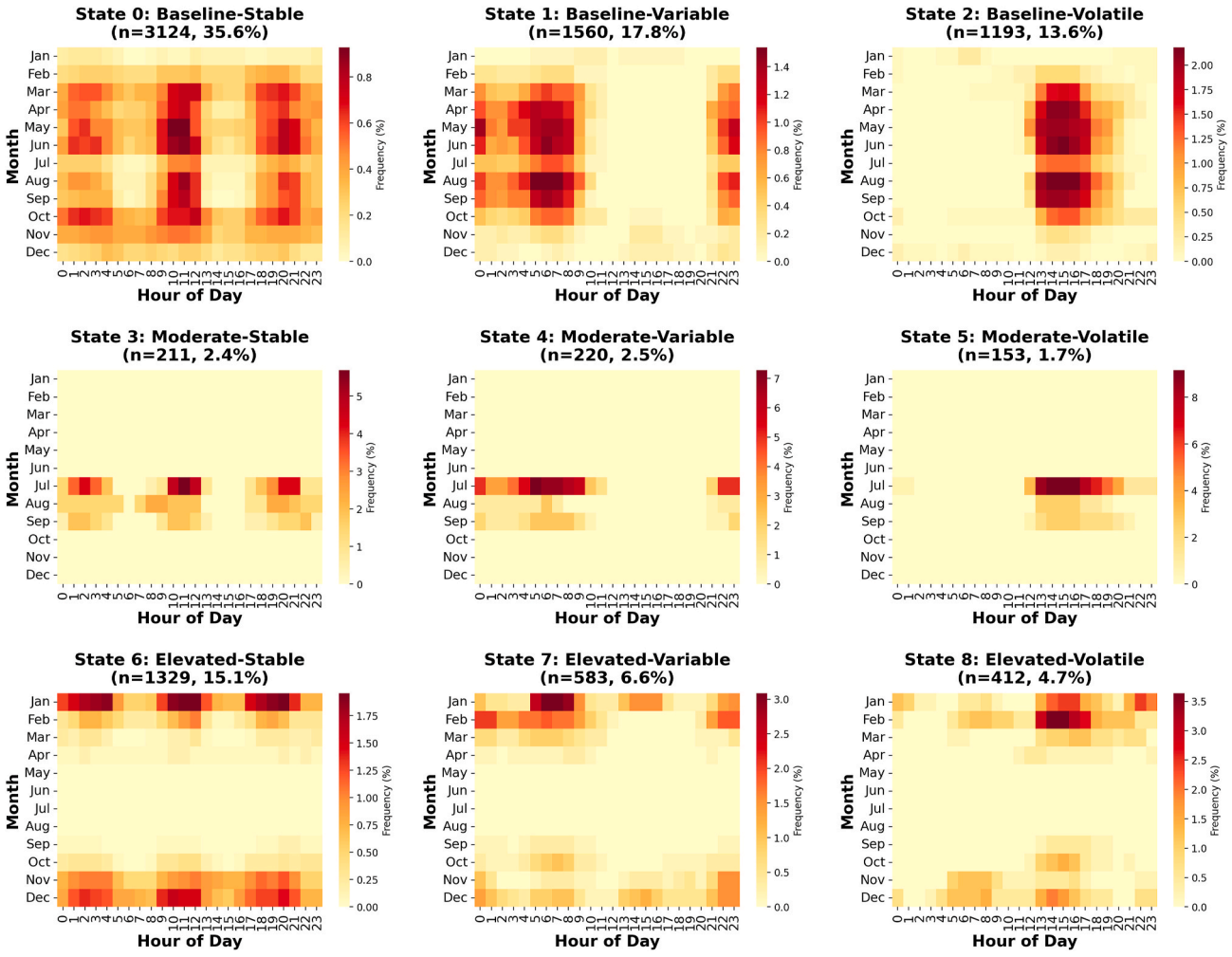

Fig. 5. Temporal Fingerprint Heatmaps of the Nine Derived Environmental Semantic States. The charts visualize the distinct physical fingerprints of the identified clusters based on standardized feature values. It visualizes the occurrence frequency of each semantic state across months (y-axis, 1–12) and hours of the day (x-axis, 0–23). Darker colours indicate higher frequency.

while slightly enhancing generalization capability.

Finally, the Complete Model, which simultaneously activates all constraints, exhibited a performance drop to $R^2 = 0.73$ (−5.2%). This indicates a multi-objective optimization conflict, where the gradient landscape becomes overly complex when attempting to simultaneously satisfy clustering consistency, spatial graph topology, and PDE residuals. In terms of computational efficiency, the Physics-Informed model offered an optimal trade-off, providing physical consistency and improved accuracy (+1%) with only a marginal increase in training time (4.8 min vs. 4.0 min), whereas the high-performing Temporal + Cluster model required significantly more computational resources (13.8 min). It is important to note that the modest +1% improvement in R^2 should not be interpreted as a limitation of the physics-informed framework. R^2, as a globally averaged point-prediction metric, is insensitive to spatial field structure and cannot capture the physically consistent concentration gradients enforced by the PDE constraints. The primary contribution of this configuration is therefore not scalar accuracy improvement, but the restoration of physically meaningful advection-diffusion spatial fields, a capability that has no analogue in purely data-driven approaches and is essential for downstream applications such as source attribution and regional concentration mapping. This spatial contribution is quantitatively evaluated in Section 3.5 and Table S3.

3.5. Spatial realism and transport consistency

To examine the models' spatial generalization capabilities in regions between sparse sensors, Fig. 6 compares the spatial methane concentration fields predicted by the data-driven Baseline Model and the physics-informed configuration. This visualization evaluates a specific, real-world transport-dominated event extracted from the test set with the timestamp of 2024-11-12 06:00:00 UTC, characterized by elevated emission levels and a prevailing northwest wind (315° at 3.0 m/s). To ensure a rigorous and unbiased comparison, both spatial fields were generated from the 15 Tier-1 station predictions using identical linear interpolation and Gaussian smoothing ($\sigma = 1.5$), anchored to the true observational background mean. The overlaid markers represent the identical, real-world ground truth observations at that exact timestamp, with empty circles denoting stations experiencing temporary sensor outages or data gaps. A detailed quantitative comparison of the spatial variance and concentration ranges during this specific event is provided in Table S3 in Supplementary Material.

The purely data-driven Baseline Model shown in Fig. 6a exhibits a fundamental limitation of statistical models when extrapolating across unsampled spatial coordinates: without physical transport constraints, the model collapses all Tier-1 station predictions toward the global temporal mean with around 2.11 ppm, producing a spatially homogeneous background field that fails to resolve local concentration gradients or emission-driven heterogeneity. In contrast, the physics-informed

Table 1
Definition and environmental interpretation of the nine semantic cluster combinations.

State ID	Semantic Definition	Frequency	Physical Interpretation
State 0	Baseline-Stable	35.60%	**Clean Background:** Typical clear days with minimal industrial impact. Observed year-round with a slight Spring/Fall dominance.
State 1	Baseline-Variable	17.80%	**Minor Disturbance:** Low background pollution with moderate meteorological fluctuations.
State 2	Baseline-Volatile	13.60%	**Transient Plume:** Low background interrupted by sudden passing emission spikes, showing a distinct afternoon peak (15:00).
State 3	Moderate-Stable	2.40%	**Summer Accumulation:** Rare transition state, strongly concentrated in July (47-69%), indicating seasonal photochemical stagnation.
State 4	Moderate-Variable	2.50%	**Active Transition:** Changing weather patterns, mixing moderate levels of pollution, predominantly in summer.
State 5	Moderate-Volatile	1.70%	**Turbulent Dispersion:** Moderate pollution driven by strong turbulent transport during summer months.
State 6	Elevated-Stable	15.10%	**Winter Inversion (Persistent Stagnation):** High accumulation of pollutants trapped by a stable atmosphere. Strictly winter-dominant (68.2% in Dec-Feb) with peaks in early morning/night, consistent with nocturnal boundary layer effects.
State 7	Elevated-Variable	6.60%	**Active Transport:** High pollution plume arriving with changing wind conditions.
State 8	Elevated-Volatile	4.70%	**Critical Operational Event:** Intense episodes of high concentration and rapid fluctuation. Peaks sharply at 15:00, correlating with daytime industrial activity and convective mixing.

model successfully reconstructs the region's spatial heterogeneity and localized emission plumes. Guided by the partial differential equation (PDE) constraints embedded in its loss function, the model captures localized diffusion processes occurring around active emission sites, generating realistic concentration gradients that disperse outward in a physically consistent manner. Notably, the localized high-concentration region predicted in the southern oil sands area, which is near Anzac station, AMS 23, is physically consistent with known emission sources in that region, including the Surmont and Long Lake in-situ facilities, demonstrating the model's ability to resolve sub-regional spatial heterogeneity that the baseline cannot capture.

The difference map as shown in Fig. 6c quantifies the spatial correction introduced by the physics-informed constraints. Positive values indicate areas where the PINN model predicts higher concentrations consistent with emission plume transport from active source regions. Minor negative corrections reflect the model's conservative adjustment in areas distant from major emission sources, consistent with the atmospheric background concentration. This spatial correction pattern highlights the regularization effect of the PDE loss: it restores physically meaningful advection-diffusion gradients that the purely statistical baseline model fails to resolve.

3.6. Operational robustness and uncertainty quantification

While visual inspection suggests the physics-informed model produces coherent fields, we employed quantitative diagnostics to verify adherence to atmospheric constraints. First, we assessed spatial realism by analyzing the roughness of the predicted fields relative to observations shown in Fig. S4 in the supplementary material. The predicted field's spatial roughness, with a mean of 0.0452 ppm^2 is lower than the observational baseline of 0.0701 ppm^2, yielding a roughness ratio of 0.644. This reduction indicates that the physics-informed constraints effectively filter high-frequency observational noise while retaining necessary structural complexity.

Beyond static structure, the model demonstrates strong alignment with dynamic transport mechanisms, as shown in Fig. S5. In a wind-aligned consistency test involving 1307 test samples, we evaluated whether predicted concentration gradients align with observed wind vectors. Under eligible transport conditions, the predicted downwind-minus-upwind concentration gradients (Δ_{pred}) correlate strongly with observations (Spearman $\rho = 0.817$) shown in Fig. S5a. The model correctly identifies the direction of plume propagation in 85.3% of cases and statistically favours positive downwind gradients. These results confirm that the PINN module successfully encodes advection-diffusion principles, ensuring that spatial allocation is driven by wind dynamics rather than mere proximity interpolation.

3.7. Station dropout resilience and predictive uncertainty calibration

Reliability under network stress is a precondition for deployment in the harsh environment of the oil sands. We evaluated the framework's resilience through a station dropout stress test, as shown in Fig. 7. The results demonstrate remarkable stability: even when 50% of the monitoring stations are randomly masked, the model's performance remains robust, with the RMSE increasing by only ~13% (from 0.0549 to 0.0620 ppm) and the R^2 remaining above 0.80. The performance degradation is approximately linear (RMSE slope = +0.0036 ppm per 10% dropout), suggesting that the model effectively leverages the learned spatial correlations to reconstruct missing nodes, minimizing the impact of individual sensor failures. Furthermore, we assessed the trustworthiness of the model's predictions with uncertainty quantification. While the raw deep ensemble initially exhibited overconfidence (under-dispersion), the application of a post-hoc scaling calibration of 9.4 significantly restored probabilistic validity. The calibrated 95% prediction intervals achieved a coverage probability of 0.916 on the test set, closely matching the nominal confidence level, with a physically meaningful mean interval width of pm ±215 ppb. This calibrated uncertainty enables operators to distinguish between confident predictions in well-monitored conditions and high-uncertainty estimates during complex transport events, facilitating risk-informed decision-making.

3.8. Fidelity in extreme emission events

Detecting acute emission spikes is critical for environmental compliance and leak detection. A stratified analysis of the test set, as shown in Fig. 8, revealed the inherent challenge of modelling rare extremes. For high-concentration events (>95th percentile of 2.401 ppm), the model initially exhibited a conservative bias, with RMSE increasing by a factor of roughly 6.8 compared to normal conditions. This attenuation is a known trade-off in stability-oriented loss functions. However, our diagnostic analysis confirms that the signal is recoverable. By applying a monotonic isotonic correction derived strictly from the validation set, we observed a 55.5% improvement in the amplitude recovery slope for extreme anomalies (P95–P99 range). This correction reduced the maximum negative bias by 33% (from −161.9 ppb to −108.1 ppb) without compromising the model's baseline performance. This finding indicates that while the physics-informed backbone prioritizes stability, it captures the timing and location of extreme events accurately enough that, with minimal post-processing, it can also recover their magnitude, serving as an effective alert system for fugitive emissions.

4. Conclusion and future research

This study proposed a physics-informed spatiotemporal modelling framework designed to bridge the gap between data-driven

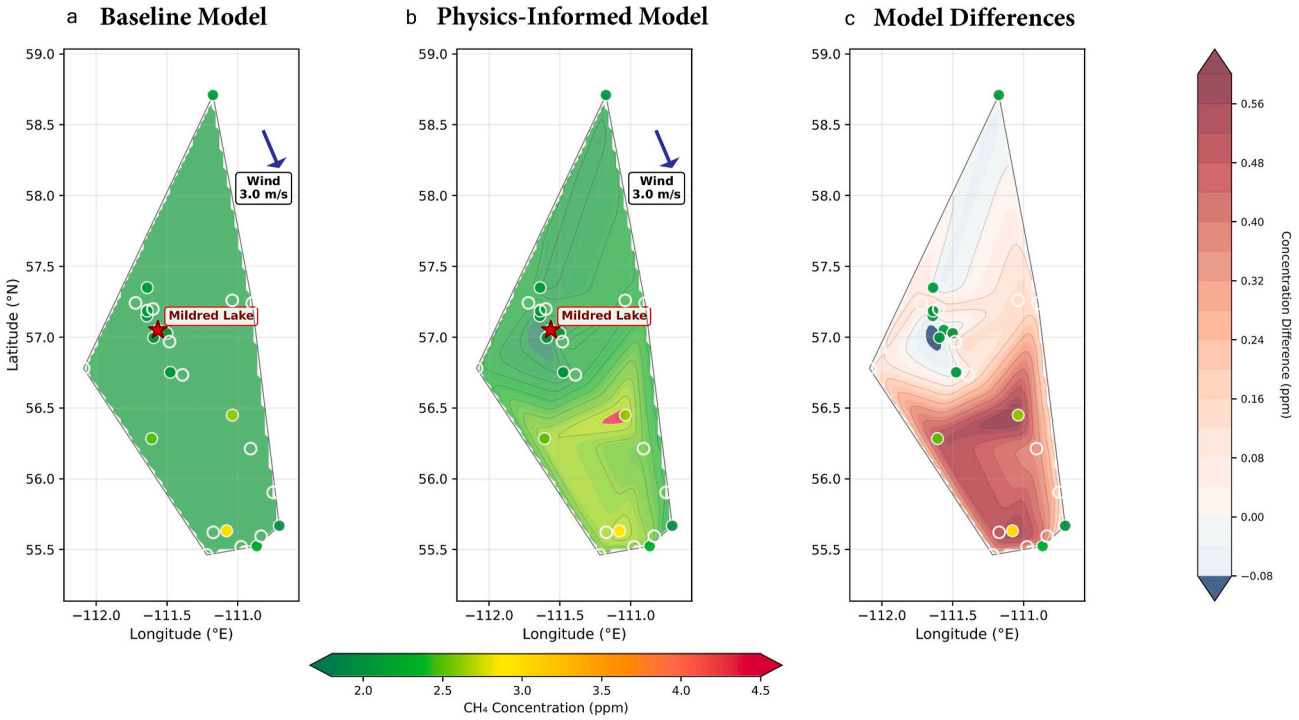

Fig. 6. Spatial Methane Concentration Fields for a Real-World Transport-Dominated Event Extracted from the Test Set. (a) Spatial concentration field predicted by the Baseline Model (corresponding to the "Baseline" entry in Table S1, $R^2 = 0.77$), which collapses toward a spatially uniform field due to the absence of physical transport constraints. (b) Spatial concentration field predicted by the Physics-Informed Model (corresponding to the "Physics-Informed" entry in Table S1, $R^2 = 0.78$), which incorporates PDE-based advection-diffusion constraints and reconstructs spatially heterogeneous concentration gradients consistent with the emission geography of the Athabasca Oil Sands region. The localized high-concentration region in the southern area near AMS 23, reflects emission contributions from the Surmont and Long Lake in-situ facilities. (c) Difference map showing the magnitude and spatial distribution of model corrections, with red indicating regions where the Physics-Informed Model predicts higher concentrations consistent with emission plume transport (up to +0.56 ppm), and blue indicating conservative adjustments in areas distant from major source regions (up to −0.08 ppm). Across all panels, solid markers represent identical ground truth observations from the 15 WBEA Tier-1 stations at the selected timestamp; empty circles denote stations with missing data. Continuous spatial fields were generated using linear interpolation anchored to Tier-1 station predictions, with identical Gaussian smoothing applied to both models ($\sigma = 1.5$).

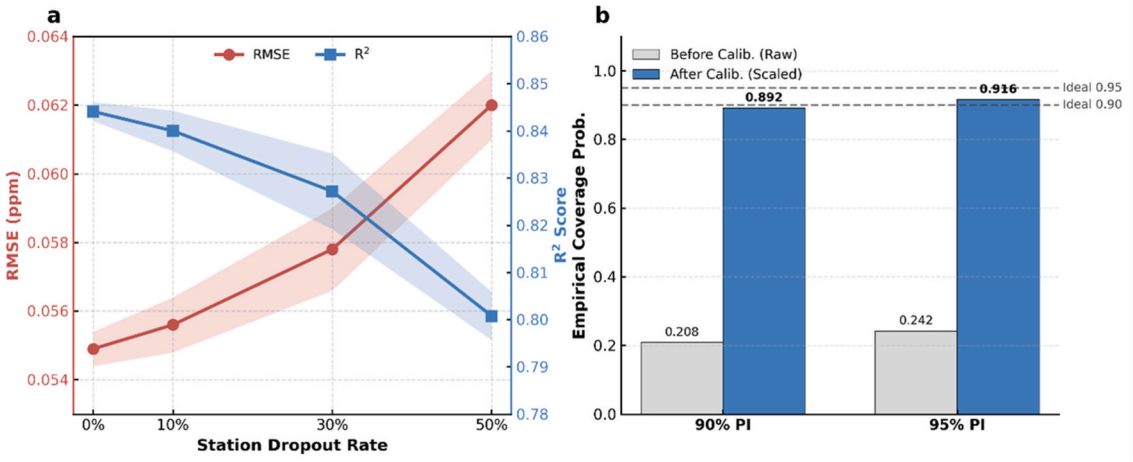

Fig. 7. Operational Reliability Assessment. (a) Robustness to Station Dropout: The degradation of model performance (Left Axis: RMSE, red; Right Axis: R^2, blue) as a function of station unavailability. Shaded bands denote the standard deviation across 20 independent trials. The linear degradation trend demonstrates the model's structural resilience even under high data loss scenarios. **(b) Uncertainty Calibration:** Comparison of prediction interval coverage probability (PICP) before (grey) and after (blue) post-hoc calibration. The calibrated model achieves near-nominal coverage (0.892 for 90% PI and 0.916 for 95% PI), ensuring trustworthy risk estimates.

interpolation and physical dispersion modelling for regional methane monitoring. By integrating ground-based observations with satellite-derived background context and transport-aware regularization, the framework addresses the critical challenge of estimating methane emissions in complex, heterogeneously monitored industrial environments like the Athabasca oil sands. Our comprehensive diagnostic evaluation yields three key scientific implications for environmental monitoring. First, physical constraints are essential for spatial realism.

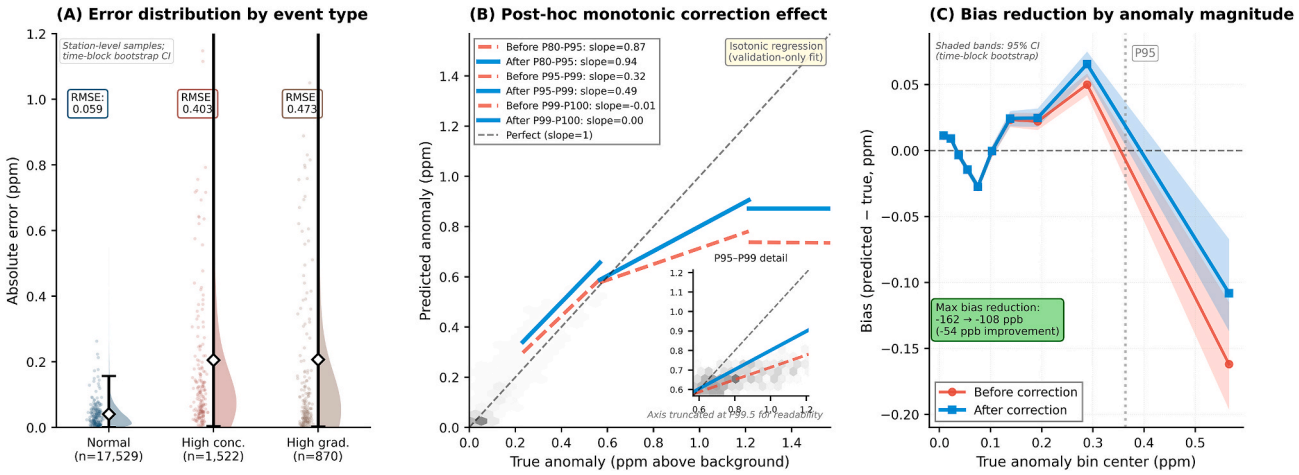

Fig. 8. Diagnostic Evaluation of Fidelity in Extreme Emission Events. (A) Error stratification by event type: Comparison of absolute prediction errors across normal conditions (n = 17,529) versus high-concentration (>P_{95}) and high-gradient events. The significant increase in RMSE (from 0.059 to 0.403 ppm) highlights the inherent challenge of stability-oriented optimization in capturing rare peaks. (B) Amplitude preservation and correction: The relationship between predicted and true anomalies relative to the background. The pre-correction model (dashed red lines) exhibits conservative attenuation (slope <1.0) in the tail. The post-hoc isotonic correction (solid blue lines), fitted strictly on validation data, improves the recovery slope in the critical P95–P99 regime by 55.5%, effectively restoring peak magnitudes. (C) Systematic bias reduction: Bias profiles conditioned on anomaly magnitude. The correction significantly mitigates the negative bias in the extreme tail (reducing maximum underestimation from −162 to −108 ppb) while preserving the unbiased baseline performance for smaller anomalies (left side). Shaded bands indicate 95% confidence intervals via time-block bootstrapping.

Quantitative diagnostics confirm that embedding advection-diffusion principles significantly reduces unphysical noise (spatial roughness ratio of 0.64) and ensures that reconstructed concentration gradients align with wind-driven transport dynamics ($\rho > 0.81$). This confirms that the model does not merely fit data points but effectively encodes the region's atmospheric dispersion mechanisms. Second, the system demonstrates operational resilience. Stress tests reveal that the framework maintains robust performance with $R^2 > 0.8$ even under severe sensor network decay (50% station dropout), while calibrated uncertainty estimates provide operators with trustworthy confidence intervals crucial for risk-informed decision-making. Third, we identified and mitigated a fundamental trade-off in measuring extreme events. While stability-oriented loss functions inherently attenuate rare, high-concentration spikes, our findings demonstrate that these signals are recoverable. The application of a monotonic diagnostic correction successfully restores the magnitude of extreme emission events by over 55%, proving that physics-informed stability need not come at the expense of sensitivity to extreme concentration enhancements. The primary contribution of this work lies in demonstrating that interpretable, physics-grounded AI can serve as a reliable surrogate for continuous regional monitoring. Rather than replacing sparse physical measurements, the proposed framework acts as a physically consistent 'soft sensor' surrogate for regional monitoring that fills spatiotemporal gaps with physically consistent estimates. Future research should address several remaining challenges to facilitate full operational deployment. (1) High-Resolution Spatial Validation: While our transport consistency was validated against point sensors, the integration of emerging high-resolution methane imagery, for example, from GHGSat, MethaneSAT, or airborne campaigns, will allow for pixel-level validation of the reconstructed plumes. (2) Near-Real-Time Inference for Operational Monitoring: Transitioning from retrospective analysis to real-time inference on edge computing devices would enable immediate alert generation for fugitive leaks. (3) Multi-Pollutant Generalization: Extending the transport-consistent regularization to co-emitted pollutants could further constrain the source attribution process, creating a holistic "digital twin" of the regional atmospheric environment.

CRediT authorship contribution statement

Yang Xu: Writing – review & editing, Writing – original draft, Visualization, Methodology, Conceptualization. **Hao Wang:** Writing – review & editing, Validation, Supervision, Funding acquisition. **Jude D. Kong:** Writing – review & editing, Supervision, Methodology, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) through an NSERC Alliance Missions grant (577242) "Predicting Methane Emissions from Alberta Oil Sands Territories Using a Holistic Model and Monitoring System" (HW as the PI and JDK as a co-PI). HW acknowledges support from the Natural Sciences and Engineering Research Council of Canada (Individual Discovery Grant RGPIN-2020-03911 and Discovery Accelerator Supplement Award RGPAS-2020-00090) and the Canada Research Chairs program (Tier 1 Canada Research Chair Award). JDK acknowledges support from Canada's International Development Research Centre (IDRC) (Grant No. 109981-001), IDRC and the Foreign, Commonwealth and Development Office (FCDO) (Grant No. 110554-001), NSERC Discovery Grant (Grant No. RGPIN-2022-04559), NSERC Discovery Launch Supplement (Grant No: DGECR-2022-00454), New Frontier in Research Fund- Exploratory (Grant No. NFRFE-2021-00879) and Canada Research Chair in Community-Oriented Artificial Intelligence and Mathematical Modelling of Infectious Diseases (Award ID: CRC-2023-00234.)

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jenvman.2026.129618.

Data availability

The Zenodo record associated with this study is publicly accessible under DOI: https://doi.org/10.5281/zenodo.18063417. Due to data size and third-party usage constraints, the datasets, model code, and analysis scripts hosted in this record are provided under controlled access. Access can be granted to qualified researchers upon reasonable request through the Zenodo platform. Ground-based methane monitoring data are originally sourced from the Wood Buffalo Environmental Association (WBEA) and are publicly available through the WBEA open data portal https://wbea.org/data/network-map-station-data/. Satellite-derived methane background data are processed from the GHGSat SPECTRA platform and are available to registered users via the SPECTRA Basic service https://spectra-basic.ghgsat.com/.

References

Alamu, R., Karkala, S., Hossain, S., Krishnapatnam, M., Aggarwal, A., Zahir, Z., Pandhare, H.V., Shah, V., 2025. Physics-Informed Neural Networks for Climate Modeling: Bridging Machine Learning and Physical Laws.

BOE Report, 2025. Energy projects occupy less than three per cent of Alberta's oil sands region, a report says. https://boereport.com/2025/05/05/energy-projects-occupy-less-than-three-per-cent-of-albertas-oil-sands-region-report-says/.

Brantley, H.L., Thoma, E.D., Squier, W.C., Guven, B.B., Lyon, D., 2014. Assessment of methane emissions from oil and gas production pads using Mobile measurements. Environ. Sci. Technol. 48 (24), 14508–14515. https://doi.org/10.1021/es503070q.

Cuomo, S., Di Cola, V.S., Giampaolo, F., Rozza, G., Raissi, M., Piccialli, F., 2022. Scientific machine learning through physics-informed neural networks: where we are and what's next. J. Sci. Comput. 92 (3), 88. https://doi.org/10.1007/s10915-022-01939-z.

Daw, A., Karpatne, A., Watkins, W.D., Read, J.S., Kumar, V., 2022. Physics-guided neural networks (PGNN): an application in Lake temperature modeling. In: Knowledge Guided Machine Learning. Chapman and Hall/CRC, pp. 353–372.

Dumont Le Brazidec, J., Vanderbecken, P., Farchi, A., Broquet, G., Kuhlmann, G., Bocquet, M., 2023. Deep learning applied to CO_2 power plant emissions quantification using simulated satellite images. Geosci. Model Dev. Discuss. (GMDD) 2023, 1–30. https://doi.org/10.5194/gmd-17-1995-2024.

Fox, T.A., Barchyn, T.E., Risk, D., Ravikumar, A.P., Hugenholtz, C.H., 2019. A review of close-range and screening technologies for mitigating fugitive methane emissions in upstream oil and gas. Environ. Res. Lett. 14 (5), 053002. https://doi.org/10.1088/1748-9326/ab0cc3.

GHGSat, 2024. Spectra BASIC. https://spectra-basic.ghgsat.com/.

Jacob, D.J., Turner, A.J., Maasakkers, J.D., Sheng, J., Sun, K., Liu, X., Chance, K., Aben, I., McKeever, J., Frankenberg, C., 2016. Satellite observations of atmospheric methane and their value for quantifying methane emissions. Atmos. Chem. Phys. 16 (22), 14371–14396. https://doi.org/10.5194/acp-16-14371-2016.

Jacob, D.J., Varon, D.J., Cusworth, D.H., Dennison, P.E., Frankenberg, C., Gautam, R., Guanter, L., Kelley, J., McKeever, J., Ott, L.E., Poulter, B., 2022. Quantifying methane emissions from the global scale down to point sources using satellite observations of atmospheric methane. Atmos. Chem. Phys. 22 (14), 9617–9646. https://doi.org/10.5194/acp-22-9617-2022.

Jiang, Y., Zhang, L., Zhang, X., Cao, X., 2024. Methane retrieval algorithms based on satellite: a review. Atmosphere 15 (4), 449. https://doi.org/10.3390/atmos15040449.

Jin, Z., He, J., Wang, W., 2024. Monitoring methane concentrations with high spatial resolution over China by using random forest model. Remote Sens. 16 (14), 2525. https://doi.org/10.3390/rs16142525.

Kong, J.D., Wang, H., Siddique, T., Foght, J., Semple, K., Burkus, Z., Lewis, M.A., 2019. Second-generation stoichiometric mathematical model to predict methane emissions from oil sands tailings. Sci. Total Environ. 694, 133645. https://doi.org/10.1016/j.scitotenv.2019.133645.

Liao, Q., Zhu, M., Wu, L., Pan, X., Tang, X., Wang, Z., 2020. Deep learning for air quality forecasts: a review. Curr. Pollut. Rep. 6 (4), 399–409. https://doi.org/10.1007/s40726-020-00159-z.

Meng, X., Chang, H., Wang, X., 2022. Methane concentration prediction method based on deep learning and classical time series analysis. Energies 15 (6), 2262. https://doi.org/10.3390/en15062262.

Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis, vol. 4. Cambridge university press. https://doi.org/10.1017/CBO9780511841040.

Raissi, M., Perdikaris, P., Karniadakis, G.E., 2019. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J. Comput. Phys. 378, 686–707. https://doi.org/10.1016/j.jcp.2018.10.045.

Saha, E., Wang, O., Chakraborty, A.K., Garcia, P.V., Milne, R., Wang, H., 2025. Dispersion based recurrent neural network model for methane monitoring in Albertan tailings ponds. J. Environ. Manag. 395, 127748. https://doi.org/10.1016/j.jenvman.2025.127748.

Sahoo, G.R., Freed, J.H., Srivastava, M., 2024. Optimal wavelet selection for signal denoising. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3377664.

Sysoeva, L., Bouderbala, I., Kent, M.H., Saha, E., Zambrano-Luna, B.A., Milne, R., Wang, H., 2025. Decoding methane concentration in Alberta oil sands: a machine learning exploration. Ecol. Indic. 170, 112835. https://doi.org/10.1016/j.ecolind.2024.112835.

Tiemann, E., Zhou, S., Kläser, A., Heidler, K., Schneider, R., Zhu, X.X., 2024. Machine learning for methane detection and quantification from space-a survey. arXiv Preprint. https://doi.org/10.48550/arXiv.2408.15122 arXiv:2408.15122.

WBEA, 2024. Ambient air monitoring station site documentation reports. Wood Buffalo Environmental Association, Fort McMurray, Alberta. Available at: https://wbea.org/resources-section/wbea-site-documents/. (Accessed 15 March 2026).

Wei, C., Jafari Raad, S.M., Hassanzadeh, H., 2023. Estimation of natural methane emissions from the largest oil sand deposits on Earth. PNAS Nexus 2 (9). https://doi.org/10.1093/pnasnexus/pgad260 pgad260.

Wesselkamp, M., Moser, N., Kalweit, M., Boedecker, J., Dormann, C.F., 2024. Process-informed neural networks: a hybrid modelling approach to improve predictive performance and inference of neural networks in ecology and beyond. Ecol. Lett. 27 (11), e70012. https://doi.org/10.1111/ele.70012.

WMO, 2024. Greenhouse gas concentrations surge again to new record in 2023. In: WMO Greenhouse Gas Bulletin No. 20. World Meteorological Organization, Geneva. Available at: https://wmo.int/news/media-centre/greenhouse-gas-concentrations-surge-again-new-record-2023.

WMO, 2025. WMO confirms 2024 as warmest year on record at about 1. In: 55°C Above Pre-industrial Level. WMO Press Release No. 01102025. World Meteorological Organization, Geneva. Available at: https://wmo.int/news/media-centre/wmo-confirms-2024-warmest-year-record-about-155degc-above-pre-industrial-level.

Xu, Y., Wang, H., Kong, J.D., 2026. Environmental semantic clustering-guided multimodal fusion for enhanced interpretability in methane concentration prediction. J. Environ. Manag. 401, 128845. https://doi.org/10.1016/j.jenvman.2026.128845.

Yazdinejad, A., Wang, H., Kong, J.D., 2025. Advanced AI-Driven methane emission detection, quantification, and localization in Canada: a hybrid multi-source fusion framework. https://doi.org/10.1016/j.scitotenv.2025.180142.

Zhu, S., Tang, J., Zhou, X., Li, P., Liu, Z., Zhang, C., Zou, Z., Li, T., Peng, C., 2023. Research progress, challenges, and prospects of $PM_{2.5}$ concentration estimation using satellite data. Environ. Rev. 31 (4), 605–631. https://doi.org/10.1139/er-2022-0125.