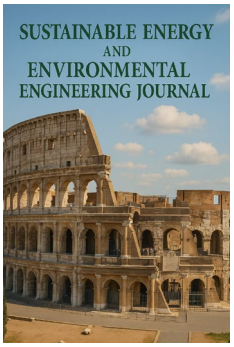
UNIVERSITY OF WEST ATTICA
eISSN:3136-8174

Sustainable Energy and Environmental Engineering Journal

Vol. 4, No. 1, January, 2026



Topology Aware Neural ODE Transformer for Hydrogen Sulfide Risk Prediction in Urban Pipe Networks

Milena Nikolić[1,*], Vasilije Šumaher[1], Marija Jovanović[1]

[1] *Faculty of Environmental Engineering, University of Niš, Niš, 18000, Serbia*

Received 16 November 2025 Revised 27 February 2026; Accepted 02 March 2026; Published 06 March 2026

Abstract

Hydrogen sulfide accumulation in urban pipe networks creates forecasting problems in which irregular observations, directional connectivity, and ventilation changes interact. This study proposes a topology aware neural ordinary differential equation and Transformer framework for concentration forecasting and exceedance risk assessment. A directed graph encoder combines verified connectivity, geometric attributes, and observation reliability. Continuous latent evolution connects measurement events, while causal temporal attention integrates operating histories available at the forecast origin. The benchmark configuration comprises 120 monitoring locations, five operating regimes, and forecast horizons of 5, 15, and 30 minutes. At the 15-minute horizon, the reported mean absolute error is 1.84 parts per million, the root mean squared error is 2.67 parts per million, and the area under the precision–recall curve is 0.914. Event recall is 0.902 at the specified advisory threshold. The separate 30-minute event head yields a median warning lead time of 18.6 minutes. Removing topology increases mean absolute error to 2.45 parts per million. Across the missing-observation sensitivity settings, error increases from 1.72 to 2.91 parts per million as missingness rises from zero to 40 percent. The reported calibration gap decreases from 0.109 to 0.019. These comparisons describe the stated benchmark conditions. The contribution lies in coordinating connectivity, continuous state evolution, and temporal context for hydrogen sulfide forecasting. Interpretation of deployment suitability requires traceable acquisition records, complete implementation settings, and prospective operational assessment.

Keywords: Topology Aware Learning, Hazardous Gas Forecasting, Neural Ordinary Differential Equations, Temporal Attention, Urban Pipe Networks

1. Introduction

Reviews of sewage odor examine microbial processes and the release of malodorous compounds in wastewater collection and treatment infrastructure [1]. Gas–liquid mass-transfer studies examine hydrogen sulfide exchange under turbulent mixing conditions [2]. Occupational monitoring studies compare measurement strategies for characterizing hydrogen sulfide exposure in water and wastewater work [3]. Sewer inspection research reviews robotic monitoring methods and the practical constraints of assessing underground infrastructure [4]. Concentration records can therefore vary between nearby access points as local geometry and ventilation conditions change. A forecasting system for these environments should represent the observation location, measurement age, and potential influence of connected monitoring sites. Research on distributed environmental sensing has investigated observation reliability in the presence of unreliable communication channels [5]. Network risk mapping methods have also considered the connection of infrastructure to the area of hazard occurrence [6]. These findings motivate a prediction task that relates concentration dynamics to the topology of the monitored infrastructure. The engineering objective is to identify locations where gas

* Corresponding author: milena.ni@ftn.ni.ac.rs

https://doi.org/10.68406/seeej.2026.vol4iss1nz5:54-67

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

concentrations may cross an operational trigger before the next inspection or ventilation adjustment. Station histories characterize local temporal behavior, while the network representation provides context for disturbances originating in connected branches. The two information sources differ in their observation requirements and in the delayed responses they can represent. Their integration is particularly relevant at junctions where changing ventilation and branch-specific accumulation produce spatially uneven conditions.

Existing forecasting methods differ in how they represent relationships between observations across space and time. Statistical time-series methods have been used in the field of environmental monitoring to incorporate seasonality and lagged covariates for regular operating patterns [7]. The predictions are less reliable in the event of a disturbance to a branch-specific transport route or a change in the sampling interval. Water-quality forecasting research combines statistical prediction with recurrent networks to model linear structure and nonlinear residual behavior [8]. Graph learning extends this representation by conditioning information exchange between locations on a connectivity structure [9]. Spatially adjacent stations can occupy separate transport pathways and exhibit different concentration histories. Continuous-time neural models have been developed to address the problem of latent trajectories with irregular observations [10]. The latent trajectory spans unobserved intervals and remains conditioned on information available at the latest observation. Attention-based sequence models can also select relevant distant observations when a short local window fails to capture the changes in operating conditions [11]. Integrating these methods requires a common representation of spatial connectivity and observation timing. A graph model with a fixed time step may provide the incorrect time context for the delayed packet. Continuous trajectories require explicit transport constraints to preserve physically plausible relationships between stations. Temporal attention can retrieve relevant historical events while remaining sensitive to degraded sensor inputs. The research gap concerns the coordination of verified physical connections, continuous state evolution, and observation reliability within an auditable prediction pipeline.

The target application is hydrogen sulfide forecasting within a connected urban wastewater and utility atmosphere monitoring network. Methane, oxygen, humidity, temperature, ventilation status, and liquid flow indicators provide contextual inputs when available. Hydrogen sulfide concentration provides a common target for regression and event detection, while auxiliary channels retain their distinct physical meanings. A concentration trigger defines the engineering event label, while a separate probability threshold governs advisory issuance. The benchmark trigger is a study setting whose suitability requires site-specific assessment. The proposed architecture represents verified pipe connections as graph support, incorporates directional attributes to distinguish potential upstream and downstream influence, and includes missingness and measurement age in the encoder. A neural ordinary differential equation describes latent state evolution between incoming observations. A temporal attention module extracts forecast information from the sequence of event states. The concentration and event probability heads share a representation and use separate training objectives. The separate objectives allow concentration accuracy and warning performance to be assessed independently, including cases where moderate concentration error accompanies an ineffective advisory policy. The mathematical specification is accompanied by an evaluation protocol covering concentration accuracy, event detection, probability reliability, and inference requirements. Separate assessment of these outcomes identifies the conditions under which the architecture is effective and the assumptions requiring further investigation.

Developing a dependable forecasting service for connected underground infrastructure requires explicit links between model assumptions, information available at each forecast origin, and engineering evaluation objectives. The first research question concerns whether restricting graph information exchange to documented connections improves prediction at poorly ventilated nodes. The second concerns the contribution of continuous state evolution when observations are missing or arrive after unequal delays. The third examines whether extended temporal context improves event prediction while preserving probability reliability. The contribution concerns the coordination of established model components under verified connectivity, unequal observation times, and causal control availability. Its scope is assessed through concentration samples, accumulation episodes, and operating regimes. Model interpretation and physical explanation require separate assessment because attention weights represent computational allocations, whereas gas flux is a physical quantity. The scientific argument comprises five sections, followed by declarations and references. Section 2 reviews topology-based environmental learning and continuous-time forecasting, focusing on research gaps relevant to underground monitoring. Section 3 introduces the graph representation, continuous-state dynamics, temporal attention and learning objectives, and the two diagrams show the flow of information and the forecasting process. Section 4 presents the benchmark configuration and analyzes comparative errors, episode trajectories, component removal, missing observations, probability calibration, and inference resource trade-offs. Concentration and event metrics are interpreted alongside the timing and operating conditions of each forecast. Section 5 states the contributions, examines the limitations of the formulation, and identifies future research directions. This organization links architectural choices to the engineering questions motivating their evaluation.

2. Related Work

2.1. Topology based learning for infrastructure atmospheres

Spatiotemporal multi-graph forecasting for electric vehicle charging illustrates how connectivity can structure information exchange [12]. Its transfer to underground atmospheres requires separate evaluation of transport pathways and sensor reliability. For underground infrastructure, adjacency can encode a documented conduit, a plausible exchange pathway, or a statistical association inferred from observations. These meanings should be distinguished during evaluation because a model that uses unrestricted similarity can exploit recurring operational coincidences that disappear after maintenance. Pipe length, junction type, ventilation openings, and flow direction offer additional context for the strength and orientation of information exchange. Estimating a gas transport coefficient requires physical information beyond graph connectivity. Learned neighborhood updates represent predictive associations whose underlying physical mechanisms require separate identification. Engineering evaluation should assess forecast performance under operating changes and examine the physical interpretability of learned graph weights separately. Sparse connections also require special treatment at isolated monitoring locations, where a residual local pathway is necessary to prevent the graph from erasing the station history.

Traffic forecasting research uses graph attention to adapt neighborhood influence to changing observations [13]. This offers a methodological analogy for gas forecasting, with application-specific validation still required. This flexibility is valuable when a ventilation intervention weakens an otherwise persistent relationship. It also creates an ambiguity between genuine changes in transport and changes in measurement quality. A station with a sharp sensor spike may attract attention even when the spike reflects drift or packet corruption. Incorporating reliability indicators into attention can reduce this ambiguity; reliability estimates are restricted to information available at the forecast origin. Traffic forecasting methods explicitly represent propagation delays when modeling spatiotemporal dependencies [14]. Such methods motivate representations that incorporate edge direction and measurement age alongside network connectivity. Their applicability still depends on whether the available topology captures ventilation connections and temporary bypasses. Topology adaptation approaches have examined learned edges as a means of accommodating incomplete network descriptions [15]. In a hazardous atmosphere application, unrestricted edge creation would complicate interpretation and allow leakage through correlations with future operating states. A conservative alternative is to learn weights within verified support and record unrepresented pathways as a source of model uncertainty.

2.2. Continuous time forecasting and calibrated warnings

Neural ordinary differential equation and Transformer components have already been combined for temporal video understanding [16]. This architectural precedent motivates an application-specific contribution based on verified gas-network support and observation reliability. This formulation separates continuous state evolution from the times at which predictions are requested. The formulation supports asynchronous monitoring by allowing station queries between observation events. The correspondence between latent smoothness and physical concentration continuity requires separate assessment. Abrupt valve closure, ventilation activation, or maintenance openings can introduce discontinuities requiring explicit observation or control updates. Continuous-time graph representation learning models jointly evolving node embeddings and graph structure [17]. These models incorporate elapsed time into inference while preserving actual observation intervals in the event representation. Event-driven assimilation distinguishes newly received measurements from latent estimates propagated during communication gaps.

Transformer models employ attention to aggregate information in sequences via learned compatibility weights, and thus can model the dependencies between observations at different times [18]. In infrastructure monitoring, attention can retrieve historical information about operating transitions or ventilation cycles beyond the latest local window. Its computational cost and sensitivity to shortcut variables require careful evaluation. Probability calibration research examines the distinction between predictive ranking and numerical confidence, motivating joint evaluation of event detection and reliability [19]. For accumulation warnings, systematic probability overestimation can produce excessive advisories even when event ranking is accurate. Calibration should be evaluated alongside precision, recall, and episode warning lead time. The proposed framework combines topology-constrained encoding, continuous hidden evolution, and causal temporal attention with a separate probability calibration stage. The framework remains sensitive to topology errors, unobserved discontinuities, and changes in event prevalence that can weaken probability calibration.

3. Topology Aware Continuous Time Forecasting Model

3.1. Directed network representation and observation encoding

The monitored infrastructure is represented as a directed graph whose nodes correspond to instrumented chambers or conduit locations. Edges indicate verified connections capable of supporting atmospheric exchange. Each edge stores conduit length, effective diameter, orientation, and a ventilation connectivity flag when those attributes are available.

Edge direction represents a time dependent predictor of atmospheric exchange and can change with the operating context. Reversible pathways are represented by two directed edges with separately updated contextual attributes. Observations combine the primary gas concentration with auxiliary channels, an observation mask, elapsed time since the last valid reading, and recorded operating controls. Continuous variables are standardized using training partition statistics only. The topology remains fixed within a forecast window; documented changes create a new graph version at the next observation event.

Figure 1 presents the end-to-end forecasting architecture as a single flow diagram. It connects timestamped sensor packets and verified pipe attributes to reliability aware graph encoding, continuous latent evolution, causal temporal attention, and separate concentration and exceedance outputs. The operational trigger enters only the event label and forecast interpretation pathways. A distinct validation branch fits the probability calibrator before deployment.

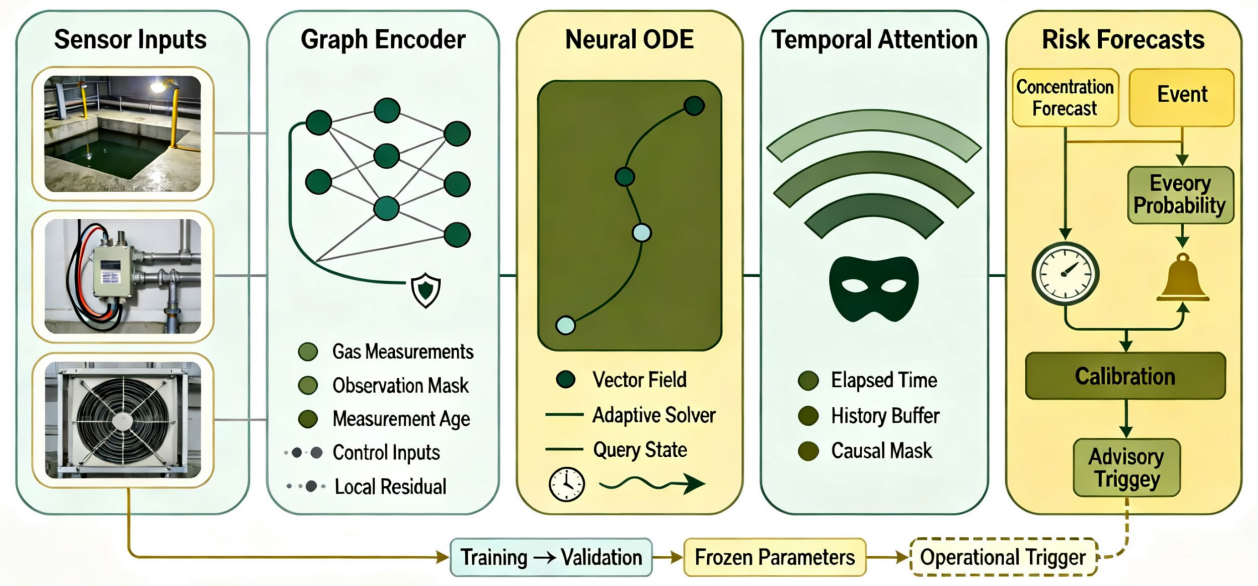


Figure 1. Information flow in the topology aware neural ODE Transformer architecture

For node i and observation event k, the encoder first constructs an input vector x_i from the standardized measurement channels and quality indicators. A learned edge descriptor e_{ji} supplies attributes of the supported connection from neighboring node j. The neighborhood set $\mathcal{N}_i$ includes only verified incoming edges. A scalar edge score combines the two node descriptors, the edge attributes, and elapsed time. Normalization is restricted to that neighborhood, while a separate residual branch preserves the local station information. The encoded event state follows the neighborhood aggregation equation below.

$$z_i(t_k) = \sigma\left(W_0 x_i + \sum_{j \in \mathcal{N}_i} \alpha_{ji} W_1 x_j\right) \tag{1}$$

The state z_i has a common latent dimension across stations, and W_0 and W_1 are trainable projection matrices. The coefficients α_{ji} are nonnegative normalized edge weights, while σ denotes the activation function. Self information enters through the first term and is retained even when the incoming neighborhood is empty. Reliability masks adjust inputs and edge scores while retaining the station representation when an auxiliary channel is missing.

3.2. Continuous latent evolution and causal temporal attention

An integral equation describes the continuous evolution of the latent state between observation events. Its neural vector field is conditioned on the verified graph and known control inputs. The full network state z contains all node states, enabling supported neighbor interactions inside the vector field. Controls u describe recorded ventilation and maintenance status, with changes applied at their known activation times. The parameters θ define the trainable vector field. Integrating from the latest assimilated event time t_k to the requested query time t incorporates unequal elapsed intervals directly into the state update.

$$z(t) = z(t_k) + \int_{t_k}^{t} f_\theta(z(s), G, u(s))ds \tag{2}$$

An adaptive solver uses validation-selected relative and absolute tolerances. The solver identity, numerical tolerances, and gradient implementation are required for reproducibility; their numerical settings are unspecified here. A control change is represented by an explicit split in the integration interval. Observation availability is determined by receipt time. Acquisition time describes measurement age; processing an older packet requires a documented late-arrival policy that preserves the information set at each forecast origin. The reliability gate blends the available graph encoded observation with the propagated state. A valid, recently calibrated gas reading receives a different gate input from a stale packet or a missing channel. Solver accuracy and forecasting accuracy are evaluated separately: the former concerns numerical integration, while the latter depends on how well the learned dynamics represent concentration evolution.

Figure 2 describes the event driven forecasting procedure in one flow diagram. Each incoming packet passes timestamp and quality checks before the supported graph update. The state is integrated to the packet time, assimilated, appended to a causal history buffer, and queried at requested forecast horizons. The buffer includes elapsed time encodings and masks for unavailable stations. A training branch computes concentration and event losses, whereas a deployment branch applies the frozen calibrator and stores the advisory probability.

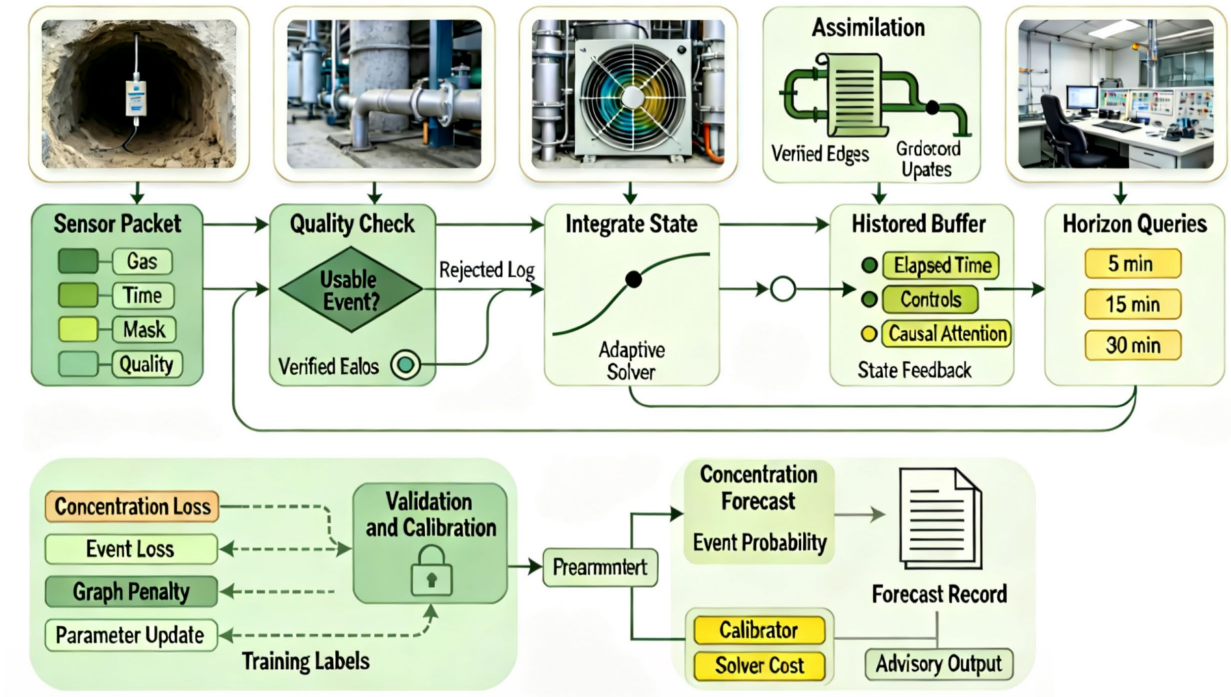


Figure 2. Event driven training and inference procedure

Temporal attention operates on the sequence of assimilated event states for each station, enriched by elapsed time and control embeddings. Query, key, and value matrices Q, K, and V are formed from that sequence. The key dimension d scales the compatibility scores. The mask M assigns an excluded score to positions occurring after the forecast origin or lacking a usable state. The attention output H aggregates the permitted historical information.

$$H = softmax\left(\frac{QK^{\top}}{\sqrt{d}} + M\right)V \tag{3}$$

In this attention equation, allowed positions receive a zero additive mask and excluded positions receive negative infinity before normalization. A fully unavailable history requires an explicit local-state fallback. Event labels remain supervision targets, while receipt-time availability determines access to observations. A bounded history buffer controls sequence cost and prevents a very long record from dominating recent event information. Separate attention heads can emphasize local accumulation, ventilation transitions, or longer recurring patterns, with their weights serving as computational diagnostics. Elapsed time encodes temporal position according to the actual duration between observation events. This distinction matters when two adjacent packets are separated by different durations. At a new forecast origin, the temporal representation is combined with the integrated query state for each requested horizon.

3.3. Forecast outputs and learning objectives

The output equation defines separate concentration and event prediction heads using the shared temporal representation H. A smooth positive activation produces a nonnegative concentration forecast, while a logistic transformation produces the raw probability of an exceedance within the specified horizon. For horizon τ, the parameters a, b, w, and c control the two projections.

$$\hat{y}_\tau = softplus(a_\tau^\top H + b_\tau), \hat{p}_\tau = sigmoid(w_\tau^\top H + c_\tau) \quad (4)$$

The concentration output $\hat{y}$ is expressed directly in parts per million; only the input channels are standardized. The event output $\hat{p}$ is subsequently transformed by the validation fitted probability calibrator. A horizon embedding distinguishes the requested future forecast lead times; interpolation concerns historical gaps between available observations. The event probability and thresholded point prediction characterize distinct aspects of risk: predictive uncertainty can imply a plausible exceedance while the point estimate remains below the concentration trigger. Event weights address scarcity within training data only; probability calibration subsequently uses an unweighted validation sample with its natural prevalence.

The supervised objective combines a robust concentration loss L_c with a binary event loss L_e. Robust regression limits the influence of isolated sensor spikes while retaining the contribution of sustained high concentration episodes. A graph penalty L_g discourages excessive latent differences across weakly separated, verified connections while respecting directional edge weights. Its role is a modeling prior that permits concentration differences at adjacent nodes.

$$L = L_c + \lambda_e L_e + \lambda_g L_g \quad (5)$$

This loss equation balances concentration accuracy, event prediction, and spatial regularization through the nonnegative weights λ_e and λ_g, which are selected using validation performance. The graph penalty compares latent states aligned to common query times, accounting for unequal observation ages. A normalized edge weight ω_{ji} reflects available connection attributes and excludes nonexistent edges. The graph regularization equation averages the weighted state differences over the verified directed edge set E.

$$L_g = \frac{1}{|E|} \sum_{(j,i)\subset E} \omega_{ji} \left\| z_i - z_j \right\|_2^2 \quad (6)$$

The squared-distance graph penalty is a latent smoothness prior. Its pairwise discrepancy is symmetric even when edge weights are directed; transport orientation enters the graph encoder and vector field. Strong regularization can suppress localized accumulation near stagnant branches. Training stops according to a predefined validation criterion, and model selection is restricted to the training and validation partitions. Equal tuning budgets and documented solver settings are evaluation requirements. Architectural dimensions, optimization settings, stopping rules, and tuning counts remain necessary to reproduce the numerical comparisons. Forecast records retain the observation age, graph version, query horizon, and calibrated probability so that concentration errors and warning behavior can be traced to a common prediction origin.

4. Benchmark Evaluation and Performance Analysis

4.1. Benchmark design and evaluation protocol

The benchmark configuration comprises 120 monitoring locations and 168 directed connections. Independent verification requires acquisition dates, network identity or an anonymized descriptor, sensor specifications, and quality-control records; these metadata are outside the reported benchmark description. The evaluation schedule spans 90 days at a nominal five-minute interval, corresponding to 25,920 timestamp positions and 3,110,400 potential node observations before missingness. Five operating regimes distinguish routine flow, rainfall intrusion, weak ventilation, maintenance intervention, and communication outage. Their episode counts are 96, 72, 84, 60, and 48, respectively. The concentration trigger is 10 parts per million, while an exceedance probability above 0.50 defines the advisory rule. These thresholds distinguish concentration conditions from forecast confidence. The event probability concerns an exceedance inside the future query horizon. A crossing-based lead time requires the first transition from below to at least the trigger. Records already above the trigger require separate persistence labels and a documented episode-reset rule. Blocked temporal evaluation addresses dependence between nearby environmental observations by separating contiguous periods [20]. The chronological 60/20/20 allocation corresponds to 15,552 training positions and 5,184 positions in each subsequent partition, before boundary exclusions: 54, 18, and 18 days, respectively. Windows whose target horizons extend beyond a partition boundary are excluded. Normalization statistics and graph attributes are restricted to information available before the corresponding forecast origin. This chronology limits leakage from future measurements and keeps the warning task aligned with the information that a deployed service would possess.

Figure 3(a) describes the event concentration distribution using six bins with edges at 0, 2, 5, 10, 15, 20, and 25 parts per million; their counts are 90, 96, 54, 60, 42, and 18. The three bins above the trigger contain 120 of the 360 event samples, whereas the two lowest bins contain 186. These 360 summarized samples differ in scope from the full node–time record. Their proportions describe the displayed distribution; estimating test-event prevalence requires the selection rule and test-set denominator. Figure 3(b) presents observation delays across the five operating regimes, with medians of 1.1, 2.8, 3.6, 4.9, and 6.5 minutes and 72 observations per regime. The outage distribution has the greatest median

delay and a longer upper tail, indicating that observation age varies systematically with operating conditions. A delayed packet can describe an earlier accumulation stage even when its concentration value remains within the valid sensor range. Delay and concentration therefore need separate quality indicators. The delay distributions are displayed over a 15-minute range, and the missingness analysis covers 0, 10, 20, 30, and 40 percent of nominal observation positions. The regime comparison helps distinguish a model that tolerates isolated missing values from one that can remain informative during sustained communication degradation.

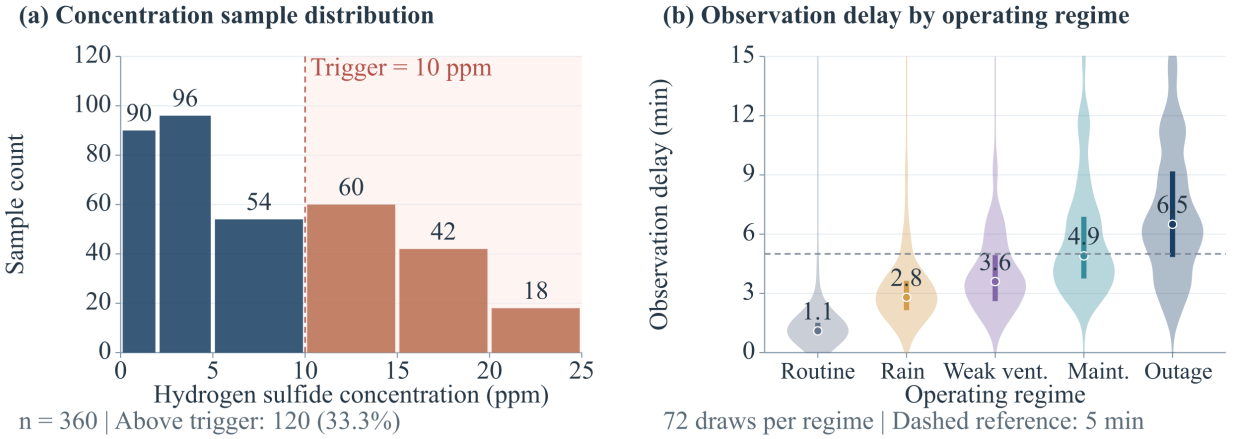


Figure 3. Concentration composition and observation timing of the benchmark. (a) Histogram of 360 event concentration samples with a trigger at 10 ppm. (b) Observation delay distributions in five operating regimes, with median labels and 72 observations per regime

Forecast horizons are 5, 15, and 30 minutes, with a 120-minute causal history specified for every comparator. The baseline families comprise an autoregressive model, a long short term memory network, a temporal Transformer, and a graph Transformer. The proposed architecture combines spatial and temporal representation with continuous state evolution and reliability aware encoding. Fair comparison protocols emphasize equal access to information and comparable optimization effort across architectures [21]. The evaluation design therefore specifies identical channel availability, temporal partitions, event definitions, and model selection budgets. Five training seeds and paired episode resampling form the replication protocol. Each run retains a separate forecast record, while paired forecasts use identical outage masks. Common outage masks align observation conditions across architectures. The displayed point estimates alone establish neither seed variability nor statistical significance; those assessments require retained paired run and episode records. Hyperparameter selection also preserves the evaluation horizon, because tuning only short forecasts can favor configurations that deteriorate during longer accumulation episodes. Episode resampling keeps connected stations together when they are affected by the same disturbance. Lead time uses the 30-minute event head and is defined from the first advisory to the first trigger crossing, with missed events retained as a separate outcome. This event record distinguishes timely detection, late detection, and the complete absence of a warning.

Table 1 provides the 15-minute concentration and event comparison, with lead time evaluated through the 30-minute head. Mean absolute error is 1.84 parts per million for the proposed model and 2.32 for the graph Transformer, a relative difference of 20.7 percent. Root mean squared error is 2.67 against 3.21, corresponding to a smaller relative difference of 16.8 percent. The numerical pattern favors a larger improvement in typical deviations than in the most severe misses. Precision recall area increases from 0.872 to 0.914, while event recall rises from 0.861 to 0.902. Concentration accuracy and event ranking improve concurrently and provide complementary assessments of the forecasts. Interval coverage ranges from 0.842 to 0.913 across the comparator set. The coverage entries are descriptive containment fractions. Calibration interpretation requires the nominal interval level, construction method, width, and evaluation denominator, which are unspecified in this comparison. The autoregressive and recurrent comparators provide reference points for separating the influence of nonlinear temporal representation from that of connected station information. Comparing their error distributions by regime would identify whether the aggregate differences arise primarily during routine operation or during the less frequent accumulation episodes.

Table 1. Concentration and event comparison at the 15-minute horizon

Model	MAE ppm	RMSE ppm	PR area	Recall	Lead min	Coverage
Autoregressive	3.86	5.24	0.742	0.701	8.2	0.842
LSTM	2.96	4.12	0.813	0.789	11.4	0.871
Transformer	2.78	3.89	0.836	0.817	13.0	0.887

Graph Transformer	2.32	3.21	0.872	0.861	15.8	0.901
Proposed	1.84	2.67	0.914	0.902	18.6	0.913

4.2. Comparative forecasts and episode behavior

Figure 4(a) contrasts the 5 and 30-minute absolute errors across the five model families. Short horizon errors are 2.92, 2.21, 2.08, 1.79, and 1.42 parts per million, while long horizon errors are 5.10, 3.89, 3.66, 3.08, and 2.53. Every comparator becomes less accurate as the forecast horizon expands. The proposed model increases by 1.11 parts per million, compared with 1.29 for the graph Transformer and 2.18 for the autoregressive model. Figure 4(b) relates 15-minute concentration errors to median warning lead times of 8.2, 11.4, 13.0, 15.8, and 18.6 minutes. Lower concentration error accompanies earlier warning in this comparison; establishing a causal relationship between these changes would require a separate intervention. Figure 4(c) presents cumulative absolute residual distributions with means of 3.86, 2.96, 2.78, 2.32, and 1.84. Figure 4(d) gives the distribution of episode errors for 30 episodes per model family, with medians at the same five central values. Precision recall analysis addresses detection behavior that aggregate accuracy can conceal when events are sparse [22]. The residual means and episode-error medians summarize different statistical units. Equal displayed central values do not establish identical distributions or independent replication. Interpretation also requires event recall and the frequency of advisories without a subsequent crossing.

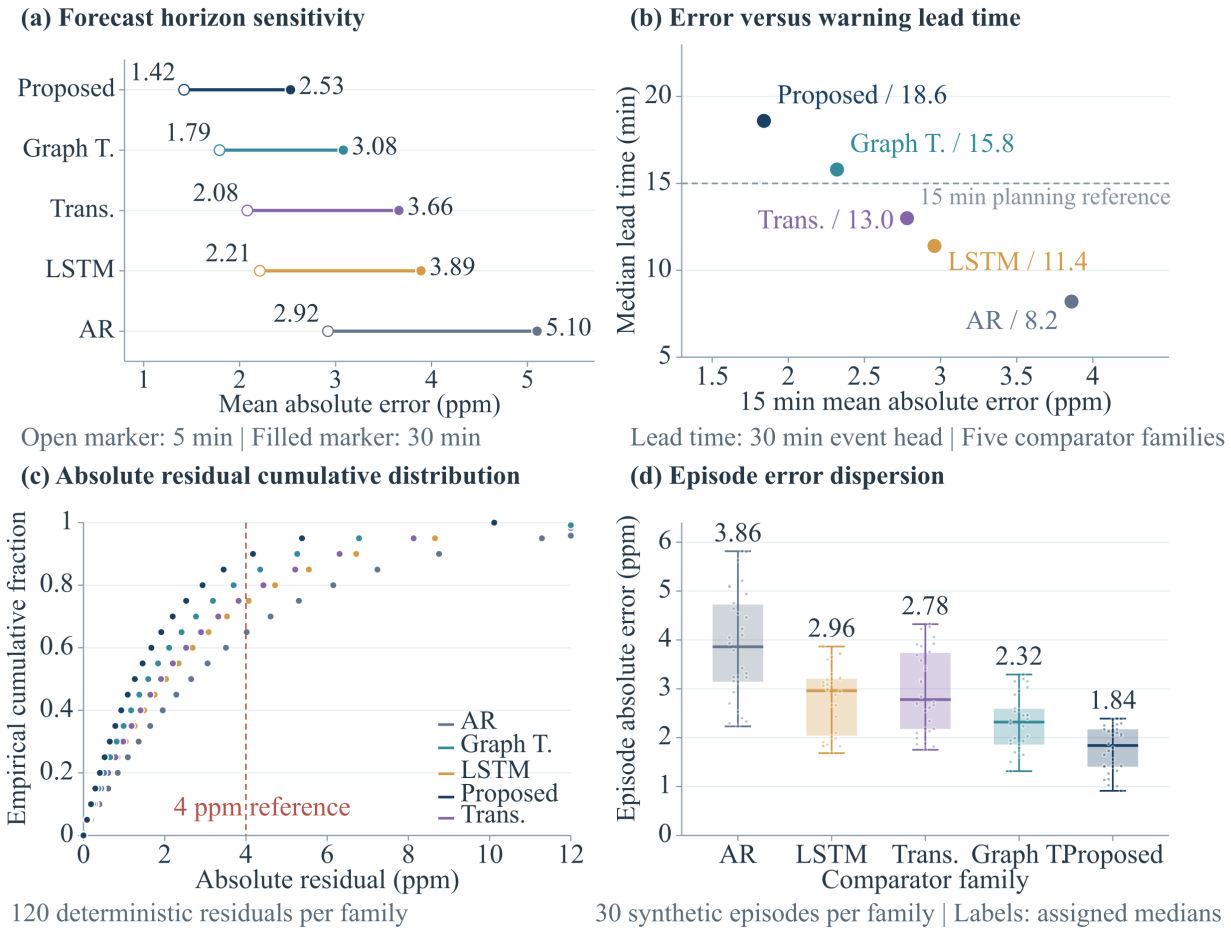


Figure 4. Comparator outcomes and error distributions. (a) Dumbbell comparison of 5 and 30-minute mean absolute errors. (b) Error versus median warning lead time for five model families. (c) Empirical cumulative distributions of absolute residuals. (d) Box plots of 30-episode errors per family with displayed medians

The engineering value of a concentration forecast depends on the timing of its errors around an accumulation event. Time aware event scoring methods evaluate the alignment of predicted anomalies with complete event intervals [23]. The benchmark distinguishes a threshold crossing, the first advisory associated with that crossing, and advisories without a subsequent event inside the forecast horizon. Underestimation during ascent can postpone the advisory even when the average error over the full episode is moderate. Positive residuals during recovery have a different consequence: the service may continue to indicate elevated risk after conditions have improved. Ventilation interventions require additional context because an advisory can be followed by a concentration decline after an operator act. Evaluating that sequence as an ordinary false positive would confuse a changed operating condition with an incorrect estimate of the

earlier state. The forecast target must therefore distinguish the trajectory under recorded controls from a counterfactual trajectory under unchanged controls. The present formulation conditions on controls available at the forecast origin. Unscheduled future interventions remain outside that information set; recorded post-origin actions are relevant to retrospective interpretation, rather than predictive inputs. Subsequent interventions belong in the event record so that their effects can be examined separately from the concentration and warning errors preceding them.

Figure 5(a) shows a 120-minute accumulation episode at 13 ten-minute timestamps. The concentration reaches 12.2 parts per million at minute 70, peaks at 18.4 at minute 90, and falls to 7.2 by minute 120. The proposed trajectory has an episode mean absolute error of 0.44 parts per million, compared with 1.32 for the graph Transformer trajectory. These episode errors characterize one selected sequence. Representativeness depends on the episode-selection rule and its partition membership; aggregate benchmark performance requires the full evaluation set. An advisory at minute 50 precedes the threshold crossing at minute 70 by 20 minutes. The shaded envelope has a half width of 1.8 parts per million and provides a pointwise sensitivity reference. Figure 5(b) presents the 13 signed residuals against concentration for both forecasts. The graph Transformer residual reaches negative 3.4 parts per million during the rising portion, whereas the proposed residuals remain between negative 0.8 and positive 1.0. Forecast visualization methods use uncertainty and operational context to clarify the implications of temporal errors [24]. The trigger line, advisory marker, peak annotation, and ventilation marker at minute 95 distinguish delayed ascent from delayed recovery. Their relative positions demonstrate the value of temporal diagnostics alongside aggregate episode error when evaluating the advisory interval.

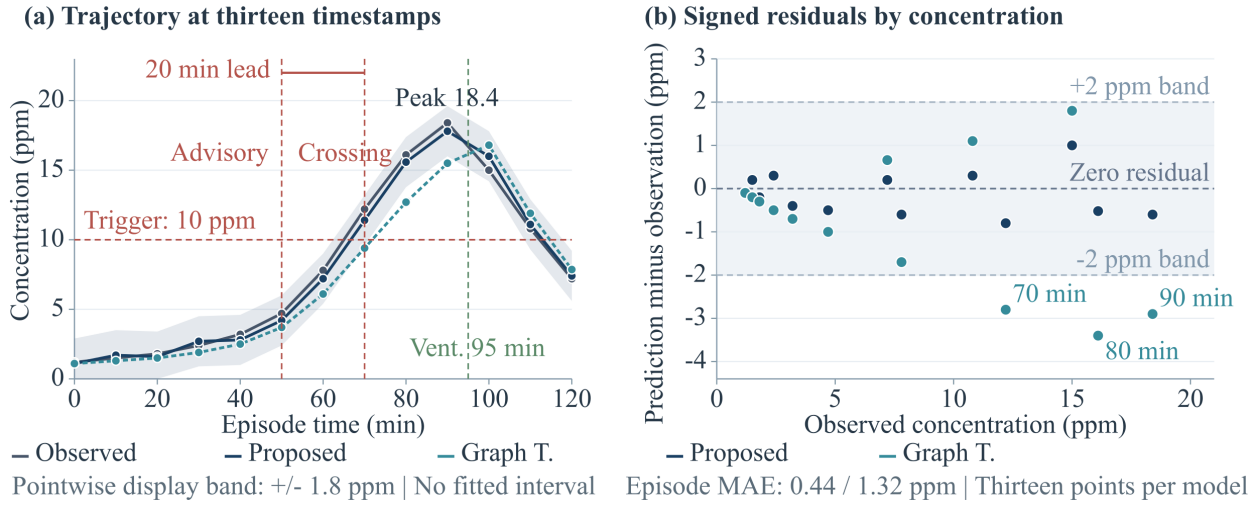


Figure 5. Accumulation episode trajectory and signed forecast residuals. (a) Concentration trajectory, proposed forecast, graph Transformer forecast, pointwise sensitivity envelope, trigger crossing, advisory, and ventilation intervention. (b) Residuals for the same 13 timestamps with a zero-error reference and a plus or minus 2 ppm comparison band

Error sign retains operational information that mean absolute error suppresses. Negative residuals near the peak indicate underestimation, while positive residuals during recovery can sustain an advisory after concentration has declined. Probabilistic forecasting evaluation distinguishes predictive reliability from the sharpness of an uncertainty distribution [25]. Accordingly, a forecast interval needs assessment for both containment and width, while the constant width sensitivity envelope serves a separate diagnostic role. During the rising portion of the episode, successive negative residuals exhibit temporal persistence. An advisory policy requiring repeated evidence before activation could accumulate that delay across several forecast origins. During recovery, the same persistence requirement may delay withdrawal of an advisory. Evaluating both transitions helps separate early detection from unnecessarily prolonged warnings. Station specific residual histories offer another diagnostic: errors concentrated near one junction may indicate inadequate local attributes, whereas errors across the network after a ventilation transition may indicate insufficient control representation. The distinction determines whether the next modeling change should target graph metadata, observation quality, or the temporal response to a known intervention.

4.3. Component sensitivity calibration and resource tradeoffs

Figure 6(a) compares five component configurations: the complete model, removal of topology, removal of continuous evolution, removal of temporal attention, and removal of the graph penalty. Their 15-minute absolute errors are 1.84, 2.45, 2.31, 2.22, and 2.05 parts per million. Figure 6(b) gives corresponding recalls of 0.902, 0.834, 0.851, 0.863, and 0.884. Topology removal has the largest numerical penalty in both measures, while removal of the graph penalty produces the smallest difference. Figure 6(c) examines missingness at five levels. Proposed errors are 1.72, 1.84, 2.08, 2.43, and 2.91; graph Transformer errors are 2.14, 2.32, 2.74, 3.25, and 3.93; temporal Transformer errors are 2.57, 2.78, 3.26, 3.94, and 4.78. Figure 6(d) presents regime and horizon effects, ranging from 1.11 during routine flow at five

minutes to 3.12 during communication outage at 30 minutes. Missing data research distinguishes random deletion from informative missingness associated with operating conditions [26]. A stratified interpretation is therefore necessary when communication degradation coincides with accumulation. The regime pattern motivates separate analysis of observation age and outage duration. The missingness percentages alone leave the deletion mechanism, block length, and synchronization across nodes unspecified, limiting interpretation of persistent communication failures.

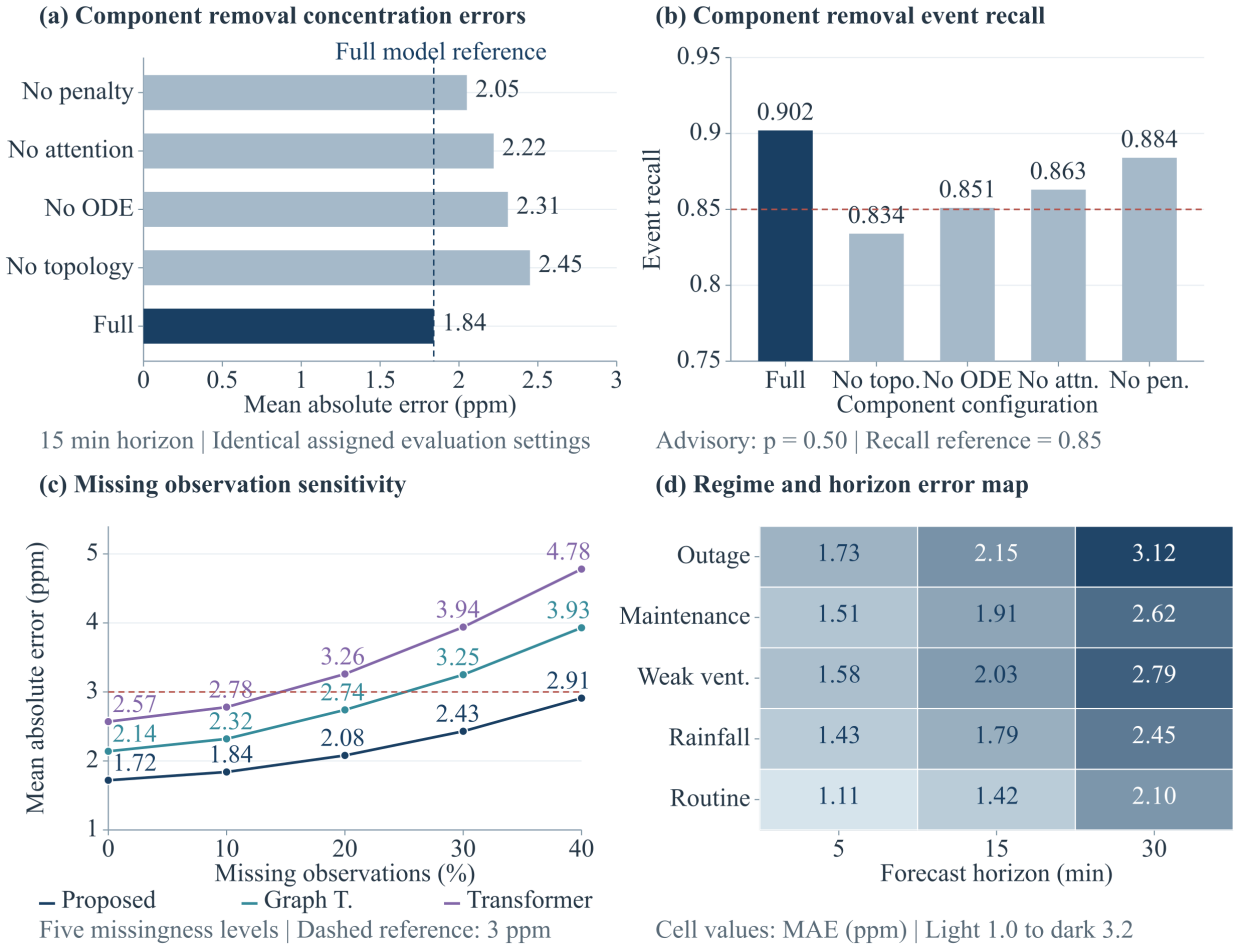


Figure 6. Component and observation sensitivity. (a) Concentration error for five component configurations. (b) Event recall for the same configurations with a 0.85 reference. (c) Error at five missingness levels for three model families. (d) Regime by horizon heatmap containing 15 concentration error values

The missingness trend increases proposed model error by 1.19 parts per million between the two extreme levels, compared with 1.79 for the graph Transformer and 2.21 for temporal attention alone. Continuous state evolution describes the latent trajectory during observation gaps, while observation age indicates the strength of the constraint supplied by recent measurements. Post-hoc calibration research examines probability adjustment using held-out data [27]. Preservation of event ranking depends on the fitted mapping; input-dependent scaling requires its own ranking assessment. Figure 7(a) compares raw and calibrated reliability at eight bins centered from 0.10 to 0.90. Calibrated event frequencies are 0.08, 0.17, 0.29, 0.41, 0.52, 0.64, 0.77, and 0.88, with 100 observations per bin. The reported weighted calibration gap is 0.019, compared with 0.109 before calibration. Reproduction requires bin boundaries, mean predicted probabilities, and memberships. The 800 displayed bin entries and the 360 concentration samples serve separate summaries whose overlap is unspecified. Figure 7(b) presents precision at five recall positions from 0.20 to 0.98; the proposed sequence is 0.97, 0.96, 0.94, 0.89, and 0.68. These selected operating points describe the warning tradeoff. Reproducing the reported area requires the full ranked prediction–label record and integration convention, rather than integration of these five displayed points. The decline in precision near the highest recall position makes the cost of a more sensitive advisory rule explicit.

Resource requirements affect whether a forecast can reach an operator within the intended advisory interval. Solver analyses relate adaptive integration cost to numerical tolerances and the behavior of the learned vector field [28]. Figure 8(a) compares per window latency budgets of 2, 8, 12, 19, and 28 milliseconds for the five model families, paired with their 15-minute concentration errors. Bubble area represents parameter counts of 0.02, 0.18, 0.42, 0.61, and 0.83 million. The comparison describes reported timing and error values. Processor, batch size, software configuration, timing scope, and latency distribution are unspecified, so portability and deadline reliability remain unverified. Relative to the graph

Transformer, the proposed configuration adds 9 milliseconds and 0.22 million parameters while reducing mean absolute error by 0.48 parts per million. The suitability of this computational trade-off depends on available processing capacity, forecast update frequency, and the number of locations processed concurrently. Hardware timing should separate preprocessing, continuous integration, temporal aggregation, and output calibration. A single aggregate latency obscures which stage is responsible when a forecast approaches its delivery deadline.

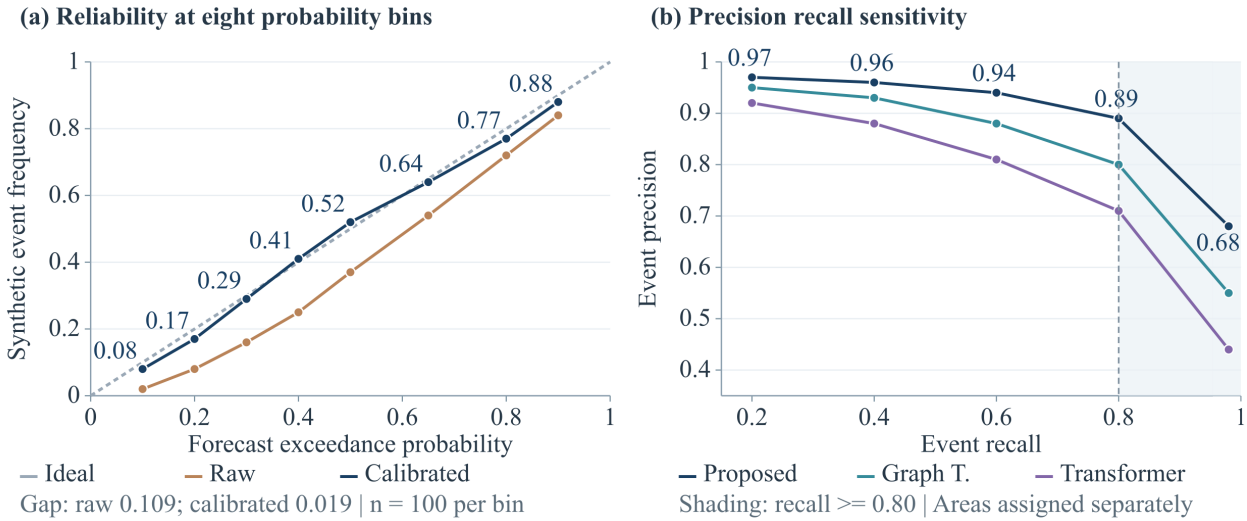


Figure 7. Probability reliability and warning tradeoffs. (a) Eight bin raw and calibrated reliability curves with 100 observations per bin and weighted calibration gaps. (b) Precision recall operating points for three model families at five recall positions, with the high recall region marked

A vertical 30 millisecond threshold marks the planning budget. The reported 28 millisecond value leaves a 2 millisecond planning margin. Deadline compliance requires a tail-latency assessment under representative concurrent load. Figure 8(b) presents regime recall centers of 0.940, 0.910, 0.900, 0.870, and 0.840, with range half widths of 0.025, 0.030, 0.035, 0.040, and 0.045. Uncertainty assessment for dependent observations preserves episode structure when constructing resampling intervals [29]. The displayed ranges are descriptive dispersion bars; a confidence interpretation requires the interval construction and sampling unit. Temporal distribution shift methods examine deterioration when deployment conditions differ from the training record [30]. The outage recall center of 0.840 falls below the displayed 0.85 reference. A stale-data status and a separately validated fallback procedure are therefore operational requirements. A deployment assessment should record processor type, batch size, software version, and whether preprocessing is included in the timing interval. Vector field evaluation counts can then reveal whether long observation gaps or abrupt controls increase solver work. Graph version transitions require paired forecasts around the change using the same event history, allowing the influence of revised connectivity to be separated from concurrent changes in the incoming observations.

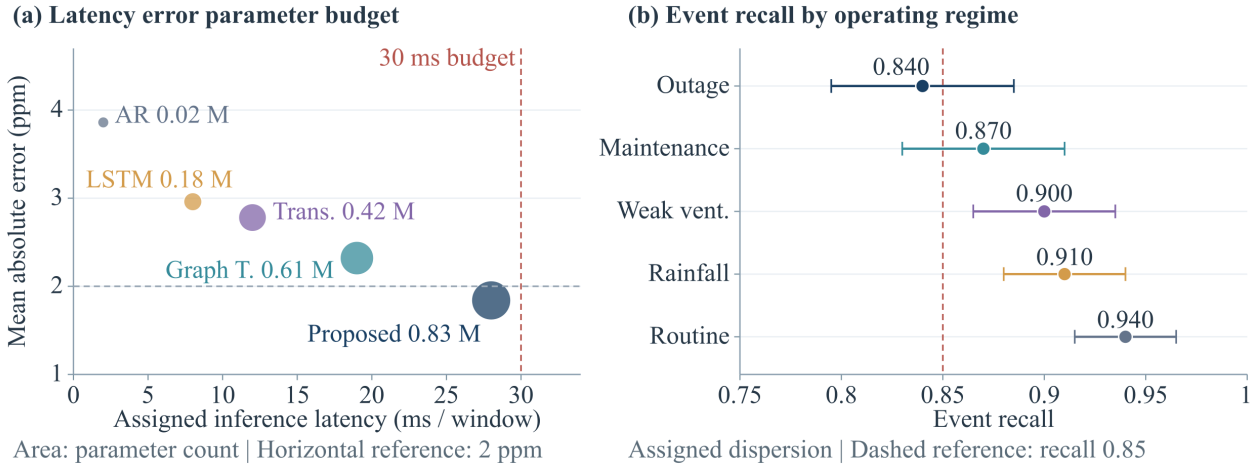


Figure 8. Deployment budgets and regime sensitivity. (a) Latency and error bubble plot with parameter counts and a 30 ms planning threshold. (b) Recall point ranges for five operating regimes with dispersion bars and a 0.85 reference

5. Conclusion

This study develops a topology aware Neural ODE Transformer for hydrogen sulfide forecasting in urban pipe networks; the reported benchmark comparisons support further evaluation of the integrated architecture under the stated conditions.

1. Forecast accuracy and warning capability. At the 15-minute horizon, the proposed model achieves a mean absolute error of 1.84 ppm, a root mean squared error of 2.67 ppm, and a precision recall area of 0.914. Event recall reaches 0.902, while the 30-minute event head provides a median warning lead time of 18.6 minutes. These metrics address concentration deviation, event detection, and the time available for an advisory response.

2. Comparative performance and observation sensitivity. The proposed model reduces mean absolute error by 20.7% relative to the graph Transformer. Removing topology increases error from 1.84 to 2.45 ppm within the reported configuration. This comparison indicates component sensitivity; interpreting a causal mechanism requires capacity-matched controls and repeated-run uncertainty. At 40% missing observations, its error reaches 2.91 ppm, compared with 3.93 ppm for the graph Transformer. Performance nevertheless deteriorates as observations become less complete, highlighting the continuing dependence on timely sensor information.

3. Applicability and limitations. The framework depends on accurate connectivity records and sufficiently informative control inputs. Undocumented ventilation changes, sensor bias, and shifts in event prevalence can weaken state estimation and probability calibration. Continuous evolution supports inferences between observations; during long interruptions, the constraints of recent measurements on potential trajectories become weaker. Acquisition traceability, complete implementation settings, and prospective field validation remain necessary to establish reproducibility and operational suitability.

Future work should examine graph uncertainty, persistent communication failures, and calibration under changing operating conditions, followed by a controlled engineering evaluation of advisory timing and operator response.

6. Declarations

6.1 Author Contributions

Milena Nikolić contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Vasilije Šumaher contributes to validation, analysis, investigation, data collection, draft preparation. Marija Jovanović contributes to conceptualization, validation, analysis. All authors have read and agreed with the manuscript before its submission and publication.

6.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

6.3 Institutional Review Board Statement

Not applicable.

6.4 Informed Consent Statement

Not applicable.

6.5 Date Availability

Not applicable.

6.6 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.

6.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

7. References

[1] Felföldi, T. (2024). Microbiological aspects of sewage odor problems in the urban environment - a review. Biologia Futura, 75(3), 371–377. https://doi.org/10.1007/s42977-024-00242-2

[2] Teuber, K., Dixit, A., & Hinkelmann, R. (2025). CFD simulation of turbulent mass transfer of H2S and O2 in a stirring tank. Water Science & Technology, 91(1), 69–82. https://doi.org/10.2166/wst.2024.406

[3] Austigard, Å. D., Smedbold, H. T., & von Hirsch Svendsen, K. (2024). Comparison of 3 methods characterizing H2S exposure in water and wastewater management work. Annals of Work Exposures and Health, 68(7), 725–736. https://doi.org/10.1093/annweh/wxae043

[4] Villinger, G., & Reiterer, A. (2025). Robotic systems for sewer inspection and monitoring tasks: Overview and novel concepts. IEEE Sensors Journal, 25(5), 8786–8796. https://doi.org/10.1109/jsen.2025.3528109

[5] Xiaoxiang, S., Yan, G., Ning, L., & Bing, R. (2024). A practical approach for missing wireless sensor networks data recovery. China Communications, 21(5), 202–217. https://doi.org/10.23919/jcc.ea.2021-0283.202401

[6] Shen, C., Xia, H., Fu, X., Wang, X., & Wang, W. (2024). Identifying risk components using a sewer-road integrated urban stormwater model. Water Resources Management, 38(8), 3049–3070. https://doi.org/10.1007/s11269-024-03804-0

[7] Taoussi, B., Boudia, S. M., & Mazouni, F. S. (2025). Wind speed forecasting using univariate and multivariate time series models. Stochastic Environmental Research and Risk Assessment, 39(2), 547–579. https://doi.org/10.1007/s00477-024-02881-2

[8] Wang, T., Chen, W., & Tang, B. (2024). Water quality prediction using ARIMA-SSA-LSTM combination model. Water Supply, 24(4), 1282–1297. https://doi.org/10.2166/ws.2024.060

[9] Jin, G., Liang, Y., Fang, Y., Shao, Z., Huang, J., Zhang, J., & Zheng, Y. (2024). Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. IEEE Transactions on Knowledge and Data Engineering, 36(10), 5388–5408. https://doi.org/10.1109/tkde.2023.3333824

[10] Losada, I. B., & Terranova, N. (2024). Bridging pharmacology and neural networks: A deep dive into neural ordinary differential equations. CPT: Pharmacometrics & Systems Pharmacology, 13(8), 1289–1296. https://doi.org/10.1002/psp4.13149

[11] Li, W., & Law, K. L. E. (2024). Deep learning models for time series forecasting: A review. IEEE Access, 12, 92306–92327. https://doi.org/10.1109/access.2024.3422528

[12] Shi, J., Zhang, W., Bao, Y., Gao, D. W., & Wang, Z. (2024). Load forecasting of electric vehicle charging stations: Attention based spatiotemporal multi–graph convolutional networks. IEEE Transactions on Smart Grid, 15(3), 3016–3027. https://doi.org/10.1109/tsg.2023.3321116

[13] Vashistha, S., & Kumar, N. (2026). Selective state-focused graph attention networks for multihorizon traffic forecasting. IEEE Transactions on Computational Social Systems, 13(2), 2030–2046. https://doi.org/10.1109/tcss.2025.3605930

[14] Yingran, Z., Chao, L., & Rui, S. (2025). Enhancing traffic flow forecasting with delay propagation: Adaptive graph convolution networks for spatio-temporal data. IEEE Transactions on Intelligent Transportation Systems, 26(1), 650–660. https://doi.org/10.1109/tits.2024.3485794

[15] Su, Y., Ma, T., Rong, H., & Wahab, M. M. A. (2025). Adaptive graph structure learning neural rough differential equations for multivariate time series forecasting. IEEE Transactions on Big Data, 11(5), 2710–2723. https://doi.org/10.1109/tbdata.2025.3552334

[16] Artham, S., & Shaikh, S. H. (2024). A neural ODE and transformer-based model for temporal understanding and dense video captioning. Multimedia Tools and Applications, 83(23), 64037–64056. https://doi.org/10.1007/s11042-023-17809-1

[17] Qin, T., Walker, B., Lyons, T., Yan, H., & Li, H. (2026). Learning dynamic graph embeddings with neural controlled differential equations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 48(2), 2096–2103. https://doi.org/10.1109/tpami.2025.3617660

[18] Saravana, M. K., Roopa, M. S., Arunalatha, J. S., & Venugopal, K. R. (2026). Transformers for multivariate time series forecasting: Comprehensive analysis, challenges, research opportunities, and future prospects. IEEE Access, 14, 11424–11457. https://doi.org/10.1109/access.2026.3654408

[19] Penso, C., Frenkel, L., & Goldberger, J. (2024). Confidence calibration of a medical imaging classification system that is robust to label noise. IEEE Transactions on Medical Imaging, 43(6), 2050–2060. https://doi.org/10.1109/tmi.2024.3353762

[20] Liu, S., & Zhou, D. J. (2024). Using cross-validation methods to select time series models: Promises and pitfalls. British Journal of Mathematical and Statistical Psychology, 77(2), 337–355. https://doi.org/10.1111/bmsp.12330

[21] Altin, M., Cakir, A., & Toreyin, B. U. (2026). Time-series forecasting and evaluation on city sensing data: Pedestrian and microclimate benchmarks. IEEE Access, 14, 22618–22629. https://doi.org/10.1109/access.2026.3662469

[22] Wen, P., Xu, Q., Yang, Z., He, Y., & Huang, Q. (2024). Algorithm-dependent generalization of AUPRC optimization: Theory and algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(7), 5062–5079. https://doi.org/10.1109/tpami.2024.3361861

[23] Sayyaf, M. I., Pascacio, P., Zhu, N., & Renaudin, V. (2025). Time-series anomaly detection for sensor data: Models, metrics, and methodologies—A review. IEEE Sensors Journal, 25(24), 43603–43619. https://doi.org/10.1109/jsen.2025.3616395

[24] Leffrang, D., & Müller, O. (2025). Visualizing uncertainty in time series forecasts: The impact of uncertainty visualization on users' confidence, algorithmic advice utilization, and forecasting performance. Journal of Forecasting, 44(4), 1235–1246. https://doi.org/10.1002/for.3222

[25] Phipps, K., Heidrich, B., Turowski, M., Wittig, M., Mikut, R., & Hagenmeyer, V. (2024). Generating probabilistic forecasts from arbitrary point forecasts using a conditional invertible neural network. Applied Intelligence, 54(8), 6354–6382. https://doi.org/10.1007/s10489-024-05346-9

[26] Richter, A., Ijaradar, J., Wetzker, U., Jain, V., & Frotzscher, A. (2024). A survey on multivariate time series imputation using adversarial learning. IEEE Access, 12, 148167–148189. https://doi.org/10.1109/access.2024.3473540

[27] Balanya, S. A., Maroñas, J., & Ramos, D. (2024). Adaptive temperature scaling for robust calibration of deep neural networks. Neural Computing and Applications, 36(14), 8073–8095. https://doi.org/10.1007/s00521-024-09505-4

[28] Ouala, S., Debreu, L., Chapron, B., Collard, F., Gaultier, L., & Fablet, R. (2026). Enhanced computational complexity in continuous-depth models: Neural ordinary differential equations with trainable numerical schemes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 48(1), 1045–1051. https://doi.org/10.1109/tpami.2025.3599629

[29] Poskitt, D. S. (2025). Bootstrapping non-stationary and irregular time series using singular spectral analysis. Journal of Time Series Analysis, 46(1), 81–112. https://doi.org/10.1111/jtsa.12759

[30] Lukats, D., Zielinski, O., Hahn, A., & Stahl, F. (2025). A benchmark and survey of fully unsupervised concept drift detectors on real-world data streams. International Journal of Data Science and Analytics, 19(1), 1–31. https://doi.org/10.1007/s41060-024-00620-y