Journal of Applied Computational Intelligence

eISSN: 3136-6937

Vol. 4, No. 1, January, 2026

Humidity Threshold Constrained Temporal Fusion Transformer for Airport Fog Visibility Prediction

Hugo Polák[1,*], Kamil Hora[1], Miloš Příhoda[2]

[1] *Faculty of Computer Science, University of Bohemia, Plzeň, 301 00, Czech Republic*
[2] *Faculty of Information Technology, University of Moravia, Brno, 602 00, Czech Republic*

Received 18 November 2025; Revised 15 Feburary 2026; Accepted 20 Feburary 2026; Published 24 Feburary 2026

Abstract

Airport fog forecasting requires both continuous visibility estimates and reliable warnings for rare low-visibility hours. This study evaluates a humidity-threshold-constrained Temporal Fusion Transformer using 24 hours of observations to predict visibility one, two and three hours ahead. Variable-specific embeddings, state-dependent selection, recurrent encoding and masked temporal attention combine observed meteorology with known UTC calendar variables. An auxiliary future-humidity estimate supplies a smooth increasing gate at relative humidity 0.90; a conditional penalty relates this gate to the predicted probability of reported visibility at or below 1 km. The model jointly learns ordered 0.1, 0.5 and 0.9 visibility quantiles. Public hourly reports from Fresno Airport between 2021 and 2025 yield 24,430 training windows, 4,220 checkpoint-selection windows, 4,259 calibration windows and 8,572 final test windows. Test mean absolute errors at one, two and three hours are 0.462, 0.590 and 0.681 km. One-hour warning recall is 0.7576 and F1 is 0.617, with false-positive rate 0.00814 and 69 false alarms. Post-fit probability calibration reduces one-hour Brier score from 0.01045 to 0.00596. The nominal 80% interval covers 94.67% of all one-hour targets but 75.85% of targets below the 16.09344 km reporting ceiling, which contains 79.13% of the one-hour test targets. Persistence has lower one-hour absolute error and higher one-hour F1, whereas the constrained model has lower absolute error at two and three hours. Matched ablations separate humidity regularization, variable selection and attention; standardized noise and held recent observations expose distinct sensitivities. These results qualify continuous forecast accuracy, event detection and interval reliability within a chronological single-airport evaluation.

Keywords: Airport Visibility; Humidity Threshold; Temporal Fusion Transformer; Quantile Calibration; Low Visibility Warning

1. Introduction

Multivariate sequence learning supports short horizon forecasting from interacting environmental observations [1]. Attention based prediction provides mechanisms for linking recent changes with longer observation histories [2]. Conditional learning can incorporate domain relationships through penalties that depend on the available state [3]. Probabilistic forecasting adds information about the range and reliability of future outcomes [4]. These computational capabilities are relevant to airport fog, where visibility can decrease rapidly while temperature, humidity and wind evolve at different rates. A high humidity observation may accompany several visibility states because aerosol content, mixing and the recent cooling trajectory also influence the observed conditions. A useful forecasting model therefore needs a representation of the preceding meteorological evolution and a clear distinction between a continuous forecast

* Corresponding author: polak.h@bohemiau.cz

http://dx.doi.org/10.68406/jaci.2026.vol4iss1no7:83-95

➢ 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.

and the decision to issue a low visibility warning. The connection between these outputs determines how prediction errors translate into missed deteriorations and unnecessary alerts.

Airport visibility evaluation depends on the frequency and temporal clustering of reduced-visibility states. Imbalanced learning research [5] motivates reporting rare-event recall together with ordinary-condition error. Most archive targets lie at the reporting ceiling, so a low global forecast error can coexist with missed fog hours. Calibration studies [6] assess whether warning probabilities correspond to observed event frequency, while interval assessment [7] requires coverage to be separated by ceiling status. Chronological validation methodology [8] assigns network selection, post-fit adjustment and final scoring to distinct time periods when nearby forecast origins share much of their history.

This study evaluates a humidity-threshold-constrained temporal fusion model for one-, two- and three-hour visibility forecasts at Fresno Airport. A 24-hour observed history and known UTC calendar features support ordered visibility quantiles, event probabilities and auxiliary future-humidity estimates. Section 2 reviews temporal visibility forecasting, meteorological constraints and probabilistic reliability. Section 3 defines issue-time alignment, variable selection, recurrent and attention processing, the humidity penalty and post-fit calibration. Section 4 reports matched forecast errors, hour-level warnings, interval coverage, ablations and observation perturbations. Section 5 summarizes three quantitative findings and the scope of the station-specific evaluation.

2 Related Work

2.1 Airport Visibility Forecasting with Temporal Models

Environmental Time Series Prediction Combines Lagged Measurements and Known Future Descriptors. Recurrent Sequence Models [9] are local modifications that update a state at each step. For the visibility prediction, this state can keep a recent decrease, a stable increase, or a change in wind even if the latest measurement is normal. Dependency is one-way; observations before the forecast start time are used to conduct the prediction, and the future verification record serves as the goal. Therefore, a forecast several hours in advance needs to specify the horizon of the forecast and the exact time of the data used. Observation histories are used in this study to measure the evaluation up to the origin, and calendar descriptors for the future target hours are already known. This way, the different models can be compared based on the same issue time of the leads. Then, the change in performance can be attributed to the extended distance from the last measurement point to the verification target.

Temporal Fusion Research [10] integrates recurrent processing, variable selection and attention in a shared forecasting architecture. Variable Selection assigns state-dependent weights to the heterogeneous descriptors, and Attention finds the temporal representations of each output. Gated-regression studies [11] introduce residual pathways to regulate the flow of transformed features in a shared space. The Scales of observation for airports and the distribution of humidity, temperature, wind, etc., are different. Feature representations need to keep the original identities of objects and change according to their circumstances. Compact Recurrent and Attention Layers are employed in this implementation along with a shared quantile output structure and an auxiliary humidity target. They have the same observation windows and target definitions for ablations, so the contributions of different representation components can be observed directly.

2.2 Meteorological Constraints in Predictive Learning

Constraint-based learning [12] introduces a constraint to stabilise the predictor in the presence of an uneven empirical training distribution. A humidity threshold is a type of relationship for fog-oriented risk estimation that can be conditioned on the forecast state and the uncertainty of future humidity. Multi-task learning [13] can be employed to establish a shared representation via an auxiliary continuous objective. Predicted relative humidity serves as the input for a differentiable gate here, and observed future humidity acts as an auxiliary output during training. A penalty will be imposed for a high-visibility risk near the gate. It has limitations, so the fitted probability can be used for both event monitoring and humidity analysis. This construction provides a computable role for the constraint via matched training without the coefficient.

2.3 Quantile Calibration and Low Visibility Warnings

Quantile learning [14] estimates several conditional target levels using an asymmetric error objective. Ordered outputs are the median forecast and a central interval of a certain width based on the state. Research on probability calibration [15] uses the reliability of an event score as an independent fitting function. Therefore, a low visibility probability can be adjusted after network training without affecting its rank under a positive temperature transformation. Time-dependent evaluation studies [16] can also be used for chronological verification of neighbouring observations. The above split will be used to set the probability adjustment and interval expansion. Threshold Selection Research [17] connects an alarm rule to the desired balance of detected events and false alarms. A fixed probability threshold is used before the final year's score to obtain confusion counts, which are then stored with the overall forecast data. The

proportion of the target event will be different each year. An alert rule set in a particular period will have an observable transfer effect and will be carried over in the final year's confusion counts and probability reliability analysis.

3 Threshold Constrained Temporal Fusion and Quantile Prediction

3.1 Forecast Targets and Humidity Regime Construction

Each issue uses 24 hourly states with nine observed descriptors: reported visibility, air temperature, temperature minus dew point, relative humidity, wind speed, sine and cosine of wind direction, pressure and a direction-availability indicator. Visibility is converted from miles to kilometres and capped at 16.09344 km, the retained reporting ceiling. Temperature, wind speed and altimeter pressure are expressed in degrees Celsius, metres per second and hectopascals. A logarithmic target transformation gives finer resolution near low visibility while limiting the influence of ceiling observations. The same transformation is applied to historical visibility inputs.

$$y_h = ln(1 + min(V_h, C)) \tag{1}$$

In this expression V is reported visibility in kilometres, C is the 16.09344 km ceiling and y is the transformed target. Verification occurs one, two or three hours after the UTC issue time. Each hourly state uses the latest routine observation timestamped within its preceding hour, retaining its delay relative to issue. Complete input and target windows stay inside one chronological partition. A window enters the analysis when its required continuous observations are finite and visibility and wind speed satisfy their physical domains. This alignment defines information available at issue and the future records reserved for verification.

$$e_h = 1[V_h \leq 1km] \tag{2}$$

The binary event target e equals one when reported visibility is at or below 1 km, independently of the weather code. Event probability p is trained jointly with continuous visibility quantiles. Relative humidity is represented as a fraction, and an auxiliary head estimates its future value at each lead time. The smooth humidity gate has threshold 0.90 and transition scale 0.03; its logistic value increases as predicted humidity passes the threshold. The gate acts as a conditional regularizer for event probability, with observed visibility supplying the event label. Future humidity and visibility retain distinct supervised targets.

$$g_h = \sigma(\frac{R_h - \tau}{s}) \tag{3}$$

The logistic sigmoid maps predicted future relative humidity to the smooth gate g. At humidity near 0.90 the transition scale 0.03 gives a gradual change; higher predicted humidity yields a larger gate value and permits higher fog-event probability under the conditional penalty. Figure 1 shows the hourly observation history, issue-time alignment and three future targets. Temperature, humidity, wind and pressure histories feed the shared forecast path, while the auxiliary humidity output supplies the gate for each horizon. Continuous visibility quantiles, event probability and future humidity are supervised by their respective observed targets.

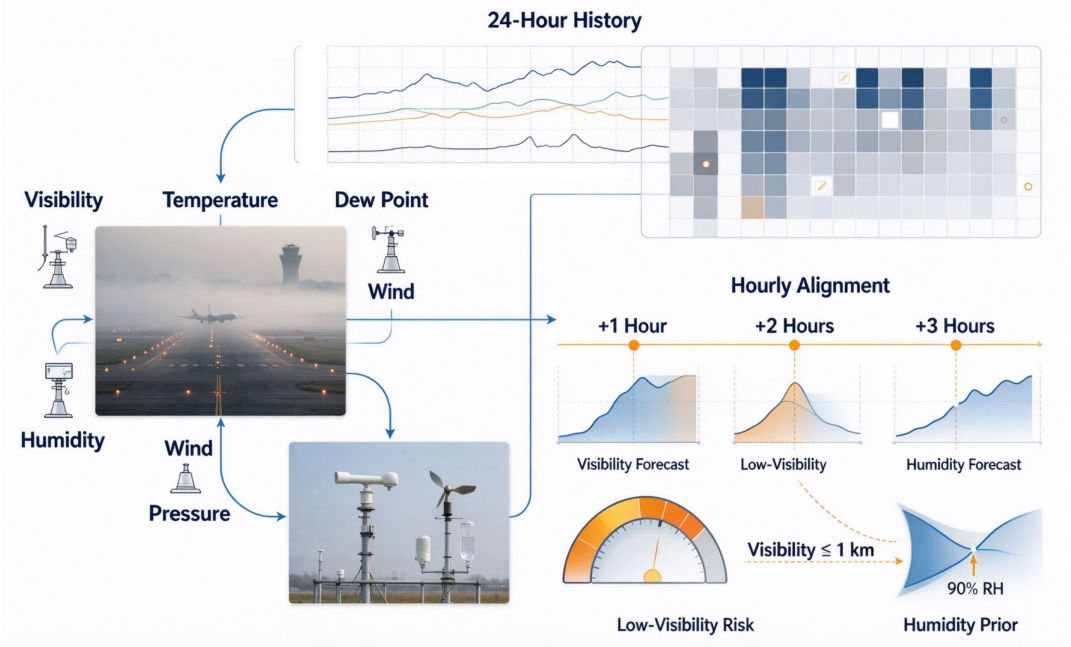


Figure 1. Airport observation histories humidity states and multi horizon forecast targets.

Training uses observations from 2021-2023. The first half of 2024 selects the network checkpoint, the second half calibrates probabilities and intervals, and 2025 supplies final testing. Descriptor means and standard deviations are fitted on finite training observations and then fixed. Missing or variable wind direction has zero directional components plus an availability flag; complete-window screening applies to the other required channels. UTC hour and day-of-year sine-cosine pairs are known for each future verification time. These calendar descriptors provide future context while measured future meteorology remains outside the model input.

3.2 Variable Selection and Temporal Evidence Fusion

Each observed scalar receives a variable-specific linear embedding of width sixteen, followed by a shared gated residual transform. A two-layer selection network processes the nine standardized descriptors at each historical hour, uses sixteen hidden units and produces nine softmax weights. Their weighted sum enters the recurrent encoder. State-dependent weights allow a humidity value to have a different internal contribution during stable clear weather and a developing visibility reduction. Their fitted values describe model computation and require outcome-based evaluation before meteorological interpretation.

$$z_t = \Sigma_j \alpha_{tj} \phi_j(x_{tj}) \tag{4}$$

Selected state z combines nine transformed descriptors phi with nonnegative weights alpha that sum to one. A one-layer long short-term memory encoder with sixteen hidden units processes the 24 observed states. A second recurrent layer processes three known future-calendar vectors from the retained encoder state. Gated residual processing and normalization connect these representations into 27 positions: 24 historical states and three horizon-specific calendar states. Each future query consequently has a defined issue-time information set, and measured future weather remains reserved for target verification.

Two attention heads use eight-dimensional queries and keys with a shared sixteen-dimensional value transformation. Each horizon query accesses the historical positions and calendar positions permitted by its causal mask. Averaged attended features join the corresponding decoder state before another gated residual block. Five outputs comprise a bounded median, positive lower and upper offsets, an event logit and an auxiliary humidity logit. The offsets form ordered 0.1, 0.5 and 0.9 visibility quantiles within the physical reporting range. This construction keeps continuous forecast, warning probability and humidity estimate distinct.

Figure 2 traces variable selection, recurrent encoding and temporal attention into three output families. The observed weather branch supplies past conditions; the known calendar branch supplies issue-time information for each future horizon. The ordered quantile head yields the median and interval, while the event and auxiliary humidity heads support the conditional probability penalty. The schematic also identifies post-fitting probability and interval calibration after checkpoint selection. Raw network logits and quantiles remain identifiable for comparison with adjusted outputs. Fitted selection and attention weights can summarize internal calculations, while forecast errors and warning scores evaluate prediction quality directly.

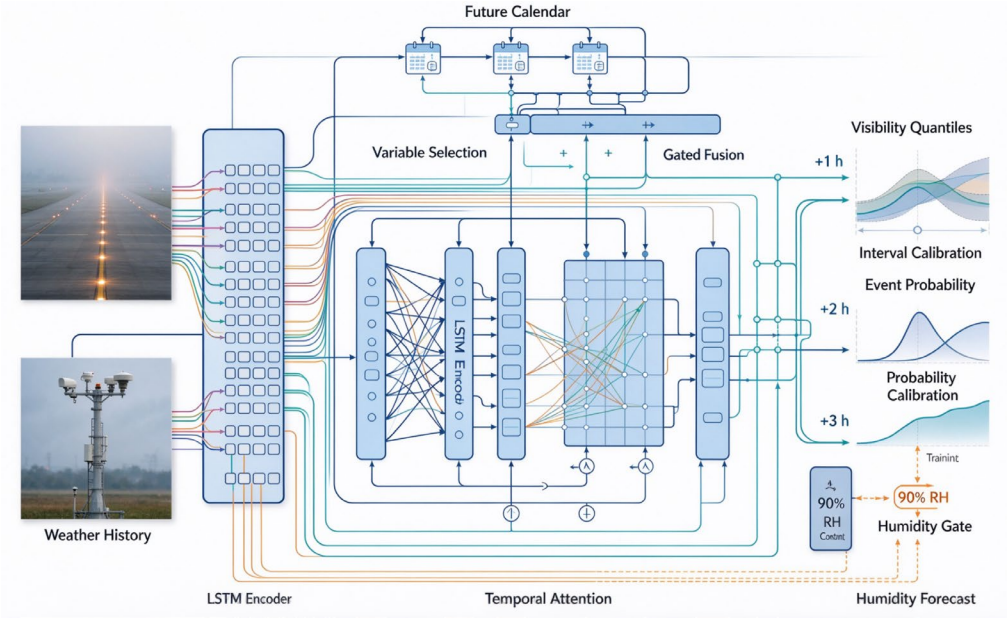


Figure 2. Variable selection temporal fusion humidity regularisation and calibrated forecast outputs.

3.3 Conditional Regularisation and Probability Calibration

The humidity-consistency term is the mean squared positive excess of event probability above the gate calculated from predicted future humidity. Its gradient activates when event probability exceeds that gate and reaches both the event and auxiliary-humidity heads. Observed future humidity supervises the latter, anchoring the gate to a measured meteorological target. The increasing logistic gate offers greater admissible event probability at higher predicted humidity. A matched ablation sets only the consistency coefficient to zero while retaining architecture, continuous visibility targets and training budget. This comparison isolates the regularizer within the declared implementation.

$$L_H = \langle [p - g]_+^2 \rangle \tag{5}$$

The training objective combines visibility quantile loss Q, event classification loss E, auxiliary humidity loss R and consistency loss H. Q averages pinball loss at levels 0.1, 0.5 and 0.9 in transformed visibility units. E is binary cross entropy with positive-event weight eight, and R is mean squared error of forecast relative humidity. Coefficients for E, R and H are 0.15, 0.05 and 0.20 in the complete model. Each output therefore has a specified supervised role; the humidity term additionally couples the two auxiliary forecast heads through a smooth conditional relation.

$$L = Q + 0.15E + 0.05R + 0.20H \tag{6}$$

AdamW optimisation uses a learning rate of 0.002, weight decay of 0.0001, batches of 256 and a gradient norm limit of one. Each configuration receives 12 complete passes through the training windows. The checkpoint with the smallest mean quantile loss on the selection period is retained. The principal comparisons use random seed 298. The models include the complete constrained network, the same network without humidity regularisation, an attention removal, uniform variable weights and a recurrent comparator. The recurrent comparator receives the same nine historical descriptors and future calendar features. Persistence uses the current visibility and calibration period residual quantiles, with event rates conditioned on fixed current visibility bins. The bins have edges at zero, one, two, four, eight, sixteen and seventeen kilometres. Counts are smoothed by adding one event and one nonevent per bin. This supplies a simple state conditioned warning reference using the same calibration observations.

Positive temperature T and intercept b calibrate each horizon event logit after network fitting. Candidate temperatures are 0.5, 0.75, 1.0, 1.5 and 2.0; intercepts range from minus four to two in increments of 0.25. The selected pair minimizes binary cross entropy on the second half of 2024. Positive temperature preserves event-score ranking while changing probability scale and base rate. A warning threshold is then chosen from 0.01-0.50 in 0.01 increments to maximize calibration-period F1. The same fitted map and threshold are carried into final 2025 scoring.

$$p_{cal} = \sigma(\frac{l}{T} + b) \tag{7}$$

The nominal central 80% interval receives a nonnegative expansion derived from calibration conformity scores. Each score is the larger of the lower quantile minus observed visibility and observed visibility minus the upper quantile. The finite-sample quantile index is the ceiling of 0.8 times calibration count plus one, and final bounds are clipped to the retained reporting range. Forecast error is computed in kilometres after inverse transformation. Coverage, width, Brier score, ten-bin probability calibration error and warning confusion counts remain separate metrics. An inclusive comparison tolerance of 0.00001 km accounts for floating-point conversion at the reporting ceiling.

4 Multi Horizon Forecasting and Low Visibility Decision Performance

4.1 Meteorological Coverage and Event Separation

The archive contains 43,739 routine reports from Fresno Airport, station FAT, obtained through the Iowa Environmental Mesonet ASOS service over 2021-2025. Data quality research [18] motivates retaining observation timestamps and availability fields during hourly screening. Figure 3 compares retained weather states and target frequencies. Figure 3(a) relates relative humidity to reported visibility, showing that high humidity spans multiple visibility outcomes and serves as a conditional signal. Figure 3(b) shows 317 low-visibility training targets and 99 final-test targets. This imbalance motivates separate reporting of warning recall, false alarms and forecast errors.

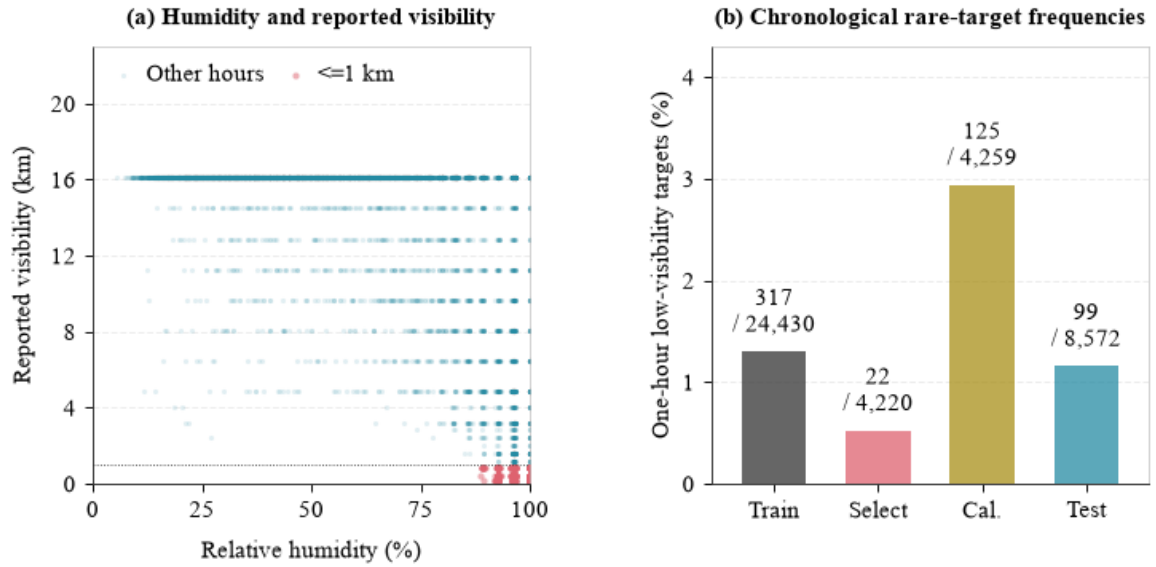


Figure 3. Meteorological coverage and low visibility target distribution. (a) Humidity and visibility relationship. (b) Target frequencies by chronological partition.

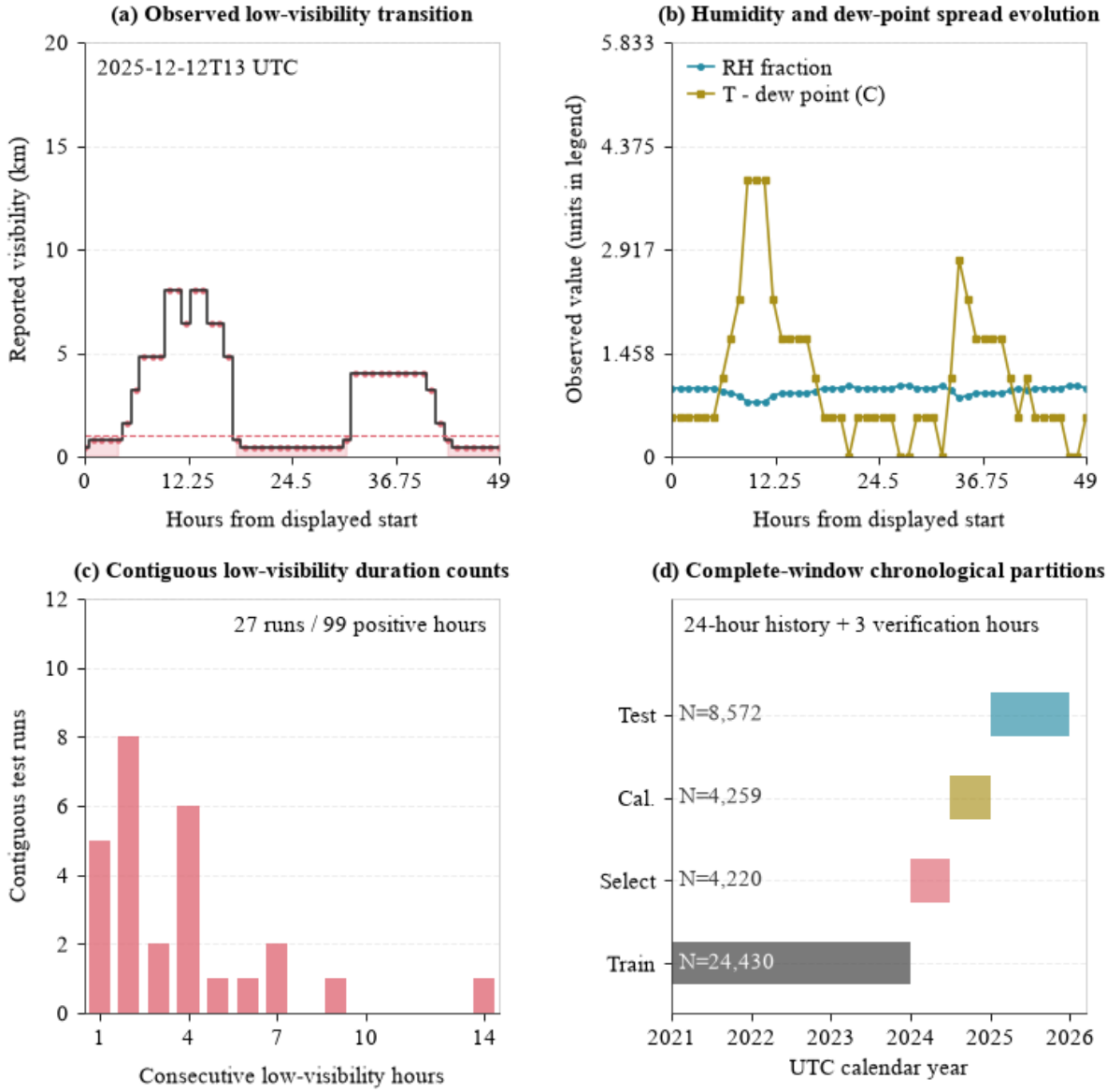


Figure 4. Event evolution and chronological evaluation structure. (a) Visibility trajectories. (b) Humidity and temperature spread transitions. (c) Low visibility run durations. (d) Partition boundaries and complete windows.

The chronological partitions yield 24,430 training, 4,220 checkpoint-selection, 4,259 calibration and 8,572 final-test windows. Chronological validation studies [19] support this separation of fitting and post-fit adjustment. The 2025 test period contains 1,789 one-hour targets below the 16.09344 km reporting ceiling; 79.13% of targets are at the ceiling. At issue, humidity is below 85% in 7,215 windows, 85%-95% in 1,141 and at least 95% in 216. The calibration period has 125 one-hour positive targets and the checkpoint-selection period 22, giving different rare-event populations for model choice and probability adjustment.

Figure 4 traces event evolution and the chronological organization of forecast windows. Figure 4(a) shows recent reported visibility around extended reductions and short minima; each forecast remains linked to its issue time and later verification hour. Figure 4(b) relates humidity and temperature spread to the same transitions. Temporal representation research [20] motivates preserving joint histories because similar final humidity values can arise through different cooling or mixing trajectories. The ordered 24-hour input therefore supplies context beyond the latest observed state.

Figure 4(c) groups the 99 positive test target hours into 27 contiguous runs lasting one to fourteen hours. These clusters give multiple hour-level verification opportunities within fewer weather episodes. Figure 4(d) shows the complete 24-hour history and three target horizons relative to partition boundaries. Dependence-aware evaluation research [21] motivates distinguishing counts of valid forecast origins from counts of independent fog events. Every retained window has the required historical and verification records within its assigned chronological partition; reported recall and F1 refer to target hours.

4.2 Forecast Error and Quantile Reliability

Figure 5 compares forecast errors by lead time and weather condition. Figure 5(a) reports complete-model mean absolute errors 0.462, 0.590 and 0.681 km at one, two and three hours, against persistence 0.406, 0.605 and 0.755 km. Matched forecast comparison research [22] concerns common issue and verification times. Persistence has lower one-hour error, whereas the constrained model has lower two- and three-hour error. Figure 5(b) separates difficult observed regimes. Conditional evaluation studies [23] relate these to 1.931 km one-hour error across 99 low-visibility targets and 0.262 km when origin humidity is below 85%.

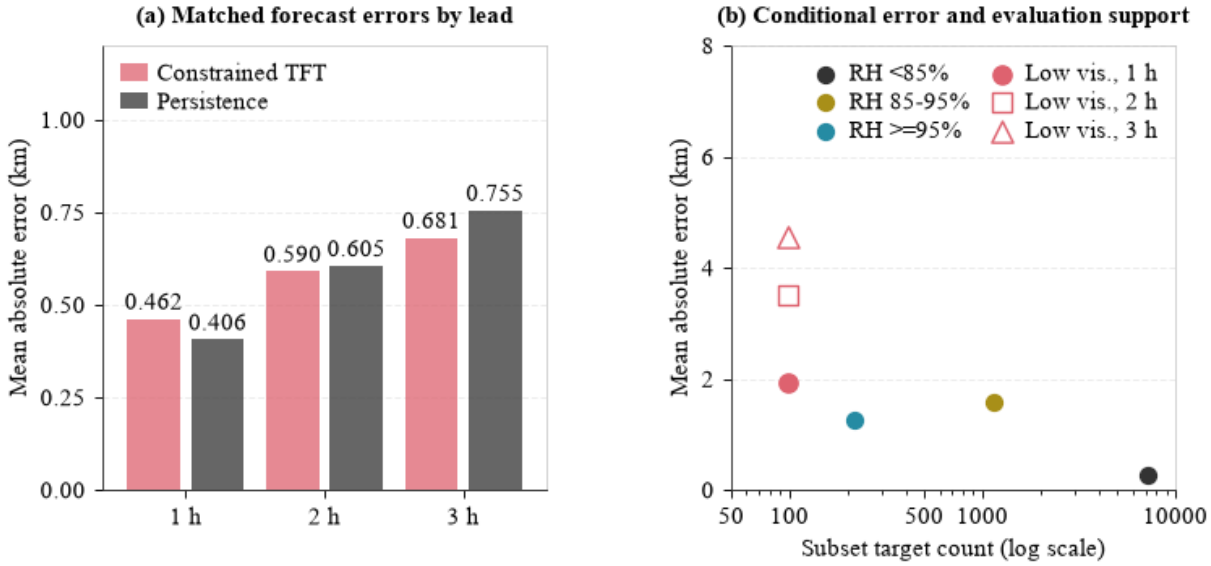


Figure 5. Forecast errors across horizons and weather conditions. (a) Matched model comparison by lead time. (b) Conditional errors by humidity and low visibility state.

The one-hour Root Mean Square Error of the complete model is 1.427 km, which is lower than 1.446 km for persistence and 1.477 km for the recurrent comparator. Robust Scoring Research [24] supports the examination of this indicator along with absolute error, as the two scores place different weights on large deviations. The total model has an absolute error of 1.580 km for the 85%-95% origin humidity group and 1.253 km for the highest humidity group. The intermediate group has many changes; the top group has relatively few changes. The above conditional values are the effect of the frequency of clear periods and the concentration of difficult transitions on the global error. The low-visibility subset gradually deteriorates with increasing horizon, reaching absolute errors of 3.516 and 4.561 km at two and three hours, respectively. As the median is based on log visibility, kilometre errors are also in the inverse-transformed and bounded output range. Both the logarithmic quantile loss for checkpoint selection and the physical unit errors for evaluation have been kept to distinguish between the fitted objective and the final engineering score.

Figure 6 reports interval behavior and calibrated warning probabilities. Figure 6(a) gives overall coverage of the nominal 80% interval: 94.67%, 94.14% and 93.95% over the three horizons. Interval assessment research [25] motivates examining the ceiling-heavy target distribution alongside these values. Figure 6(b) shows mean interval widths 1.268, 1.594 and 1.898 km. Longer lead times therefore have wider reported intervals. The selected nonnegative calibration expansion is zero at every horizon, leaving the original ordered quantile bounds unchanged after this adjustment.

Figure 6(c) compares raw and calibrated event probabilities against observed low-visibility frequency. Probability reliability research [26] motivates retaining proper scores and ten-bin calibration error. At one hour, Brier score decreases from 0.01045 to 0.00596 and ten-bin calibration error from 0.01621 to 0.00407. Figure 6(d) separates all-target interval coverage from below-ceiling coverage. For 1,789 one-hour targets below the reporting cap, coverage is 75.85%; it is 73.63% and 72.75% at two and three hours. The selected temperature is one and intercept minus 1.5 at each horizon, shifting probability level while preserving score ranking. These conditional interval values qualify the high all-target coverage.

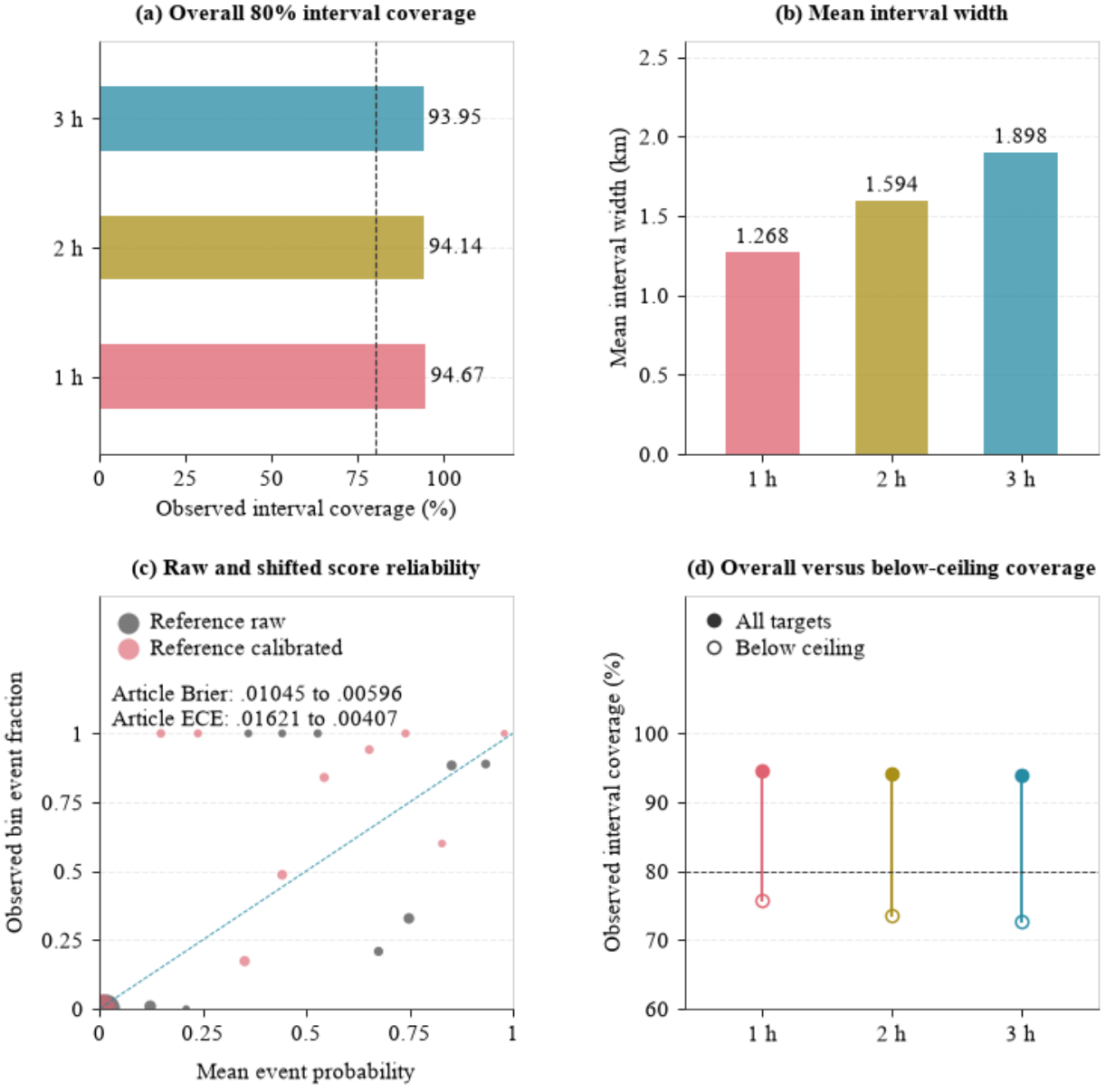


Figure 6. Interval behaviour and probability calibration. (a) Overall interval coverage. (b) Mean interval width. (c) Raw and adjusted probability reliability. (d) Coverage below the reporting ceiling.

4.3 Warning Performance and Observation Sensitivity

Table 1 shows the matched warning and forecasting results for the retained configurations. Decision-oriented assessment research [27] relates these numerical outputs to the actual effects of a probability threshold. The rows are: Lead time, Absolute Error, Interval Coverage, Width, Event Recall, F1 Score and Brier Score. There are a total of 8572 test origins and 99 positive targets in the one-hour network comparison, and each method has its own probability adjustment and warning threshold fitted during the calibration period. The full model has 75 positive targets and

produces 69 false alarms; the unconstrained model has 79 positive targets and 87 false alarms. Persistence has achieved the highest F1 score of 0.727 for one hour, and the complete model has increased the one-hour recall to 75.76% from 72.73%. These values show the various proportions of missed events, false alarms and continuous forecast errors.

Table 1. Forecast interval and warning comparison on the common test population.

Model	Lead (h)	MAE (km)	Coverage (%)	Width (km)	Recall (%)	F1	Brier
Persistence	1	0.406	86.28	0.000	72.73	0.727	0.0060
Recurrent model	1	0.500	93.50	1.322	68.69	0.574	0.0074
Uniform selection	1	0.498	94.00	1.423	71.72	0.612	0.0073
Without attention	1	0.462	93.98	1.233	79.80	0.572	0.0065
Without constraint	1	0.462	94.63	1.280	79.80	0.596	0.0057
Constrained TFT	1	0.462	94.67	1.268	75.76	0.617	0.0060
Constrained TFT	2	0.590	94.14	1.594	63.64	0.477	0.0090
Constrained TFT	3	0.681	93.95	1.898	49.49	0.362	0.0110

Figure 7 reports hour-level warning counts and threshold response. Figure 7(a) gives calibration-selected probability thresholds 0.32, 0.24 and 0.29 for one, two and three hours. Rare-event evaluation research [28] motivates retaining detection counts: the complete model identifies 75, 63 and 49 of 99 positive target hours. Figure 7(b) shows precision-recall changes as thresholds vary. At the selected points, false-positive rates are 0.814%, 1.204% and 1.452%, with F1 scores 0.617, 0.477 and 0.362. Multiple alerts may refer to the same contiguous fog run, so the published counts describe forecast hours.

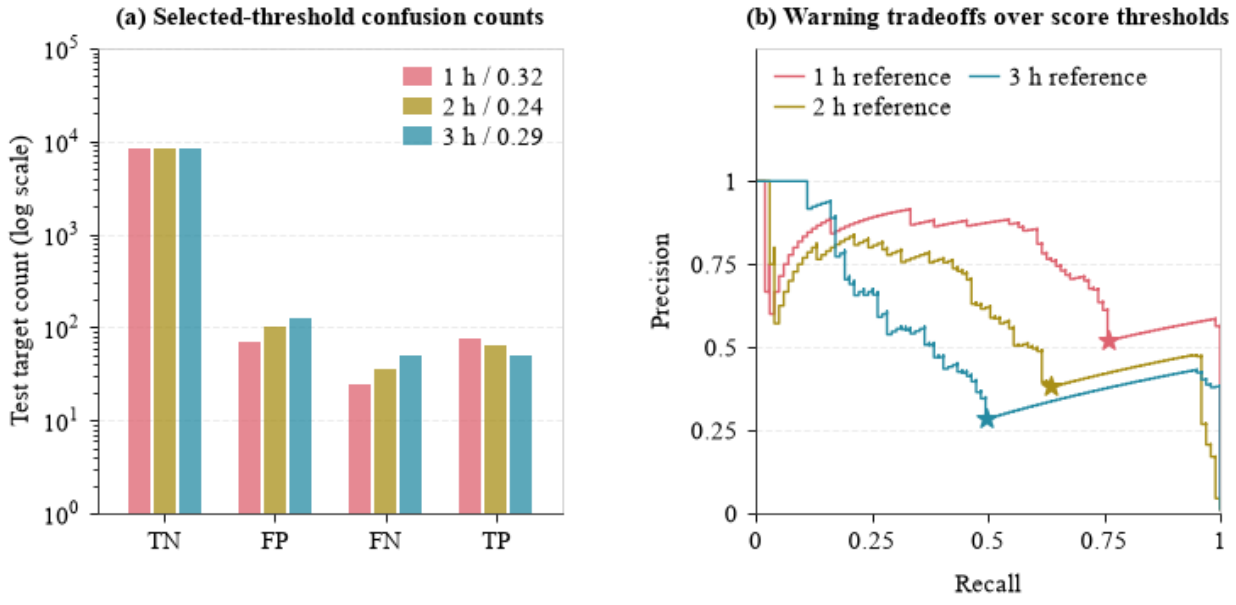


Figure 7. Low visibility warning results and threshold response. (a) Confusion counts recall and precision. (b) Warning tradeoffs across probability thresholds.

Figure 8 assesses observation disturbances and representation components. Figure 8(a) adds independent Gaussian perturbations of standard deviation 0.05, 0.10 or 0.20 in standardized input units to the first eight channels. Robustness assessment research [29] distinguishes this from missing or stale reports. One-hour mean absolute error rises from 0.462 km to 0.466, 0.473 and 0.506 km. Figure 8(b) holds the last six observations of one selected channel at their earlier values. Holding visibility raises error to 0.997 km and reduces recall to 14.14%; holding humidity yields 0.460 km and 73.74% recall. Labels remain fixed across these input interventions.

Figure 8(c) summarizes fitted variable-selection weights for all test windows and low-visibility targets. Representation interpretation research [30] concerns these internal computations. Mean historical-visibility weights are 0.401 and 0.566 in the two groups, while humidity weights are 0.119 and 0.182. Figure 8(d) compares matched removal of humidity regularization, temporal attention and adaptive variable selection. Controlled ablation studies [31] motivate common source windows and training budget. Uniform variable weights yield one-hour error 0.498 km; attention removal yields 0.462 km. Removing the humidity penalty yields one-hour error 0.462 km and F1 0.596, compared with full-model F1 0.617. At three hours, penalty removal gives F1 0.379 versus full-model 0.362, showing horizon-specific trade-offs.

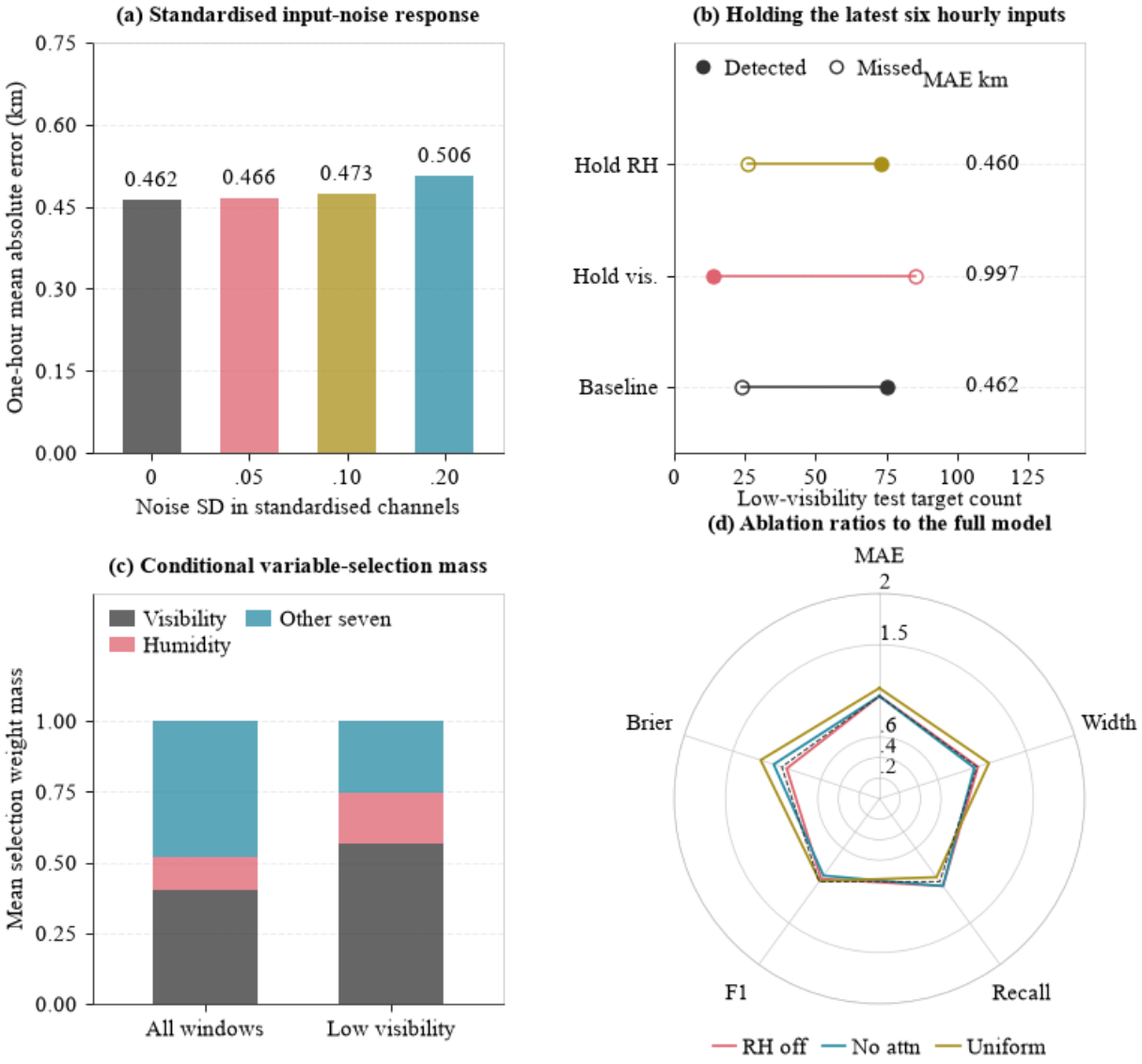


Figure 8. Observation sensitivity and representation ablations. (a) Standardised sensor perturbations. (b) Six-hour channel holding. (c) Conditional variable selection weights. (d) Matched architecture and constraint comparisons.

The complete network has 10,414 registered parameters and the recurrent comparator 3,151. Reproducible learning research [32] motivates retaining source records, partition indices, training-derived standardization, selected weights and calibration coefficients. The main network comparison uses one initialization, and hourly screening excludes incomplete required observations. The station-specific 2025 test assesses chronological transfer under the retained reporting ceiling. Continuous error, ceiling-conditional interval coverage, hour-level confusion counts and alert burden should therefore remain separate criteria when choosing a warning operating point.

5 Conclusion

This study evaluated a humidity-threshold-constrained Temporal Fusion Transformer for airport visibility one, two and three hours after issue. Nine observed meteorological descriptors from a 24-hour history enter variable selection, recurrent processing and temporal attention; known UTC calendar features identify each future horizon. Ordered quantiles predict continuous visibility, while event and auxiliary-humidity heads support a smooth conditional warning

penalty. Chronological periods separate network fitting, checkpoint selection, probability and interval calibration, and final testing at Fresno Airport.

First, across 8,572 test origins, mean absolute errors are 0.462, 0.590 and 0.681 km at one, two and three hours; persistence reaches 0.406, 0.605 and 0.755 km, respectively. Second, one-hour warning recall is 75.76% with F1 0.617 and 69 false alarms, while calibration reduces Brier score from 0.01045 to 0.00596. Persistence retains the higher one-hour F1; removal of the humidity penalty lowers model F1 to 0.596 at one hour but raises it to 0.379 versus 0.362 at three hours. Third, nominal 80% intervals cover 94.67% of all one-hour targets and 75.85% of below-ceiling targets, reflecting the reporting-limit concentration.

The 16.09344 km reporting ceiling contains 79.13% of one-hour test targets, and 99 positive target hours occur in 27 contiguous low-visibility runs. These features affect interval interpretation and the relationship between hour-level alerts and weather episodes. Holding the most recent six visibility observations raises one-hour error to 0.997 km and reduces warning recall to 14.14%, emphasizing dependence on current reports. Independent stations, differing reporting ceilings, observation delays and operational alert policies remain appropriate subjects for external validation.

6. Declarations

6.1 Author Contributions

Hugo Polák contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Kamil Hora contributes to validation, analysis, investigation, data collection, draft preparation. Miloš Příhoda 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 DeepL 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] Yi, K., Zhang, Q., He, H., Shi, K., Hu, L., An, N., & Niu, Z. (2024). Deep coupling network for multivariate time series forecasting. ACM Transactions on Information Systems, 42(5), Article 127, 1–28. https://doi.org/10.1145/3653447

[2] Ying, C., & Lu, J. (2024). TFEformer: Temporal feature enhanced transformer for multivariate time series forecasting. IEEE Access, 12, 153694–153708. https://doi.org/10.1109/ACCESS.2024.3480953

[3] Goyal, K., Dumancic, S., & Blockeel, H. (2024). DeepSaDe: Learning neural networks that guarantee domain constraint satisfaction. Proceedings of the AAAI Conference on Artificial Intelligence, 38(11), 12199–12207. https://doi.org/10.1609/aaai.v38i11.29109

[4] Feng, S., Miao, C., Xu, K., Wu, J., Wu, P., Zhang, Y., & Zhao, P. (2024). Multi-scale attention flow for probabilistic time series

forecasting. IEEE Transactions on Knowledge and Data Engineering, 36(5), 2056–2068. https://doi.org/10.1109/TKDE.2023.3319672

[5] Chen, W., Yang, K., Yu, Z., Shi, Y., & Chen, C. L. P. (2024). A survey on imbalanced learning: Latest research, applications and future directions. Artificial Intelligence Review, 57, 137. https://doi.org/10.1007/s10462-024-10759-6

[6] Dong, J., Jiang, Z., Pan, D., Chen, Z., Guan, Q., Zhang, H., Gui, G., & Gui, W. (2025). A survey on confidence calibration of deep learning-based classification models under class imbalance data. IEEE Transactions on Neural Networks and Learning Systems, 36(9), 15664–15684. https://doi.org/10.1109/TNNLS.2025.3565159

[7] Jensen, V., Bianchi, F. M., & Anfinsen, S. N. (2024). Ensemble conformalized quantile regression for probabilistic time series forecasting. IEEE Transactions on Neural Networks and Learning Systems, 35(7), 9014–9025. https://doi.org/10.1109/TNNLS.2022.3217694

[8] Apicella, A., Isgrò, F., & Prevete, R. (2025). Don't push the button! Exploring data leakage risks in machine learning and transfer learning. Artificial Intelligence Review, 58, 339. https://doi.org/10.1007/s10462-025-11326-3

[9] Raj, S., Deo, R. C., Sharma, E., Prasad, R., Dinh, T., & Salcedo-Sanz, S. (2024). Atmospheric visibility and cloud ceiling predictions with hybrid IIS-LSTM integrated model: Case studies for Fiji's aviation industry. IEEE Access, 12, 72530–72543. https://doi.org/10.1109/ACCESS.2024.3401091

[10] Liu, K., Ding, K., Cheng, X., Xu, G., Hu, X., Liu, T., Feng, S., Cai, B., Chen, J., Lin, H., Song, J., & Zhu, C. (2025). Airport delay prediction with temporal fusion transformers. In Proceedings of the 17th ACM SIGSPATIAL International Workshop on Computational Transportation Science (pp. 5–11). Association for Computing Machinery. https://doi.org/10.1145/3681772.3698212

[11] Cao, C., Huang, J., Wu, M., Lin, Z., & Sun, Y. (2024). A multivariate time series prediction method based on convolution-residual gated recurrent neural network and double-layer attention. Electronics, 13(14), 2834. https://doi.org/10.3390/electronics13142834

[12] Rudolph, M., Kurz, S., & Rakitsch, B. (2024). Hybrid modeling design patterns. Journal of Mathematics in Industry, 14, 3. https://doi.org/10.1186/s13362-024-00141-0

[13] Abdelsamie, M. M., Azab, S. S., & Hefny, H. A. (2026). Deep multi-task learning: A review of concepts, methods, and cross-domain applications. International Journal of Data Science and Analytics, 21, 77. https://doi.org/10.1007/s41060-025-00892-y

[14] Song, M., Yang, D., Lerch, S., Xia, X., Yagli, G. M., Bright, J. M., Shen, Y., Liu, B., Liu, X., & Mayer, M. J. (2024). Non-crossing quantile regression neural network as a calibration tool for ensemble weather forecasts. Advances in Atmospheric Sciences, 41(7), 1417–1437. https://doi.org/10.1007/s00376-023-3184-5

[15] 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

[16] Schlembach, F., Smirnov, E., Koprinska, I., & Winands, M. H. M. (2025). Conformal multistep-ahead multivariate time-series forecasting. Machine Learning, 114, 165. https://doi.org/10.1007/s10994-024-06722-9

[17] von Stackelberg, P., Goedhart, R., Birbil, Ş. İ., & Does, R. J. M. M. (2024). Comparison of threshold tuning methods for predictive monitoring. Quality and Reliability Engineering International, 40(1), 499–512. https://doi.org/10.1002/qre.3436

[18] Li, N., Qi, Y., Li, C., & Zhao, Z. (2024). Active learning for data quality control: A survey. Journal of Data and Information Quality, 16(2), Article 11, 1–45. https://doi.org/10.1145/3663369

[19] Chen, M., Shen, L., Fu, H., Li, Z., Sun, J., & Liu, C. (2024). Calibration of time-series forecasting: Detecting and adapting context-driven distribution shift. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 341–352). Association for Computing Machinery. https://doi.org/10.1145/3637528.3671926

[20] Liu, W., He, Y., Guan, J., & Zhou, S. (2024). Multivariate time series forecasting with causal-temporal attention network. In ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 5735–5739). IEEE. https://doi.org/10.1109/ICASSP48485.2024.10448031

[21] Adin, A., Krainski, E. T., Lenzi, A., Liu, Z., Martínez-Minaya, J., & Rue, H. (2024). Automatic cross-validation in structured models: Is it time to leave out leave-one-out? Spatial Statistics, 62, 100843. https://doi.org/10.1016/j.spasta.2024.100843

[22] Beck, N., Dovern, J., & Vogl, S. (2025). Mind the naive forecast! A rigorous evaluation of forecasting models for time series with low predictability. Applied Intelligence, 55, 395. https://doi.org/10.1007/s10489-025-06268-w

[23] Cerqueira, V., Roque, L., & Soares, C. (2025). Modelradar: Aspect-based forecast evaluation. Machine Learning, 114, 229. https://doi.org/10.1007/s10994-025-06877-z

[24] Terven, J., Cordova-Esparza, D.-M., Romero-González, J.-A., Ramírez-Pedraza, A., & Chávez-Urbiola, E. A. (2025). A

comprehensive survey of loss functions and metrics in deep learning. Artificial Intelligence Review, 58, 195. https://doi.org/10.1007/s10462-025-11198-7

[25] Sousa, M., Tomé, A. M., & Moreira, J. (2024). A general framework for multi-step ahead adaptive conformal heteroscedastic time series forecasting. Neurocomputing, 608, 128434. https://doi.org/10.1016/j.neucom.2024.128434

[26] Lane, R. O. (2026). A comprehensive review of classifier probability calibration metrics. Data Science, 9(1), 24518492261436231. https://doi.org/10.1177/24518492261436231

[27] Yang, M., & Bi, X. (2025). Cost-aware calibration of classifiers. INFORMS Journal on Data Science, 4(2), 101–113. https://doi.org/10.1287/ijds.2024.0038

[28] Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. Transactions of the Association for Computational Linguistics, 12, 820–836. https://doi.org/10.1162/tacl_a_00675

[29] Inácio, R., Cerqueira, V., Barandas, M., & Soares, C. (2025). Mast: Interpretable stress testing via meta-learning for forecasting model robustness evaluation. Machine Learning, 114, 251. https://doi.org/10.1007/s10994-025-06881-3

[30] Su, Q., Kloukinas, C., & d'Avila Garcez, A. (2024). FocusLearn: Fully-interpretable, high-performance modular neural networks for time series. In 2024 International Joint Conference on Neural Networks (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN60899.2024.10651481

[31] Shao, Z., Wang, F., Xu, Y., Wei, W., Yu, C., Zhang, Z., Yao, D., Sun, T., Jin, G., Cao, X., Cong, G., Jensen, C. S., & Cheng, X. (2025). Exploring progress in multivariate time series forecasting: Comprehensive benchmarking and heterogeneity analysis. IEEE Transactions on Knowledge and Data Engineering, 37(1), 291–305. https://doi.org/10.1109/TKDE.2024.3484454

[32] Semmelrock, H., Ross-Hellauer, T., Kopeinik, S., Theiler, D., Haberl, A., Thalmann, S., & Kowald, D. (2025). Reproducibility in machine-learning-based research: Overview, barriers, and drivers. AI Magazine, 46(2), e70002. https://doi.org/10.1002/aaai.70002