AIRPORTMIST: A MULTIMODAL BENCHMARK FOR AIRPORT WEATHER-STATE FORECASTING WITH RARE ADVERSE CONDITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Airport operations depend on timely forecasts of operationally meaningful local states, including rare adverse conditions such as thunderstorms and degraded flight conditions, yet existing weather and aviation benchmarks primarily focus on continuous variables, gridded atmospheric fields, or individual hazards. We introduce AirportMIST, a multimodal benchmark for multi-airport, multi-horizon categorical airport weather-state forecasting. AirportMIST covers 37 major civil airports in China during JJA 2020 and JJA 2025 and integrates high-frequency AWOS observations, irregular METAR reports, regional WAFS forecast fields in GRIB format, and static airport metadata. It defines precipitation/thunderstorm and flight-condition forecasting at 6-, 12-, and 24-hour horizons, together with chronological splits, causality-preserving multimodal alignment, and rare-event evaluation. We further provide MISTNet as a reproducible reference model for heterogeneous local and regional inputs. Experiments with operational, statistical, and time-series baselines reveal that rare-event ranking, multiclass argmax detection, and operational forecast behavior can lead to different conclusions: several learned models retain nonzero AP while making no Thunder predictions under multiclass argmax, whereas risk-aware TAF forecasts operate at a substantially higher predicted-positive rate. These results motivate evaluating airport weather-state forecasting beyond aggregate classification performance. AirportMIST provides a standardized and reproducible foundation for research on multimodal airport weather-state forecasting and rare-event prediction. Our code is available at `https://anonymous.4open.science/r/MIST-1FA1/`.

1 INTRODUCTION

Airport operations depend on timely assessments of local weather conditions that directly affect visibility, ceiling, precipitation, convection, and other operational states. In practice, aviation-weather guidance is communicated through products such as Terminal Aerodrome Forecasts (TAFs), while many downstream planning and safety-related decisions depend on whether an airport will enter operationally meaningful states, particularly rare adverse conditions such as Thunder and IFR/LIFR. Forecasting these states is challenging because high-impact conditions are often rare, their prevalence varies substantially across airports and forecast horizons, and the relevant evidence is distributed across heterogeneous observational and numerical sources. A systematic machine-learning benchmark for multi-airport, multi-horizon airport weather-state forecasting could therefore support reproducible development and comparison of data-driven forecasting methods.

Existing airport-weather studies have investigated visibility, ceiling, fog, precipitation, and related hazards using airport observations, numerical weather prediction products, statistical post-processing, and machine-learning models Cordeiro et al. (2021); Wolters et al. (2025); Boneh et al. (2015). More broadly, recent data-driven weather forecasting methods have demonstrated the value of learning from large-scale atmospheric fields and local observations Bi et al. (2022); Pathak et al. (2022); Nguyen et al. (2023); Schmude et al. (2024), while multimodal time-series models and benchmarks provide mechanisms for integrating heterogeneous sources with irregular sampling, missingness, and cross-modal interactions Liu et al. (2024a); Chang et al. (2025); Zhong et al.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

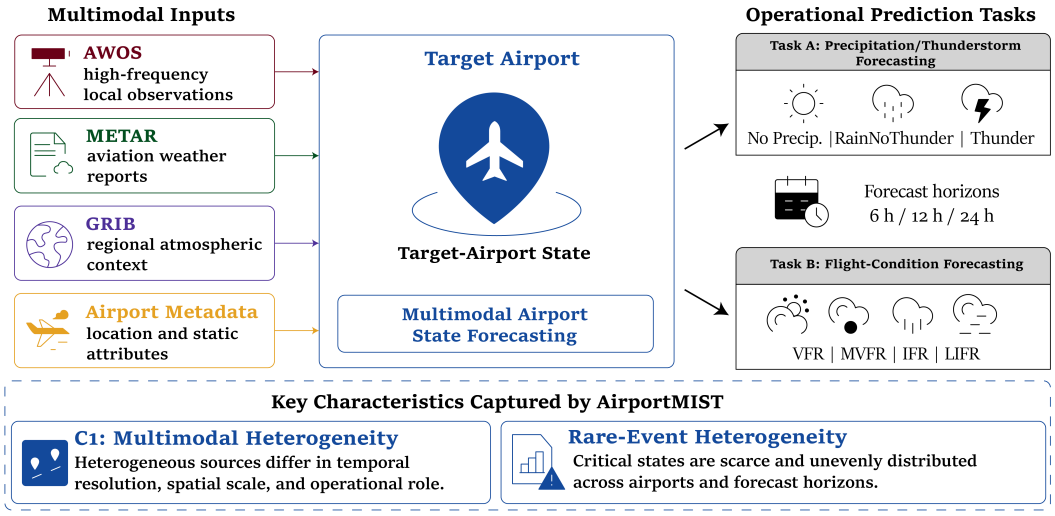

Figure 1: Overview of AirportMIST, a multimodal benchmark for multi-airport, multi-horizon airport weather-state forecasting with heterogeneous inputs and rare-event targets.

(2025); Huang et al. (2025). Together, these studies provide a strong methodological foundation for data-driven airport-weather forecasting.

Despite this progress, existing airport-weather research remains fragmented across individual hazards, meteorological variables, airports, and evaluation settings. There is still no standardized benchmark for jointly studying high-frequency airport observations, irregular aviation weather reports, regional numerical forecast fields, and static airport information under a common multi-airport, multi-horizon categorical forecasting protocol. In addition, operationally important events are strongly imbalanced and spatially heterogeneous, so aggregate multiclass metrics alone may not reveal meaningful differences in rare-event ranking and detection behavior. These gaps make systematic method comparison difficult.

To address these gaps, we introduce **AirportMIST** (**M**ultimodal **I**nputs for **S**tate **T**asks at Airports), a multimodal benchmark for multi-airport, multi-horizon categorical airport weather-state forecasting across 37 major civil airports in China. AirportMIST integrates airport-local observations, aviation weather reports, regional numerical forecasts, and static airport metadata under standardized task definitions, causality-preserving temporal alignment, chronological evaluation, and rare-event analysis.

We further provide **MISTNet** as a reproducible reference model for the benchmark. Experiments with operational, statistical, and time-series baselines reveal distinct failure modes across model families, including rare-event decision collapse under multiclass argmax and different operating points for TAF-based forecasts. These findings highlight the need for benchmark evaluation beyond a single aggregate metric.

Our main contributions are summarized as follows:

AirportMIST benchmark. We introduce a multimodal benchmark for multi-airport, multi-horizon airport weather-state forecasting, with standardized operational labels, causality-preserving multimodal alignment, chronological splits, and rare-event evaluation.

Reference modeling and benchmark analysis. We provide MISTNet as a reproducible reference model and conduct systematic comparisons that characterize multiclass performance, rare-event ranking and detection behavior, and operational TAF trade-offs.

2 Related Work

2.1 Weather, Aviation, and Multimodal Forecasting Benchmarks

Existing weather and climate benchmarks primarily target continuous meteorological variables or gridded atmospheric fields. Reanalysis products such as ERA5 and benchmark suites such as WeatherBench support global or regional continuous forecasting, while recent data-driven weather models including Pangu-Weather Bi et al. (2022), FourCastNet Pathak et al. (2022), ClimaX Nguyen et al. (2023), and Prithvi WxC Schmude et al. (2024) are evaluated in such settings. These resources do not define airport-level categorical operational states or the cross-modal alignment required when local observations, aviation reports, and regional forecast fields are combined.

Multimodal time-series benchmarks such as Time-MMD Liu et al. (2024a) and Time-IMM Chang et al. (2025) address heterogeneous sources, irregular sampling, and missingness, but focus on general multimodal time-series learning rather than airport-specific operational thresholds and rare events. Aviation-weather studies have investigated visibility, ceiling, fog, and related hazards using airport observations and numerical forecast products Cordeiro et al. (2021); Wolters et al. (2025); Boneh et al. (2015), yet typically focus on individual variables, hazards, or limited airport settings. AeroWF Li et al. (2026) further studies multimodal aerodrome representation learning using aligned airport-local AWOS and METAR data across four airports, with downstream forecasting, imputation, classification, and few-shot transfer tasks. AirportMIST extends this line of research toward a broader benchmark setting by incorporating regional WAFS forecast fields, expanding airport coverage, and standardizing multi-horizon categorical airport weather-state forecasting with explicit forecast-time causality and dedicated rare-event evaluation. To our knowledge, no existing public benchmark combines AWOS, METAR, GRIB, and airport metadata for multi-airport, multi-horizon categorical airport weather-state forecasting with explicit forecast-time causality. AirportMIST is designed to fill this gap.

2.2 Multimodal Spatiotemporal Representation Learning

Spatiotemporal forecasting methods model temporal dynamics together with structured spatial dependencies, with representative approaches including DCRNN, STGCN, Graph WaveNet, and MTGNN Li et al. (2018); Yu et al. (2018); Wu et al. (2019; 2020). General time-series models such as DLinear, PatchTST, iTransformer, and FreTS provide alternative architectures based on linear decomposition, patching, variable-centric attention, and frequency-domain modeling Zeng et al. (2023); Nie et al. (2023); Liu et al. (2024b); Yi et al. (2023). Recent multimodal time-series methods further study heterogeneous-source interaction, irregular observations, and missingness Liu et al. (2024a); Chang et al. (2025); Zhong et al. (2025); Huang et al. (2025).

2.3 Rare-Event Learning

Rare-event forecasting further introduces severe class and group imbalance. Existing strategies include resampling, focal and class-balanced objectives, and group-aware optimization Lin et al. (2020); Cui et al. (2019); Sagawa et al. (2019). In AirportMIST, event prevalence varies across airports and forecast horizons; MISTNet therefore includes reliability-aware group reweighting as one reference strategy, while the benchmark evaluates rare-event behavior independently through dedicated event-oriented metrics.

3 The AirportMIST Benchmark

AirportMIST is a multimodal benchmark for multi-airport, multi-horizon categorical airport weather-state forecasting. It integrates heterogeneous airport-local observations, aviation weather reports, regional numerical forecast fields, and static airport information, and standardizes sample construction, forecasting tasks, chronological data splits, causality-preserving multimodal alignment, and evaluation protocols for both overall classification and rare-event forecasting. Table 1 summarizes the core properties of the benchmark.

Table 1: Summary of the AirportMIST benchmark.

Property	Value	Property	Value
Airports	37	Period	JJA 2020 and JJA 2025
Input history	12 h	Forecast horizons	6 / 12 / 24 h
Modalities	AWOS, METAR, GRIB, metadata	Tasks	Precip./Thunder; Flight condition
Total anchors	205,496	Geographic scope	China
Train	135,776	Validation	35,059
Test	34,661	Rare events	Thunder; IFR-or-worse
AWOS	720×35	METAR	up to 24×19
GRIB	$12 \times 15 \times 15$	Valid labels	614,425 (P) / 614,419 (F)

P and F denote the precipitation/thunderstorm and flight-condition tasks, respectively.

3.1 Data and Sample Construction

AirportMIST covers 37 major civil airports in China during the boreal summers (JJA) of 2020 and 2025. The airport set is determined by the availability of sufficiently complete cross-modal data and matching airport metadata across both study periods. The temporal scope is likewise constrained by the availability of operational data with reliable alignment among AWOS, METAR, GRIB, and airport metadata. After preprocessing and deduplication, the dataset contains 9,400,318 observed airport-minute AWOS records and 213,449 METAR/SPECI reports.

The basic benchmark unit is a forecast anchor, identified by an airport s and a METAR report time t. Each anchor contains a 12-hour multimodal history ending at t and may be associated with up to three horizon-specific targets at $t + h$, where $h \in \{6, 12, 24\}$ hours. AWOS provides high-frequency airport-local observations, METAR contributes irregular aviation weather reports, and regional WAFS forecast fields distributed in GRIB format provide airport-centered atmospheric context. Static airport metadata provide location and elevation information, while modality-specific missingness and validity masks are retained explicitly. Table 1 summarizes the core dimensions; full preprocessing details are provided in Appendix A.

To ensure forecast-time causality, all timestamped inputs are restricted using forecast-time causal eligibility rules. For GRIB, eligibility is determined by the forecast initialization/reference time, which must not exceed the anchor time. Among eligible records, AirportMIST selects the record whose valid time is closest to the anchor within a 3-hour tolerance, with ties resolved in favor of the most recent initialization.

For each forecast horizon, the nominal target time is $t + h$. The nearest METAR report within ± 15 minutes of this time is used to construct the target. AirportMIST therefore defines a point-forecast setting: each label represents the airport state reported near the corresponding valid time, rather than whether an event occurs at any time within the preceding forecast interval. Horizon-specific targets that do not satisfy the matching or validity requirements are excluded from training and evaluation. Detailed feature definitions, tensor organization, anchor construction, and multimodal alignment rules are provided in Appendix A.

Chronological split. AirportMIST uses JJA 2020 and June 2025 for training, July 2–31, 2025 for validation, and August 2–31, 2025 for testing, with July 1 and August 1 reserved as temporal buffers. The resulting split contains 135,776 training anchors, 35,059 validation anchors, and 34,661 test anchors. All 37 airports are present in the standard chronological split. Feature-normalization statistics are estimated using the training split only; hyperparameter selection and early stopping use validation data, while the test split is reserved exclusively for final evaluation.

3.2 Forecasting Tasks and Labels

AirportMIST defines two categorical forecasting tasks at each horizon $h \in \{6, 12, 24\}$ hours. The precipitation/thunderstorm task uses the three-class target $\{\mathrm{NoRain}, \mathrm{RainNoThunder}, \mathrm{Thunder}\}$. Labels are derived from METAR present-weather codes, with thunderstorm indicators taking precedence over ordinary precipitation. Models are trained on the full three-class target, while **Thunder** is treated as the rare event for event-oriented evaluation.

Table 2: Rare-event prevalence and spatial heterogeneity in AirportMIST.

Statistic	Thunder	IFR-or-worse
Overall prevalence	1.399%	6.697%
Test prevalence	1.435–1.539%	2.775–2.825%
Airport range (train)	0.141–3.542%	0.377–21.680%
Max/min airport ratio	$\sim 25\times$	$\sim 57.5\times$

Table 3: Evaluation measures used in AirportMIST.

Measure	Interpretation
Macro-F1	Overall multiclass
AP	Rare-event ranking
POD / CSI	Argmax detection
Pred.+	Prediction frequency

The flight-condition task uses the four-class target $\{\mathrm{VFR}, \mathrm{MVFR}, \mathrm{IFR}, \mathrm{LIFR}\}$, constructed from METAR visibility and ceiling information according to the most restrictive applicable category. Models are trained on all four classes, while IFR and LIFR are combined only for rare-event evaluation as **IFR-or-worse**. Complete weather-code mappings and flight-condition label definitions are provided in Appendix A.4.

3.3 Dataset Statistics

AirportMIST is characterized by severe rare-event imbalance, substantial spatial heterogeneity, and temporal distribution shift. Table 2 summarizes the prevalence of the two rare events. Across all splits and forecast horizons, Thunder accounts for only 1.399% of valid precipitation/thunderstorm labels, while IFR-or-worse accounts for 6.697% of valid flight-condition labels. More importantly, their prevalence varies substantially across airports: in the training split, the maximum-to-minimum prevalence ratio is approximately $25\times$ for Thunder and $57.5\times$ for IFR-or-worse.

The benchmark also exhibits temporal base-rate shift. For the flight-condition task, IFR-or-worse prevalence decreases from approximately 8.49% in training to 2.78–2.83% in testing. These properties expose complementary challenges in learning globally rare events and maintaining stable behavior under changing class prevalence. They motivate the use of both overall classification metrics and dedicated rare-event ranking and detection measures.

3.4 Evaluation Protocol

AirportMIST evaluates model behavior from complementary perspectives. Macro-F1 summarizes overall multiclass performance, while average precision (AP) evaluates rare-event ranking quality before the multiclass argmax decision. Probability of detection (POD) and critical success index (CSI) characterize rare-event detection under the common argmax operating point, and the predicted-positive rate (Pred.+) indicates how frequently a model predicts the rare class. The observed event prevalence is summarized in Section 3.3.

For probabilistic models, AP is computed from continuous rare-event scores: $p(\mathrm{Thunder})$ for precipitation/thunderstorm forecasting and $p(\mathrm{IFR}) + p(\mathrm{LIFR})$ for flight-condition forecasting. Persistence and TAF produce deterministic categorical predictions, so their AP values are computed from binary event decisions and are reported as hard-decision references.

3.5 Data Availability and Extensibility

AirportMIST separates task-neutral anchor metadata from horizon-specific targets, while historical inputs are constructed at runtime from the underlying timestamped source data. This organization decouples input construction from target definition and supports alternative input-history lengths without rebuilding the benchmark. The official benchmark uses a 12-hour input history and forecast horizons of 6, 12, and 24 hours.

The complete AirportMIST benchmark, including the processed multimodal data, splits, labels, metadata, and supporting construction and evaluation resources, will be publicly released upon acceptance under the applicable licenses and access conditions.

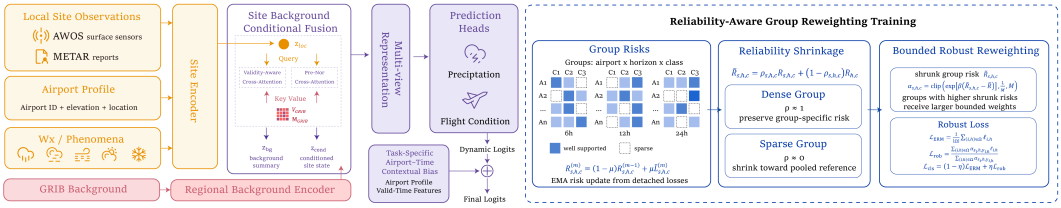

Figure 2: Overview of MISTNet, the reference model used in AirportMIST. MISTNet separately encodes airport-local observations and regional GRIB fields, integrates regional atmospheric context conditioned on the local airport state, and performs multi-horizon prediction. During training, reliability-aware group reweighting is applied over airport–horizon–class groups.

4 METHODOLOGY

4.1 OVERVIEW

Figure 2 illustrates **MISTNet**, the reference model used to instantiate the AirportMIST benchmark. Given the information available at airport s and forecast anchor time t, MISTNet separately models airport-local observations and regional atmospheric context, integrates the two representations, and performs horizon-specific categorical prediction. The architecture reflects the asymmetric roles of the benchmark modalities: AWOS and METAR provide direct evidence about the target airport, whereas GRIB fields describe the surrounding atmospheric background. For the benchmark's heterogeneous rare-event setting, training additionally uses reliability-aware group reweighting and auxiliary supervision. These training-specific components are not required at inference.

4.2 MISTNET ARCHITECTURE

Local and regional encoding. MISTNet first constructs an airport-local representation from AWOS, METAR, and static airport information. AWOS and METAR are encoded separately to preserve their different temporal resolutions, semantics, and missingness patterns. For each source, standardized observations are augmented with explicit missingness, observation-age, and source/type information, followed by source-specific temporal encoding and masked aggregation. Static airport attributes and sparse weather-phenomenon indicators are subsequently incorporated into the local representation.

Abstracting these source-specific encoders, the local and regional representations can be written as $\mathbf{z}^{\mathrm{loc}}_{s,t} = \mathrm{LocalEnc}\left(\mathbf{X}^{\mathrm{AWOS}}_{s,t}, \mathbf{X}^{\mathrm{METAR}}_{s,t}, \mathbf{x}^{\mathrm{meta}}_{s}\right)$, and $\mathbf{V}^{\mathrm{grid}}_{s,t} = \mathrm{GridEnc}\left(\mathbf{X}^{\mathrm{GRIB}}_{s,t}, \mathbf{M}^{\mathrm{grid}}_{s,t}\right)$, where $\mathbf{z}^{\mathrm{loc}}_{s,t}$ summarizes information directly available at the target airport, and $\mathbf{V}^{\mathrm{grid}}_{s,t}$ denotes the spatially distributed regional GRIB tokens. The mask $\mathbf{M}^{\mathrm{grid}}_{s,t}$ prevents invalid or unavailable grid locations from contributing to the regional representation.

Site–background fusion. Airport-local observations provide direct evidence about the target airport, whereas GRIB fields provide broader atmospheric context. MISTNet therefore conditions regional information extraction on the current airport-local state by using the local representation as a query over the regional GRIB tokens: $\mathbf{z}^{\mathrm{reg}}_{s,t} = \mathrm{Attn}\left(\mathbf{z}^{\mathrm{loc}}_{s,t}, \mathbf{V}^{\mathrm{grid}}_{s,t}; \mathbf{M}^{\mathrm{grid}}_{s,t}\right)$. In implementation, MISTNet uses two independently parameterized attention branches with residual refinement to obtain complementary regional summaries. These regional representations are subsequently combined with the airport-local representation for prediction.

Multi-horizon prediction. For each forecast horizon $h \in \mathcal{H}$, the local and regional representations are combined with a learned horizon embedding $\mathbf{e}_h$ to form a horizon-specific prediction state. The prediction process can be summarized as $\mathbf{q}_{s,t,h} = \mathrm{Fuse}\left(\mathbf{z}^{\mathrm{loc}}_{s,t}, \mathbf{z}^{\mathrm{reg}}_{s,t}, \mathbf{e}_h\right)$ and $\mathbf{p}_{s,t,h} = \mathrm{Softmax}\left(f_{\mathrm{task}}(\mathbf{q}_{s,t,h}) + \mathbf{b}^{\mathrm{ctx}}_{s,t,h}\right), \qquad \hat{y}_{s,t,h} = \arg\max_c p_{s,t,h,c}$, where $\mathbf{p}_{s,t,h}$ denotes the predicted class-probability vector, $\hat{y}_{s,t,h}$ is the corresponding categorical prediction, f_{task} denotes the task-specific classifier, and $\mathbf{b}^{\mathrm{ctx}}_{s,t,h}$ is a contextual bias derived from static airport information and cyclic forecast-valid-time features. The shared local and regional representations are

reused across forecast horizons, while the horizon embedding allows the prediction head to adapt to horizon-specific forecasting difficulty.

4.3 Reliability-Aware Training for Rare Events

AirportMIST exhibits substantial variation in rare-event prevalence and empirical difficulty across airports and forecast horizons. A single global class weight cannot capture this heterogeneity, while directly emphasizing fine-grained airport–horizon–class groups may be unstable when some groups contain only a small number of observations. As a reference strategy for this setting, MISTNet combines tempered inverse-frequency sampling with reliability-aware group reweighting.

We first use tempered inverse-frequency sampling to increase the exposure of rare classes during training without fully equalizing the class distribution. For group reweighting, each training group is defined as $g = (s, h, c)$, where s, h, and c denote airport, forecast horizon, and class, respectively. Let $R_{s,h,c}$ denote the running empirical risk of a group, $R_{h,c}$ the corresponding pooled class–horizon risk, and $n_{s,h,c}$ its training support. To reduce the influence of noisy empirical risks from sparsely observed groups, we shrink the group-specific estimate toward the pooled reference using $\rho_{s,h,c} = n_{s,h,c}/(n_{s,h,c} + \tau)$ and $\widetilde{R}_{s,h,c} = \rho_{s,h,c} R_{s,h,c} + (1 - \rho_{s,h,c}) R_{h,c}$.

Well-supported groups therefore retain more of their group-specific risk, whereas sparse groups rely more strongly on the pooled class–horizon estimate. The shrunk risks are converted into bounded exponential group weights and combined with the standard multiclass classification objective. This allows training to emphasize empirically difficult groups while reducing sensitivity to unreliable risk estimates from very small groups.

MISTNet additionally uses auxiliary binary objectives that reflect nested operational states within the two benchmark tasks. For precipitation/thunderstorm forecasting, the auxiliary supervision first distinguishes NoRain from the two non-NoRain states and then distinguishes Thunder from Rain-NoThunder within the non-NoRain subset. For flight-condition forecasting, an auxiliary objective models IFR-or-worse conditional on degraded flight conditions. These auxiliary heads operate on the same horizon-specific representation $\mathbf{q}_{s,t,h}$ as the main multiclass classifiers and are optimized jointly during training. They are discarded at inference and therefore do not increase inference-time complexity. Complete architectural, training, and implementation details are provided in Appendix C.

5 Experiments

5.1 Baselines and Implementation Details

Baselines. We compare MISTNet with three groups of reference methods. First, **Persistence** predicts future airport states using the most recently observed state and serves as a simple operational baseline. We also include two TAF-based baselines. **TAF prevailing** uses the prevailing forecast condition valid at the target time, whereas **TAF risk-aware** additionally incorporates relevant temporary, probabilistic, or transition groups when they indicate a more adverse state. The resulting TAF conditions are mapped to the corresponding AirportMIST target categories. Because TAF availability is incomplete at some forecast horizons, coverage is reported together with predictive performance.

Second, we evaluate representative general-purpose time-series forecasting models, including DLinear Zeng et al. (2023), FreTS Yi et al. (2023), PatchTST Nie et al. (2023), iTransformer Liu et al. (2024b), TimeMixer Wang et al. (2024), and TQNet Lin et al. (2025). Third, we include LightGBM Ke et al. (2017) as a strong non-neural tabular baseline. All learnable baselines receive AWOS, METAR, GRIB, and airport metadata; generic time-series models use a shared early-fusion adapter, while MISTNet preserves source-specific and regional spatial structure. Detailed baseline input adaptation and architecture-specific configurations are provided in Appendix B.3.

Implementation details. All learnable methods use the same chronological training, validation, and test splits defined in Section 3. We train a separate model for each forecasting task, while each model jointly predicts the 6-, 12-, and 24-hour horizons. Hyperparameters and checkpoints are selected using validation data only, and the test set is reserved for final evaluation. Unless otherwise

Table 4: Main benchmark results for both forecasting tasks. Macro-F1, POD, CSI, AP and Pred.+ for learnable methods are averaged over seeds 42, 43, and 44. Best and second-best results among full-coverage learnable methods are shown in **bold** and underlined, respectively. Operational baselines are reported for reference and are not included in this ranking. Pred.+ denotes the predicted positive rate and is not a metric to be maximized. TAF coverage at 24 h is 60.1%; all other reported methods have full coverage.

Task	Method	6 h					12 h					24 h				
		M-F1	POD	CSI	AP	Pred.+	M-F1	POD	CSI	AP	Pred.+	M-F1	POD	CSI	AP	Pred.+
Flight	Persistence	.3471	.1960	.1038	.0578	.0301	.2923	.0805	.0403	.0318	.0303	.3204	.1638	.0859	.0487	.0303
	TAF Prevailing	.1615	.4413	.0740	.0515	.1500	.1567	.3630	.0601	.0422	.1516	.1481	.3502	.0550	.0419	.1801
	TAF Risk-Aware	.1631	.5136	.0767	.0560	.1724	.1568	.4236	.0637	.0457	.1705	.1465	.3659	.0546	.0420	.1913
	DLinear	.3618	.1897	.1296	.1326	.0276	.3113	.1464	.0821	.0871	.0516	.3076	.0571	.0405	.0825	.0410
	FreTS	.2920	.0849	.0508	.0759	.0210	.2856	.1172	.0609	.0798	.0292	.2547	.0296	.0187	.0531	.0173
	PatchTST	.3149	.2338	.1019	.1105	.0424	.2647	.0000	.0000	.0805	.0000	.2302	.0000	.0000	.0582	.0000
	iTransformer	.3445	.1761	.1105	.1511	.0214	.3205	.1464	.0810	.1047	.0268	.3125	.0793	.0535	.0750	.0159
	TimeMixer	.3271	.1384	.0882	.1293	.0196	.2914	.0847	.0543	.1018	.0181	.2734	.1184	.0625	.0893	.0286
	TQNet	.3302	.2704	.1524	.1979	.0290	.3038	.0481	.0374	.1134	.0094	.3081	.1459	.0848	.1055	.0245
	LightGBM	**.3854**	.2369	.1707	**.2651**	.0173	.3386	.1297	.0984	.1644	.0126	.3225	.1300	.0810	.1264	.0208
	MISTNet	.3847	**.3906**	**.1901**	.2542	.0363	**.3672**	**.3585**	**.1553**	**.1931**	.0466	**.3367**	**.3042**	**.1321**	**.1550**	.0608
Precip.	Persistence	.4480	.0662	.0291	.0176	.0206	.4003	.0348	.0144	.0147	.0207	.3636	.0350	.0147	.0149	.0206
	TAF Prevailing	.4544	.0624	.0386	.0201	.0105	.4446	.0634	.0373	.0187	.0109	.4074	.0181	.0126	.0169	.0076
	TAF Risk-Aware	.4827	.4045	.0804	.0461	.0683	.4636	.3272	.0678	.0354	.0596	.4501	.2175	.0614	.0300	.0454
	DLinear	.3810	.0095	.0074	.0333	.0041	.3649	.0061	.0044	.0229	.0068	.3359	.0267	.0167	.0193	.0091
	FreTS	.3380	.0000	.0000	.0169	.0000	.3192	.0000	.0000	.0167	.0000	.3171	.0000	.0000	.0149	.0000
	PatchTST	.4076	.0000	.0000	.0177	.0000	.3170	.0000	.0000	.0140	.0000	.3171	.0000	.0000	.0160	.0000
	iTransformer	.3191	.0000	.0000	.0256	.0000	.3170	.0000	.0000	.0253	.0000	.3171	.0000	.0000	.0196	.0000
	TimeMixer	.3167	.0000	.0000	.0413	.0000	.3170	.0000	.0000	.0326	.0000	.3171	.0000	.0000	.0236	.0000
	TQNet	.3481	.0019	.0019	.0256	.0003	.3278	.0000	.0000	.0206	.0000	.3276	.0000	.0000	.0231	.0000
	LightGBM	.4274	.0000	.0000	.0602	.0001	.3673	.0000	.0000	**.0365**	.0000	.3327	.0000	.0000	.0320	.0000
	MISTNet	**.4851**	**.1594**	**.0646**	**.0656**	.0240	**.4596**	**.1316**	**.0525**	.0353	.0197	**.4162**	**.1262**	**.0387**	**.0334**	.0200

specified, results are averaged over three random seeds. To ensure a consistent treatment of class imbalance, all learnable baselines use focal classification loss with $\gamma = 2$ together with tempered inverse-frequency sampling. MISTNet uses the same base classification setup and additionally applies the reliability-aware group reweighting and auxiliary supervision described in Section 4. All methods are evaluated under the common protocol defined in Section 3.4.

5.2 Benchmark Results

Table 4 summarizes the standard chronological benchmark results. For flight-condition forecasting, MISTNet achieves the highest Macro-F1 among full-coverage learnable methods at 12 and 24 h, while LightGBM is marginally higher at 6 h and attains the highest 6-h AP. For precipitation/thunderstorm forecasting, MISTNet achieves the highest Macro-F1 among the full-coverage learnable methods at all three horizons and maintains nonzero Thunder detection throughout. Several other models retain nonzero AP while producing no Thunder predictions under multiclass argmax, showing that rare-event ranking and categorical decisions can yield different assessments. Such ranking performance should nevertheless be interpreted relative to event prevalence.

The operational TAF baselines exhibit a different operating regime. At the fully covered 6- and 12-hour horizons, the risk-aware TAF baseline achieves higher Thunder POD and CSI than MISTNet, together with a higher predicted-positive rate, indicating a more aggressive adverse-event operating point. At 24 h, TAF coverage is only 60.1%, so its results are evaluated on a different subset and are not directly comparable with the full-coverage methods. These contrasting behaviors motivate reporting multiclass performance, rare-event ranking, and decision-based detection jointly rather than relying on a single metric.

5.3 Cross-Airport Generalization

As an additional analysis, we evaluate spatial transfer using a five-fold leave-airports-out protocol. In this setting, the held-out airports are never used for model training or checkpoint selection. Learned airport-identity embeddings are disabled, so MISTNet relies only on transferable static air-

Table 5: Leave-airports-out evaluation on precipitation/thunderstorm forecasting.

Method	Macro-F1	Prec.	POD	CSI
Persistence	.4024	.0296	.0407	.0175
w/o Group Reweight.	.3785	.0459	.0219	.0142
w/o Airport Profile	.4299	.0472	.0922	.0303
MISTNet	**.4303**	**.0581**	**.1144**	**.0381**

Table 6: Component ablations of MISTNet on both forecasting tasks.

Variant	Flight		Precip.	
	F1	CSI	F1	CSI
MISTNet	**.2739**	**.1592**	**.0984**	**.0519**
w/o S–B Fusion	.2487	.1424	.0971	.0512
w/o GRIB	.2406	.1377	.0783	.0410
w/o Group Reweight.	.2501	.1432	.0413	.0212
w/o Airport-Time	.2285	.1294	.0769	.0401

port metadata when applied to unseen airports. Table 5 shows that cross-airport transfer remains challenging, particularly for rare-event detection. MISTNet retains nonzero Thunder detection on unseen airports and achieves higher rare-event precision, POD, and CSI than the evaluated ablations. Removing group reweighting substantially degrades both overall and rare-event performance, while removing the transferable airport profile primarily reduces rare-event detection quality. These results suggest that group-aware rare-event training and transferable airport context support cross-airport performance in this setting.

Unless otherwise specified, results are averaged over forecast horizons and seeds 42, 43, and 44; leave-airports-out results are additionally averaged over the five held-out-airport folds. S–B denotes Site–Background. Horizon-wise results and the complete leave-airports-out protocol are provided in Appendix D.2.

5.4 Reference Model Analysis

Table 6 examines the main components of the reference model. Here, F1 denotes binary rare-event F1, computed on IFR-or-worse for flight-condition forecasting and Thunder for precipitation/thunderstorm forecasting, rather than multiclass Macro-F1. Regional information and airport-time context are particularly important for flight-condition forecasting, whereas group reweighting has the largest effect on Thunder detection, reducing Thunder F1 from 0.0984 to 0.0413 when removed. Additional comparisons with unshrunk reweighting and canonical Group DRO are provided in Appendix D.1.

6 Conclusion

We introduced AirportMIST, a multimodal benchmark for categorical airport weather-state forecasting across 37 airports at 6-, 12-, and 24-hour horizons. By combining heterogeneous airport-local observations, regional numerical forecasts, and static airport information under standardized tasks and evaluation protocols, AirportMIST supports systematic comparison under severe rare-event imbalance and temporal distribution shifts. Experiments with MISTNet and operational and data-driven baselines show that multiclass performance, rare-event ranking, and detection under the multiclass argmax rule can lead to different assessments of forecasting quality. MISTNet serves as a reproducible learned reference, achieving the highest Macro-F1 among full-coverage learnable baselines at the 12- and 24-hour flight-condition horizons and across all three precipitation/thunderstorm horizons, while maintaining nonzero Thunder detection at every horizon. The risk-aware TAF baseline, derived from operational forecaster-issued TAFs, achieves higher Thunder POD and CSI at the fully covered 6- and 12-hour horizons, highlighting a different operational decision regime. Leave-airports-out evaluation further highlights the difficulty of transferring rare-event forecasts to unseen airports. These findings motivate evaluation with complementary ranking and decision-based metrics, as well as continued research on multimodal representation learning and rare-event forecasting. The current benchmark is limited to Chinese airports, two summer seasons, and point-state targets; broader geographic and seasonal coverage and event-onset forecasting remain important directions for future work.

AI USE STATEMENT

We used generative AI tools to assist with methodological and experimental feedback, interpretation and presentation of experimental results, and drafting and language editing of the manuscript. AI tools were also used to assist with code editing and debugging where applicable. All model designs, experimental protocols, data-processing procedures, quantitative results, and scientific claims were reviewed and verified by the authors. AI-generated suggestions were not treated as experimental evidence, and the reported results were obtained from the authors' own implementations and evaluation pipelines. The authors take full responsibility for the final content of the paper, including all AI-assisted text, code, analyses, and claims.

ETHICS STATEMENT

AirportMIST is intended for research on airport-weather forecasting and is not designed or validated for direct operational or safety-critical decision making. The benchmark uses meteorological and aviation-weather data rather than personal or human-subject information. All data are used and will be released subject to their applicable licenses and access conditions. Users of the benchmark should independently assess model reliability and regulatory requirements before any operational deployment.

REPRODUCIBILITY STATEMENT

Detailed benchmark construction, preprocessing, multimodal alignment, label definitions, data splits, class distributions, and training specifications are provided in the appendices. During review, the benchmark code is available through the anonymous repository. The complete AirportMIST benchmark and supporting resources will be publicly released upon acceptance under the applicable licenses and access conditions.

REFERENCES

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *CoRR*, abs/2211.02556, 2022. doi: 10.48550/ARXIV.2211.02556. URL https://doi.org/10.48550/arXiv.2211.02556.

Tal Boneh, Gary T. Weymouth, Peter Newham, Rodney J. Potts, John Bally, Ann E. Nicholson, and Kevin B. Korb. Fog forecasting for melbourne airport using a bayesian decision network. *Weather and Forecasting*, 30(5):1218–1233, 2015. doi: 10.1175/WAF-D-15-0005.1.

Ching Chang, Jeehyun Hwang, Yidan Shi, Haixin Wang, Wen-Chih Peng, Tien-Fu Chen, and Wei Wang. Time-imm: A dataset and benchmark for irregular multimodal multivariate time series. *CoRR*, abs/2506.10412, 2025. doi: 10.48550/ARXIV.2506.10412. URL https://doi.org/10.48550/arXiv.2506.10412.

Fabricio Magalhaes Cordeiro, Gutemberg Borges Franca, Francisco Leite de Albuquerque Neto, and Ismail Gultepe. Visibility and ceiling nowcasting using artificial intelligence techniques for aviation applications. *Atmosphere*, 12(12):1657, 2021. doi: 10.3390/atmos12121657.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 9268–9277. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00949. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Cui_Class-Balanced_Loss_Based_on_Effective_Number_of_Samples_CVPR_2019_paper.html.

Yuting Huang, Ziquan Fang, Zhihao Zeng, Lu Chen, and Yunjun Gao. Causal spatio-temporal prediction: An effective and efficient multi-modal approach. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-

Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/b71169b4db5448c494c91db848f060be-Abstract-Conference.html.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 3146–3154, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

Xinyu Li, Yunyi Huang, Quan Fang, Yang Yang, Can Zhao, and Kaiquan Cai. Aerowf: A geometric spectral-temporal dual-stream learning framework for aerodrome weather forecasting. In Won Kim, Jae-Gil Lee, Kyuseok Shim, James Caverlee, David F. Gleich, Xiangliang Zhang, Qi He, Faisal Farooq, Rómer Rosales, Sayan Ranu, Haixun Wang, Yan Liu, Carl Yang, Leman Akoglu, Nitesh V. Chawla, and Vagelis Papalexakis (eds.), *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD 2026, Jeju Island, Korea, August 9-13, 2026*, pp. 2814–2825. ACM, 2026. doi: 10.1145/3770855.3817822. URL https://doi.org/10.1145/3770855.3817822.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=SJiHXGWAZ.

Shengsheng Lin, Haojun Chen, Haijie Wu, Chunyun Qiu, and Weiwei Lin. Temporal query network for efficient multivariate time series forecasting. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/lin25e.html.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):318–327, 2020. doi: 10.1109/TPAMI.2018.2858826. URL https://doi.org/10.1109/TPAMI.2018.2858826.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. Time-mmd: Multi-domain multimodal dataset for time series analysis. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024a. URL http://papers.nips.cc/paper_files/paper/2024/hash/8e7768122f3eeec6d77cd2b424b72413-Abstract-Datasets_and_Benchmarks_Track.html.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024b. URL https://openreview.net/forum?id=JePfAI8fah.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K. Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 25904–25938. PMLR, 2023. URL https://proceedings.mlr.press/v202/nguyen23a.html.

Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=Jbdc0vTOcol.

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, and Animashree Anandkumar. Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators. *CoRR*, abs/2202.11214, 2022. URL https://arxiv.org/abs/2202.11214.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *CoRR*, abs/1911.08731, 2019. URL http://arxiv.org/abs/1911.08731.

Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, Christopher Phillips, Romeo Kienzler, Daniela Szwarcman, Vishal Gaur, Rajat Shinde, Rohit Lal, Arlindo Da Silva, Jorge Luis Guevara Diaz, Anne Jones, Simon Pfreundschuh, Amy Lin, Aditi Sheshadri, Udaysankar Nair, Valentine Anantharaj, Hendrik F. Hamann, Campbell D. Watson, Manil Maskey, Tsengdar J. Lee, Juan Bernabé-Moreno, and Rahul Ramachandran. Prithvi wxc: Foundation model for weather and climate. *CoRR*, abs/2409.13598, 2024. doi: 10.48550/ARXIV.2409.13598. URL https://doi.org/10.48550/arXiv.2409.13598.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. Timemixer: Decomposable multiscale mixing for time series forecasting. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=7oLshfEIC2.

Dirk Wolters, Maurice Schmeits, and Kirien Whan. Probabilistic forecasting of cloud-base height and visibility using quantile regression forests, based on nwp and observation features. *Weather and Forecasting*, 40(4):543–559, 2025. doi: 10.1175/WAF-D-23-0231.1.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In Sarit Kraus (ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 1907–1913. ijcai.org, 2019. doi: 10.24963/IJCAI.2019/264. URL https://doi.org/10.24963/ijcai.2019/264.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In Rajesh Gupta, Yan Liu, Jiliang Tang, and B. Aditya Prakash (eds.), *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pp. 753–763. ACM, 2020. doi: 10.1145/3394486.3403118. URL https://doi.org/10.1145/3394486.3403118.

Kun Yi, Qi Zhang, Wei Fan, Shoujin Wang, Pengyang Wang, Hui He, Ning An, Defu Lian, Longbing Cao, and Zhendong Niu. Frequency-domain mlps are more effective learners in time series forecasting. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/f1d16af76939f476b5f040fd1398c0a3-Abstract-Conference.html.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In Jérôme Lang (ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pp. 3634–3640. ijcai.org, 2018. doi: 10.24963/IJCAI.2018/505. URL https://doi.org/10.24963/ijcai.2018/505.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 11121–11128. AAAI Press, 2023. doi: 10.1609/AAAI.V37I9.26317. URL https://doi.org/10.1609/aaai.v37i9.26317.

Siru Zhong, Weilin Ruan, Ming Jin, Huan Li, Qingsong Wen, and Yuxuan Liang. Time-vlm: Exploring multimodal vision-language models for augmented time series forecasting. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/zhong25a.html.

A Benchmark Construction Details

A.1 Airport Coverage

AirportMIST contains 37 major civil airports in China during JJA 2020 and JJA 2025. The airport set is determined by the availability of sufficiently complete and alignable AWOS, METAR, GRIB, and airport-metadata records across both study periods. The standard chronological benchmark retains the same airport set in training, validation, and testing, while the dedicated leave-airports-out protocol evaluates transfer to airports excluded from model training and checkpoint selection.

Table 7 lists the included airports together with geographic metadata, the number of training anchors, and training-set prevalence of the two rare-event targets. The statistics show substantial spatial variation in both Thunder and IFR-or-worse prevalence across airports.

A historical site transition affects ZSQD in the 2020 portion of the benchmark: records from that period correspond to Qingdao Liuting Airport, whereas the static airport metadata use the later Qingdao Jiaodong location. The two sites are geographically close, and at the 1.25° GRIB resolution used in AirportMIST they map to the same nearest grid point, so the associated regional GRIB extraction is unchanged.

A.2 Multimodal Input Specification

Table 8 summarizes the processed representation of each input source under the official 12-hour history configuration.

AWOS. For each anchor, the preceding 12 hours of AWOS observations are mapped onto the native one-minute grid, yielding 720 time steps. Reindexing onto the complete minute grid does not imply temporal interpolation: missing observations remain missing and are accompanied by feature-level missingness masks and row-level observation masks. The 35 processed variables cover wind and gust information, visibility and runway visual range, temperature, relative humidity, dew point, pressure, cloud-base height, precipitation intensity and amount, pavement temperature, and encoded weather-phenomenon indicators.

METAR. METAR reports within $(t - 12\mathrm{h}, t]$ are retained in their original temporal order and are not resampled. Up to 24 reports are represented per anchor; shorter sequences are padded and accompanied by observation and validity masks. The 19 encoded inputs include continuous meteorological variables together with structured indicators for precipitation, rain, shower, Thunder, fog/mist, haze/smoke, dust/sand, nearby weather, and weather intensity. Variables derived directly from the prediction targets, including flight-category and precipitation-regime identifiers, are excluded from the model inputs.

WAFS forecast fields (GRIB). The regional background contains temperature, zonal wind, meridional wind, and relative humidity at 850, 700, and 500 hPa, yielding 12 channels. For each anchor,

Table 7: Airports included in AirportMIST. "Anchors" denotes the number of training anchors. Thunder and IFR$^-$ are the percentages of valid training labels after pooling the 6/12/24-h horizons. IFR$^-$ denotes IFR-or-worse, combining IFR and LIFR; a dash indicates missing metadata.

ICAO	IATA	Airport	Lat.	Lon.	Elev. (ft)	Anchors	Thunder (%)	IFR$^-$ (%)
ZBAA	PEK	Beijing Capital International Airport	40.077	116.597	116	5,687	1.477	6.647
ZBAD	PKX	Beijing Daxing International Airport	39.501	116.414	98	5,699	1.123	9.639
ZBHH	HET	Hohhot Baita International Airport	40.850	111.825	3,556	2,831	2.543	1.778
ZBSJ	SJW	Shijiazhuang Zhengding International Airport	38.281	114.697	233	2,832	0.812	17.326
ZBTJ	TSN	Tianjin Binhai International Airport	39.124	117.346	10	5,672	1.093	19.752
ZBYN	TYN	Taiyuan Wusu International Airport	37.747	112.628	2,575	5,704	1.245	11.734
ZGGG	CAN	Guangzhou Baiyun International Airport	23.392	113.299	50	5,773	3.505	1.836
ZGHA	CSX	Changsha Huanghua International Airport	28.189	113.220	217	2,830	0.895	5.866
ZGKL	KWL	Guilin Liangjiang International Airport	25.220	110.040	570	2,833	3.542	4.824
ZGNN	NNG	Nanning Wuxu International Airport	22.598	108.182	421	2,829	1.296	0.919
ZGOW	SWA	Jieyang Chaoshan International Airport	23.552	116.503	–	2,830	0.860	1.095
ZGSZ	SZX	Shenzhen Bao'an International Airport	22.639	113.803	13	2,831	1.554	0.636
ZHCC	CGO	Zhengzhou Xinzheng International Airport	34.526	113.849	495	2,830	0.483	16.678
ZHHH	WUH	Wuhan Tianhe International Airport	30.775	114.214	113	2,831	1.978	7.783
ZJHK	HAK	Haikou Meilan International Airport	19.935	110.459	75	2,833	2.024	1.224
ZJSY	SYX	Sanya Phoenix International Airport	18.303	109.412	92	2,833	1.694	1.024
ZLLL	LHW	Lanzhou Zhongchuan International Airport	36.515	103.620	6,388	2,831	0.471	0.966
ZLXY	XIY	Xi'an Xianyang International Airport	34.447	108.752	1,572	2,827	0.141	16.071
ZPPP	KMG	Kunming Changshui International Airport	25.110	102.937	6,903	2,831	1.213	1.849
ZSAM	XMN	Xiamen Gaoqi International Airport	24.544	118.127	59	2,829	1.084	0.377
ZSFZ	FOC	Fuzhou Changle International Airport	25.929	119.673	46	2,830	0.942	5.230
ZSHC	HGH	Hangzhou Xiaoshan International Airport	30.236	120.429	23	5,722	1.188	18.269
ZSNB	NGB	Ningbo Lishe International Airport	29.827	121.462	13	2,828	1.756	9.536
ZSNJ	NKG	Nanjing Lukou International Airport	31.735	118.866	49	2,829	1.084	19.618
ZSOF	HFE	Hefei Xinqiao International Airport	31.988	116.977	207	2,829	1.720	21.680
ZSPD	PVG	Shanghai Pudong International Airport	31.143	121.805	13	5,699	0.602	17.214
ZSQD	TAO	Qingdao Jiaodong International Airport	36.362	120.088	30	2,825	0.861	14.330
ZSSS	SHA	Shanghai Hongqiao International Airport	31.198	121.334	10	5,722	0.833	11.703
ZUCK	CKG	Chongqing Jiangbei International Airport	29.712	106.652	1,365	2,832	0.330	9.769
ZUGY	KWE	Guiyang Longdongbao International Airport	26.542	106.804	3,736	2,835	1.775	1.423
ZUUU	CTU	Chengdu Shuangliu International Airport	30.558	103.946	1,625	2,829	2.215	4.030
ZWSH	KHG	Kashgar Laining International Airport	39.542	76.020	4,529	2,293	0.890	3.326
ZWWW	URC	Urumqi Tianshan International Airport	43.914	87.479	2,125	5,683	0.481	0.968
ZYCC	CGQ	Changchun Longjia International Airport	43.996	125.685	706	2,829	1.249	4.737
ZYHB	HRB	Harbin Taiping International Airport	45.623	126.250	457	2,833	1.047	5.342
ZYTL	DLC	Dalian Zhoushuizi International Airport	38.966	121.538	107	5,675	0.628	12.235
ZYTX	SHE	Shenyang Taoxian International Airport	41.640	123.484	198	5,687	1.377	4.384

a 15×15 patch is extracted around the grid point nearest to the airport. Boundary positions, unavailable channels, and non-finite grid values are represented through explicit spatial-validity and missingness masks.

For each GRIB record, the initialization/reference time t_{init} is derived from the GRIB `dataDate` and `dataTime` fields, and the valid time is computed as

$$t_{\text{valid}} = t_{\text{init}} + \Delta t_{\text{lead}}.$$

At an anchor time t, only forecast records satisfying

$$t_{\text{init}} \leq t$$

are considered eligible. Among eligible records, AirportMIST selects the record whose valid time is closest to the anchor within the configured tolerance, with ties resolved in favor of the latest initialization time. We refer to this protocol as nominal-cycle causal GRIB selection.

Airport metadata and forecast-valid time. The static airport profile contains normalized latitude, longitude, elevation, and an elevation-missing indicator. Missing elevation values are median-filled before normalization while retaining the missingness indicator. The data loader additionally derives annual and diurnal sine/cosine features at each forecast-valid time $t + h$.

A.3 Anchor Construction and Causal Alignment

Anchor candidates are generated from METAR report times. A candidate at airport s and time t is retained only if an observed AWOS row is available within the preceding 5 minutes and a causally eligible GRIB record is available.

For GRIB, only records whose initialization time is no later than the forecast anchor are eligible. Among eligible records, AirportMIST selects the record whose valid time is closest to t, subject to a maximum valid-time mismatch of 3 hours; ties are resolved using the most recent initialization.

Table 8: Processed representation of one AirportMIST anchor under the official 12-hour input-history configuration.

Source	Native organization	Processed shape	Information retained
AWOS	One-minute local observations	720×35	Surface weather, visibility/RVR, wind, precipitation, and encoded weather phenomena.
METAR	Irregular local reports	up to 24×19	Continuous variables and encoded aviation-weather semantics.
GRIB	1.25° regional grid	$12 \times 15 \times 15$	Temperature, zonal wind, meridional wind, and relative humidity at 850, 700, and 500 hPa.
Airport metadata	Static per airport	4	Latitude, longitude, elevation, and elevation-missing indicator.

For forecast horizon h, the nominal target time is $t + h$. The nearest METAR report within ± 15 minutes of this time is used to construct the horizon-specific target. This defines a point-forecast target rather than an "event occurs at any time within the next h hours" target. If no report satisfies the matching tolerance, the corresponding sample–horizon pair is marked invalid and excluded from the associated loss and metric computation. Flight-condition targets are additionally marked invalid when neither visibility nor ceiling information is available.

A.4 Operational Label Construction

Precipitation/thunderstorm state. The task contains three mutually exclusive classes, $\{\mathrm{NoRain}, \mathrm{RainNoThunder}, \mathrm{Thunder}\}$. METAR present-weather codes are normalized before label construction. Reports containing `TS` are assigned to Thunder with highest priority. Otherwise, reports containing a recognized precipitation code are assigned to RainNoThunder, and all remaining valid reports are assigned to NoRain. Recognized precipitation codes include `DZ`, `RA`, `SN`, `SG`, `IC`, `PL`, `GR`, `GS`, `UP`. RainNoThunder is a dataset class identifier and should not be interpreted as rain-only; it includes liquid, solid, and mixed precipitation reported without `TS`.

Flight condition. The flight-condition task contains four classes: $\{\mathrm{VFR}, \mathrm{MVFR}, \mathrm{IFR}, \mathrm{LIFR}\}$. Visibility is converted to statute miles (SM), and ceiling is defined as the lowest reported `BKN`, `OVC`, or vertical-visibility (`VV`) cloud layer. The most restrictive condition triggered by either visibility v or ceiling c is assigned:

$$\mathrm{LIFR}: \quad v < 1 \text{ SM} \ \vee \ c < 500 \text{ ft}, \tag{1}$$

$$\mathrm{IFR}: \quad v < 3 \text{ SM} \ \vee \ c < 1000 \text{ ft}, \tag{2}$$

$$\mathrm{MVFR}: \quad v \leq 5 \text{ SM} \ \vee \ c \leq 3000 \text{ ft}, \tag{3}$$

$$\mathrm{VFR}: \quad \text{otherwise}. \tag{4}$$

The IFR rule is evaluated only after excluding LIFR, and the MVFR rule is evaluated after excluding both more restrictive categories. A report is considered valid when at least one of visibility or ceiling is available. For rare-event evaluation, IFR and LIFR are merged into IFR-or-worse, while model training uses the original four-class target.

A.5 Chronological Split and Leakage Prevention

The official benchmark uses chronological rather than random splitting. JJA 2020 and June 2025 form the training set, July 2–31, 2025 the validation set, and August 2–31, 2025 the test set. July 1 and August 1 are excluded as anchor dates and serve as temporal buffers. A horizon-specific target from an anchor in the preceding split may fall inside a buffer day, but that day is never used as an anchor period for the subsequent split.

Feature-normalization statistics are estimated exclusively from the training split. Finite continuous values are standardized using the corresponding training-set mean and standard deviation; missing positions are subsequently zero-filled for tensor construction while their modality-specific missingness and observation masks remain available. Validation and test labels are not used to compute normalization statistics, sampling weights, model parameters, or training objectives. Hyperparameter selection and early stopping use validation data only, and the test set is reserved for final evaluation.

Table 9: Class distribution (%) by split and forecast horizon. N_P and N_F denote the numbers of valid precipitation/thunderstorm and flight-condition labels, respectively. "Ord." denotes non-Thunder precipitation. Percentages may not sum to exactly 100 due to rounding.

Split	Horizon	N_P	Precipitation state (%)			N_F	Flight condition (%)			
			No precip.	Ord. precip.	Thunder		VFR	MVFR	IFR	LIFR
Train	6 h	135,770	89.082	9.635	1.284	135,768	73.235	18.245	7.659	0.861
Train	12 h	135,767	89.123	9.587	1.290	135,765	73.344	18.190	7.600	0.866
Train	24 h	135,767	89.220	9.497	1.283	135,765	73.322	18.208	7.615	0.855
Validation	6 h	35,059	90.037	8.146	1.817	35,059	79.951	16.441	3.346	0.262
Validation	12 h	35,059	90.219	8.001	1.780	35,059	80.042	16.367	3.340	0.251
Validation	24 h	35,059	90.596	7.704	1.700	35,059	80.424	16.136	3.192	0.248
Test	6 h	34,373	90.493	7.968	1.539	34,373	85.314	11.911	2.534	0.241
Test	12 h	34,079	90.675	7.890	1.435	34,079	85.152	12.043	2.565	0.241
Test	24 h	33,492	90.687	7.862	1.451	33,492	85.325	11.851	2.571	0.254

A.6 Class Distribution and Rare-Event Prevalence

Table 9 reports the complete class distribution for each split and forecast horizon, using the number of valid labels for the corresponding task–horizon pair as the denominator. Across all splits and horizons, Thunder accounts for 8,597 of 614,425 valid precipitation/thunderstorm labels (1.399%), while IFR-or-worse accounts for 41,146 of 614,419 valid flight-condition labels (6.697%). In the test split, Thunder prevalence ranges from 1.435% to 1.539% across horizons, whereas IFR-or-worse ranges from 2.775% to 2.825%.

The similar class frequencies across training horizons should not be interpreted as duplicated targets, because each horizon is independently matched to its own forecast-valid time. The reported counts represent horizon-specific label instances rather than independent meteorological episodes, so a persistent event may contribute valid labels to multiple neighboring anchors or forecast horizons.

A.7 Data Organization

AirportMIST separates task-neutral anchor metadata, horizon-specific targets, and modality-specific source data. The benchmark release will include:

- `metadata/airports.parquet`: airport identifiers, names, coordinates, elevation, and facility type;
- `metadata/feature_schema.json`: feature definitions, sources, types, and units;
- `samples/anchors_base.parquet`: task-neutral anchor metadata and causal GRIB matching information;
- `samples/targets_6h_12h_24h.parquet`: horizon-specific labels and validity flags in long format;
- `airportmist/awos/`: per-year and per-airport AWOS tables;
- `airportmist/metar/`: per-year and per-airport parsed METAR/SPECI tables;
- `airportmist/grib/`: yearly GRIB-derived background fields and record indices;
- `metadata/normalization_stats_multihorizon.json`: normalization statistics estimated from the training split.

Historical input windows are constructed at runtime from the timestamped source data rather than materialized separately for every anchor. Horizon-specific targets are stored independently, with one record for each sample–horizon pair. This organization avoids redundant storage, allows target validity and multimodal alignment to be audited independently of model implementation, and enables alternative input-history lengths to be studied without reconstructing the benchmark from the original raw sources.

B Evaluation and Baseline Details

B.1 Evaluation Metrics

For a rare event, let TP, FP, and FN denote true positives, false positives, and false negatives under the common multiclass argmax decision rule. Probability of detection (POD) and critical success index (CSI) are defined as

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{CSI} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{5}$$

The predicted-positive rate is

$$\mathrm{Pred.+} = \frac{\mathrm{TP} + \mathrm{FP}}{N}, \tag{6}$$

where N is the number of valid evaluation instances for the corresponding task and forecast horizon.

Average precision (AP) is computed directly from the rare-event prediction scores before applying the multiclass argmax decision rule and therefore measures ranking quality independently of a particular classification operating point. Macro-F1 is computed over all classes of the corresponding multiclass task and is used to summarize overall classification performance.

Rare-event scores. For the Flight task, IFR-or-worse is defined as the union of the IFR and LIFR classes. For probabilistic models, its ranking score is

$$s_{\mathrm{IFR^-}}(x) = p(y = \mathrm{IFR} \mid x) + p(y = \mathrm{LIFR} \mid x).$$

For the precipitation/thunderstorm task, the corresponding ranking score is

$$s_{\mathrm{Thunder}}(x) = p(y = \mathrm{Thunder} \mid x).$$

Average precision is computed from these continuous event scores against the corresponding binary event labels. Argmax-based POD and CSI are computed separately from the original multiclass predictions.

Persistence and the two TAF baselines produce deterministic categorical predictions rather than continuous class probabilities. For these methods, the event score is defined as

$$s(x) = \mathbb{1}[\hat{y} \in \mathcal{Y}_{\mathrm{event}}],$$

where $\mathcal{Y}_{\mathrm{event}} = \{\mathrm{IFR}, \mathrm{LIFR}\}$ for Flight and $\mathcal{Y}_{\mathrm{event}} = \{\mathrm{Thunder}\}$ for precipitation/thunderstorm forecasting. Their AP values therefore correspond to a two-valued hard-decision score and are reported as reference quantities rather than as directly comparable probabilistic ranking measures.

B.2 TAF-Based Operational Baselines

TAF forecasts are aligned to the nominal valid time of each benchmark target. For a sample anchored at time t, the availability time of a TAF bulletin is defined as

$$t_{\mathrm{available}} = \max(t_{\mathrm{issue}}, t_{\mathrm{arrival}}),$$

where t_{arrival} denotes the time at which the bulletin became available in the source archive. Only bulletins satisfying $t_{\mathrm{available}} \leq t$ are eligible. For each airport, eligible bulletins are ordered by $t_{\mathrm{available}}$, and the most recently available bulletin whose validity interval covers the requested target time is selected. Ties are resolved deterministically by source-file ingestion order. This procedure prevents the use of TAF information unavailable at the prediction anchor time.

TAF visibility, cloud, and weather fields are parsed separately. Visibility may be reported in statute miles or metres; `CAVOK` and `P6SM` are treated as non-restrictive for flight-condition category mapping. Ceiling is defined as the lowest valid `BKN`, `OVC`, or `VV` layer. Flight-condition severity follows

$$\mathrm{VFR} < \mathrm{MVFR} < \mathrm{IFR} < \mathrm{LIFR},$$

with the most restrictive category implied by visibility or ceiling being selected. If only one of visibility and ceiling is available, classification uses the available field; if neither is available, no flight-condition prediction is produced.

For precipitation, any weather group containing `TS` is mapped to Thunder. Otherwise, groups containing one of `DZ`, `RA`, `SN`, `SG`, `IC`, `PL`, `GR`, `GS`, or `UP` are mapped to ordinary precipitation, while explicit `NSW` or `CAVOK` conditions are mapped to no precipitation. Thunder takes precedence over ordinary precipitation.

Table 10: Interpretation of TAF change groups in the operational baselines.

Group	Prevailing	Risk-aware
Initial state	Used	Used
`FM` after start	New prevailing state	Same
`BECMG` during transition	Previous state	Alternative scenario
`BECMG` after transition	New prevailing state	Same
`TEMPO` while active	Ignored	Alternative scenario
`PROB30/40` while active	Ignored	Alternative scenario
`PROB30/40 TEMPO`	Ignored	One compound scenario

TAF prevailing. The prevailing baseline represents the deterministic condition expected to prevail at the target time. The initial state remains active until superseded by an applicable change group. `FM` updates the prevailing state from its start time, while `BECMG` updates it only after the end of its transition interval. During the transition interval, the preceding state remains prevailing. Change groups are applied as field-wise overlays, with unspecified fields inherited from the preceding prevailing state. Temporary and probabilistic groups do not alter the prevailing prediction.

TAF risk-aware. The risk-aware baseline starts from the same prevailing state but additionally evaluates each adverse scenario applicable at the target time. Active `TEMPO`, `PROB30`, and `PROB40` groups are applied individually as field-wise overlays. A `BECMG` group defines an alternative scenario during its transition interval and becomes prevailing after the transition ends. Compound `PROB30 TEMPO` and `PROB40 TEMPO` groups are parsed as single probabilistic scenarios. Distinct scenarios are evaluated independently rather than cumulatively merged, and `PROB30`/`PROB40` probabilities are not used as numerical weights.

For each task, the prevailing and alternative categories are compared and the most adverse category is retained:

$$\hat{y}_{\text{risk}} = \max\left(\hat{y}_{\text{prev}}, \hat{y}_1, \ldots, \hat{y}_K\right).$$

The maximization is performed independently for the flight-condition and precipitation/thunderstorm tasks. The risk-aware baseline should therefore be interpreted as a conservative scenario-based forecast interpretation rather than a calibrated probabilistic forecast.

Coverage. For each task and horizon, the coverage denominator consists of all test instances with valid target labels at that horizon. An instance contributes to the numerator only when its task label is valid and a leakage-safe TAF prediction can be constructed. No extrapolation is performed outside the TAF validity interval. The 6- and 12-hour horizons attain full coverage under this definition, whereas the 24-hour horizon covers 60.1% of the benchmark test instances. The 24-hour metrics are therefore computed on the TAF-covered subset and should be interpreted jointly with coverage.

B.3 Input Adaptation for Learnable Baselines

To provide a common multimodal protocol across heterogeneous forecasting architectures, the six neural time-series baselines—DLinear, FreTS, PatchTST, iTransformer, TimeMixer, and TQNet—share a unified multimodal sequence adapter, while LightGBM uses a separate fixed-length tabular representation constructed from the same source modalities.

Neural time-series baselines. For DLinear, FreTS, PatchTST, iTransformer, TimeMixer, and TQNet, the heterogeneous inputs are converted into a unified sequence representation $\mathbf{X}_{\text{enc}} \in \mathbb{R}^{B\times T\times C}$, where T follows the AWOS minute-level timeline and C denotes the concatenated feature dimension.

AWOS retains its native minute-level history and directly defines the common temporal length T. Missing positions are zero-filled after applying the corresponding missingness masks.

METAR reports within the history window are initially represented using their original sparse report sequence together with padding and validity masks. Missing and padded positions are set to zero, after which the METAR sequence is linearly interpolated to length T. This enables dense time-series backbones to process METAR jointly with AWOS, although the original irregular report timing is not preserved as a separate temporal structure.

For GRIB, each channel of the airport-centered patch is reduced using a mask-aware spatial mean over valid and observed grid locations. The resulting regional vector is repeated across the T time steps. Static airport features are similarly repeated along the temporal dimension. The adapted modalities are then concatenated along the feature dimension:

$$\mathbf{X}_{\text{enc}} = \text{Concat}\left(\mathbf{X}^{\text{AWOS}}, \widetilde{\mathbf{X}}^{\text{METAR}}, \widetilde{\mathbf{X}}^{\text{GRIB}}, \widetilde{\mathbf{X}}^{\text{meta}}\right). \tag{7}$$

Thus, the six neural baselines preserve a shared temporal axis and use early feature-level fusion, while the internal spatial arrangement of the GRIB patch is not retained.

Backbone-specific aggregation. The six neural baselines use the same adapted multimodal sequence but differ in how they encode and aggregate it into a fixed-dimensional representation. DLinear uses its decomposition-based encoder and concatenates the temporal mean with the final time-step representation. FreTS applies frequency-domain encoding followed by global temporal aggregation. PatchTST partitions each variable into temporal patches and uses the resulting patch representation through its classification head. iTransformer treats variables as tokens and aggregates the encoded variable-token representation. TimeMixer applies multiscale temporal mixing before projecting the resulting representation for classification. TQNet additionally incorporates cyclic position information derived from the anchor time and aggregates the encoded sequence using a mean-and-last representation before projection.

LightGBM. LightGBM does not use the neural sequence adapter. Instead, it constructs a fixed-length tabular feature vector directly from AWOS, METAR, GRIB, and airport/time metadata.

For each AWOS variable, six temporal statistics are computed over the complete input history: latest value, mean, standard deviation, minimum, maximum, and valid fraction. Missing records are excluded using the AWOS missingness and observation masks, and variables with no valid observations receive zero-valued summary statistics.

For each METAR variable, the same six statistics are computed directly from the observed and padded report sequence. Padding positions do not contribute to the statistics, and no temporal interpolation is applied.

For each GRIB channel, six spatial statistics are extracted from the airport-centered patch: center-grid value, spatial mean, standard deviation, minimum, maximum, and valid fraction. All spatial statistics are computed over valid grid locations only.

The resulting statistical summaries are concatenated with normalized latitude, normalized longitude, normalized elevation, an elevation-missing indicator, airport identity, GRIB lead time, and GRIB age:

$$\mathbf{x}_{\text{tab}} = \text{Concat}\left(\phi_{\text{AWOS}}, \phi_{\text{METAR}}, \phi_{\text{GRIB}}, \mathbf{x}_{\text{meta}}\right), \tag{8}$$

where ϕ_{AWOS}, ϕ_{METAR}, and ϕ_{GRIB} denote the corresponding statistical summaries.

The resulting dimensionality is

$$D = 6C_{\text{AWOS}} + 6C_{\text{METAR}} + 6C_{\text{GRIB}} + 7. \tag{9}$$

Airport identity is treated as a categorical LightGBM feature rather than as a continuous numerical variable. Overall, the neural baselines share a common dense sequence adapter, whereas LightGBM uses a modality-specific statistical representation. All baseline inputs are constructed only from information available at the forecast anchor and exclude target-derived variables.

C MISTNet Implementation Details

This section provides the implementation details of **MISTNet**, the reference model used in the AirportMIST benchmark. MISTNet follows the local–regional modeling structure described in the main paper: airport-local observations are encoded separately from regional GRIB fields, after which the local state conditions the extraction of relevant regional atmospheric context. The resulting representations are used for multi-horizon categorical prediction. During training, reliability-aware group reweighting and auxiliary objectives are used to address heterogeneous rare-event difficulty.

C.1 Local and Regional Encoding

AWOS and METAR contain variables with different physical meanings, reporting frequencies, and missingness patterns. MISTNet therefore encodes the two sources separately before integrating them into the airport-local representation.

For each source, standardized observations are augmented with feature-level missingness indicators, row-level availability, observation age, and learned variable, type, and source embeddings. A type-specific input projection is followed by a shared temporal convolutional encoder and masked temporal aggregation, producing one token for each modeled variable. The resulting AWOS and METAR tokens are then jointly modeled using a variable-level Transformer.

The resulting local tokens are combined with an airport-profile token derived from airport identity and static geographic metadata. Sparse weather-phenomenon signals, including precipitation, Thunder, fog, and haze indicators, are encoded separately and incorporated through gated residual modulation. The resulting airport-local representation is denoted by $\mathbf{z}_{s,t}^{\mathrm{loc}}$.

In parallel, the airport-centered GRIB fields are encoded together with their missingness and spatial-validity masks into regional grid tokens $\mathbf{V}_{s,t}^{\mathrm{grid}}$, where each token corresponds to a relative spatial location around the target airport.

C.2 Site–Background Fusion

Airport-local observations provide direct evidence about the target airport, whereas GRIB fields provide regional atmospheric context. MISTNet therefore conditions regional spatial aggregation on the airport-local state.

The implementation uses two independently parameterized regional interactions. The first obtains a site-referenced background representation:

$$\begin{aligned}\mathbf{a}_{s,t}^{\mathrm{bg}} &= \mathrm{MHA}_{\mathrm{bg}}\left(\mathbf{W}_q\mathbf{z}_{s,t}^{\mathrm{loc}}, \mathbf{V}_{s,t}^{\mathrm{grid}}, \mathbf{V}_{s,t}^{\mathrm{grid}}; \mathbf{M}_{s,t}^{\mathrm{grid}}\right),\\ \mathbf{z}_{s,t}^{\mathrm{bg}} &= \mathrm{LN}\left(\mathbf{z}_{s,t}^{\mathrm{loc}} + \mathbf{a}_{s,t}^{\mathrm{bg}}\right).\end{aligned} \tag{10}$$

A second independently parameterized interaction obtains a complementary background-conditioned local representation:

$$\begin{aligned}\mathbf{a}_{s,t}^{\mathrm{cond}} &= \mathrm{MHA}_{\mathrm{cond}}\Big(\mathrm{LN}_q(\mathbf{z}_{s,t}^{\mathrm{loc}}), \mathrm{LN}_g(\mathbf{V}_{s,t}^{\mathrm{grid}}), \mathrm{LN}_g(\mathbf{V}_{s,t}^{\mathrm{grid}}); \mathbf{M}_{s,t}^{\mathrm{grid}}\Big),\\ \mathbf{u}_{s,t}^{\mathrm{cond}} &= \mathrm{LN}_1\left(\mathbf{z}_{s,t}^{\mathrm{loc}} + \mathbf{a}_{s,t}^{\mathrm{cond}}\right),\\ \mathbf{z}_{s,t}^{\mathrm{cond}} &= \mathrm{LN}_2\left(\mathbf{u}_{s,t}^{\mathrm{cond}} + \mathrm{FFN}(\mathbf{u}_{s,t}^{\mathrm{cond}})\right).\end{aligned} \tag{11}$$

Both branches mask invalid GRIB tokens. The first branch provides a site-referenced summary of the regional background, while the second applies additional normalization and nonlinear refinement. The original airport-local representation remains explicit in the prediction input so that regional context complements rather than replaces local evidence.

C.3 Multi-Horizon Prediction and Contextual Bias

For each forecast horizon $h \in \mathcal{H}$, MISTNet combines $\mathbf{z}_{s,t}^{\mathrm{loc}}$, $\mathbf{z}_{s,t}^{\mathrm{bg}}$, $\mathbf{z}_{s,t}^{\mathrm{cond}}$, and a learned horizon embedding to obtain a shared horizon-specific prediction representation $\mathbf{q}_{s,t,h}$.

For each task, the classification logits are

$$\boldsymbol{\ell}_{s,t,h}^{\mathrm{task}} = \mathbf{W}_{\mathrm{cls}}^{\mathrm{task}}\mathbf{q}_{s,t,h} + \bar{\mathbf{b}}_{s,t+h}^{\mathrm{task}}, \tag{12}$$

where $\bar{\mathbf{b}}_{s,t+h}^{\mathrm{task}}$ is a class-centered contextual bias generated from the airport profile and cyclic forecast-valid-time features. The contextual-bias layers are initialized to zero and learned jointly with the forecasting model.

C.4 Reliability-Aware Group Reweighting

Rare-event difficulty varies substantially across airports and forecast horizons. MISTNet therefore complements global class balancing with group-aware training over airport–horizon–class groups, while reducing the instability of empirical risk estimates for sparsely observed groups.

Tempered rare-class sampling. We first apply tempered inverse-frequency sampling to increase rare-class exposure during training. Compared with full inverse-frequency sampling, tempering increases the representation of rare events without allowing extremely sparse targets to dominate mini-batch construction. The sampling weight of an anchor is determined from the rarest class among its valid horizon-specific targets.

Reliability-adjusted group risk. For each task, a group is defined as $g = (s, h, c)$, where s, h, and c denote airport, forecast horizon, and task-specific class, respectively. Let $\widehat{L}^{(m)}_{s,h,c}$ denote the mean loss of group (s, h, c) in update step m. Its running empirical risk is maintained using an exponential moving average:

$$R^{(m)}_{s,h,c} = (1-\mu)R^{(m-1)}_{s,h,c} + \mu\widehat{L}^{(m)}_{s,h,c}, \tag{13}$$

where μ is the EMA update rate.

To reduce the influence of noisy estimates from sparsely observed groups, the group-specific risk is shrunk toward the pooled class–horizon risk:

$$\rho_{s,h,c} = \frac{n_{s,h,c}}{n_{s,h,c}+\tau}, \qquad \widetilde{R}_{s,h,c} = \rho_{s,h,c}R_{s,h,c} + (1-\rho_{s,h,c})R_{h,c}, \tag{14}$$

where $n_{s,h,c}$ is the number of valid training sample–horizon pairs assigned to group (s, h, c), $R_{h,c}$ is the pooled risk for class c at horizon h, and τ controls the amount of reliability shrinkage. Well-supported groups therefore retain more of their group-specific risk, whereas sparse groups rely more strongly on the pooled reference.

The resulting shrunk risks are converted into bounded group weights:

$$\alpha_{s,h,c} = \mathrm{clip}\left(\exp\left[\beta\left(\widetilde{R}_{s,h,c} - \bar{R}\right)\right], \frac{1}{M}, M\right), \tag{15}$$

where $\bar{R}$ is the mean shrunk risk across available groups, β controls sensitivity to relative risk differences, and M bounds the maximum reweighting factor. The resulting group weights are detached from gradient computation before being applied to the classification loss.

Let Ω denote the set of valid sample–horizon pairs for the current task, and let $\ell_{i,h}$ denote the base focal classification loss for sample i at horizon h. The unweighted empirical risk is

$$\mathcal{L}_{\mathrm{ERM}} = \frac{1}{|\Omega|}\sum_{(i,h)\in\Omega}\ell_{i,h}, \tag{16}$$

while the group-weighted risk is

$$\mathcal{L}_{\mathrm{grp}} = \frac{\sum_{(i,h)\in\Omega}\alpha_{s_i,h,y_{i,h}}\ell_{i,h}}{\sum_{(i,h)\in\Omega}\alpha_{s_i,h,y_{i,h}}}. \tag{17}$$

The multiclass classification objective interpolates between the two:

$$\mathcal{L}_{\mathrm{cls}} = (1-\eta)\mathcal{L}_{\mathrm{ERM}} + \eta\mathcal{L}_{\mathrm{grp}}, \tag{18}$$

where $\eta \in [0, 1]$ controls the contribution of reliability-aware group reweighting.

C.5 Auxiliary Objectives

MISTNet additionally uses binary auxiliary objectives that reflect nested operational states. For precipitation/thunderstorm forecasting, the first auxiliary objective distinguishes NoRain

from $\{\mathrm{RainNoThunder}, \mathrm{Thunder}\}$, while the second is optimized over the subset with labels in $\{\mathrm{RainNoThunder}, \mathrm{Thunder}\}$ and distinguishes Thunder from RainNoThunder. For flight-condition forecasting, the auxiliary objective models IFR-or-worse conditional on degraded flight conditions. Each auxiliary head shares the horizon-specific prediction representation $\mathbf{q}_{s,t,h}$ and is optimized only over its eligible subset.

For auxiliary objective k, let $\Omega_k \subseteq \Omega$ denote its eligible sample–horizon pairs, $a_{i,h}^{(k)}$ the auxiliary prediction, and $u_{i,h}^{(k)}$ the corresponding binary target. The auxiliary loss is

$$\mathcal{L}_{\mathrm{aux}}^{(k)} = \frac{1}{|\Omega_k|} \sum_{(i,h)\in\Omega_k} \mathrm{BCE}\left(a_{i,h}^{(k)}, u_{i,h}^{(k)}\right). \tag{19}$$

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \frac{\lambda_{\mathrm{aux}}}{K} \sum_{k=1}^{K} \mathcal{L}_{\mathrm{aux}}^{(k)}, \tag{20}$$

where K is the number of auxiliary objectives and λ_{aux} controls their aggregate contribution. The auxiliary heads and group-risk machinery are used only during training; inference uses only the task-specific multiclass classifiers.

C.6 Computational Complexity and Efficiency

We analyze the inference-path complexity of MISTNet; reliability-aware group reweighting, class-balanced sampling, and auxiliary objectives are used only during training. Let T_A and T_M denote the AWOS and METAR history lengths, V_A and V_M the corresponding numbers of encoded variables, $V = V_A + V_M$, $G = P^2$ the number of spatial tokens in a $P \times P$ GRIB patch, C_G the number of GRIB channels, D the hidden dimension, L the number of cross-variable Transformer layers, and H the number of forecast horizons.

The local encoder applies lightweight temporal processing independently to each variable before cross-variable interaction, giving a temporal cost of

$$\mathcal{O}\left(D^2(V_A T_A + V_M T_M)\right). \tag{21}$$

Self-attention is then applied over the compact variable-token set rather than over all historical time steps, with cost

$$\mathcal{O}\left(L(V^2 D + V D^2)\right). \tag{22}$$

Thus, the high-frequency history length enters linearly rather than through temporal self-attention of order $\mathcal{O}((T_A + T_M)^2 D)$.

For the regional branch, convolutional encoding of the GRIB patch requires

$$\mathcal{O}\left(G(C_G D + D^2)\right), \tag{23}$$

up to constant kernel factors. Site–background fusion uses the airport representation as a query over the G regional tokens, resulting in linear dependence on the number of grid tokens rather than full pairwise self-attention over the regional grid. The multi-horizon prediction module adds $\mathcal{O}(HD^2)$.

The overall inference complexity is therefore

$$\mathcal{O}\left(D^2(V_A T_A + V_M T_M) + L(V^2 D + V D^2) + G(C_G D + D^2) + HD^2\right). \tag{24}$$

Reliability-aware group reweighting maintains training-time statistics over airport–horizon–class groups and does not introduce additional inference parameters or inference-time computation.

Measured efficiency. MISTNet contains 1.10M trainable parameters. On a single NVIDIA RTX 4090, it jointly produces predictions for the 6-, 12-, and 24-hour horizons with an average latency of 9.27 ms per sample and a P95 latency of 9.44 ms under the reported single-sample inference setting. With batch size 32, throughput reaches 1,697.7 samples/s with 539.4 MB peak GPU memory. These measurements characterize the reference implementation and are intended primarily as reproducibility information rather than hardware-independent performance guarantees.

Table 11: Component ablations and training-objective comparisons for MISTNet. Results are averaged over the 6-, 12-, and 24-hour horizons and over seeds 42, 43, and 44.

Variant	Flight		Precip.	
	F1 ↑	CSI ↑	F1 ↑	CSI ↑
MISTNet	**0.2739**	**0.1592**	**0.0984**	**0.0519**
w/o Site–Background Fusion	0.2487	0.1424	0.0971	0.0512
w/o GRIB	0.2406	0.1377	0.0783	0.0410
w/o Group Reweighting	0.2501	0.1432	0.0413	0.0212
Group Reweighting w/o Shrinkage	0.2584	0.1486	0.0966	0.0509
Canonical Group DRO	0.2451	0.1401	0.0854	0.0447
w/o Airport-Time Bias	0.2285	0.1294	0.0769	0.0401

D Additional Experimental Results

D.1 Reference Model Ablations

Table 11 evaluates the main architectural and training components of MISTNet using rare-event F1 and CSI. For flight-condition forecasting, removing GRIB or the airport-time contextual bias produces the largest degradations, indicating that regional atmospheric information and airport/time-specific context are particularly useful for this task. For precipitation/thunderstorm forecasting, removing group reweighting causes the largest reduction in rare-event performance, with F1 decreasing from 0.0984 to 0.0413 and CSI from 0.0519 to 0.0212. Removing reliability shrinkage results in a small reduction in the reported mean scores relative to the full configuration, while canonical Group DRO also remains below the full MISTNet configuration on both tasks.

D.2 Leave-Airports-Out Evaluation

Protocol. The 37 airports are partitioned into five mutually exclusive folds, with 7–8 airports held out in each fold. For each fold, MISTNet is trained using only the training split of the remaining 29–30 source airports, and checkpoint selection is performed exclusively on the validation split of those source airports. The selected checkpoint is then evaluated directly on the test split of the held-out target airports. No observations from held-out airports are used for model training or checkpoint selection, and no target-airport fine-tuning, calibration, or parameter update is performed before evaluation. The experiment is conducted on the precipitation/thunderstorm task using the same 6-, 12-, and 24-hour horizons as the standard benchmark.

Transferable airport profile. The standard MISTNet airport profile contains both a learned airport-identity embedding and encoded static airport attributes. Because a learned identity embedding is unavailable for airports excluded from training, the airport-ID component is set to zero in all leave-airports-out experiments. The remaining profile contains only transferable static attributes: latitude, longitude, elevation, and an elevation-missing indicator. Consequently, predictions for held-out airports do not depend on target-specific learned identity parameters.

Training configuration. The leave-airports-out experiments use the same model architecture and rare-event learning configuration as the corresponding standard precipitation/thunderstorm experiments. Reliability-aware group reweighting uses $\tau = 10$, selected using source-airport validation data and fixed across all held-out folds. Experiments are repeated with seeds 42, 43, and 44 using identical cross-airport partitions. Aggregate leave-airports-out results are reported in the main paper; Table 12 provides the corresponding horizon-wise results.

Cross-airport performance decreases as the forecast horizon increases, particularly for rare-event precision and CSI. Nevertheless, MISTNet retains nonzero Thunder detection at all three horizons, indicating that the learned representation transfers beyond the airports observed during training, although the longer-horizon setting remains substantially more difficult.

Table 12: Horizon-wise leave-airports-out performance of MISTNet on precipitation/thunderstorm forecasting. Values are mean ± standard deviation over seeds 42, 43, and 44 after averaging the five held-out-airport folds within each seed.

Horizon	Macro-F1	Prec.	POD	F1	CSI
6 h	0.4631 ± 0.0023	0.0774 ± 0.0091	0.1340 ± 0.0218	0.0926 ± 0.0067	0.0490 ± 0.0039
12 h	0.4337 ± 0.0021	0.0533 ± 0.0010	0.1090 ± 0.0158	0.0682 ± 0.0018	0.0356 ± 0.0008
24 h	0.3940 ± 0.0023	0.0436 ± 0.0064	0.1000 ± 0.0133	0.0575 ± 0.0076	0.0298 ± 0.0041

Table 13: Hyperparameter search values and final MISTNet configurations.

Parameter	Evaluated values	Flight	Precip.
Learning rate	$\{3\times10^{-4}, 5\times10^{-4}, 10^{-3}\}$	3×10^{-4}	3×10^{-4}
Dropout	$\{0.10, 0.20\}$	0.20	0.10
p_s	$\{0.25, 0.50, 0.75\}$	0.25	0.50
η	$\{0.15, 0.25, 0.50, 0.75\}$	0.25	0.75
β	$\{2.0, 3.5, 5.0, 7.0\}$	2.0	5.0
τ	$\{10, 25, 50, 100\}$	50	10
M	$\{2, 3, 4\}$	3	3

D.3 Hyperparameter Selection

All hyperparameters are selected using validation data only, and the test set is never used for hyperparameter ranking, model selection, or parameter tuning. Initial joint searches are performed using seed 42, after which competitive configurations are re-evaluated over seeds 42, 43, and 44 before selecting the final task-specific settings. Table 13 summarizes the evaluated values and final configurations; complete trial-level search logs are omitted for compactness.

Here, p_s denotes the exponent used in tempered inverse-frequency sampling, η controls interpolation between empirical and group-weighted objectives, β controls sensitivity to risk differences, τ controls reliability shrinkage, and M specifies the maximum reweighting factor.

D.4 Final MISTNet Training Configuration

Table 14 summarizes the configurations used for the final Flight and Precipitation models. The two tasks are trained independently with AdamW and no learning-rate scheduler. Both use a learning rate of 3×10^{-4}, weight decay of 10^{-4}, and a mini-batch size of 32. We retain the PyTorch AdamW defaults $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, and clip gradients to a maximum global norm of 1.0. Training runs for at most 50 epochs. Checkpoints are selected by the highest validation Macro-F1 averaged across the 6-, 12-, and 24-h horizons, with early stopping after 10 consecutive epochs without improvement.

The shared hidden dimension is $D = 96$. The cross-variable Transformer contains two layers with four attention heads, and the same number of heads is used in local-token fusion and site–background cross-attention. The weather-phenomenon encoder contains one Transformer layer. All models use a 12-h input history and a 15×15 GRIB patch. Continuous normalization statistics are fitted using the training split only.

The final implementation follows the training objective in Equation 20, with $\mathcal{L}_{\text{cls}}$ defined in Equation 18. The auxiliary objectives are optimized with binary cross-entropy on their corresponding eligible subsets. The task-specific values of λ_{aux} are reported in Table 14.

Table 14: Configuration of the final MISTNet models. Parameters not shown as task-specific are shared by Flight and Precipitation.

Configuration	Flight	Precipitation
Optimizer	AdamW	AdamW
Learning rate	3×10^{-4}	3×10^{-4}
Weight decay	10^{-4}	10^{-4}
Batch size	32	32
Maximum epochs	50	50
Early-stopping patience	10	10
Validation criterion	Mean Macro-F1	Mean Macro-F1
Hidden dimension D	96	96
Transformer layers / heads	2 / 4	2 / 4
Dropout	0.20	0.10
History / GRIB patch	12 h / 15×15	12 h / 15×15
Focal-loss γ	2.0	2.0
Sampler exponent p_s	0.25	0.50
EMA rate μ	0.05	0.05
Group-weighting strength η	0.25	0.75
Risk sensitivity β	2.0	5.0
Shrinkage τ	50	10
Maximum reweighting factor M	3.0	3.0
Auxiliary weight λ_{aux}	0.25	0.50
Random seeds	$\{42, 43, 44\}$	