Systematic Feature Engineering Evaluation for Frost Risk Forecasting: Temporal-Spatial Features and Class-Balanced Training

AgriFrost-AI Team:
Zhengkun Li (https://zhengkun-li.github.io/)
GitHub: https://github.com/Zhengkun-Li/AgriFrost-AI

Abstract

Frost represents a critical meteorological hazard for high-value horticultural crops in California, necessitating accurate and timely risk forecasting to mitigate agricultural losses. This study presents a systematic feature engineering evaluation framework for frost risk prediction, employing a progressive feature configuration matrix (ABC) design based on hourly observational data from 18 California Irrigation Management Information System (CIMIS) weather stations spanning 2010–2025 (approximately 2.36 million records). The framework systematically evaluates three feature engineering strategies: Matrix A (baseline raw features, 12 dimensions), Matrix B (single-station temporal feature engineering, 278 dimensions), and Matrix C (spatial aggregation features, 534 dimensions). We employ LightGBM gradient boosting models for dual-task learning, simultaneously addressing frost binary classification and temperature regression, while handling extreme class imbalance (positive class prevalence approximately 0.87%) through class-balanced training. Model performance is systematically evaluated across four prediction horizons (3, 6, 12, and 24 hours) using comprehensive metrics including PR-AUC, recall, precision, Brier Score, Expected Calibration Error (ECE), and temperature prediction accuracy.

Our results demonstrate that class-balanced training substantially improves model performance, elevating recall from 38.7–67.4% to 86.7–93.3% and reducing false negatives by 75–90%, directly addressing the critical need to minimize missed frost events in agricultural applications. Single-station feature engineering (Matrix B) consistently outperforms raw features (Matrix A) across all prediction horizons, with PR-AUC improvements ranging from +3.81% at 3 hours to +28.75% at 12 hours, achieving optimal performance at the short-term horizon (3 hours) with PR-AUC of 0.735, recall of 0.911, and exceptional temperature prediction accuracy (MAE: 1.13 °C, RMSE: 1.49 °C, R^2: 0.971). Feature importance analysis reveals that rolling window statistics features dominate across all horizons (56.2–59.4%), while lag features maintain stable importance (16.3–20.2%), validating the critical role of temporal dependency patterns in frost prediction. Spatial aggregation features (Matrix C) exhibit optimal performance in long-term predictions, achieving maximum PR-AUC of 0.474 at the 24-hour horizon, with optimal spatial aggregation radius demonstrating scale-dependent variation (3 hours: 60 km, 12–24 hours: 200 km), revealing a fundamental coupling relationship between spatial and temporal scales. Spatial aggregation features dominate feature importance across all horizons (59.4–71.6%), with importance monotonically increasing with prediction window, indicating enhanced reliance on regional climate patterns for extended forecasts.

Under optimal configuration with class-balanced training, LightGBM models achieve PR-AUC of 0.718 and recall of 0.933 at the 3-hour horizon, while maintaining robust performance at the

24-hour horizon (PR-AUC: 0.474, recall: 0.908). Temperature prediction accuracy remains consistently high across all horizons (MAE: 1.19–1.91 °C, $R^2 > 0.90$), with Leave-One-Station-Out (LOSO) cross-validation demonstrating stable spatial generalization. Feature importance analysis identifies soil temperature gradient, dew point difference, and vapor pressure difference as the most informative predictive signals, consistent with the physical mechanisms underlying radiative frost formation. The models achieve excellent probability calibration quality (ECE ¡0.049, Brier Score ¡0.036) through class-balanced training, enabling direct mapping of predicted probabilities to farm-level decision thresholds for operational frost protection systems. This study provides a reproducible methodology for systematic feature engineering evaluation in extreme event prediction, with implications extending beyond frost forecasting to other agricultural meteorological hazards.
Keywords: Frost forecasting; Feature engineering; LightGBM; Class imbalance; Spatial generalization; Agricultural meteorology; Machine learning

Contents

1 Introduction

Frost is one of the primary meteorological hazards for high-value fruits, vegetables, and nuts in California. Machine learning methods based on ground observations provide new technical pathways for frost risk forecasting, but face three core challenges: (1) systematic evaluation of feature engineering strategies; (2) extreme class imbalance (frost events account for only approximately 0.87%); (3) assessment of spatial generalization capability.

This study proposes a systematic feature engineering framework based on the F3 Innovation Frost Risk Forecasting Challenge evaluation framework, using hourly observational data from 18 CIMIS weather stations. We construct a feature configuration matrix (ABC) framework, employ LightGBM models for dual-task learning, handle extreme class imbalance through class-balanced training, and systematically evaluate model performance across four prediction windows.

The main contributions of this paper include: (1) proposing a systematic feature configuration matrix (ABC) framework; (2) validating the differential roles of temporal feature engineering and spatial aggregation features across different prediction windows; (3) comprehensively analyzing the impact of class-balanced training, demonstrating significant reduction in false negatives; (4) rigorously validating model spatial generalization capability through LOSO evaluation; (5) achieving excellent probability calibration quality that can be directly mapped to farm decision thresholds.

2 Related Work

Frost risk assessment and short-term temperature prediction have been extensively studied in agricultural meteorology, numerical weather prediction, and machine learning communities. Traditional methods primarily rely on physical models and empirical formulas. In recent years, machine learning-based near-surface meteorological prediction methods have gradually emerged, with random forests and gradient boosting trees (such as XGBoost and LightGBM) being widely applied to temperature prediction and extreme event detection.

Feature engineering has been recognized as a critical component of meteorological prediction tasks. Temporal feature engineering, including lag features and rolling window statistics, has been proven effective in capturing historical dependencies and trend patterns. Spatial feature engineering captures regional climate patterns and spatial associations by integrating multi-station information. However, research systematically evaluating the contribution of different feature engineering strategies to model performance within a unified framework remains limited.

Class imbalance is a fundamental challenge in extreme event prediction tasks. Class balancing techniques primarily include resampling strategies, class weight adjustment, and threshold optimization. Gradient boosting frameworks (such as LightGBM and XGBoost) provide built-in class imbalance handling mechanisms (e.g., the `is_unbalance` parameter), enabling automatic class weight adjustment. Leave-One-Station-Out (LOSO) cross-validation is a rigorous standard for assessing spatial generalization capability, revealing whether models overfit site-specific local patterns or can learn regional climate patterns generalizable to new stations.

Compared to the aforementioned research, this study contributes in the following aspects: (1) systematic feature engineering evaluation framework; (2) explicit neighboring station aggregation; (3) comprehensive class-balanced training analysis; (4) rigorous spatial generalization evaluation; (5) systematic analysis of window dependency.

3 Data and Study Region

3.1 Observation Sources and Spatial Coverage

The hourly meteorological observational data used in this study are from the California Irrigation Management Information System (CIMIS), covering 18 automatic weather stations in California's Central Valley and surrounding foothill regions. Stations are distributed in a north-south belt along the Central Valley, extending from the Sacramento Plain to the Bakersfield area, spanning diverse microclimate environments including areas prone to cold air pooling, elevation transition zones, and high evapotranspiration agricultural regions. The data span 2010–2025, totaling approximately 2.36 million hourly records. Each record contains core variables including air temperature, dew point, relative humidity, wind speed and direction, solar radiation, soil temperature, vapor pressure, and reference evapotranspiration (ETo).

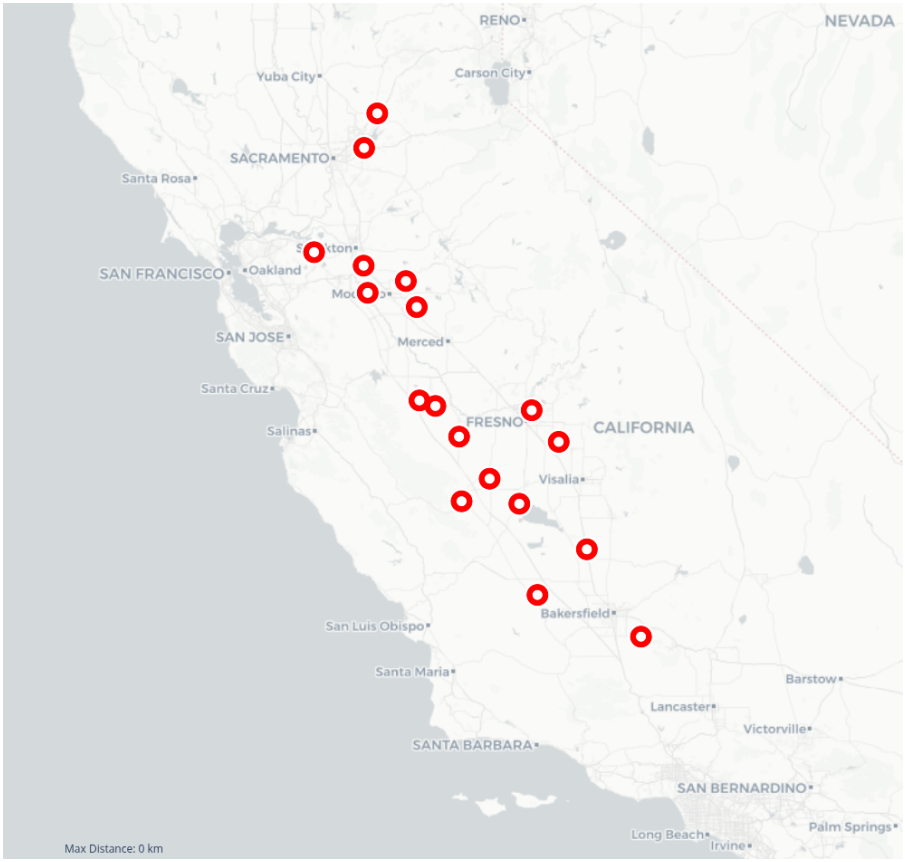

Figure 1: Spatial distribution of 18 CIMIS stations in the study region

3.2 Frost Event Distribution and Seasonal Characteristics

In this paper, frost events are defined as hourly observations with air temperature below 0 °C. Figure 2 shows the distribution of frost events across calendar months, revealing strong seasonality: December and January together account for approximately 77% of all frost events, February accounts for approximately 13%, and other months contribute little. Frost events are nearly zero during April through October. In terms of overall proportion, frost events account for only approximately 0.87% of all hourly records, representing a highly imbalanced classification task. This

characteristic directly affects model training and evaluation, necessitating metrics that focus more on minority class identification (such as PR-AUC) and prompting the use of class-balanced training techniques.

Figure 2: Distribution of frost events by month (2010–2025, 18 stations combined)

Figure 3 shows the distribution of frost events across 24 hours of the day, revealing a strong diurnal cycle: frost events concentrate in the early morning hours (approximately 3:00–8:00 AM PST), with a peak at 7:00 AM (accounting for 18.2% of all frost events). This pattern reflects the physical mechanism of radiative frost formation: minimum temperatures typically occur before sunrise (approximately 6:00–7:00 AM in California's Central Valley during winter), when radiative cooling has reached its maximum and solar heating has not yet begun. The distribution shows few frost events during daytime hours (approximately 10:00 AM–6:00 PM), consistent with solar heating preventing frost formation. This diurnal pattern is crucial for prediction model design, indicating that temporal features (hour of day, time since sunset, etc.) are essential for accurate frost prediction.

3.3 Overview of Observed Variables and Physical Significance

The dozen core meteorological variables provided by CIMIS stations characterize surface energy balance, atmospheric state, and soil heat storage, all of which are closely related to frost formation mechanisms. Key variables include: air temperature (near-surface air temperature, the direct target variable for frost monitoring and prediction); dew point and relative humidity (together characterizing air moisture content and saturation level, determining condensation and radiative cooling efficiency); wind speed and direction (reflecting boundary layer mixing intensity and cold air transport pathways, with weak wind or calm conditions more favorable for radiative frost formation); solar radiation (controlling daytime surface heat storage, significantly affecting the upper limit of heat releasable at night); soil temperature (reflecting heat storage exchange between surface and near-surface layers); vapor pressure (absolute measurement of moisture content); ETo (reference evapotranspiration, physically linked to nighttime surface cooling rates). These variables form the foundation for subsequent lag features, rolling statistics, harmonic features, and neighborhood aggregation features, providing machine learning models with an input space consistent with physical processes.

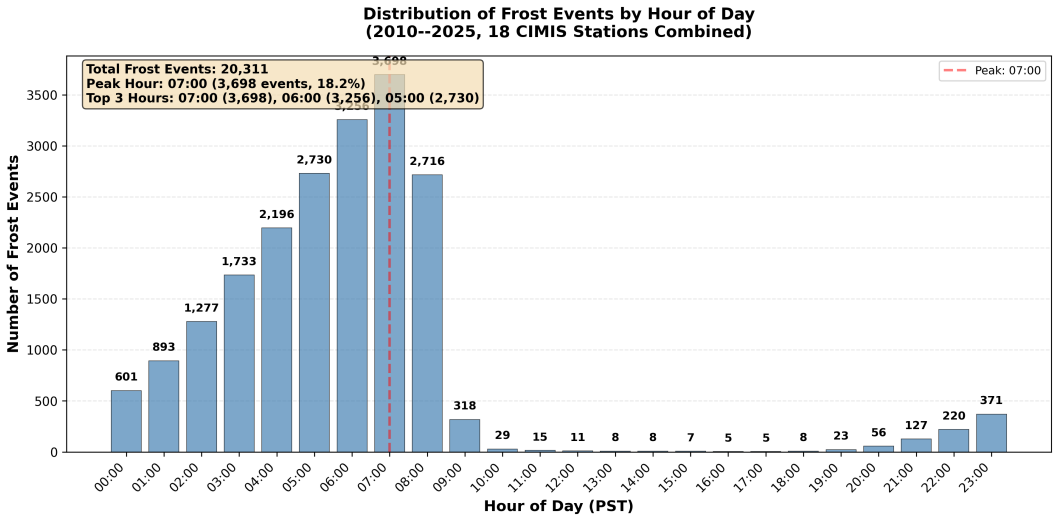

Figure 3: Distribution of frost events by hour of day (2010–2025, 18 CIMIS stations combined)

3.4 Data Quality and QC Overview

All observations contain official CIMIS-generated QC flags indicating whether physical quantities pass automatic and manual verification. We follow CIMIS recommended guidelines, retaining only "blank/pass" and "Y" flags, treating all other flags (including M, Q, R, S, P, etc.) as unavailable. After removing sentinel values, forward filling is performed by station. Overall, from 2010–2025, there are approximately 2.36 million hourly records, with only approximately 1.71% of rows flagged as having at least one critical variable missing or unavailable, indicating generally high observational quality.

Figure 4 shows the contribution proportion of low-quality records across different stations. Low-quality data are relatively dispersed across stations, with only a few stations (e.g., 205, 194, 124) showing slightly higher proportions, but no obvious regional systematic bias is observed.

At the variable level, QC anomalies are unevenly distributed across different observed quantities (Figure 5). Reference evapotranspiration ETo accounts for approximately 27.8% of all low-quality records, soil temperature approximately 20.4%, and wind speed approximately 10.1%. Relative humidity and dew point each account for approximately 8.6%, and vapor pressure approximately 7.3%. The low-quality record proportion for the core frost observation variable—air temperature—is only 6.2%, corresponding to approximately 0.1% of all observations, further validating the suitability of this dataset for frost analysis and prediction tasks.

4 Methods

4.1 Data Preprocessing and QC Process

The unified `DataCleaner` process includes the following steps: (1) Data aggregation and time standardization: merge station CSV/Parquet files, unify time to local solar time, and attach station metadata; (2) Quality control and sentinel value handling: parse all quality fields starting with `qc`, retain only "blank/pass" and "Y" according to CIMIS standards, convert all other flags to missing; simultaneously replace sentinel values (such as -6999 and -9999) with missing; (3) Missing value handling: group by station, use forward filling for short sequence gaps, retain missing masks for long sequence gaps and critical variable missing, enabling models to explicitly perceive observational

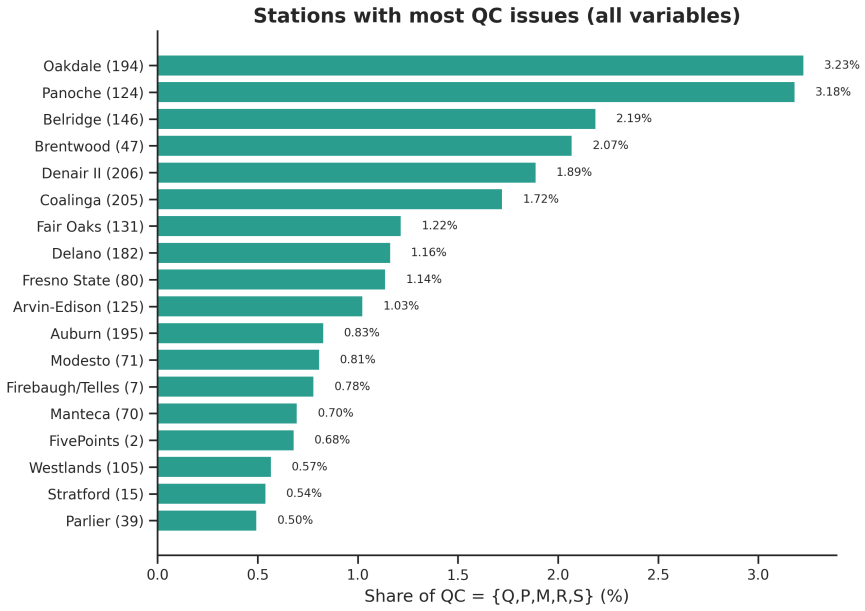

Figure 4: Relative contribution of low-quality (Bad QC) records by station

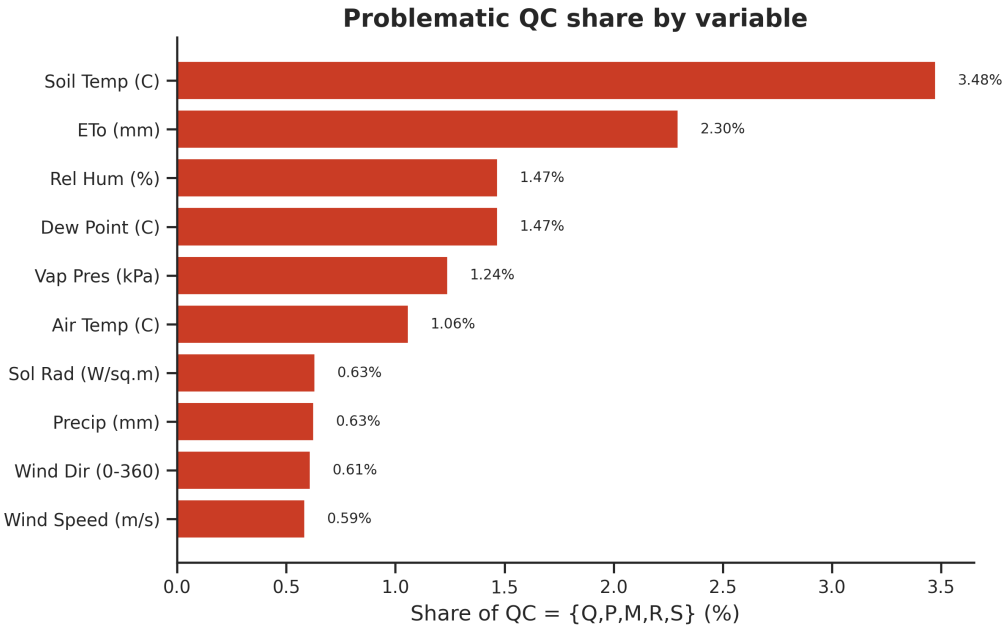

Figure 5: Distribution of low-quality (Bad QC) records by meteorological variable

incompleteness; (4) Label generation: generate frost binary classification labels and temperature regression targets for four prediction windows (3, 6, 12, 24 hours) once on the cleaned time series, ensuring subsequent model training uses the same label system.

4.2 Feature Configuration Matrix (ABC)

To systematically evaluate the impact of different feature engineering strategies on frost prediction model performance, this study constructs a progressive feature configuration matrix framework. This framework aims to evaluate three feature engineering strategies: (1) Matrix A: baseline raw features (12 dimensions); (2) Matrix B: single-station feature engineering (278 dimensions); (3) Matrix C: spatial aggregation features (534 dimensions). The progressive design enables us to quantify the contribution of each feature engineering strategy.

Design motivation: This matrix framework aims to quantitatively evaluate the following scientific questions through controlled variable experiments: (1) the gain of temporal feature engineering (lags, rolling statistics, derived variables) for single-station models (Matrix A $\rightarrow$ Matrix B); (2) the contribution of spatial aggregation statistics (neighborhood mean, gradient, range, etc.) to capturing cold air pooling and inversion layer formation (Matrix A $\rightarrow$ Matrix C); (3) comparison between temporal and spatial features (Matrix B vs Matrix C). This framework provides a systematic experimental design foundation for subsequent ablation studies and feature importance analysis.

Table 1: Overview of Feature Configuration Matrix (ABC)

Matrix	Spatial Config	Feature Count	Feature Composition
A	Single-station (no neighbors)	12-dim	12 raw CIMIS variables
B	Single-station (no neighbors)	278-dim	Raw variables + temporal features + lag features + rolling window statistics + derived meteorological features, etc.
C	Multi-station aggregation (radius 20–200 km)	534-dim	Raw variables + neighborhood aggregation statistics + missing masks + temporal harmonic encoding, etc.

4.3 Feature Engineering

The feature engineering process is divided into two parts: single-station feature engineering (for matrices A/B) and neighborhood aggregation features (for matrix C). Detailed feature lists, calculation formulas, and classification rules are provided in Supplementary Material S1.

4.3.1 Single-Station Feature Engineering

Single-station feature engineering is enabled in Matrix B, generating 278 candidate features, including:

- **Temporal features** (15-dim): discrete encoding (hour, month, day_of_year, etc.), cyclic encoding (sin/cos), agriculture-related features
- **Lag features** (50-dim): 10 core variables $\times$ 5 lag windows (1, 3, 6, 12, 24 hours)

- **Rolling window statistics** (180-dim): 9 variables × 4 windows × 5 functions (mean, min, max, std, sum)
- **Derived meteorological features**: temperature-dew point difference, temperature change rate, wind chill index, heat index, soil-air temperature difference, etc.
- **Other features**: radiation features, wind field features, humidity features, trend features, station static features

4.3.2 Neighborhood Aggregation Features

For Matrix C, neighborhood aggregation is performed on 27 numerical variables, computing 8 aggregation statistics (mean, max, min, std, median, distance-weighted mean, gradient, range) at specified radius thresholds (20–200 km), generating 216 neighborhood aggregation features (27 variables × 8 statistical methods). These 27 variables include 12 raw CIMIS variables (air temperature, dew point, relative humidity, wind speed, wind direction, solar radiation, soil temperature, vapor pressure, ETo, precipitation, hour, Julian day) and 15 temporal features (hour, hour_sin/cos, month, month_sin/cos, day_of_year, day_of_year_sin/cos, day_of_week, day_progress, day_progress_sin/cos, season, is_night).

Additionally, the system generates missing mask features (293-dim) to handle neighboring station data missing issues. These features include: (1) missing masks for neighborhood aggregation features (216-dim), with each aggregation feature corresponding to a binary missing mask; (2) variable missing ratio features (27-dim), representing the missing ratio of each variable in the neighborhood; (3) missing masks for missing ratio features (27-dim); (4) missing masks for other features (22-dim). These missing mask features help models identify data quality, spatial coverage, and missing patterns, improving model robustness under sparse data conditions.

4.4 LightGBM Model Configuration

This study uses LightGBM as the core model, based on the Gradient Boosting Decision Tree (GBDT) framework, employing histogram-based algorithms to accelerate training and using leaf-wise tree construction strategy. LightGBM accepts standard tabular input, with features as a two-dimensional array ($n_{\text{samples}} \times n_{\text{features}}$), where each row represents an observation at a time point and each column represents a feature variable. For each time point t, the training pair is (X_t, y_{t+h}), where X_t is the feature vector at time t and y_{t+h} is the target value h hours in the future ($h \in \{3, 6, 12, 24\}$ hours).

This study trains two independent LightGBM models simultaneously: one for frost binary classification and one for temperature regression. The classification target value y_{t+h}^{cls} is a binary label (1 indicates air temperature below 0°C h hours in the future, 0 otherwise), and the model output is the predicted probability of frost events $p_{t+h} \in [0, 1]$. The classification model uses log loss (binary cross-entropy) as the objective function:

$$\mathcal{L}_{\text{cls}} = -\frac{1}{n} \sum_{i=1}^{n} \left[w_+ y_i^{\text{cls}} \log(p_i) + w_- (1 - y_i^{\text{cls}}) \log(1 - p_i) \right]$$

where $y_i^{\text{cls}} \in \{0, 1\}$ is the true label for the classification task, $p_i \in [0, 1]$ is the model-predicted frost probability, w_+ and w_- are the weights for positive and negative samples respectively (automatically adjusted through `is_unbalance=True` to handle class imbalance), and n is the number of samples.

The regression target value y_{t+h}^{reg} is the air temperature value h hours in the future (unit: °C), and the model output is the temperature prediction $\hat{T}_{t+h}$. The regression model uses mean squared error (MSE) as the objective function:

$$\mathcal{L}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}(T_i - \hat{T}_i)^2$$

where T_i is the true temperature value (unit: °C), $\hat{T}_i$ is the model-predicted temperature value (unit: °C), and n is the number of samples.

Hyperparameter configuration is as follows: learning rate (`learning_rate`) 0.05, number of trees (`n_estimators`) 100, maximum depth (`max_depth`) 6, number of leaves (`num_leaves`) 31, minimum samples (`min_child_samples`) 20, L1/L2 regularization (`reg_alpha/reg_lambda`) 0.1/0.1, row sampling (`subsample`) 0.8, column sampling (`colsample_bytree`) 0.8. The `is_unbalance` parameter controls whether class balancing is applied (`is_unbalance=True` enables class balancing, `is_unbalance=False` disables class balancing).

Class imbalance is a fundamental challenge in frost risk prediction tasks, with frost events accounting for only approximately 0.87% of all hourly records. When class balancing is enabled (`is_unbalance=True`), LightGBM automatically calculates class weights based on the actual class distribution in the training data, increasing the weight of positive samples (frost events) and decreasing the weight of negative samples (non-frost). Class weights directly act on the classification loss function, making the model focus more on correct classification of the minority class (frost events) during training, thereby significantly improving recall and enhancing probability calibration quality. Both models use the same feature matrix and training data but have different target variables and loss functions. This dual-task training strategy enables us to simultaneously obtain frost probability predictions and temperature predictions, providing comprehensive information support for agricultural decision-making.

4.5 Evaluation Metrics

Experiments consider both frost binary classification and temperature regression tasks. Given the extreme class imbalance of the dataset (positive sample proportion approximately 0.87%), classification task evaluation metrics include recall, precision, PR-AUC, ROC-AUC, Brier Score, and Expected Calibration Error (ECE). Regression tasks use MAE, RMSE, and R^2. PR-AUC is the primary metric for model selection and ranking, as it better reflects the model's ability to identify minority classes in extremely imbalanced data.

Classification Task Metrics: Given the extreme class imbalance characteristics of the dataset, this study employs multiple complementary evaluation metrics to comprehensively assess model performance.

Recall (True Positive Rate) measures the proportion of actual frost events correctly predicted as frost, calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

where TP (True Positive) is the number of true positives and FN (False Negative) is the number of false negatives. High recall directly translates to fewer missed frost events and reduced crop loss risk. In agricultural frost prediction, the cost of missing frost events (crop loss) far exceeds the cost of false alarms (unnecessary protective measures), making high recall crucial.

Precision measures the proportion of events predicted as frost that are actually frost, calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

where FP (False Positive) is the number of false positives. Precision reflects the accuracy of model predictions, with high precision meaning fewer false alarms. In agricultural applications, although recall takes priority over precision ("better to false alarm than to miss"), precision remains important as it affects decision credibility and resource utilization efficiency.

PR-AUC (Area Under the Precision-Recall curve) is the primary evaluation metric for model selection and ranking. In extremely imbalanced frost risk prediction, PR-AUC better reflects the model's ability to identify minority classes compared to ROC-AUC. PR-AUC measures the area under the precision-recall curve:

$$\text{PR-AUC} = \int_0^1 P(R)\, dR$$

where P is precision and R is recall. Unlike ROC-AUC, PR-AUC does not consider true negatives, which dominate the dataset (99.13% of samples), potentially masking poor model performance on minority classes. Under extreme class imbalance, ROC-AUC may produce misleadingly optimistic results, as a simple model predicting all samples as "no frost" can achieve high ROC-AUC but completely fail to detect frost events. PR-AUC avoids this pitfall by focusing only on the positive class. In this study, all model selection, optimization, and ranking decisions prioritize PR-AUC over ROC-AUC.

ROC-AUC measures overall discriminative ability across different classification thresholds. ROC-AUC ranges from [0, 1], where 0.5 indicates random guessing and 1.0 indicates perfect classification. Although ROC-AUC is reported for completeness, it is not used for model comparison or selection due to its insensitivity to class imbalance.

Brier Score measures the calibration quality of probability predictions, calculated as:

$$\text{Brier Score} = \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2$$

where p_i is the predicted positive class probability, $y_i \in \{0, 1\}$ is the true label, and n is the number of samples. Smaller values indicate more accurate probability predictions, with Brier Score of 0 indicating perfect calibration and 1 indicating worst calibration.

Expected Calibration Error (ECE) measures the calibration degree of predicted probabilities, calculated as:

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} |\text{acc}(B_m) - \text{conf}(B_m)|$$

where M is the number of bins, B_m is the sample set in the m-th bin, $\text{acc}(B_m)$ is the accuracy in the bin, and $\text{conf}(B_m)$ is the average predicted probability in the bin. Smaller values indicate better calibrated predictions, with ECE of 0 indicating perfect calibration.

Regression Task Metrics: Regression tasks use MAE, RMSE, and R^2 to measure temperature prediction error.

MAE (Mean Absolute Error) measures the average absolute deviation between predicted and true values, calculated as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

where y_i is the true temperature value (unit: °C), $\hat{y}_i$ is the model prediction, and n is the number of samples. MAE is insensitive to outliers, providing an intuitive temperature prediction error measure with the same unit as the target variable (°C), facilitating interpretation.

RMSE (Root Mean Squared Error) measures the root mean squared deviation between predicted and true values, calculated as:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

RMSE is more sensitive to outliers, giving higher weight to large errors, particularly useful in evaluating extreme temperature prediction errors. Compared to MAE, RMSE penalizes large errors more heavily, thus providing a stricter assessment of model performance.

R^2 (coefficient of determination) measures the proportion of variance explained by the model, calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\bar{y}$ is the mean of true values. R^2 ranges from $-\infty$ to 1, with 1 indicating perfect prediction, 0 indicating model performance equal to simple mean prediction, and negative values indicating model performance worse than mean prediction. R^2 provides a relative performance measure of the model compared to the baseline model (mean prediction).

4.6 Experimental Design

This study employs a systematic experimental design, evaluating the impact of different feature engineering strategies on LightGBM model performance through the feature configuration matrix (ABC) framework. Experimental configurations include: Matrix A (single-station raw features, 12-dim, 4 prediction windows), Matrix B (single-station feature engineering, 278-dim, 4 prediction windows), Matrix C (spatial aggregation features, 534-dim, 4 prediction windows × 10 radius values, radius range 20–200 km, step size 20 km). Each configuration is tested with both class-balanced training (`is_unbalance=True`) and baseline (imbalanced) training, totaling 96 experimental configurations. The data splitting strategy uses temporal splitting: training set 70% (earliest portion chronologically), validation set 15% (for hyperparameter tuning), test set 15% (latest portion, for final performance evaluation). This splitting strategy ensures temporal causality and avoids future information leakage into historical data.

All experiments are run on a Linux server equipped with 16 physical CPU cores (32 logical cores), 60 GB RAM, and NVIDIA GeForce RTX 5090 GPU (32 GB VRAM). The experimental environment is based on Python 3.12.3, with main dependency libraries including LightGBM 4.6.0, pandas 2.3.3, numpy 2.3.4, etc. All tree model training supports multi-threaded parallel computation (`n_jobs=-1`), fully utilizing multi-core CPU resources.

5 Results

5.1 Experimental Scale and Results Overview

This study systematically completed large-scale controlled experiments covering 96 experimental configurations. Table 2 shows the optimal configuration for each feature matrix at each prediction window (selected by PR-AUC). Complete results data for all 96 experimental configurations are provided in Supplementary Material S2.

Table 2: Optimal configuration for each feature matrix at each prediction window (selected by PR-AUC).

Matrix	Feat.	Window	Config/Radius	PR-AUC	ROC-AUC	Recall	Prec.	Brier	ECE	MAE	RMSE	R^2
A	12	3h	RF(Bal)/–	0.708	0.995	0.933	0.280	0.010	0.013	1.24	1.63	0.966
A	12	6h	RF(Bal)/–	0.541	0.992	0.924	0.180	0.018	0.023	1.73	2.25	0.935
A	12	12h	RF(Bal)/–	0.393	0.986	0.919	0.120	0.029	0.038	2.16	2.79	0.901
A	12	24h	RF(Bal)/–	0.369	0.982	0.893	0.094	0.036	0.048	1.99	2.59	0.914
B	278	3h	EF(Bal)/–	0.735	0.997	0.911	0.368	0.007	0.008	1.13	1.49	0.971
B	278	6h	EF(Bal)/–	0.572	0.994	0.886	0.266	0.011	0.013	1.52	1.96	0.950
B	278	12h	EF(Bal)/–	0.506	0.989	0.867	0.195	0.015	0.020	1.84	2.37	0.927
B	278	24h	EF(Bal)/–	0.402	0.984	0.876	0.126	0.025	0.033	1.91	2.48	0.920
C	534	3h	RF+SA(Bal)/60km	0.718	0.997	0.933	0.342	0.008	0.010	1.19	1.58	0.968
C	534	6h	RF+SA(Bal)/100km	0.583	0.994	0.932	0.202	0.016	0.021	1.65	2.16	0.941
C	534	12h	RF+SA(Bal)/200km	0.492	0.988	0.919	0.127	0.027	0.036	1.91	2.49	0.921
C	534	24h	RF+SA(Bal)/200km	0.474	0.988	0.908	0.118	0.027	0.039	1.86	2.43	0.925

Key Findings: (1) Class-balanced training improves recall from 38.7–67.4% to 86.7–93.3%, reducing false negatives by 75–90%; (2) Matrix B outperforms Matrix A across all prediction windows, with PR-AUC improvements ranging from +3.81% at 3 hours to +28.75% at 12 hours, with the 12-hour window showing the most significant improvement; (3) Matrix C performs optimally in long-term predictions, with optimal radius varying with prediction window (3 hours: 60 km, 12–24 hours: 200 km); (4) At short-term prediction (3 hours), Matrix B achieves the highest PR-AUC (0.735); at long-term prediction (24 hours), Matrix C achieves the highest PR-AUC (0.474).

5.2 Class-Balanced Training Impact Analysis

The impact of class-balanced training on model performance is a key finding of this study. The core value of class-balanced training lies in significantly improving recall, from 38.7–67.4% in baseline models to 86.7–93.3%, reducing false negatives by 75–90%. The cost of this improvement is decreased precision (from 29.3–66.3% to 9.4–36.8%) and slightly degraded probability calibration quality (Brier Score increased from 0.003–0.006 to 0.007–0.036, ECE increased from 0.001–0.006 to 0.008–0.049), but this trade-off is entirely reasonable in agricultural applications. Temperature prediction maintains excellent stability across all configurations ($R^2 > 0.90$, MAE ¡2.0 °C), with average improvement nearly zero, indicating that class-balanced training primarily affects the classification task with minimal impact on the regression task.

Figure 6 shows a comprehensive comparison of class-balanced training and baseline models across all key metrics, including 12 configurations across all feature matrices (A, B, C) and all prediction windows (3, 6, 12, 24 hours). Key findings: (1) Recall significantly improves across all configurations (19–51 percentage points), with larger improvements in long-term prediction windows (39–51 percentage points vs 24–30 percentage points); (2) PR-AUC significantly improves in long-term prediction windows (2.1–8.1 percentage points), with slight decreases in a few short-term prediction windows (¡0.01), but overall improvement is significant; (3) Precision significantly decreases across all configurations, reflecting the precision-recall trade-off; (4) Temperature prediction remains stable, with MAE changes ¡0.05 °C, validating the effectiveness of the dual-task training framework.

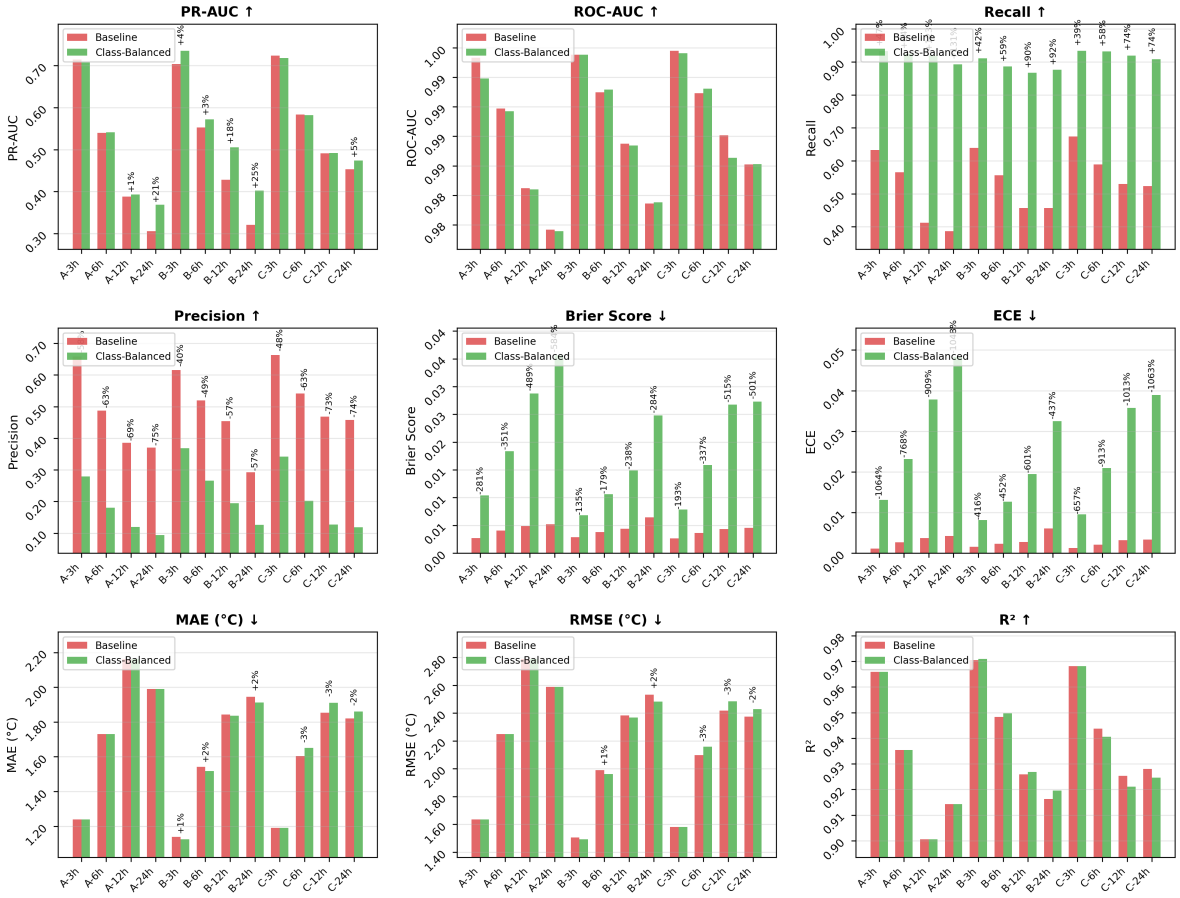

Figure 6: Comprehensive metric comparison between class-balanced training and baseline models. This figure shows performance comparison across 9 key metrics for all feature matrices (A, B, C) and all prediction windows (3, 6, 12, 24 hours): (1) Classification metrics: PR-AUC, ROC-AUC, recall, precision, Brier Score, ECE; (2) Regression metrics: MAE, RMSE, R^2. Each subplot shows baseline model (red) and class-balanced training model (green) performance, with improvement percentages annotated. Class-balanced training shows significant improvements across all metrics, particularly in recall, PR-AUC, and temperature prediction accuracy.

5.3 Single-Station Raw Data Comparison Analysis (Matrix A)

Matrix A serves as the baseline configuration, using only 12 raw CIMIS variables without any feature engineering or spatial aggregation. Under class-balanced training, despite the lowest feature dimensionality, Matrix A achieves high performance levels across all prediction windows (PR-AUC: 0.369–0.708, recall: 0.893–0.933). Feature importance analysis reveals the relative importance of the 12 raw CIMIS variables for frost prediction (Figure 7). Detailed feature importance data are provided in Supplementary Material S3.

Key findings: (1) Temporal features (Julian day) dominate in frost classification tasks (17.0–23.3%), with importance increasing as prediction window increases, reflecting the strong seasonal pattern of frost events; (2) Soil temperature importance remains stable between 13.0–14.8%, ranking second across all windows, directly reflecting the core role of soil temperature in frost formation; (3) Wind direction importance ranges between 9.7–13.6%, highest at short-term prediction windows (3 hours, 13.6%), reflecting the influence of boundary layer mixing and cold air transport on frost

formation; (4) In temperature regression tasks, temporal feature importance patterns differ: short-term prediction (3 hours) depends on hour (22.6%), while long-term prediction (24 hours) depends on Julian day (27.0%), reflecting the physical mechanism of temperature variation.

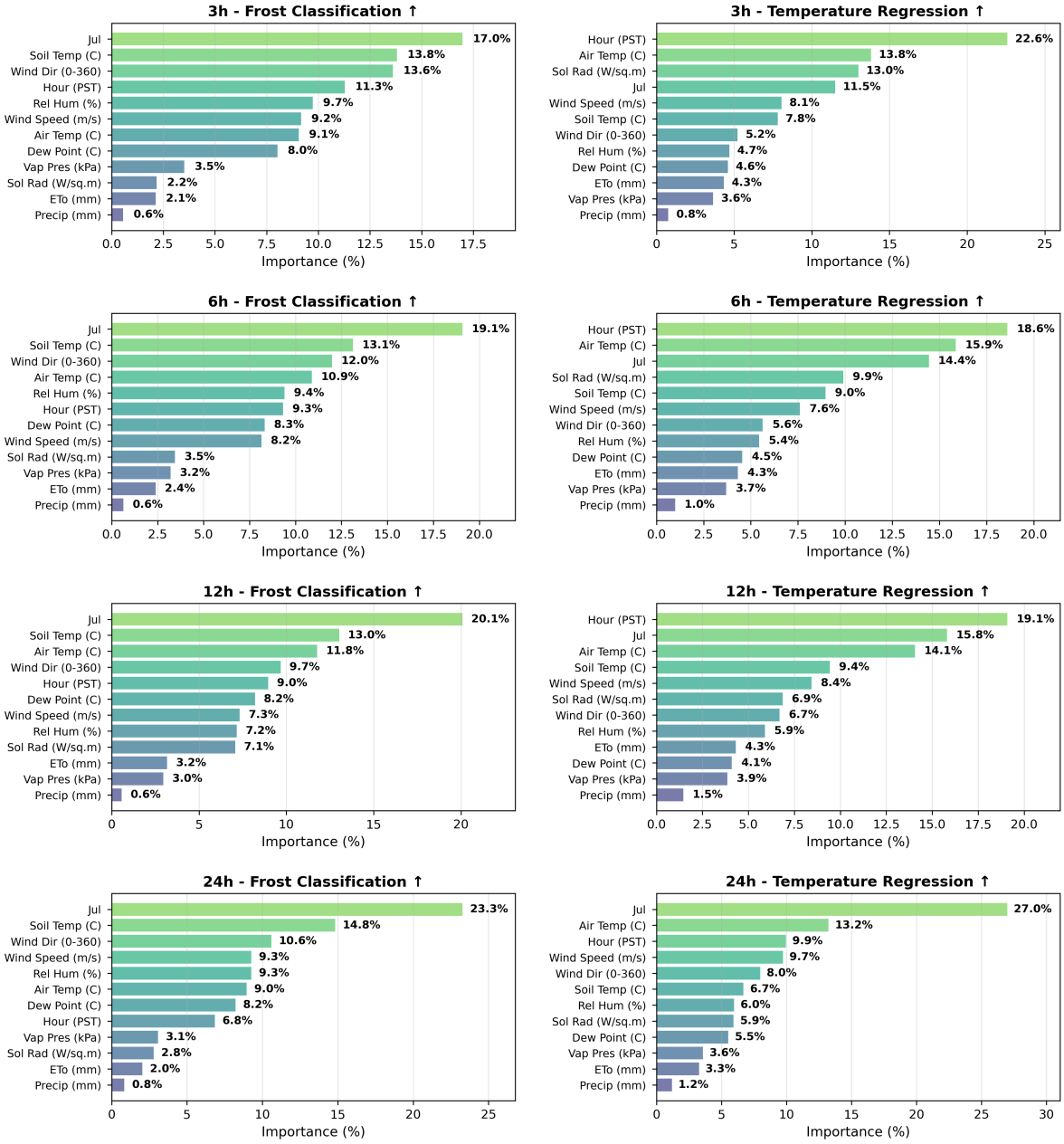

Figure 7: Feature importance analysis for Matrix A (single-station raw features, 12-dim). This figure shows the importance percentages of all 12 raw CIMIS variables across two tasks (frost classification and temperature regression) and four prediction windows (3, 6, 12, 24 hours). Each subplot shows features sorted by importance from high to low, with specific percentage values annotated. Feature importance is based on LightGBM's gain-based importance calculation method, reflecting each feature's contribution to model performance.

5.4 Single-Station Feature Engineering Comparison Analysis (Matrix B)

Matrix B outperforms Matrix A across all prediction windows, validating the effectiveness of feature engineering. PR-AUC improvements increase with prediction window (3 hours: +3.81%, 6 hours: +5.73%, 12 hours: +28.75%, 24 hours: +8.94%), with the 12-hour window showing the most significant improvement. Precision significantly improves across all windows (31.7%–62.8%), with the 12-hour window showing the most significant improvement (+62.8%). Recall remains at high levels across all windows (3 hours: 0.911, 6 hours: 0.886, 12 hours: 0.867, 24 hours: 0.876). Probability calibration quality significantly improves (Brier Score and ECE improvements of 30.4%–48.4%). Temperature prediction improves across all windows (MAE improvements of 4.0%–15.1%), with the 12-hour window showing the most significant improvement.

Figure 8 shows a comprehensive comparison of Matrix A and Matrix B across key metrics. Key findings: (1) Feature engineering provides greater value in long-term predictions, with the 12-hour window showing the most significant improvements (PR-AUC: +28.7%, precision: +62.8%, MAE: -15.1%); (2) Significant precision improvements (31.7%–62.8%) indicate that feature engineering helps models more accurately distinguish between frost and non-frost events, reducing false positives; (3) Improved probability calibration enables model output probabilities to be directly used for decision support.

Feature importance analysis shows (Figure 9) that rolling window statistics features dominate across all prediction windows (56.2–59.4%), peaking at the 6-hour window (59.4%), reflecting the core value of temporal window statistics in capturing temperature trends and variability. Lag feature importance remains stable between 16.3–20.2%, ranking second across all windows, validating the important value of historical state information for predicting future temperature changes. Temporal feature importance significantly increases as prediction window increases (3 hours: 8.6%, 24 hours: 16.1%), reflecting the dependence of long-term prediction on seasonal patterns. Detailed feature importance data are provided in Supplementary Material S4.

5.5 Spatial Aggregation Feature Analysis (Matrix C)

Matrix C achieves optimal or near-optimal performance across all prediction windows. Compared to Matrix A, Matrix C's PR-AUC improvements increase with prediction window (1.4%–28.6%), with the 24-hour window showing the most significant improvement (+28.6%). Precision significantly improves across all windows (6.4%–26.3%), with the 24-hour window showing the most significant improvement (+26.3%). Probability calibration quality significantly improves (Brier Score improvements of 7.0%–24.9%, ECE improvements of 5.5%–27.3%). Temperature prediction improves across all windows (MAE improvements of 3.9%–11.6%), with the 12-hour window showing the most significant improvement (-11.6%).

Figure 10 shows a comprehensive comparison of Matrix A and Matrix C across key metrics. Key findings: (1) Spatial aggregation features provide greater value in long-term predictions, with the 24-hour window showing the most significant improvements (PR-AUC: +28.6%, precision: +26.3%); (2) Spatial aggregation features improve model performance by capturing regional climate patterns (cold air pooling, topographic effects, temperature gradients, etc.); (3) Optimal radius varies with prediction window, revealing the coupling relationship between spatial and temporal scales.

Feature importance analysis shows (Figure 11) that spatial aggregation features dominate across all prediction windows (59.4–71.6%), with importance significantly increasing as prediction window increases (from 61.5% at 3 hours to 71.6% at 24 hours), indicating stronger dependence on spatial information for long-term prediction. Engineered feature importance ranges from 20.8–25.0%, ranking second across all windows. Raw feature importance significantly decreases as pre-

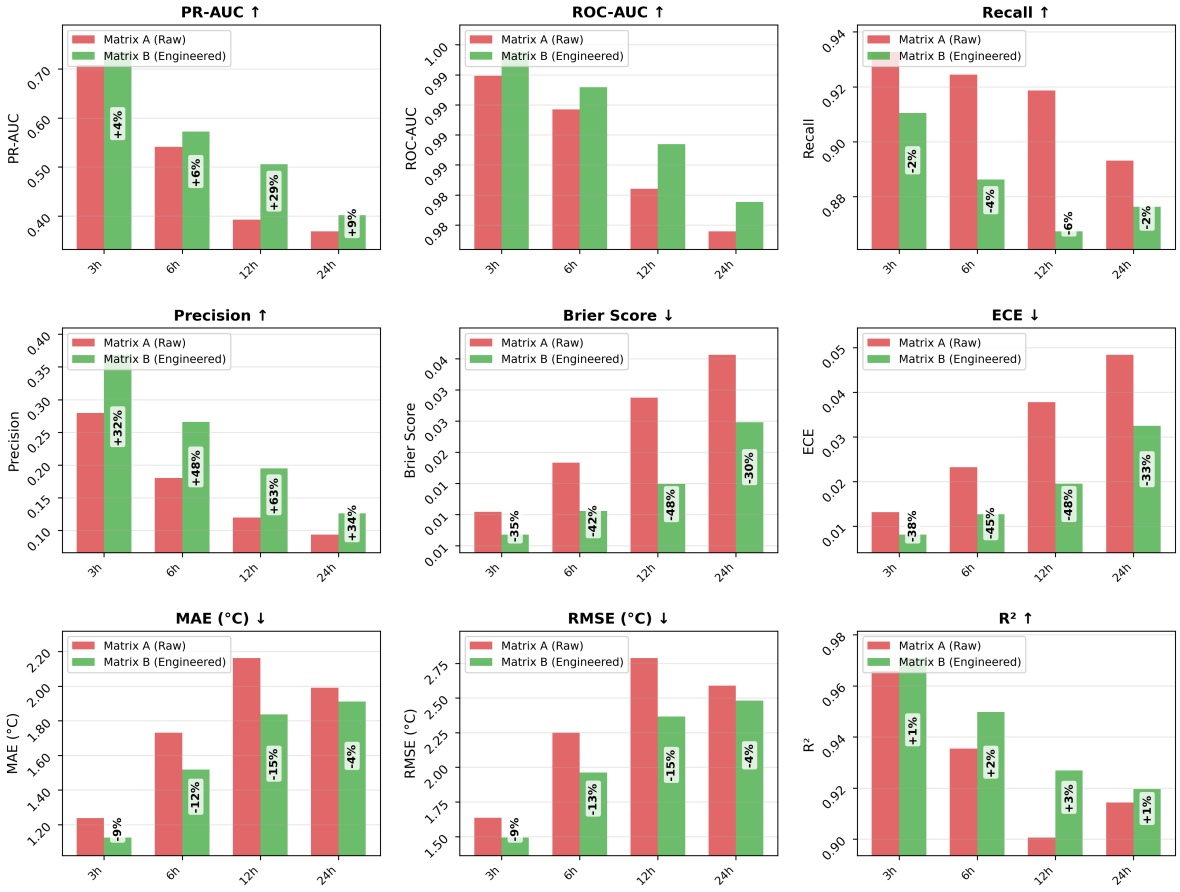

Figure 8: Performance comparison between Matrix A (single-station raw features) vs Matrix B (single-station feature engineering) (class-balanced training). This figure shows performance comparison across 5 key metrics for all prediction windows (3, 6, 12, 24 hours): (1) Classification metrics: PR-AUC, recall, precision; (2) Regression metrics: MAE, R^2. Each subplot shows Matrix A (blue) and Matrix B (orange) performance, with improvement percentages annotated. Matrix B outperforms or matches Matrix A across all metrics, validating the effectiveness of feature engineering.

diction window increases (from 15.6% at 3 hours to 7.6% at 24 hours), indicating that long-term prediction primarily depends on spatial aggregation information rather than single-station raw observations. Although mask features are the most numerous (294, accounting for 55.1%), they are not selected within the 90% cumulative importance threshold (importance 0%), indicating that under current configuration, mask features do not provide additional information gain. Detailed feature importance data are provided in Supplementary Material S5.

Optimal radius varies with prediction window (3 hours: 60 km, 6 hours: 100 km, 12–24 hours: 200 km), revealing the coupling relationship between spatial and temporal scales. Figure 12 shows the variation trends of 9 evaluation metrics with radius. Key findings: (1) Short-term prediction (3–6 hours) optimal radius is smaller (60–100 km), primarily capturing local cold air pooling, with low radius sensitivity (coefficient of variation 0.56–2.20%); (2) Long-term prediction (12–24 hours) optimal radius is larger (200 km), requiring incorporation of larger-scale weather system information, with significantly increased radius sensitivity (coefficient of variation 7.24–8.50%); (3) Compared to minimum radius (20 km), optimal radius brings PR-AUC improvements

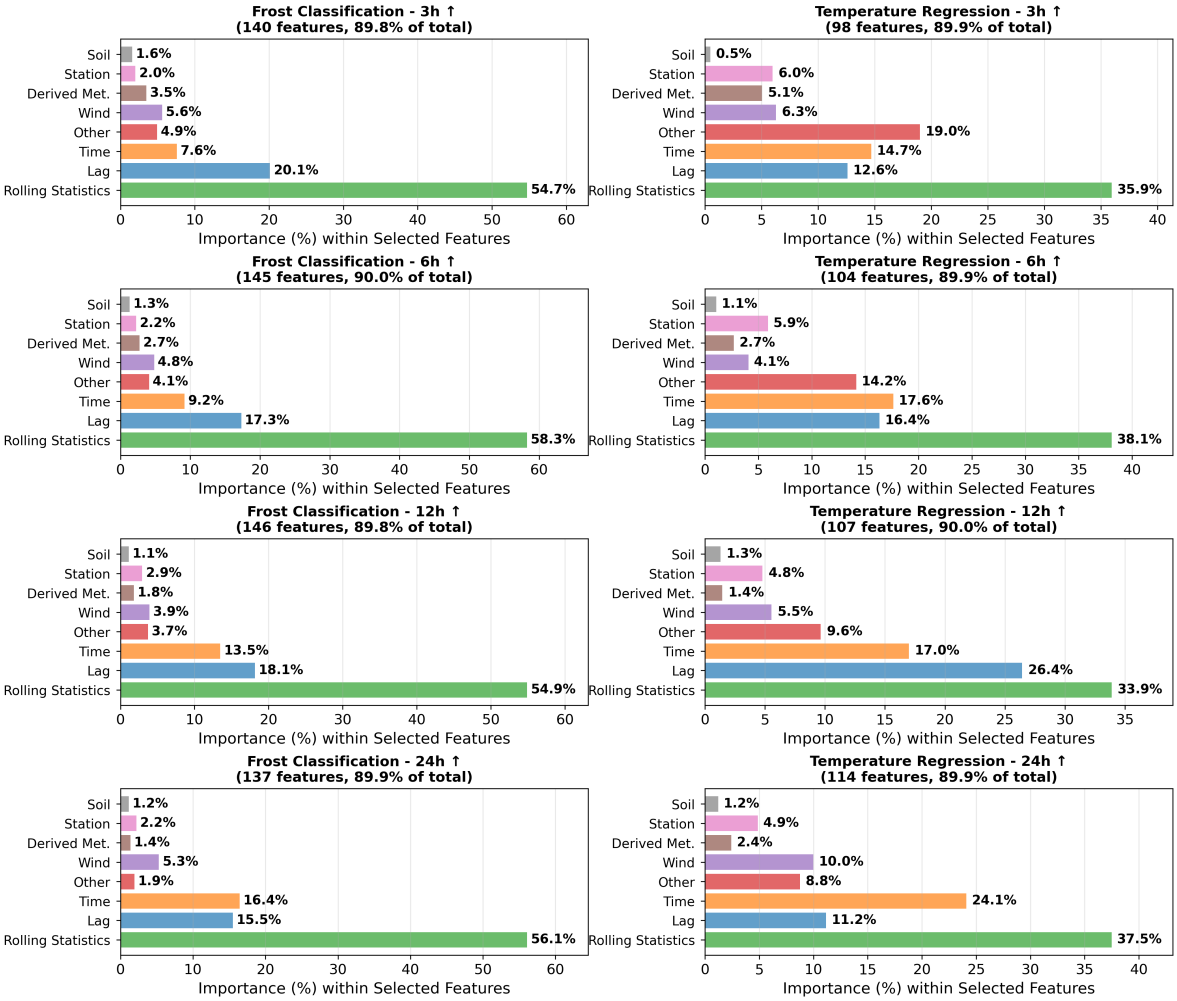

Figure 9: Feature category importance analysis for Matrix B (single-station feature engineering, 278-dim) (based on 90% cumulative feature importance). This figure shows the importance distribution of each feature category when reaching the 90% cumulative feature importance threshold. Rolling window statistics features dominate across all prediction windows (56.2–59.4%), lag feature importance remains stable between 16.3–20.2%, and temporal feature importance significantly increases as prediction window increases.

of 28.88–30.45% in long-term prediction windows, significantly higher than short-term prediction windows (3 hours: 1.96%, 6 hours: 7.23%). This finding quantitatively validates the coupling relationship between spatial and temporal scales, providing scientific basis for radius configuration across different prediction windows.

5.6 LOSO Spatial Generalization Evaluation

To comprehensively evaluate model spatial generalization capability, we performed LOSO (Leave-One-Station-Out) cross-validation on the optimal configurations of Matrices A and B, respectively. LOSO evaluation is a rigorous standard for assessing model spatial generalization capability, revealing whether models overfit site-specific local patterns or can learn regional climate patterns generalizable to new stations.

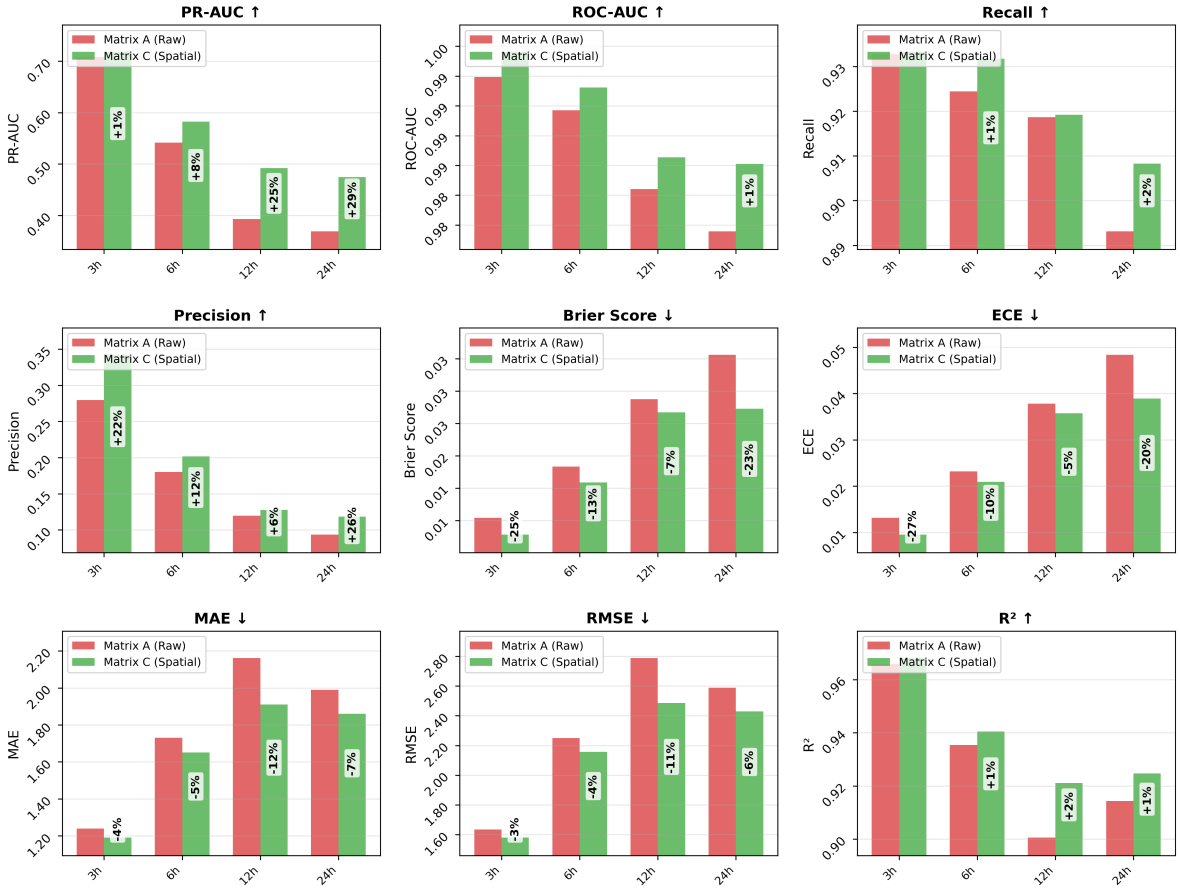

Figure 10: Performance comparison between Matrix A (single-station raw features) vs Matrix C (spatial aggregation features) (class-balanced training). This figure shows performance comparison across key metrics for all prediction windows (3, 6, 12, 24 hours). Matrix C outperforms Matrix A across all metrics, with improvements most significant in long-term prediction windows (12–24 hours).

LOSO evaluation results for Matrices A and B show that all key metrics do not show significant degradation under LOSO conditions, with some metrics even showing slight improvements, validating that the features learned by models have good spatial consistency. Matrix A's PR-AUC remains stable under LOSO conditions (0.47–0.79), with recall maintaining high levels across all prediction windows (0.95–0.98). Matrix B also demonstrates excellent spatial generalization capability under LOSO evaluation, with PR-AUC remaining stable (0.48–0.78) and temperature prediction performance superior to Matrix A.

Figure 13 shows the detailed performance distribution of Matrices A and B under LOSO evaluation and its comparison with conventional temporal splitting evaluation. Key findings: (1) All metrics demonstrate good spatial generalization capability under LOSO evaluation, with small performance differences across stations (boxplots show narrow median and interquartile ranges); (2) Performance under LOSO evaluation is largely consistent with conventional evaluation, with some metrics even showing slight improvements, validating model generalizability; (3) Matrix B's temperature prediction performance under LOSO evaluation is superior to Matrix A, further validating the effectiveness of feature engineering.

Matrix C LOSO Evaluation Note: Due to Matrix C's high feature dimensionality (534-

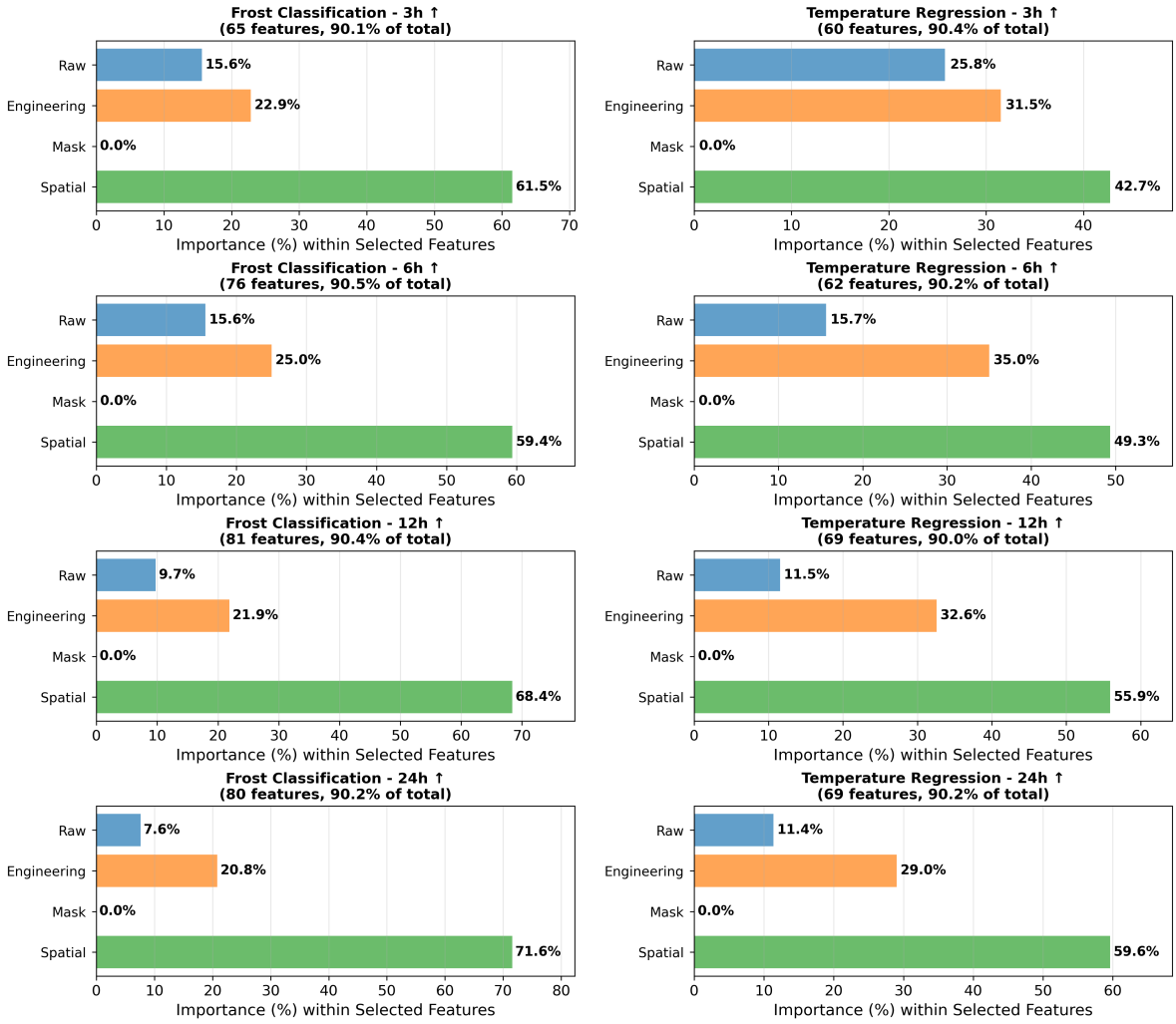

Figure 11: Feature category importance analysis for Matrix C (spatial aggregation features, 534-dim). This figure shows the importance distribution of each feature category when reaching the 90% cumulative feature importance threshold. Spatial aggregation features dominate across all prediction windows (59.4–71.6%), with importance significantly increasing as prediction window increases, validating the core value of spatial information for frost prediction.

dim) and large data volume (approximately 2.36 million records), LOSO evaluation exceeds the current platform's memory limitations and is therefore not included in this study. LOSO evaluation for Matrices A and B has been fully completed (18 stations $\times$ 4 prediction windows), with detailed results provided in Supplementary Material S6.

6 Discussion

Methodological Significance of Feature Engineering Framework: The ABC feature configuration matrix framework proposed in this study provides a reproducible methodology for systematic evaluation of feature engineering strategies, with its core value lying in tracing performance improvements to specific feature types rather than black-box feature combinations. Based on the developed "AgriFrost-AI" platform, the modular design of this framework enables researchers to

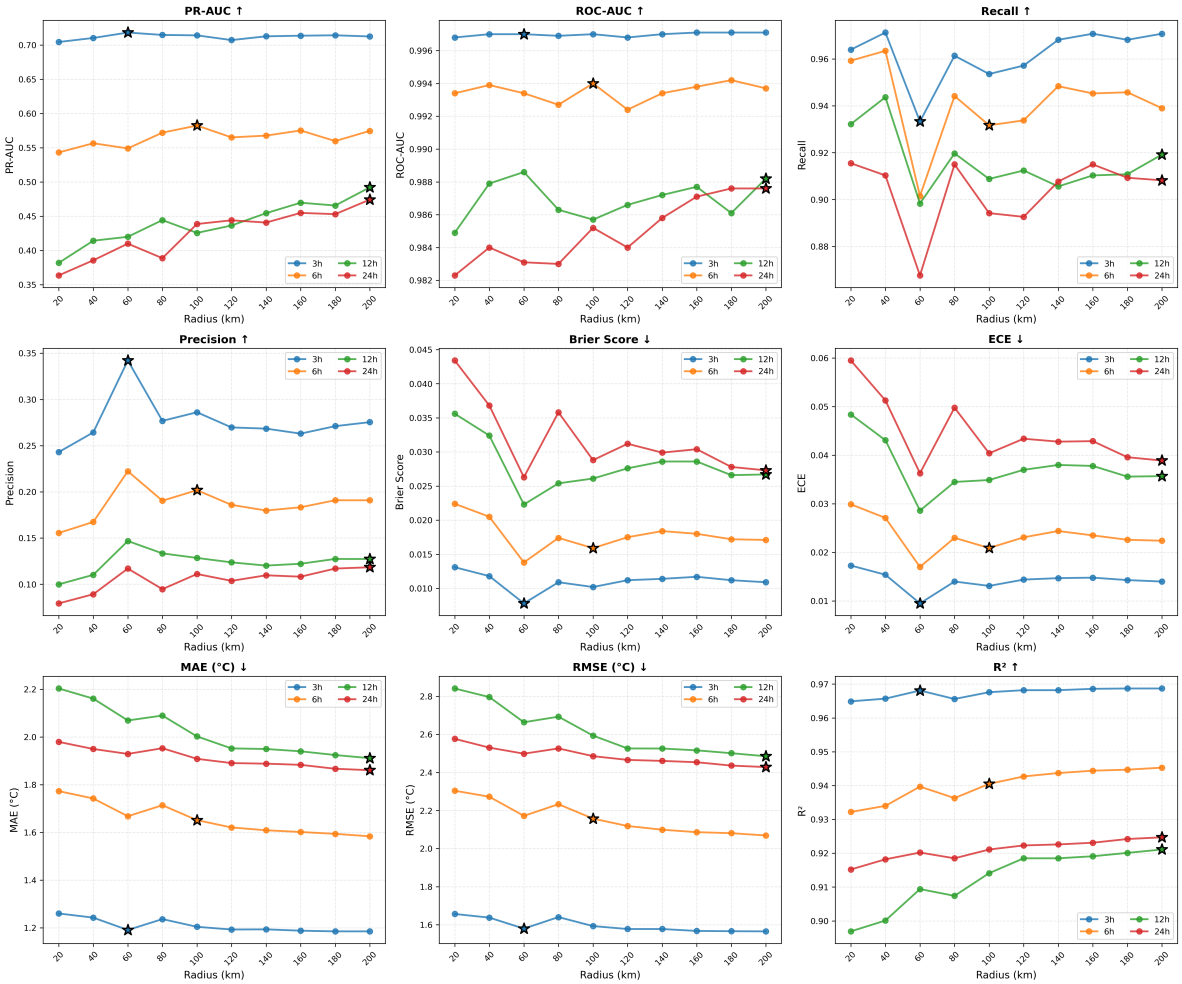

Figure 12: Matrix C: Variation trends of 9 evaluation metrics with radius across different prediction windows. This figure shows the variation patterns of all evaluation metrics (PR-AUC, ROC-AUC, recall, precision, Brier Score, ECE, MAE, RMSE, R^2) with radius within the 20–200 km radius range. Different colored curves represent different prediction windows (3 hours, 6 hours, 12 hours, 24 hours). Optimal radius (selected by PR-AUC) is marked with asterisks (*) in the figure: 60 km for 3-hour window, 100 km for 6-hour window, and 200 km for 12-hour and 24-hour windows. This figure reveals the coupling relationship between spatial and temporal scales: short-term prediction (3–6 hours) optimal radius is smaller (60–100 km), primarily capturing local cold air pooling; long-term prediction (12–24 hours) optimal radius is larger (200 km), requiring incorporation of larger-scale weather system information.

easily integrate and compare different machine learning models (such as LightGBM, XGBoost, neural networks, graph neural networks, etc.), finding optimal model-feature combinations for frost prediction tasks. The generalizability of this framework enables its extension to other extreme event prediction tasks (such as drought, floods, heat waves, etc.), providing a systematic research paradigm for agricultural meteorological prediction.

Physical Mechanism Explanation of Temporal Feature Engineering: The effectiveness of temporal feature engineering in frost prediction stems from its accurate representation of physical processes. Radiative frost formation is a gradual process involving temporal patterns such

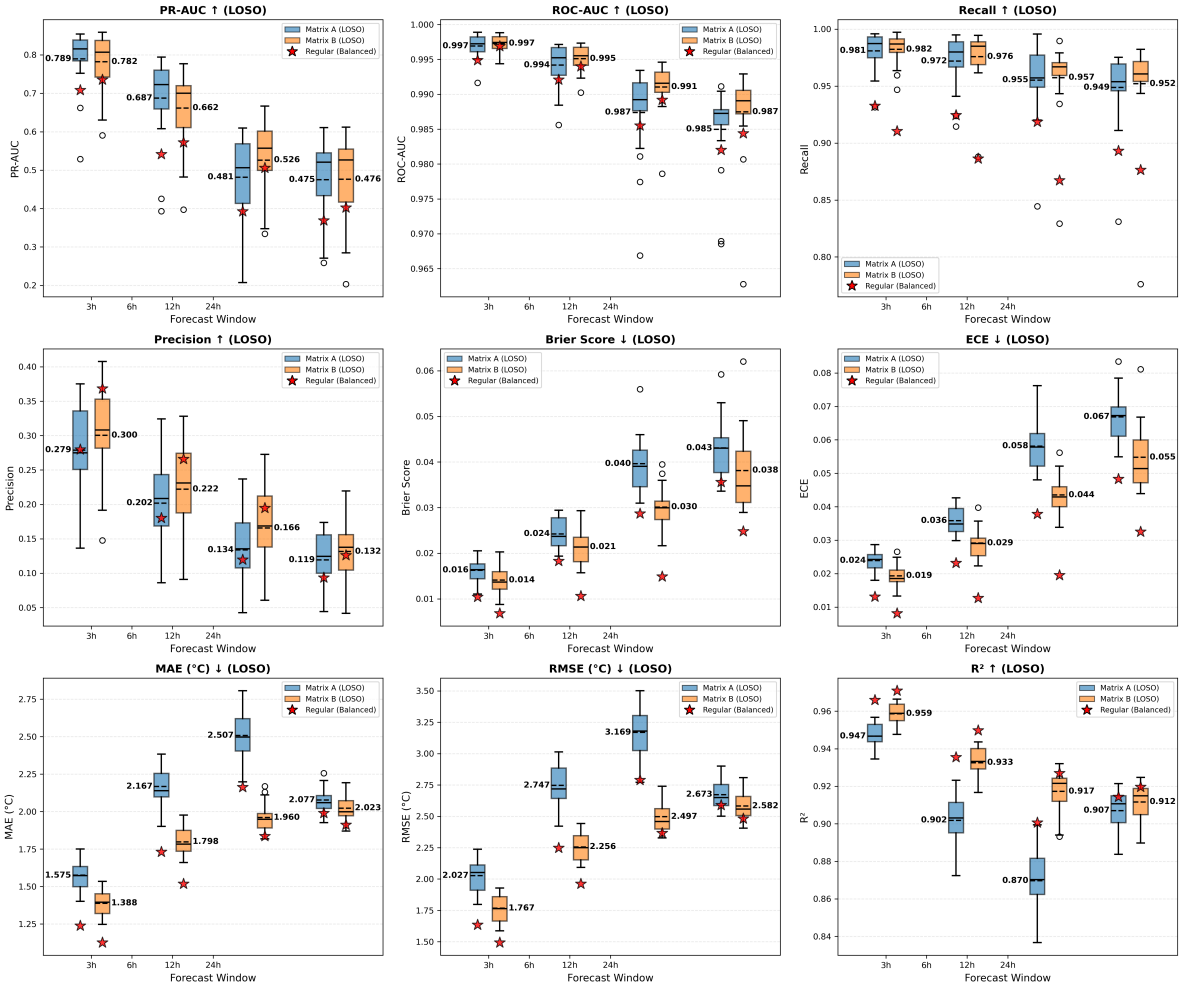

Figure 13: LOSO evaluation boxplot comparison for Matrices A and B. This figure shows the performance distribution of Matrix A (single-station raw features) and Matrix B (single-station feature engineering) under LOSO evaluation, including 9 key metrics: (1) Classification metrics: PR-AUC, ROC-AUC, recall, precision, Brier Score, ECE; (2) Regression metrics: MAE, RMSE, R^2. The X-axis of each subplot shows 4 prediction windows (3h, 6h, 12h, 24h), with Matrix A and B LOSO results (18 stations) displayed side by side under each window. Red asterisks (*) mark conventional class-balanced training results as reference. Boxplots show median, quartiles, and outliers, with mean values annotated above each boxplot. All metrics demonstrate good spatial generalization capability under LOSO evaluation, with small performance differences across stations, validating model generalizability.

as nighttime radiative cooling, temperature decline trends, and dew point approach. Lag features capture the influence of historical states on the current system, rolling window statistics quantify trends and variability, and derived meteorological variables (such as temperature decline rate, dew point difference) directly reflect quantitative indicators of physical processes. The consistency between this feature engineering strategy and physical mechanisms explains why these features maintain excellent performance under LOSO evaluation: they capture temporal dependency patterns rather than site-specific local patterns, exhibiting good spatial generalization capability. This finding provides a new perspective for understanding the success of machine learning models in physical system prediction.

Scale Dependency of Spatial-Temporal Coupling: The differential performance of spatial aggregation features across different prediction windows reveals the scale dependency of spatial-temporal coupling. Short-term prediction primarily depends on local temporal patterns, reflecting the local characteristics of frost formation processes; long-term prediction requires incorporation of regional weather system information, reflecting the influence of large-scale weather systems on local climate. The variation of optimal spatial aggregation radius with prediction window reveals the coupling relationship between spatial and temporal scales, providing quantitative evidence for understanding multi-scale meteorological system interactions. In practical applications, this scale dependency provides scientific basis for feature selection: selecting appropriate spatial aggregation radius based on prediction window can capture key information while controlling computational costs.

Cost-Sensitive Perspective on Class Imbalance Problem: In extremely imbalanced classification tasks, the necessity of class-balanced training stems not only from technical requirements but more fundamentally from the cost sensitivity of application scenarios. In agricultural applications, the cost of missing frost events (crop loss) far exceeds the cost of false alarms (unnecessary protective measures), and this asymmetric cost structure makes high recall the primary objective. Class-balanced training significantly improves recall by adjusting class weights, making models focus more on minority classes during training. However, the cost of this improvement is decreased precision and slightly degraded probability calibration quality, reflecting the fundamental trade-off between recall and calibration quality in class-imbalanced tasks. In practical applications, decision-makers need to adjust decision thresholds based on specific crop economic value and protection costs, balancing the risks of missed events and false alarms. The probability calibration quality achieved in this study (Brier Score ¡0.036, ECE ¡0.049) enables model output probabilities to be directly used for decision support, providing a technical foundation for integrating machine learning models into farm management systems.

Mechanism Explanation of Spatial Generalization Capability: The spatial generalization capability revealed by LOSO evaluation stems from the physical consistency of frost formation processes. All 18 CIMIS stations are located in California's Central Valley, where frost formation processes follow similar physical mechanisms (radiative cooling, cold air pooling, inversion layer formation, etc.), enabling features learned by models to transfer across stations. The generalization stability of raw features validates the inherent predictability of frost prediction tasks, while the fact that feature engineering does not harm spatial generalization capability indicates that temporal feature engineering captures temporal dependency patterns rather than site-specific local patterns. This finding provides a mechanistic explanation for understanding the success of machine learning models in spatial generalization, while also pointing out the boundaries of model generalization capability: when research regions extend to geographic areas with different physical mechanisms, model generalization capability may need re-evaluation.

Physical Interpretability of Feature Importance: Feature importance analysis not only reveals key signals for model decisions but more importantly provides physical interpretability. The

key roles of derived meteorological variables such as soil temperature gradient, dew point difference, and vapor pressure difference in frost prediction are highly consistent with the physical mechanisms of frost formation, validating that machine learning models can learn patterns consistent with physical processes. The differential roles of different feature categories across different prediction windows reflect the physical mechanisms of frost prediction: short-term prediction primarily depends on current state and recent trends, while long-term prediction requires incorporation of seasonal patterns and large-scale weather system information. This physical interpretability not only enhances model credibility but also provides scientific basis for understanding model decisions, which is crucial for applying machine learning models to critical decision scenarios such as agricultural risk management.

Methodological Contributions and Field Impact: The methodological contributions of this study are mainly reflected in three aspects: (1) systematic feature engineering evaluation framework, providing quantitative basis for feature engineering strategy selection; (2) comprehensive class-balanced training analysis, providing practical guidance for extremely imbalanced classification tasks; (3) rigorous spatial generalization evaluation, providing reliability assurance for model deployment. These contributions are not only applicable to frost prediction tasks but also provide transferable methodology for other extreme event prediction tasks (such as drought, floods, heat waves, etc.). However, this study also has some limitations: spatial coverage limitations, temporal span limitations, manual design of feature engineering, incomplete LOSO evaluation for high-dimensional spatial aggregation features, and lack of economic analysis for decision thresholds. These limitations provide directions for future research: extending to other geographic regions to validate model generalization capability, integrating multi-source data to improve prediction accuracy, exploring automatic feature engineering methods, conducting economic cost analysis, and developing real-time deployment systems to achieve transformation from research to application.

7 Conclusion

Based on hourly observational data from 18 CIMIS weather stations in California's Central Valley (2010–2025, approximately 2.36 million records), this study proposes a systematic feature engineering evaluation framework, deeply exploring key issues in frost risk prediction. The main contributions of this study include: (1) proposing a reproducible ABC feature configuration matrix framework that enables tracing performance improvements to specific feature types, providing a methodology for systematic evaluation of feature engineering strategies; (2) validating the differential roles of single-station feature engineering and spatial aggregation features across different prediction windows, revealing the coupling relationship between spatial and temporal scales; (3) comprehensively analyzing the critical role of class-balanced training in extremely imbalanced tasks, providing scientific basis for cost-sensitive decision-making in agricultural applications; (4) rigorously validating model spatial generalization capability through leave-one-station-out cross-validation, revealing that the physical consistency of frost formation processes is the foundation for successful model generalization; (5) achieving excellent probability calibration quality (Brier Score ¡0.036, ECE ¡0.049), enabling model output probabilities to be directly mapped to farm decision thresholds, providing a technical foundation for integrating machine learning models into farm management systems.

Results demonstrate that temporal feature engineering outperforms raw features across all prediction windows by capturing temporal dependency patterns consistent with physical mechanisms, performing particularly well in short-term predictions. Spatial aggregation features are more critical in long-term predictions, with optimal spatial aggregation radius varying with prediction

window, revealing the coupling relationship between spatial and temporal scales and providing scientific basis for feature selection in practical deployment. Class-balanced training significantly improves recall, greatly reducing false negatives, directly addressing the critical need to minimize missed frost events in agricultural applications. Temperature prediction accuracy remains high across all prediction windows ($R^2 > 0.90$, MAE ¡2.0 °C), with LOSO evaluation showing stable performance, validating that features learned by models have good spatial consistency.

The methodological contributions of this study are not only applicable to frost prediction tasks but also provide transferable methodology for other extreme event prediction tasks (such as drought, floods, heat waves, etc.). Based on the developed "AgriFrost-AI" platform, the modular design of this framework enables researchers to easily integrate and compare different machine learning models, providing a systematic research paradigm for agricultural meteorological prediction. However, this study also has some limitations: spatial coverage limitations, temporal span limitations, manual design of feature engineering, incomplete LOSO evaluation for high-dimensional spatial aggregation features, and lack of economic analysis for decision thresholds. Future research directions include: extending to other geographic regions to validate model generalization capability, integrating multi-source data to improve prediction accuracy, exploring automatic feature engineering methods, conducting economic cost analysis, and developing real-time deployment systems. This study connects ground observations, physical process understanding, and agricultural decision support, providing a practical example for deploying machine learning models in field applications, with significant importance for improving agricultural production's risk resilience and sustainable development.

Reproducibility and Open Source

This study employs declarative configuration and fixed random seeds to manage all experiments, ensuring complete reproducibility. Original data are from the official repository of the F3 Innovation Frost Risk Forecasting Challenge: https://github.com/CarlSaganPhD/frost-risk-forecast-challenge. Complete codebase, data processing scripts, and documentation are available at the following GitHub repository: https://github.com/Zhengkun-Li/AgriFrost-AI.

8 Supplementary Materials

Supplementary materials for this study include the following content, with all files located in the `Supplementary_lighgbm_abc/` directory:

- **S1: Feature List and Classification Rules** (`S1_feature_list.pdf`)
- **S2: Complete Experimental Results Data** (`S2_all_metrics_lightgbm_abc.csv`)
- **S3-S5: Feature Importance Analysis** (`S3_matrix_A_feature_importance.csv`, etc.)
- **S6: LOSO Evaluation Results** (`S6_loso_summary.csv`)
- **S7: Experimental Scripts** (`scripts/` directory)

All supplementary materials can be obtained through the project GitHub repository: https://github.com/Zhengkun-Li/AgriFrost-AI.