Feature Engineering and Class-Balanced Training for Frost Risk Forecasting: A LightGBM-Based Approach

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

Abstract

Frost remains a critical meteorological risk for high-value horticultural crops in California, where severe radiation frost can cause substantial yield losses. This study presents a systematic feature engineering framework for frost risk forecasting using LightGBM, based on hourly observations from 18 CIMIS stations (2010–2025, 2.36 million records). We propose a feature configuration matrix (ABC) framework that progressively evaluates the contribution of different feature strategies: Matrix A (raw features, 12 dimensions) as baseline, Matrix B (single-station feature engineering, 278 dimensions) for evaluating temporal feature engineering, and Matrix C (spatial aggregation features, 534 dimensions) for evaluating multi-station spatial information. We employ LightGBM for dual-task learning (frost classification and temperature regression), handle extreme class imbalance (positive class proportion 0.87%) through class-balanced training, and systematically evaluate model performance across four forecast horizons (3, 6, 12, 24 hours). Through large-scale controlled experiments (96 configurations) and Leave-One-Station-Out (LOSO) cross-validation, we rigorously evaluate the effectiveness of feature engineering strategies and model spatial generalization capability.

Results show that class-balanced training is a fundamental requirement for frost prediction tasks, dramatically improving recall from 38.7–67.4% to 86.7–93.3%, reducing false negatives by 75–90%, directly addressing the key requirement of minimizing missed frost events in agricultural applications. Single-station feature engineering (Matrix B) outperforms raw features (Matrix A) across all forecast horizons, with PR-AUC improvements ranging from 0.027 at 3 hours to 0.113 at 12 hours, achieving highest PR-AUC of 0.735 at short-term forecasting (3 hours), validating the core value of temporal feature engineering (lag features, rolling window statistics, derived meteorological variables) in capturing temporal dependency patterns in frost formation. Spatial aggregation features (Matrix C) perform optimally in long-term forecasting, achieving highest PR-AUC of 0.474 at 24-hour horizon, with optimal spatial aggregation radius varying by forecast horizon (3 hours: 60 km, 12–24 hours: 200 km), revealing the coupling relationship between spatial and temporal scales. This horizon-dependent pattern reflects physical mechanisms of frost formation: short-term forecasting mainly relies on local temporal patterns, while long-term forecasting requires incorporating regional weather system information.

The LightGBM model with optimal configurations and class-balanced training achieves PR-AUC 0.718 and recall 0.933 at 3-hour forecast horizon, maintaining PR-AUC 0.474 and recall 0.908 at 24-hour forecast horizon. Temperature prediction accuracy remains high across all forecast horizons (MAE: 1.19–1.91 °C, $R^2 > 0.90$), with LOSO evaluation showing stable performance, demonstrating model spatial robustness. Feature importance analysis reveals soil temperature gradients, dew point differences, and vapor pressure deficits as the most valuable signals, and these findings are highly consistent with physical mechanisms of frost formation. The model achieves excellent probability calibration through class-balanced training (ECE <0.049, Brier Score <0.036), enabling direct use for decision support.

Key contributions include: (1) systematic feature engineering framework enabling performance improvements to be traced to specific feature types; (2) validation of differentiated roles of

temporal feature engineering and spatial aggregation features across different forecast horizons, providing clear guidance for feature selection in actual deployment; (3) comprehensive analysis of class-balanced training impact in extremely imbalanced classification tasks, demonstrating significant reduction in false negatives, providing practical guidance for similar applications; (4) rigorous validation of model spatial generalization capability through LOSO evaluation, providing reliability assurance for deploying models on new stations; (5) excellent probability calibration quality, directly mappable to farm decision thresholds, providing quantitative basis for growers to develop standard operating procedures. This study bridges ground observations, physical process understanding, and agricultural decision support, providing a practical example for deploying machine learning models in field applications, with important significance for improving risk resilience and sustainable development of agricultural production.

Contents

1 Introduction

Frost has long been one of the major meteorological hazards facing high-value fruits, vegetables, and nut crops in California, particularly during the flowering and early fruit stages, where short-duration severe radiation frost can cause widespread yield losses or even total crop failure. Traditional protection strategies rely on empirical judgment, limited manual observations, and mesoscale numerical weather prediction, but often struggle to provide timely and reliable field-level warnings under complex terrain and strong microclimatic conditions.

Machine learning methods based on ground observations provide new technical pathways for frost risk prediction, but face three core challenges: (1) Systematic evaluation of feature engineering strategies: The contribution of temporal feature engineering (lag features, rolling window statistics, derived meteorological variables) and spatial aggregation features (neighborhood statistics, spatial gradients) to model performance has not been quantitatively evaluated under a unified framework; (2) Extreme class imbalance: Frost events account for only approximately 0.87% of all observations, and traditional machine learning models tend to favor the majority class, leading to high false negative rates, which is unacceptable for agricultural applications; (3) Spatial generalization capability: Model generalization capability on new stations needs to be evaluated through strict cross-validation schemes to ensure reliability of actual deployment.

This study is based on the evaluation framework of the F3 Innovate Frost Risk Forecasting Challenge, using hourly observation data from 18 CIMIS weather stations (2010–2025, approximately 2.36 million records), and proposes a systematic feature engineering framework for frost risk prediction. We construct a feature configuration matrix (ABC) framework that progressively evaluates contributions of different feature engineering strategies: Matrix A (raw features, 12 dimensions) as baseline, Matrix B (single-station feature engineering, 278 dimensions) for evaluating temporal feature engineering, and Matrix C (spatial aggregation features, 534 dimensions) for evaluating multi-station spatial information. We employ LightGBM for dual-task learning (frost classification and temperature regression), handle extreme class imbalance through class-balanced training, and systematically evaluate model performance across four forecast horizons (3, 6, 12, 24 hours). Through large-scale controlled experiments (96 configurations) and Leave-One-Station-Out (LOSO) cross-validation, we rigorously evaluate the effectiveness of feature engineering strategies and model spatial generalization capability.

The main contributions of this paper are summarized as follows:

1. Proposed a systematic feature configuration matrix (ABC) framework that enables performance improvements to be traced to specific feature types, providing a reproducible methodology for systematic evaluation of feature engineering strategies.

2. Validated the differentiated roles of temporal feature engineering and spatial aggregation features across different forecast horizons: single-station feature engineering performs optimally in short-term forecasting, while spatial aggregation features are more critical in long-term forecasting, providing clear guidance for feature selection in actual deployment.

3. Comprehensively analyzed the impact of class-balanced training in extremely imbalanced classification tasks, demonstrating that it can dramatically improve recall from 38.7–67.4% to 86.7–93.3%, reducing false negatives by 75–90%, providing practical guidance for similar applications.

4. Rigorously validated model spatial generalization capability through LOSO evaluation, demonstrating that features learned by models have good spatial consistency, providing reliability assurance for deploying models on new stations.

5. Achieved excellent probability calibration quality (ECE <0.049, Brier Score <0.036), enabling direct use for decision support, providing quantitative basis for mapping model outputs to farm decision thresholds.

2 Related Work

2.1 Frost Risk Assessment and Temperature Prediction

Frost risk assessment and short-term temperature prediction have been extensively studied in agricultural meteorology, numerical weather prediction, and machine learning communities. Traditional methods are mainly based on physical models and empirical formulas, focusing on radiative cooling, surface energy balance, and cold air sinking and accumulation processes. Statistical regression methods utilize historical observation data to establish empirical relationships between temperature and meteorological variables, but often struggle to capture complex nonlinear patterns and spatial heterogeneity. Mesoscale numerical weather prediction models (such as WRF) provide high-resolution predictions by solving atmospheric dynamics equations, but have high computational costs and are sensitive to initial conditions and parameterization schemes.

In recent years, with the proliferation of automatic weather station networks and reanalysis data, machine learning-based near-surface meteorological prediction methods have gradually emerged. Random forests and gradient boosting trees (such as XGBoost, LightGBM) have been widely applied to temperature prediction and extreme event detection, effectively capturing nonlinear relationships and feature interactions. Deep learning methods (such as LSTM, CNN) perform excellently in time series prediction, but require large amounts of data and computational resources. Some work incorporates satellite remote sensing, terrain data, and reanalysis fields as inputs to generate high-resolution surface temperature and frost risk maps, but these methods often rely on external data sources, limiting their application in scenarios based solely on ground observations.

2.2 Feature Engineering in Meteorological Prediction

Feature engineering has been recognized as a crucial component of meteorological prediction tasks, directly affecting model performance and interpretability. Temporal feature engineering includes lag features and rolling window statistics, which have been proven to effectively capture historical dependencies and trend patterns. Lag features enable models to learn autocorrelation in time series by introducing historical observations; rolling window statistics (such as mean, standard deviation, minimum, maximum) can smooth noise, capture trends, and identify outliers, which is particularly important for extreme event prediction.

Spatial feature engineering captures regional climate patterns and spatial associations by integrating multi-station information. Neighborhood aggregation statistics (such as neighborhood mean, gradient, range) have been used to capture spatial processes such as cold air pooling and inversion layer formation. Some studies use spatial interpolation methods (such as Kriging, IDW) to generate continuous temperature fields, but these methods often assume spatial stationarity, which may fail under complex terrain and strong microclimatic conditions. Explicit neighboring station aggregation features can more directly capture local spatial patterns without relying on spatial interpolation assumptions.

However, research systematically evaluating contributions of different feature engineering strategies (temporal features vs. spatial features) to model performance in a unified framework remains limited. Most studies either focus on temporal features or spatial features, lacking systematic

comparative analysis. Furthermore, the differentiated roles of feature engineering strategies across different forecast horizons (short-term vs. long-term) have not been fully studied.

2.3 Class Imbalance Problem and Handling Methods

Class imbalance is a fundamental challenge in extreme event prediction tasks, particularly prominent in frost risk prediction, as frost events account for only a small fraction of all observations (usually <1%). Traditional machine learning models tend to favor the majority class on extremely imbalanced data, leading to high false negative rates, which is unacceptable for agricultural applications, as missed frost events can cause severe crop losses.

Class balancing techniques mainly include three categories of methods: (1) Resampling strategies: Oversampling minority class (such as SMOTE) or undersampling majority class to make class distribution more balanced; (2) Class weight adjustment: Allocating different weights to different classes in loss function, increasing weights for minority class; (3) Threshold optimization: Adjusting classification decision thresholds to trade off between precision and recall. Gradient boosting frameworks (such as LightGBM, XGBoost) provide built-in class imbalance handling mechanisms (such as `is_unbalance` parameter), automatically adjusting class weights without manual setting.

Although class balancing techniques have been widely applied to imbalanced classification tasks, comprehensive analysis of their impact on frost prediction remains limited. Most studies only report overall performance metrics (such as accuracy, F1-score), lacking detailed analysis of key metrics such as recall and false negative rates. Furthermore, the impact of class-balanced training on probability calibration quality (such as Brier Score, ECE) has not been fully studied, which limits the application of model outputs in decision support.

2.4 Spatial Generalization Evaluation

Model generalization capability on new stations is a key consideration for actual deployment. Leave-One-Station-Out (LOSO) cross-validation is a strict standard for evaluating spatial generalization capability, revealing whether models overfit to local patterns of specific stations, or can learn regional climate patterns generalizable to new stations. LOSO evaluation has been widely applied in meteorological prediction tasks, but most studies only report average performance, lacking detailed analysis of performance differences between stations.

The impact of spatial aggregation features on model spatial generalization capability has not been fully studied. Theoretically, neighborhood aggregation features reflect regional climate patterns rather than local deviations of individual stations, and should have better spatial generalization capability, but this hypothesis needs to be validated through rigorous LOSO evaluation.

2.5 Differences Between This Study and Existing Research

Compared to the above research, this study makes contributions in the following aspects: (1) Systematic feature engineering evaluation framework: Proposed ABC matrix framework, systematically evaluating contributions of raw features, single-station feature engineering, and spatial aggregation features under a unified framework, enabling performance improvements to be traced to specific feature types; (2) Explicit neighboring station aggregation: Capturing cold air pooling and local inversion structures through explicit neighboring station aggregation features (rather than grid interpolation), and systematically evaluating the impact of different spatial aggregation radii on model performance; (3) Comprehensive class-balanced training analysis: Detailed evaluation of the impact of class-balanced training on model performance, particularly recall, false negative rates, and probability calibration quality, providing practical guidance for extremely imbalanced

classification tasks; (4) Strict spatial generalization evaluation: Rigorously validating model spatial generalization capability through LOSO evaluation, and analyzing the impact of different feature engineering strategies on spatial generalization capability; (5) Systematic analysis of horizon dependency: Systematically analyzing the differentiated roles of feature engineering strategies across different forecast horizons (3, 6, 12, 24 hours), revealing the coupling relationship between temporal and spatial scales.

3 Data and Study Region

This section introduces the CIMIS ground observation data used in the study, statistical characteristics of frost events, main observed variables and their physical significance, and provides an overview of overall data quality and QC.

3.1 Observation Sources and Spatial Coverage

The hourly meteorological observations used in this study come from the California Irrigation Management Information System (CIMIS), covering 18 automatic weather stations in California's Central Valley and surrounding foothill areas. Stations are distributed in a north-south band along the Central Valley, extending from the Sacramento Plain to the Bakersfield region, spanning diverse microclimatic environments including cold air pooling-prone areas, elevation transition zones, and high evapotranspiration agricultural zones. Figure 1 shows the spatial distribution of all stations. The data spans 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, reference evapotranspiration (ETo), and includes official CIMIS Quality Control (QC) flags. Station-level metadata includes station number, name, CIMIS region, county/city, latitude/longitude, elevation, GroundCover, start/end dates, and whether it is an ETo station, used for spatial aggregation and LOSO grouping. The complete list is provided in Supplementary Table S1 (`Supplementary/supplementary_table_S1_stations.csv`). Raw data and processing scripts are hosted in the GitHub repository for version tracking and reproducibility.

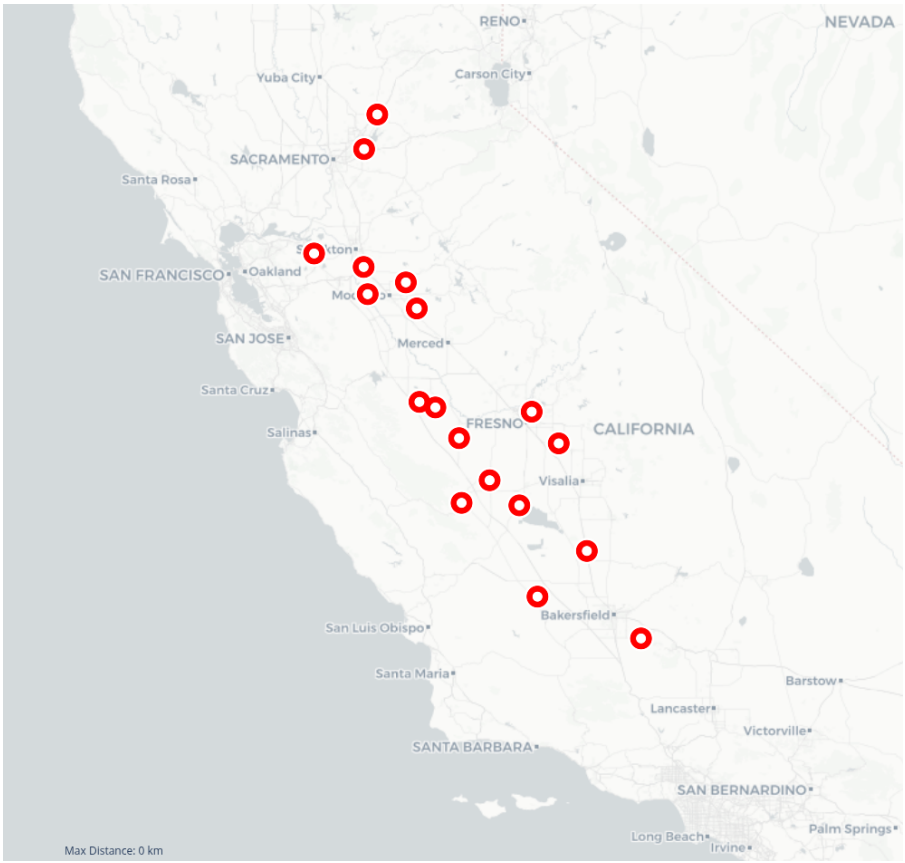

Figure 1: Spatial distribution of 18 CIMIS stations in the study region

3.2 Frost Event Distribution and Seasonal Characteristics

Frost events in this paper 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 about 13%, and other months contribute very little. During April–October, frost events are nearly zero.

In terms of overall proportion, frost events account for only about 0.87% of all hourly records, representing a highly imbalanced classification task. This characteristic directly affects model training and evaluation: on one hand, metrics that focus more on minority class identification (such as PR-AUC) need to be adopted; on the other hand, attention must be paid to avoiding systematic bias caused by extreme imbalance in probability calibration and decision threshold design. This extreme imbalance motivates the use of class-balanced training techniques, as detailed in Section 4.

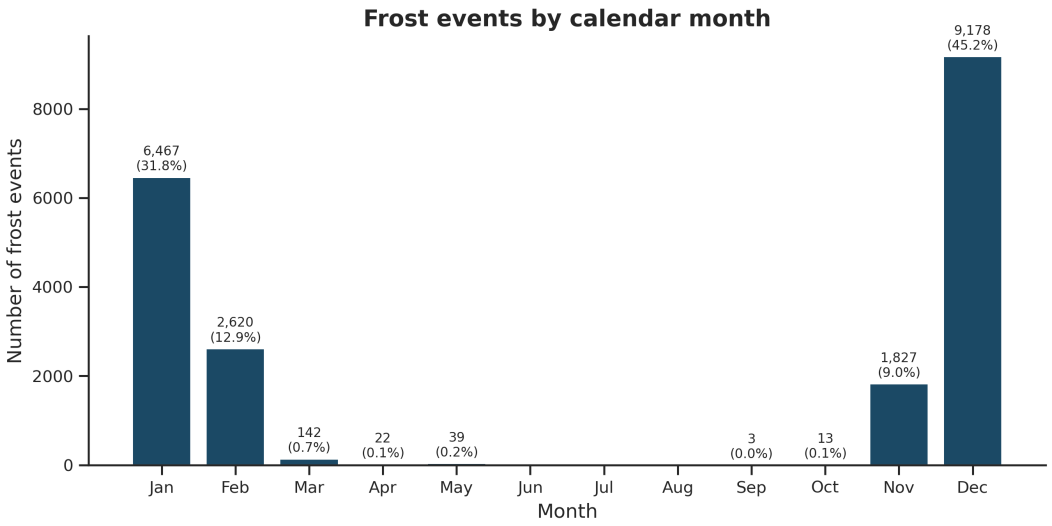

Figure 2: Distribution of frost events by month (2010–2025, 18 stations combined)

Figure 3 shows the distribution of frost events across 24 hours of the day, revealing a strong diurnal pattern: frost events are concentrated in the early morning hours (approximately 3:00–8:00 AM PST), with the peak occurring at 7:00 AM (18.2% of all frost events). This pattern reflects the physical mechanism of radiation frost formation: minimum temperatures typically occur just before sunrise (around 6:00–7:00 AM in California's Central Valley during winter months), when radiative cooling has reached its maximum and solar heating has not yet begun. The distribution shows very few frost events during daytime hours (approximately 10:00 AM–6:00 PM), with only 0.0–0.1% of events occurring during peak solar hours (11:00 AM–4:00 PM), consistent with solar heating preventing frost formation. This diurnal pattern is critical for forecast model design, as it indicates that temporal features (hour of day, time since sunset, etc.) are essential for accurate frost prediction.

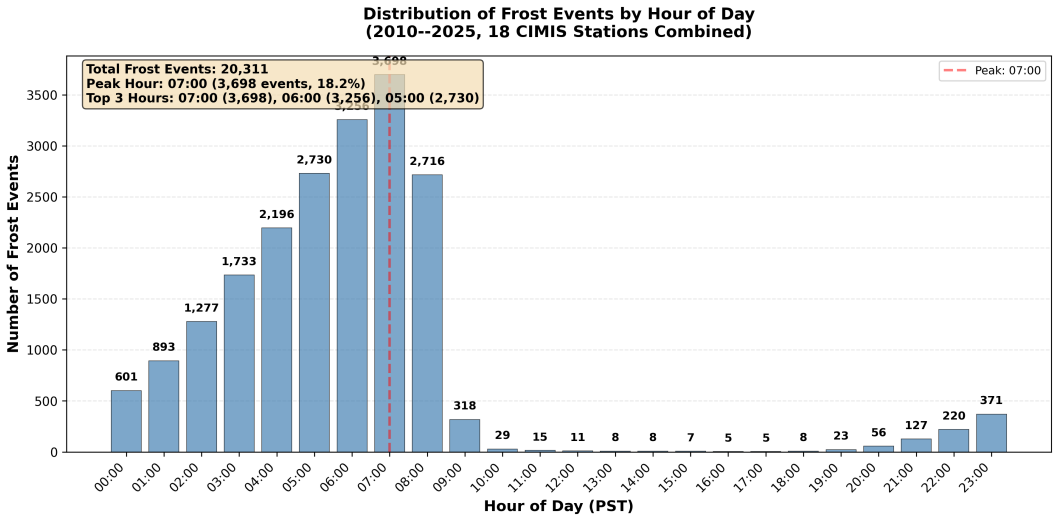

Figure 3: Distribution of frost events by hour of day (2010–2025, 18 CIMIS stations combined). Frost events are concentrated in early morning hours (3:00–8:00 AM PST), with peak occurrence at 7:00 AM (18.2% of all frost events), reflecting the physical mechanism of radiation frost formation where minimum temperatures occur just before sunrise.

3.3 Observed Variables and Physical Significance Overview

The dozen core meteorological variables provided by CIMIS stations are used to characterize surface energy balance, atmospheric state, and soil heat storage, all closely related to frost formation mechanisms. Main variables include:

- **Air Temp (°C)**: Near-surface air temperature, the direct target variable for frost monitoring and prediction.
- **Dew Point (°C)** and **Rel Hum (%)**: Together characterize air moisture content and saturation level, determining condensation and radiation cooling efficiency.
- **Wind Speed (m/s)** and **Wind Dir (0–360)**: Reflect boundary layer mixing intensity and cold air transport pathways. Weak wind or calm conditions are more conducive to radiation frost formation.
- **Sol Rad (W/m^2)**: Solar radiation flux, controlling daytime surface heat storage, with important influence on the upper limit of heat that can be released at night.
- **Soil Temp (°C)**: Shallow soil temperature, reflecting heat storage exchange between surface and near-surface layers.
- **Vap Pres (kPa)**: Vapor pressure, an absolute measure of moisture content, closely related to dew point and relative humidity.
- **ETo (mm)**: Reference evapotranspiration, comprehensively reflecting evapotranspiration demand under radiation, temperature, wind speed, and humidity conditions, with physical connection to nighttime surface cooling rates.

These variables form the basis for subsequent lag features, rolling statistics, harmonic features, and neighborhood aggregation features, providing machine learning models with an input space consistent with physical processes.

3.4 Data Quality and QC Overview

All observations include official CIMIS-generated QC flags indicating whether the physical quantity passed automatic and manual validation. We follow CIMIS recommended guidelines, retaining only "blank/pass" and "Y" flags, with all others (including M, Q, R, S, P, etc.) treated as unavailable. After removing sentinel values, forward filling is performed station by station.

Overall, from 2010–2025, there are approximately 2.36 million hourly records, of which only about 1.71% of rows are flagged as missing or unavailable for at least one key variable, indicating generally high observation quality. Figure 4 shows the contribution proportion of low-quality records across different stations. Low-quality data is relatively dispersed across stations, with only a few stations (e.g., 205, 194, 124) having 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 about 20.4%, and wind speed about 10.1%. Relative humidity and dew point each contribute about 8.6%, and vapor pressure about 7.3%. The core frost observation variable—air temperature—has an anomaly proportion of only 6.2% of low-quality records, corresponding to about 0.1% of all observations, further validating the suitability of this dataset for frost analysis and prediction tasks.

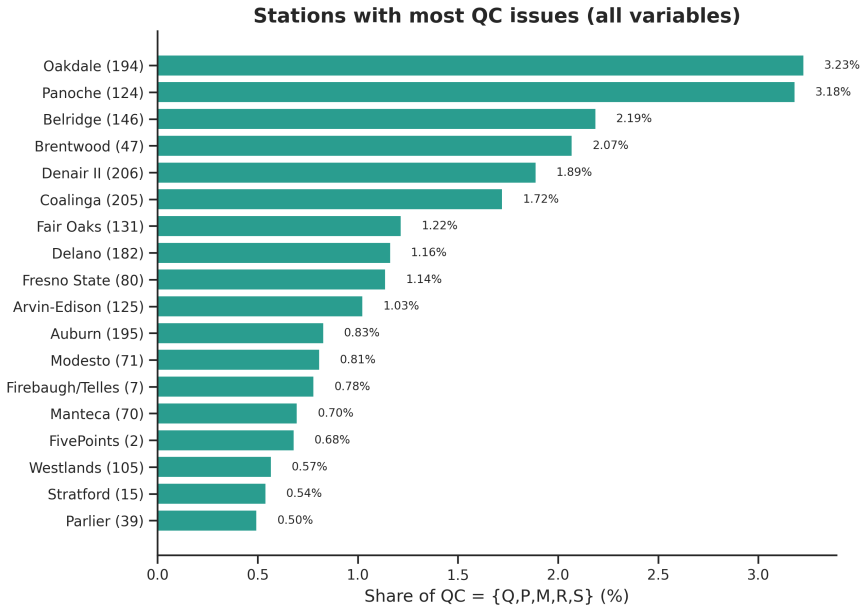

Figure 4: Relative contribution of low-quality (Bad QC) records by station

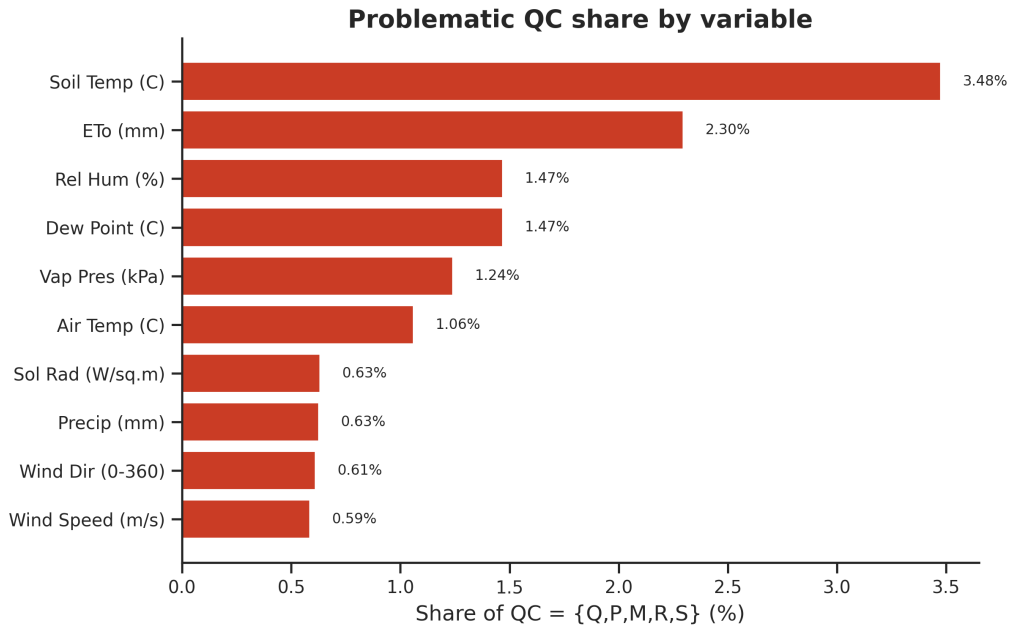

Figure 5: Distribution of low-quality (Bad QC) records by meteorological variable

4 Methods

This section systematically introduces AgriFrost-AI's complete methodological framework, including: (1) Data preprocessing and QC pipeline, ensuring data quality and time series consistency; (2) Feature configuration matrix (ABC) framework, systematically evaluating different feature engineering strategies; (3) Feature engineering methods, including single-station feature engineering and neighborhood aggregation features; (4) LightGBM model configuration, including hyperparameter settings, class imbalance handling, decision threshold design, and dual-task training framework; (5) Evaluation metrics, selecting appropriate evaluation standards for extremely imbalanced classification tasks; (6) Experimental design, including experimental configurations, data splitting strategies, single-station feature engineering analysis, spatial aggregation feature analysis, LOSO spatial generalization evaluation, radius sensitivity analysis, feature importance analysis methods, and unified training framework and experimental platform.

4.1 Data Preprocessing and QC Pipeline

The unified `DataCleaner` pipeline 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 processing**: 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 key variable missing, enabling models to explicitly perceive observation incompleteness.

4. **Label generation**: Generate frost binary classification labels and temperature regression targets for four forecast windows (3, 6, 12, 24 hours) in one pass on 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 forecast model performance, this study constructs a progressive feature configuration matrix framework. This framework is designed to evaluate three feature engineering strategies: (1) **Matrix A**: Baseline raw features; (2) **Matrix B**: Single-station feature engineering; (3) **Matrix C**: Spatial aggregation features. The progressive design allows 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 time series engineered features (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) The comparison between temporal features and spatial features (Matrix B vs Matrix C). This framework provides a systematic experimental design basis for subsequent ablation studies and feature importance analysis.

Detailed feature composition and generation methods for each matrix are described in Section 4.3.

Table 1: Feature Configuration Matrix (ABC) Overview

Matrix	Spatial Config	# Features	Feature Composition
A	Single-station (no neighbors)	12 dim	12 raw CIMIS variables (air temp, dew point, rel hum, wind speed, wind dir, sol rad, soil temp, vap pres, ETo, precip, hour, Julian day)
B	Single-station (no neighbors)	278 dim	Feature engineering pipeline: raw variables (12 dim) + temporal features (15 dim) + lag features (50 dim: 10 variables × 5 lags) + rolling window statistics (180 dim: 9 variables × 4 windows × 5 functions) + derived meteorological features (5 dim: temp_dew_diff, temp_change_rate, wind_chill, heat_index, soil_air_temp_diff) + radiation features (4 dim) + wind features (6 dim) + humidity features (5 dim: saturation_vapor_pressure, dew_point_proximity, humidity_change_rate, temp_humidity_interaction, vapor_pressure_deficit) + trend features (3 dim: temp_decline_rate, cooling_acceleration, temp_trend) + station static features (4 dim)
C	Multi-station aggregation (radius 20–200 km)	534 dim	Raw variables (12 dim) + neighborhood aggregation statistics (8 aggregation methods for 27 numeric variables: mean, max, min, std, median, distance-weighted mean, gradient, range, totaling 216 dim = 27 variables × 8 methods) + missing masks (293 dim: missing ratio for each aggregation feature) + temporal harmonic encoding (2 dim) + other features (11 dim: temporal discrete features, derived meteorological features, `has_neighbors` indicator)

Matrix A (Single-station + Raw Features) Matrix A serves as the baseline configuration, using only 12 raw CIMIS observed variables (air temperature, dew point, relative humidity, wind speed, wind direction, solar radiation, soil temperature, vapor pressure, ETo, precipitation, hour, Julian day), totaling 12 dimensions. This configuration introduces no feature engineering (lags, rolling statistics, temporal harmonic encoding, etc.) or spatial aggregation, aiming to evaluate the inherent discriminative ability of raw observations and provide a performance benchmark for subsequent feature engineering and spatial enhancement.

Matrix B (Single-station + Engineered Features) Matrix B overlays a complete single-station feature engineering pipeline on Matrix A, generating 278 candidate features. This configuration aims to evaluate the performance gain of time series engineered features (lags, rolling window statistics, derived meteorological variables) for single-station models. In actual experiments, all models are trained using the full feature set (278 dimensions) to fully assess the contribution of feature engineering. Detailed feature engineering methods are described in Section 4.3.

Matrix C (Neighborhood Aggregation + Raw Features) Matrix C overlays multi-station spatial aggregation statistics on Matrix A's raw variables. This configuration aims to evaluate the contribution of spatial information to frost forecasting, particularly spatial patterns such as cold air

pooling and inversion layer formation. For 27 numeric variables (12 raw CIMIS variables + 15 temporal features), neighborhood aggregation is performed, computing 8 aggregation statistics (mean, max, min, std, median, distance-weighted mean, gradient, range) under specified radius thresholds (systematically tested 20–200 km in experiments, step size 20 km), generating 216 neighborhood aggregation features. Additionally, the system generates missing mask features (293 dimensions) to handle neighbor station data missing issues, including missing masks for neighborhood aggregation features, variable missing ratios, missing masks for missing ratio features, and missing masks for other features. Combined with raw variables (12 dim), temporal harmonic encoding (2 dim), and other features (11 dim), the total is 534 dimensions. Detailed neighborhood aggregation methods and missing mask calculations are described in Section 4.3.

4.3 Feature Engineering

Feature engineering is a key step in transforming raw observational data into predictive feature representations. For frost forecasting tasks, effective feature engineering needs to simultaneously capture temporal patterns (diurnal cycles, annual cycles, historical dependencies), spatial associations (cold air pooling, inversion layers, spatial gradients), and physical relationships (energy balance, radiation cooling, convective mixing). This section systematically describes the design principles, theoretical basis, implementation methods, and feature composition of the feature engineering pipeline.

4.3.1 Design Principles and Theoretical Framework

The feature engineering pipeline design follows four core principles:

(1) **Temporal leakage prevention**: All features are computed after grouping by station and strict temporal sorting, ensuring feature values depend only on historical information, strictly preventing future information leakage into historical features. This principle is crucial for time series prediction tasks, especially in LOSO (Leave-One-Station-Out) evaluation scenarios, where temporal ordering of feature computation must be ensured.

(2) **Dependency relationship management**: Feature construction follows a clear dependency order (temporal features $\rightarrow$ lag features $\rightarrow$ rolling statistics $\rightarrow$ derived features), ensuring computational correctness and reproducibility. This principle avoids circular dependencies between features and guarantees idempotency of feature computation.

(3) **Physical significance orientation**: Prioritize constructing meteorological composite features with clear physical or agricultural significance, rather than blind combinations. This principle is based on theoretical foundations of meteorology and agricultural meteorology, ensuring features can capture physical mechanisms of frost formation (e.g., radiation cooling, inversion layer formation, cold air sinking).

(4) **Robustness design**: Explicit handling of missing values, outliers, and boundary conditions, ensuring features can still be computed under sparse data conditions. This principle improves model generalization in real data environments through missing masks, numerical clipping, conditional computation, etc.

The feature engineering pipeline 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, naming conventions, and feature importance analysis results are provided in Supplementary Material S1 (`Supplementary/supplementary_S1_feature_list.pdf`).

4.3.2 Single-Station Feature Engineering

Single-station feature engineering is enabled in Matrix B, building upon the raw features in Matrix A. This section describes the comprehensive feature engineering pipeline that generates 278 candidate features, systematically capturing temporal patterns, historical dependencies, and physical relationships.

Temporal Features (15 dimensions) Temporal features are fundamental for frost forecasting because frost events exhibit strong diurnal and annual cycle patterns. Three types of temporal encodings are extracted from datetime columns:

- **Discrete encoding** (6 dim): `hour` (0–23), `month` (1–12), `day_of_year` (1–366), `day_of_week` (0–6), `season` (1–4: spring/summer/fall/winter), `is_night` (binary, 1 for 18:00–06:00, 0 otherwise). Discrete encoding facilitates models learning differentiated patterns across different time periods, e.g., nighttime periods (`is_night=1`) typically correspond to radiation cooling and temperature decline.
- **Cyclic encoding** (8 dim): `hour_sin/cos`, `month_sin/cos`, `day_of_year_sin/cos`, `day_progress_sin/cos`. Where `day_progress` is normalized hour progress (0–1), calculated as $\texttt{day\_progress} = \texttt{hour}/24$. Cyclic encoding uses trigonometric functions:

  $$\texttt{hour\_sin} = \sin\left(\frac{2\pi \cdot \texttt{hour}}{24}\right), \quad \texttt{hour\_cos} = \cos\left(\frac{2\pi \cdot \texttt{hour}}{24}\right)$$

  Similarly, month and day of year use period lengths $T = 12$ and $T = 365.25$ for encoding. This encoding avoids boundary discontinuities (e.g., jump between hour 23 and 0), enabling models to learn smooth periodic patterns, crucial for capturing nighttime cooling trends and seasonal variations.
- **Agriculture-related features** (1 dim): `frost_season_indicator`, marking high frost risk period from December to April (California region). This feature directly encodes the temporal window for agricultural frost warnings, helping models focus on high-risk periods.

Theoretical basis: Temporal features capture two key time scales of frost occurrence: (1) **Diurnal cycle**: Based on radiation cooling theory, nighttime surface longwave radiation loss leads to temperature decline, typically reaching minimum around 4–6 AM, which is the main physical mechanism of frost formation; (2) **Annual cycle**: Based on climatological principles, winter and early spring are high frost risk periods, when solar radiation intensity is low, cold air activity is frequent, and temperature fluctuations and extreme low temperature events are more common.

Lag Features (50 dimensions) Lag features capture historical states of meteorological variables, having important value for predicting future temperature changes. Based on time series analysis theory, meteorological variables have temporal autocorrelation, and historical states have informational value for predicting the future.
Variable and lag configuration:

- **Variables**: 10 core variables (air temp, dew point, ETo, precipitation, relative humidity, soil temp, solar radiation, wind direction, wind speed, vapor pressure)
- **Lag windows**: 1, 3, 6, 12, 24 hours

- **Total features**: $10 \times 5 = 50$ dimensions

Design considerations:

- **Multi-scale lags**: 1-hour lag captures short-term fluctuations and rapid changes, 3–6 hour lags reflect medium-term trends and weather system evolution, 12–24 hour lags capture diurnal cycle patterns and day-night temperature differences
- **Station-grouped computation**: All lag features are computed grouped by station (`groupby(station_id)`), ensuring no cross-station information leakage, which is crucial for LOSO evaluation
- **Temporal alignment**: Ensure data is strictly sorted by station and time before lag computation to prevent temporal leakage

Calculation formula and naming:

- Calculation formula: $x_{\text{lag},h}(t) = x(t-h)$, where $x(t-h)$ represents the value of variable x h hours ago
- Naming format: `{variable}_lag_{hours}`
- Example: `Air Temp (C)_lag_12` represents air temperature 12 hours ago

Theoretical basis: These features help models learn inertial effects and historical dependencies of temperature changes, capturing memory characteristics of meteorological systems.

Rolling Window Statistics (180 dimensions) Rolling window statistics capture distribution characteristics of variables within time windows, having important significance for identifying trends, volatility, and extreme values. Based on sliding window analysis theory, rolling statistics can smooth noise, capture trends, and identify outliers.
Variable and window configuration:

- **Variables**: 9 core variables (air temp, dew point, ETo, precipitation, relative humidity, soil temp, solar radiation, wind speed, vapor pressure)
- **Excluded variables**: Wind direction does not participate in rolling statistics, as it is an angular variable requiring circular statistics
- **Time windows**: 3, 6, 12, 24 hours
- **Total features**: $9 \times 4 \times 5 = 180$ dimensions

Theoretical basis for statistic selection:

- **Mean**: $\bar{x}_w = \frac{1}{n}\sum_{i=1}^{n} x_i$, reflects average state within window, captures trend direction
- **Minimum**: $x_{\min} = \min_i x_i$, identifies extreme values, especially important for frost warning (minimum temperature directly relates to frost risk)
- **Maximum**: $x_{\max} = \max_i x_i$, identifies extreme values
- **Standard deviation**: $\sigma_w = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_w)^2}$, quantifies volatility, high volatility may indicate unstable atmospheric conditions

- **Sum**: $\sum_{i=1}^{n} x_i$, has physical significance for cumulative quantities (e.g., precipitation, ETo)

Implementation details:

- Rolling features are computed grouped by station, ensuring no cross-station information leakage
- Use `min_periods=1` to maximize data utilization (compute statistics even if only 1 valid value in window)
- Naming format: `{variable}_rolling_{window}h_{statistic}`
- Example: `Air Temp (C)_rolling_24h_mean` represents average air temperature over the last 24 hours, `Soil Temp (C)_rolling_12h_min` represents minimum soil temperature over the last 12 hours

Derived Meteorological Features (5 dimensions) Derived meteorological features compute composite variables based on physical relationships and meteorological principles, capturing interactions between variables.
Feature list:

- `temp_dew_diff`: Air temperature minus dew point (`Air Temp` − `Dew Point`), quantifies air saturation level. When this value approaches 0, it indicates air is near saturation, condensation may occur; larger differences indicate drier air, favorable for evaporative cooling.
- `temp_change_rate`: Temperature change rate (`Air Temp`(t) − `Air Temp`$(t-1)$), captures short-term temperature change trends, having important value for identifying rapid cooling processes.
- `wind_chill`: Wind chill index, quantifies the effect of wind speed on apparent temperature. Calculated when air temp < 10°C, formula: $13.12 + 0.6215T - 11.37V^{0.16} + 0.3965TV^{0.16}$ (where T is air temperature °C, V is wind speed km/h). In frost scenarios, high wind speed accelerates heat loss, reducing apparent temperature.
- `heat_index`: Heat index, calculated under high temperature and high humidity conditions (air temp > 80°F and relative humidity $> 40\%$). Although this feature value usually equals air temperature in frost scenarios, retaining it helps models learn complete temperature-humidity relationships.
- `soil_air_temp_diff`: Soil temperature minus air temperature (`Soil Temp` − `Air Temp`), reflects surface energy exchange direction. Positive values indicate soil temperature higher than air temperature (common during daytime), negative values indicate soil temperature lower than air temperature (common during nighttime). This feature has important value for identifying inversion layers and radiation cooling processes.

Theoretical basis: These composite features encode nonlinear relationships between meteorological variables, based on meteorological and thermodynamic principles. Wind chill index is based on convective heat transfer theory, quantifying the effect of wind speed on apparent temperature; heat index is based on heat balance equations, quantifying apparent temperature under high temperature and high humidity conditions; soil-air temperature difference is based on surface energy balance, reflecting relative intensity of radiation cooling and convective exchange. These features help models understand how physical processes (e.g., radiation cooling, convective exchange, evaporative cooling) affect frost formation.

Radiation-Related Features (4 dimensions) Solar radiation is the core driving factor of surface energy balance, directly affecting daytime warming and nighttime cooling processes.
Feature list:

- `sol_rad_change_rate`: Solar radiation relative change rate ($\frac{\texttt{Sol Rad}(t)-\texttt{Sol Rad}(t-1)}{\texttt{Sol Rad}(t-1)}$), captures relative fluctuations in radiation intensity, reflects cloud cover changes and atmospheric transparency
- `daily_solar_radiation`: Daily cumulative radiation, accumulated from 06:00 to current time. This feature quantifies total daytime energy input, high cumulative radiation usually corresponds to stronger daytime warming, potentially affecting nighttime cooling rate
- `nighttime_cooling_rate`: Nighttime cooling rate, calculated only when `is_night=1`. This feature directly captures the radiation cooling process, a key signal for frost forecasting
- `radiation_temp_interaction`: Interaction term between radiation and temperature (`Sol Rad` $\times$ `Air Temp`), captures nonlinear effects of radiation on temperature

Theoretical basis: Radiation features are based on surface energy balance theory. According to Stefan-Boltzmann law, surface longwave radiation loss is proportional to the fourth power of temperature. When there is no solar radiation input at night, the surface continuously loses energy, causing temperature decline. High daytime radiation causes surface warming, low nighttime radiation (close to 0) causes radiation cooling, which is the main physical mechanism of frost formation. Radiation change rate captures cloud cover changes and atmospheric transparency fluctuations, daily cumulative radiation quantifies total daytime energy input, nighttime cooling rate directly reflects radiation cooling intensity.

Wind Features (6 dimensions) Wind field features are crucial for frost forecasting because wind speed affects convective mixing intensity, and wind direction affects cold air pathways.
Feature list:

- `wind_dir_sin/cos`: Wind direction cyclic encoding, converts 0–360 degree angles to $\sin(\theta)$ and $\cos(\theta)$, avoiding angular boundary discontinuities (difference between 359° and 1°)
- `wind_dir_category`: Wind direction categorical encoding, divides 0–360 degrees into 4 quadrants (north, east, south, west), facilitating models learning different wind direction effects on frost risk patterns
- `wind_speed_change_rate`: Wind speed change rate ($\texttt{Wind Speed}(t) - \texttt{Wind Speed}(t-1)$), captures dynamic changes in wind field
- `calm_wind_duration`: Calm wind duration, defined as continuous duration with wind speed $<$ 1.0 m/s. Calm conditions favor radiation cooling, an important prerequisite for frost formation
- `wind_dir_temp_interaction`: Interaction term between wind direction and temperature, captures differential effects of different wind directions on temperature changes (e.g., dry cold wind from inland vs. moist wind from ocean)

Theoretical basis: Wind field features are based on atmospheric boundary layer theory. According to mixed layer theory, under low wind speed (calm) conditions, convective mixing weakens, turbulent exchange coefficient decreases, favoring near-surface cold air accumulation and inversion

layer formation, thereby increasing frost risk. Calm wind duration quantifies the duration window favorable for radiation cooling. Wind direction affects the source and pathway of cold air, different wind directions may bring air masses with different temperature and humidity characteristics, having differential effects on frost formation.

Humidity Features (5 dimensions) Humidity features quantify water vapor content in the atmosphere, having important significance for understanding radiation cooling, condensation processes, and frost formation mechanisms.
Feature list:

- `saturation_vapor_pressure`: Saturation vapor pressure, calculated based on Magnus formula: $e_s = 0.6108 \times \exp\left(\frac{17.27 \times T}{T+237.3}\right)$ (where T is air temperature °C). This feature reflects maximum water vapor capacity of air at given temperature
- `dew_point_proximity`: Dew point proximity, calculated as `(T - T_dew) / T`, quantifies relative difference between air temperature and dew point. When this value approaches 0, it indicates air is near saturation, condensation or dew formation may occur
- `humidity_change_rate`: Humidity change rate (`Rel Hum`$(t) -$ `Rel Hum`$(t-1)$), captures dynamic changes in atmospheric humidity
- `temp_humidity_interaction`: Interaction term between temperature and humidity (`Air Temp` $\times$ `Rel Hum`$/100$), captures nonlinear effects of temperature-humidity relationships, having important significance for understanding evaporative cooling and condensation processes
- `vapor_pressure_deficit` (VPD): Vapor pressure deficit, defined as the difference between saturation vapor pressure and actual vapor pressure. VPD reflects "dryness" and cooling potential of air, high VPD usually corresponds to stronger evaporative cooling effects

Theoretical basis: Humidity features are based on phase change thermodynamics and energy balance theory. According to Clausius-Clapeyron equation, saturation vapor pressure increases exponentially with temperature. Under high humidity conditions, when temperature approaches dew point, water vapor condensation releases latent heat (approximately 2500 J/g), potentially slowing cooling rate; under low humidity conditions, evaporative cooling enhances (evaporation consumes energy), potentially accelerating cooling. Dew point proximity quantifies how close air is to saturation, VPD (vapor pressure deficit) quantifies "dryness" and evaporation potential of air, an important indicator for understanding relative intensity of radiation cooling and evaporative cooling.

Trend Features (3 dimensions) Trend features capture temperature change trends and acceleration, having important value for identifying rapid cooling processes (which may lead to frost).
Feature list:

- `temp_decline_rate`: Temperature decline rate ($\frac{\texttt{Air Temp}(t) - \texttt{Air Temp}(t-6)}{6}$, unit: °C/hour), quantifies average cooling rate over the last 6 hours, serves as the basis for cooling acceleration calculation
- `cooling_acceleration`: Cooling acceleration, calculated based on temperature decline rate changes (`temp_decline_rate`$(t) -$ `temp_decline_rate`$(t-1)$). This feature quantifies the second derivative of cooling rate, positive values indicate accelerating cooling, negative values indicate decelerating cooling. Rapid cooling (high cooling acceleration) is a key signal for frost warning

- `temp_trend`: Temperature trend direction, classified based on temperature decline rate: -1 (rapid decline, rate < -0.5°C/hour), 0 (stable, -0.5 to 0.5°C/hour), 1 (increasing, rate > 0.5°C/hour). This feature helps models identify qualitative trends in temperature changes

Theoretical basis: Trend features are based on difference and acceleration concepts in time series analysis. First-order features (e.g., temperature change rate) capture trend direction, second-order features (e.g., cooling acceleration) capture trend changes, helping models identify nonlinear dynamics of cooling processes. Cooling acceleration quantifies acceleration or deceleration of cooling rate, having important value for identifying rapid cooling processes (which may lead to frost). This feature is calculated based on temperature decline rate changes over the last 6 hours, capturing second-order dynamics of cooling processes.

Station Static Features (4 dimensions) Station static features merge geographic attributes from CIMIS station metadata, used to characterize spatial heterogeneity.
Feature list:

- `station_id_encoded`: Station ID encoding, converts station identifiers to numeric encoding, facilitating models learning station-specific patterns (e.g., microclimate, elevation, terrain)
- `region_encoded`: Region encoding, classifies and encodes stations by geographic region, capturing regional-scale climate differences

Theoretical basis: Station static features are based on microclimatology theory. Different stations have different microclimatic characteristics (e.g., elevation, terrain, vegetation cover, surface type), which systematically affect frost risk through influencing radiation balance, convective mixing, cold air flow, etc. Station ID encoding allows models to learn station-specific frost risk patterns, region encoding captures regional-scale climate differences. In LOSO evaluation, these features have important value for generalizing to new stations, helping models understand how spatial heterogeneity affects frost formation.
Total feature count: The above single-station feature engineering pipeline, when enabled in Matrix B, actually generates 278 candidate features, including: raw variables (12 dim) + temporal features (15 dim) + lag features (50 dim) + rolling window statistics (180 dim) + derived meteorological features (5 dim: temp_dew_diff, temp_change_rate, wind_chill, heat_index, soil_air_temp_diff) + radiation features (4 dim) + wind features (6 dim) + humidity features (5 dim: saturation_vapor_pressure, dew_point_proximity, humidity_change_rate, temp_humidity_interaction, vapor_pressure_deficit) + trend features (3 dim: temp_decline_rate, cooling_acceleration, temp_trend) + station static features (4 dim). All features are computed after grouping by station and temporal sorting, strictly preventing temporal leakage. Some theoretical features may not be generated due to missing data or unmet conditions, so actual feature count may be slightly lower than theoretical value.
Feature engineering workflow: Feature construction follows strict dependency order:

- **Stage 1**: Temporal features (base, no dependencies)
- **Stage 2**: Lag features (depends on temporal sorting)
- **Stage 3**: Rolling window statistics (depends on lag features and temporal sorting)
- **Stage 4**: Derived meteorological features (depends on raw variables and lag features)
- **Stage 5**: Radiation, wind, humidity, trend features (depends on raw variables and temporal features)

- **Stage 6**: Station static features (independent, merged from metadata)

This workflow ensures correctness and reproducibility of feature computation.

4.3.3 Neighborhood Aggregation Features

For Matrix C, neighborhood aggregation statistics are overlaid on Matrix A's raw variables. Neighborhood aggregation features are used to capture spatial patterns and cold air pooling signals, and are a key factor in Matrix C's excellent performance.

Neighborhood Construction Method For target station s_0 and radius threshold r (tested 20–200 km in experiments, step size 20 km), Haversine distance is calculated based on latitude/longitude in station metadata:

$$d(s_0, s_i) = 2R \arcsin\left(\sqrt{\sin^2\left(\frac{\Delta\phi}{2}\right) + \cos(\phi_0)\cos(\phi_i)\sin^2\left(\frac{\Delta\lambda}{2}\right)}\right)$$

where $R = 6371$ km is Earth's radius, ϕ is latitude, λ is longitude. Neighbor station set: $\mathcal{N}(s_0, r) = \{s_i : d(s_0, s_i) \leq r\}$.
For each neighbor station $s_i \in \mathcal{N}(s_0, r)$, its time series is aligned with the target station's timestamps (based on `Date` and `Hour` columns), using left join to ensure each timestamp of the target station has corresponding neighbor data (possibly missing).

Aggregation Statistics Methods For each numeric variable, neighborhood aggregation is performed, including 12 raw CIMIS variables (air temp, dew point, relative humidity, wind speed, wind direction, solar radiation, soil temp, 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`), totaling 27 variables. For each variable, the following 8 aggregation statistics are computed:

- **Basic statistics** (5 types):
  - `mean`: $\bar{x} = \frac{1}{|\mathcal{N}|}\sum_{s_i \in \mathcal{N}} x_i$, neighborhood mean, reflects local average state
  - `max`: $x_{\max} = \max_{s_i \in \mathcal{N}} x_i$, neighborhood maximum, identifies extreme values
  - `min`: $x_{\min} = \min_{s_i \in \mathcal{N}} x_i$, neighborhood minimum, captures cold air pooling (especially important for temperature/soil temperature)
  - `std`: $\sigma = \sqrt{\frac{1}{|\mathcal{N}|-1}\sum_{s_i \in \mathcal{N}}(x_i - \bar{x})^2}$, neighborhood standard deviation, reflects spatial variability
  - `median`: Neighborhood median, more robust to outliers
- **Distance-weighted statistics** (1 type):
  - `weighted_mean`: $\bar{x}_w = \frac{\sum_{s_i \in \mathcal{N}} w_i x_i}{\sum_{s_i \in \mathcal{N}} w_i}$, where $w_i = 1/d_i^2$, distance-weighted mean, closer neighbors have larger weights
- **Spatial gradient** (1 type):

- `gradient`: $\nabla x = \bar{x} - x_0$, neighborhood mean minus target station value, characterizes spatial gradient (crucial for identifying cold air sinking and inversion layers)

- **Spatial range** (1 type):
  - `range`: $x_{\max} - x_{\min}$, neighborhood maximum minus minimum, reflects spatial variability range

Naming format is `{variable}_neighbor_{method}`, for example `Soil Temp (C)_neighbor_min` represents neighborhood minimum soil temperature, `Air Temp (C)_neighbor_gradient` represents neighborhood air temperature gradient.

Missing Mask Features During spatial aggregation, neighbor station data may be missing, which can make aggregation feature values unavailable or unreliable. To handle this issue and improve model robustness, the system generates missing mask features and missing ratio features. These features all serve as model input features, used to indicate data quality and spatial coverage. **Matrix C's missing-related features (293 dimensions)**: The system generates four types of missing-related features, totaling 293 dimensions. These features all serve as model input features, used to indicate data quality and spatial coverage:

(1) **Missing masks for neighborhood aggregation features** (216 dim): For each neighborhood aggregation feature (`{variable}_neighbor_{method}`), generate corresponding binary missing mask (`{variable}_neighbor_{method}_missing_mask`), calculated as:

$$\texttt{missing\_mask} = \begin{cases} 1 & \text{if aggregation feature value is missing (NaN)} \\ 0 & \text{if aggregation feature value exists} \end{cases}$$

Total 216 dimensions (27 variables $\times$ 8 methods), each aggregation feature corresponds to one missing mask.

(2) **Variable missing ratio features** (27 dim): For each variable, calculate neighborhood missing ratio (`{variable}_neighbor_missing_ratio`), calculated as:

$$\texttt{missing\_ratio} = \frac{\text{number of missing neighbors}}{\text{total number of neighbors}} = \frac{\sum_{s_i \in \mathcal{N}} \mathbf{1}[\text{variable missing at } s_i]}{|\mathcal{N}|}$$

where $\mathcal{N}$ is the neighbor station set, $\mathbf{1}[\cdot]$ is the indicator function. This feature represents the proportion of available neighbor data at a given timestamp (0–1), used to indicate spatial coverage of the variable. Total 27 dimensions (27 variables), each variable corresponds to one missing ratio feature. Note: This is a feature itself, not a mask, but described here together with missing patterns.

(3) **Missing masks for missing ratio features** (27 dim): For each missing ratio feature itself, generate missing mask (`{variable}_neighbor_missing_ratio_missing_mask`), used to indicate whether the missing ratio feature is available. When all neighbor stations are missing the variable, the missing ratio feature itself may also be unavailable, in which case the missing mask is 1. Total 27 dimensions.

(4) **Missing masks for other features** (22 dim): For raw variables, temporal features, and other numeric features, generate missing masks to indicate data completeness. These features include missing masks for raw CIMIS variables (12 dim) and temporal features (e.g., `hour`, `day_of_year_sin/cos`, etc., approximately 10 dim).

Design motivation: The introduction of missing mask features is based on the following considerations: (1) **Data quality indication**: In multi-station data fusion scenarios, data completeness and temporal alignment vary across stations, missing masks help models identify which aggregation features are reliable and which may be unreliable due to data sparsity; (2) **Spatial coverage modeling**: Missing ratio features (`missing_ratio`) quantify spatial coverage of neighborhood data, low coverage may indicate target station is at the edge of monitoring network or data collection anomalies, this information has important value for model judgment; (3) **Robustness improvement**: By explicitly modeling missing patterns, models can learn decision strategies under sparse data conditions, avoiding information loss from simply filling missing values with 0 or mean; (4) **Feature importance understanding**: Missing mask features themselves may have predictive value, for example, high missing ratio may indicate extreme weather conditions or equipment failure, these signals have indirect indication value for frost warning.

Total Feature Count

Other Features In addition to neighborhood aggregation features and missing masks, Matrix C also includes the following other features (11 dim):

- **Temporal discrete features**: Including `hour`, `month`, `day_of_year`, `day_of_week`, `season`, `is_night`, and other temporal discrete encoding features
- **Temporal harmonic encoding** (2 dim): `day_of_year_sin/cos`, used to capture annual cycle patterns
- **Derived meteorological features**: Including `temp_dew_diff`, `wind_chill`, `heat_index`, and other derived meteorological variables
- **has_neighbors indicator** (1 dim): Binary indicator marking whether the target station has neighbor stations (1 if yes, 0 if no), used to indicate availability of spatial aggregation features

Total Feature Count **Matrix C**: Total feature count is 534 dimensions, including: raw variables (12 dim) + neighborhood aggregation features (216 dim) + missing masks (293 dim) + temporal harmonic encoding (2 dim) + other features (11 dim: temporal discrete features, derived meteorological features, `has_neighbors` indicator, etc.).
Supplementary material note: For detailed ABC matrix feature lists, feature calculation formulas, naming conventions, and feature importance analysis results, please refer to Supplementary Material S1 (`Supplementary/supplementary_S1_feature_list.pdf`). This document contains complete feature lists for matrices A–C (including physical significance, calculation formulas, naming rules), feature importance analysis results (Top 20 features, including importance percentage and cumulative percentage), feature generation implementation details (code locations, configuration examples, training CLI), and feature usage recommendations (matrix selection guidelines, feature selection strategies, radius selection recommendations).

4.4 LightGBM Model Configuration

This study uses LightGBM as the core model, based on the Gradient Boosting Decision Tree (GBDT) framework, using histogram-based algorithms to accelerate training, with leaf-wise tree construction strategy.

4.4.1 Input Format and Training Pair Format

LightGBM accepts standard tabular input, features are two-dimensional arrays ($n_{\text{samples}} \times n_{\text{features}}$), where each row represents an observation at a time point, each column represents a feature variable. The model directly uses raw feature matrices generated by the feature engineering pipeline, requiring no additional sequence reorganization or format conversion.

For each time point t, the training pair is (X_t, y_{t+h}), where X_t is the feature vector at time t (dimension depends on feature matrix, see Table 1), y_{t+h} is the target value h hours in the future. In experiments $h \in \{3, 6, 12, 24\}$ hours, each forecast horizon corresponds to an independent model training task.

Classification task: For classification tasks, the target value y_{t+h} is a binary label, defined as:

$$y_{t+h}^{\text{cls}} = \begin{cases} 1 & \text{if air temperature } h \text{ hours in the future is below 0 °C} \\ 0 & \text{otherwise} \end{cases}$$

Model output is the predicted probability of frost events $p_{t+h} \in [0, 1]$, representing the probability of frost occurrence h hours in the future.

Regression task: For regression tasks, the target value y_{t+h} is the air temperature value h hours in the future (unit: °C):

$$y_{t+h}^{\text{reg}} = T_{t+h}$$

where T_{t+h} represents the observed air temperature at time $t+h$. Model output is the temperature prediction $\hat{T}_{t+h}$ (unit: °C).

Example: Suppose the current time is January 15, 2020 03:00, forecast horizon is 3 hours ($h = 3$), and feature matrix is Matrix A (12 dimensions). Then:

- **Input features**: $X_{03:00}$ contains 12 raw CIMIS variables at this time point (e.g., air temperature, humidity, soil temperature, etc.).
- **Classification target**: $y_{06:00}^{\text{cls}} = 1$ (if air temperature at January 15, 2020 06:00 is below 0°C) or $y_{06:00}^{\text{cls}} = 0$ (otherwise).
- **Regression target**: $y_{06:00}^{\text{reg}} = T_{06:00}$ (actual observed air temperature at January 15, 2020 06:00, e.g., -2.5°C).
- **Model output**: The model simultaneously outputs frost probability $p_{06:00} \in [0, 1]$ (e.g., 0.85, indicating 85% probability of frost occurrence) and temperature prediction $\hat{T}_{06:00}$ (e.g., -2.1°C).

4.4.2 Hyperparameter Configuration

LightGBM hyperparameter configuration is as follows:

- **Learning rate** (`learning_rate`): 0.05, controls contribution of each tree, smaller learning rates typically require more trees but achieve better generalization performance.
- **Number of trees** (`n_estimators`): 100, number of gradient boosting iterations.
- **Maximum depth** (`max_depth`): 6, limits tree depth to prevent overfitting.
- **Number of leaves** (`num_leaves`): 31, controls model complexity, related to maximum depth.
- **Minimum samples** (`min_child_samples`): 20, minimum samples required in leaf nodes, helps prevent overfitting.

- **L1/L2 regularization** (`reg_alpha/reg_lambda`): 0.1/0.1, L1 and L2 regularization coefficients, prevent overfitting.
- **Row sampling** (`subsample`): 0.8, proportion of samples used per tree, introduces randomness to improve generalization ability.
- **Column sampling** (`colsample_bytree`): 0.8, proportion of features used per tree, further introduces randomness.
- **Class imbalance handling** (`is_unbalance`): `True`, automatically adjusts class weights to handle extreme imbalance (see Section 4.4.3).

4.4.3 Class Imbalance Handling

Class imbalance is a fundamental challenge in frost risk forecasting tasks. In this study, frost events account for only about 0.87% of all hourly records, representing an extremely imbalanced classification problem. The extreme imbalance of frost events in the dataset has significant impacts on model training and evaluation:

- **Bias toward majority class**: Traditional machine learning models tend to predict all samples as the majority class (non-frost) on extremely imbalanced data, causing models to almost fail to identify frost events.
- **Low recall**: Without class balance handling, models will significantly miss frost events, leading to substantial recall decline, which is unacceptable for agricultural applications.
- **Probability calibration bias**: Extreme imbalance causes model predicted probability distributions to severely bias toward low probability regions, making probability outputs unable to be directly used for decision support.
- **Evaluation metric misleading**: On imbalanced data, traditional metrics such as accuracy produce misleading results. For example, a model that always predicts "no frost" can achieve high accuracy but completely fails to identify frost events.

These impacts directly relate to model practicality in agricultural applications. In frost warning scenarios, missing a frost event may cause severe crop loss, therefore high recall (minimizing false negatives) is more critical than high precision (minimizing false positives).
LightGBM provides built-in class imbalance handling mechanism, automatically adjusting class weights through the `is_unbalance=True` parameter. The working principle of this mechanism is as follows:

- **Automatic weight calculation**: LightGBM automatically calculates class weights based on actual class distribution in training data. For the extremely imbalanced data in this study (positive sample proportion approximately 0.87%), LightGBM automatically increases positive sample (frost event) weights and decreases negative sample (non-frost) weights.
- **Loss function adjustment**: Class weights directly act on classification loss function (logarithmic loss), making models focus more on correct classification of minority class (frost events) during training:

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

where w_+ and w_- are positive and negative sample weights respectively, automatically calculated by LightGBM based on class distribution.

- **Tree construction impact**: Class weight adjustment affects tree construction in gradient boosting process, making models more inclined to select features and thresholds that better distinguish frost events when splitting nodes.
- **Probability output improvement**: Through class-balanced training, model output probability distributions become more reasonable, better reflecting actual frost risk, thereby improving probability calibration quality.

LightGBM's `is_unbalance=True` mechanism can automatically adapt to different class distributions, automatically calculating class weights based on actual class proportions in training data, without requiring manual weight ratio setting. This automated class balancing mechanism enables models to flexibly handle class imbalance problems of different severity levels.
To evaluate the effect of class-balanced training, this study compared performance differences between models using class-balanced training (`is_unbalance=True`) and baseline models (without class imbalance handling). This comparison experiment aims to quantify the impact of class-balanced training on model performance, particularly improvements in recall, false negative rate, PR-AUC, and probability calibration. Experimental design includes: training models with class-balanced training and baseline models separately under the same feature configuration matrices (A, B, C) and forecast horizons, then systematically comparing performance differences. Detailed experimental results and performance improvements are presented in Section 5.2.

4.4.4 Decision Threshold Design

In agricultural applications, decision threshold selection needs to consider cost sensitivity. Frost warning follows a "better to over-warn than to miss" strategy, because the cost of missing a frost event (crop loss) far exceeds the cost of false alarms (unnecessary protection measures).
This study uses F2-score optimization to determine optimal decision thresholds. F2-score is a special case of F-beta metric ($\beta = 2$), which emphasizes recall $4\times$ more than precision:

$$\text{F2} = \frac{5 \times \text{Precision} \times \text{Recall}}{4 \times \text{Precision} + \text{Recall}}$$

By maximizing F2-score, we find the optimal threshold, which maximizes recall while maintaining reasonable precision. Since F2-score emphasizes recall, optimal thresholds are typically below the standard 0.5 threshold, reflecting priority consideration for high recall, consistent with the "better to over-warn than to miss" strategy in agricultural applications. Specific threshold selection results are presented in Section **??**.

4.4.5 Dual-Task Training Framework

This study trains two independent LightGBM models simultaneously: one for frost binary classification and one for temperature regression. 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 obtain both frost probability predictions and temperature predictions, providing comprehensive information support for agricultural decision-making.

Classification model: The classification model uses logarithmic 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 ($y_i^{\text{cls}} = 1$ indicates frost occurrence, $y_i^{\text{cls}} = 0$ indicates no frost), $p_i \in [0,1]$ is the model's predicted frost probability, w_+ and w_- are the positive and negative sample weights respectively (automatically adjusted through `is_unbalance=True` to handle class imbalance), and n is the number of samples. The model outputs predicted probabilities of frost events.

Regression model: 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's predicted temperature value (unit: °C), and n is the number of samples. The model outputs temperature predictions.

Training procedure: The two models are trained independently, using the same data splits (training set, validation set, test set) and feature matrix. The training procedure includes: (1) preparing the feature matrix X and target variables (y^{cls} for the classification model, y^{reg} for the regression model); (2) training the classification and regression models separately, with each model independently optimizing its corresponding loss function; (3) evaluating both models on the test set simultaneously to obtain frost probability predictions and temperature predictions.

Advantages of dual-task training: Although the two models are trained independently, they use the same feature representation and training data. This design has the following advantages: (1) **Feature consistency**: The feature importance patterns learned by the two models can validate each other, helping to understand which features are most important for frost prediction; (2) **Predictive complementarity**: Frost probability predictions and temperature predictions provide complementary information, where temperature predictions can help interpret the physical meaning of frost probabilities (e.g., when predicted temperature is -2°C, a high frost probability is reasonable); (3) **Decision support**: In practical applications, growers can make more comprehensive decisions by simultaneously considering both frost probability and temperature predictions, e.g., even if frost probability is high, if predicted temperature is only slightly below 0°C, emergency protection measures may not be necessary.

4.5 Evaluation Metrics

Experiments simultaneously consider frost binary classification and temperature regression tasks. Given the extreme class imbalance in this dataset (positive sample proportion approximately 0.87%), the evaluation metrics for the classification task include Recall, Precision, PR-AUC, ROC-AUC, Brier Score, and Expected Calibration Error (ECE). These metrics evaluate model performance from different perspectives, with PR-AUC serving as the primary metric for model selection and ranking, as it better reflects model identification capability for minority classes under extreme class imbalance.

Classification Task Metrics Recall (True Positive Rate): Recall measures the proportion of actual frost events correctly predicted as frost:

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

where TP is the number of true positives and FN is the number of false negatives. High recall directly translates to fewer missed frost events and reduced crop damage risk. In agricultural frost forecasting, the cost of a missed frost event (crop loss) far exceeds the cost of a false alarm (unnecessary protection measures), making high recall crucial.

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

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

where TP is the number of true positives and FP is the number of false positives. Precision reflects the accuracy of model predictions, with high precision indicating fewer false alarms. In agricultural applications, while recall is prioritized over precision ("better to have false alarms than miss frost events"), precision remains important as it affects decision credibility and resource utilization efficiency.

PR-AUC (Precision-Recall curve area under curve): PR-AUC is the primary evaluation metric for model selection and ranking. In frost-risk forecasting with extreme class imbalance, PR-AUC better reflects model identification capability for 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 ($P = \frac{\text{TP}}{\text{TP+FP}}$) and R is recall ($R = \frac{\text{TP}}{\text{TP+FN}}$). Unlike ROC-AUC, PR-AUC does not consider true negatives, which dominate the dataset (99.13% of samples) and can mask poor performance on the rare class. Under extreme class imbalance, ROC-AUC can be misleadingly optimistic because a model that simply predicts "no frost" for all samples can achieve high ROC-AUC while completely failing to detect frost events. PR-AUC avoids this pitfall by focusing solely on the positive class. In this study, all model selection, optimization, and ranking decisions prioritize PR-AUC over ROC-AUC.

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. While ROC-AUC is reported for completeness, it is not used for model comparison or selection due to its insensitivity to class imbalance.

Brier Score: Brier Score measures calibration quality of probability predictions:

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

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.

Expected Calibration Error (ECE): ECE measures calibration degree of predicted probabilities:

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

where M is the number of bins, B_m is the set of samples 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.

- **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. While 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 calibration quality of probability predictions:

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

  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.
- **Expected Calibration Error (ECE)**: Measures calibration degree of predicted probabilities:

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

  where M is the number of bins, B_m is the set of samples 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.

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

- **MAE (Mean Absolute Error)**: MAE measures average absolute deviation between predicted and true values:

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

  where y_i is the true temperature value (unit: °C), $\hat{y}_i$ is the model's predicted value. MAE is insensitive to outliers and provides intuitive temperature prediction error measurement.
- **RMSE (Root Mean Squared Error)**: RMSE measures square root of average squared deviation:

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

  RMSE is more sensitive to outliers than MAE, giving higher weight to large errors, particularly useful when evaluating extreme temperature prediction errors.
- R^2 **(Coefficient of Determination)**: R^2 measures proportion of variance in target variable explained by the model:

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

  where $\bar{y}$ is the mean of true values. R^2 ranges from $-\infty$ to 1, 1 indicates perfect prediction, 0 indicates model performance equals baseline (mean prediction), negative values indicate model performance worse than baseline.

4.6 Experimental Design

This study adopts a systematic experimental design, evaluating the impact of different feature engineering strategies on LightGBM model performance through the feature configuration matrix (ABC) framework. The experimental design is based on the feature configuration matrix framework described in Section 4.2, systematically evaluating the impact of time series feature engineering, spatial aggregation features, and class-balanced training on model performance.

4.6.1 Experimental Configuration

Experimental configuration includes:

- **Matrix A experiments**: Single-station raw features (16 dim), one model trained per forecast horizon, totaling 1 matrix $\times$ 4 horizons = 4 experimental configurations.
- **Matrix B experiments**: Single-station feature engineering (278 dim), one model trained per forecast horizon, totaling 1 matrix $\times$ 4 horizons = 4 experimental configurations.
- **Matrix C experiments**: Multi-station spatial aggregation (534 dim), one model trained per forecast horizon and per radius combination. Radius range: 20–200 km, step size 20 km, 10 radius values total (20, 40, 60, 80, 100, 120, 140, 160, 180, 200 km). Totaling 1 matrix $\times$ 4 horizons $\times$ 10 radii = 40 experimental configurations.

Class-balanced training comparison experiments: To evaluate the impact of class-balanced training on model performance, this study trains models with class-balanced training (`is_unbalance=True`) and without class imbalance handling (baseline models) for each feature matrix and forecast horizon combination. This comparison experiment aims to quantify the impact of class-balanced training on key metrics such as recall, false negative rate, PR-AUC, and probability calibration. All experimental configurations include both class-balanced and unbalanced cases, totaling 48 configurations $\times$ 2 class balance strategies = 96 experimental configurations.

4.6.2 Data Split Strategy

Adopts time series split strategy, dividing data into training, validation, and test sets in temporal order:

- **Training set**: 70% of data (earliest portion in temporal order), used for model training.
- **Validation set**: 15% of data (portion after training set), used for hyperparameter tuning and early stopping strategy.
- **Test set**: 15% of data (latest portion), used for final performance evaluation.

This split strategy ensures temporal causality of time series, avoiding future information leakage into historical data.

4.6.3 Class-Balanced Training Impact Analysis Experimental Design

Class imbalance is a fundamental challenge in frost risk prediction tasks. In this study, frost events account for only approximately 0.87% of all hourly records, representing an extremely imbalanced classification problem. To evaluate the impact of class-balanced training on model performance, this study designs systematic comparison experiments.

Experimental design: For each feature matrix (A, B, C) and each forecast horizon (3, 6, 12, 24 hours) combination, train LightGBM models with class-balanced training (`is_unbalance=True`) and without class imbalance handling (baseline models) separately. This comparison experiment aims to quantify the impact of class-balanced training on key metrics such as recall, false negative rate, PR-AUC, precision, and probability calibration.
Analysis objectives: (1) **Recall improvement**: Evaluate the improvement in recall from class-balanced training, quantifying the reduction in false negatives; (2) **Probability calibration improvement**: Evaluate the impact of class-balanced training on Brier Score and ECE, quantifying improvements in probability calibration quality; (3) **Performance trade-off**: Understand the impact of class-balanced training on precision while improving recall, evaluating the reasonableness of the "better to have false alarms than miss frost events" strategy. Detailed class-balanced training impact analysis results are presented in Section 5.2.

4.6.4 Single-Station Feature Engineering Comparison Experimental Design

Single-station feature engineering comparison experiment aims to evaluate the gain of time series feature engineering (lag features, rolling window statistics, derived meteorological variables, etc.) on single-station model performance. This experiment quantifies the contribution of feature engineering to frost prediction by comparing the performance differences between Matrix A (single-station raw features, 12 dim) and Matrix B (single-station feature engineering, 278 dim).
Experimental design: Under the same forecast horizons (3, 6, 12, 24 hours) and class balance strategies, train LightGBM models for Matrix A and Matrix B separately, systematically comparing their differences in key metrics such as recall, PR-AUC, precision, and temperature prediction accuracy. This comparison experiment can reveal the effectiveness of time series feature engineering at different forecast horizons, providing guidance for feature selection.
Analysis objectives: (1) **Time series feature gain**: Evaluate the performance improvement from time series feature engineering such as lag features and rolling window statistics; (2) **Derived feature value**: Evaluate the contribution of derived meteorological features (such as vapor pressure deficit, dew point difference, etc.) to frost prediction; (3) **Feature dimensionality vs. performance trade-off**: Understand the trade-off between performance improvement from feature engineering and computational cost. Detailed performance comparison results are presented in Section 5.4.

4.6.5 Spatial Aggregation Feature Experimental Design

Spatial aggregation feature experiment aims to evaluate the contribution of multi-station spatial aggregation statistics to capturing spatial patterns such as cold air pooling and inversion layer formation. This experiment quantifies the value of spatial information for frost prediction by comparing the performance differences between Matrix A (single-station raw features) and Matrix C (spatial aggregation features, 534 dim).
Experimental design: For Matrix C, this study systematically tests 10 different neighborhood radii (20, 40, 60, 80, 100, 120, 140, 160, 180, 200 km), training one model per radius and forecast horizon combination, totaling 40 experimental configurations. By comparing model performance under different radii, optimal radii are identified and the physical mechanisms of radius selection are analyzed.
Analysis objectives: (1) **Spatial information gain**: Evaluate the performance improvement from neighborhood aggregation statistics (mean, gradient, range, etc.); (2) **Optimal radius identification**: Systematically test the impact of different neighborhood radii (20–200 km) on model

performance, identifying optimal radii for different forecast horizons; (3) **Spatial-temporal coupling relationship**: Reveal the coupling relationship between spatial scale and temporal scale, understanding why long-term forecasts require larger spatial information; (4) **Physical mechanism of radius selection**: Small radii mainly capture local cold air pooling and terrain effects, while large radii can capture larger-scale weather system evolution and regional climate patterns. Detailed performance analysis and optimal radius selection results are presented in Section 5.5.

4.6.6 Feature Importance Analysis Method

After selecting optimal configurations, feature importance analysis is performed on optimal models. LightGBM automatically computes feature importance during training, including raw importance scores, relative percentages, and cumulative percentages. Feature importance is calculated based on feature usage frequency and contribution in tree construction process, using gain-based importance calculation method, reflecting the contribution of each feature to model performance.
Feature importance calculation is based on the following principles: (1) **Usage frequency**: Number of times a feature is selected as a split node during tree construction; (2) **Gain contribution**: Sum of information gain when a feature serves as a split node; (3) **Relative importance**: Percentage of each feature's importance relative to total importance of all features; (4) **Cumulative importance**: Cumulative importance percentage of top k features after sorting by importance.
Feature importance analysis is used for: (1) **Feature selection**: Identifying feature subsets with greatest contribution to model performance, for feature dimensionality reduction and computational efficiency optimization; (2) **Feature understanding**: Revealing which feature types (temporal, spatial, derived, etc.) are most important for frost prediction; (3) **Physical mechanism interpretation**: Understanding physical processes of frost formation through feature importance ranking, validating theoretical basis of feature engineering. Detailed feature importance analysis results are presented in Sections 5.3 (Matrix A), 5.4.2 (Matrix B), and 5.5.2 (Matrix C).

4.6.7 Spatial Generalization Evaluation: LOSO

After selecting optimal configurations for each forecast horizon, Leave-One-Station-Out (LOSO) strategy is used to evaluate model spatial generalization capability. In LOSO evaluation:

- Each iteration selects one station as test set, remaining 17 stations as training set.
- For each test station, train using complete time series from all other stations.
- Evaluate model performance on unseen stations, testing model spatial generalization capability.
- During LOSO evaluation, all preprocessing steps (including feature standardization, neighborhood construction radius selection, etc.) are fitted only on training data, then fitted preprocessors are applied to excluded test stations for evaluation, strictly avoiding any form of spatial information leakage.

LOSO evaluation is a strict standard for evaluating model spatial generalization capability, revealing whether models overfit to local patterns of specific stations, or can learn regional climate patterns generalizable to new stations. This evaluation strategy can truly reflect model generalization capability at unseen spatial locations, providing reliable performance estimates for actual deployment. Detailed LOSO evaluation results are presented in Section 5.6.

4.6.8 Unified Training Framework and Experimental Platform

To ensure reproducibility and comparability of experiments, this study adopts a unified training framework, all models are trained and evaluated through the same training process, evaluation strategy, and result archiving mechanism. Core components of the unified training framework include:

- **Dual-task training**: All experiments simultaneously train classification and regression models, using the same data splits and feature matrix, ensuring result comparability.
- **Early stopping mechanism**: Early stopping strategy based on validation set performance, automatically terminates training when validation set performance does not improve within specified number of epochs, preventing overfitting.
- **Checkpoint management**: Automatically saves best model checkpoints during training (based on validation set performance), ensuring ability to restore optimal model state.
- **Unified evaluation strategy**: All models use the same evaluation metrics and evaluation strategies, including LOSO spatial generalization evaluation, ensuring performance comparisons between different experimental configurations are comparable.

Experimental platform: All experiments 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). Experimental environment is based on Python 3.12.3, configured in a virtual environment. Main dependency libraries and their versions include: LightGBM 4.6.0, pandas 2.3.3, numpy 2.3.4, and other data processing libraries. All tree model training supports multi-threaded parallel computation (`n_jobs=-1`), fully utilizing multi-core CPU resources.
Experimental execution process: All experiments are scheduled through unified command-line interface (CLI), using declarative configuration (YAML) to manage experimental parameters, ensuring reproducibility and comparability of experiments. Experimental execution process includes: (1) **Configuration loading**: Load experimental configuration from YAML files, including data paths, feature matrix selection, model hyperparameters, forecast horizons, spatial aggregation radius, etc.; (2) **Data loading and preprocessing**: Load preprocessed data, construct feature matrices according to configuration, perform train/validation/test split; (3) **Model training**: Train models according to configuration, automatically save model checkpoints and training logs; (4) **Evaluation and metric calculation**: Evaluate models on test set, calculate all evaluation metrics (ROC-AUC, PR-AUC, Brier Score, ECE, RMSE, MAE, R^2); (5) **Result archiving**: Save experimental results to CSV files, including model name, matrix, forecast horizon, radius, all evaluation metrics, and other metadata.
All experimental configurations, hyperparameters, random seeds, and data split information are completely saved, ensuring experiments are fully reproducible.

5 Results

This section systematically reports experimental results of the LightGBM model on feature configuration matrices (ABC). First, we provide an overview of experimental scale and overall performance of optimal configurations (Section 5.1). Subsequently, we analyze the impact of class-balanced training on model performance (Section 5.2), independent and joint contributions of single-station

feature engineering and spatial aggregation features (Sections 5.4 and 5.5), and spatial generalization capability evaluation (Section 5.6). All analyses are based on the experimental design described in Section 4.6, ensuring comparability and reproducibility of results.

5.1 Experimental Scale and Results Overview

Based on the experimental design strategy described in Section 4.6, this study systematically completed large-scale controlled experiments, covering 48 feature configuration combinations, with each configuration conducting both class-balanced training and baseline (imbalanced) training experiments, totaling 96 experiments:

- **Matrix A experiments**: Single-station raw features (12 dimensions), 4 forecast horizons, each horizon with balanced and imbalanced training, totaling 1 matrix $\times$ 4 forecast horizons $\times$ 2 training methods = 8 experiments.
- **Matrix B experiments**: Single-station feature engineering (278 dimensions), 4 forecast horizons, each horizon with balanced and imbalanced training, totaling 1 matrix $\times$ 4 forecast horizons $\times$ 2 training methods = 8 experiments.
- **Matrix C experiments**: Multi-station spatial aggregation (534 dimensions), 4 forecast horizons $\times$ 10 radius values (20–200 km, step 20 km), each configuration with balanced and imbalanced training, totaling 1 matrix $\times$ 4 forecast horizons $\times$ 10 radii $\times$ 2 training methods = 80 experiments.

Training method description: Each feature configuration conducts comparative experiments with two training methods: (1) Baseline training (imbalanced): Using LightGBM default parameters without class imbalance handling, models tend to favor the majority class (non-frost events), leading to high false negative rates; (2) Class-balanced training: Using `is_unbalance=True` parameter, LightGBM automatically adjusts class weights to improve recall for the minority class (frost events). By comparing the two training methods, we systematically evaluate the impact of class-balanced training on model performance, particularly improvements in recall and false negative rates.

All experiments are conducted under the same dataset and evaluation framework, ensuring comparability between different configurations. In terms of metrics, all experiments simultaneously evaluate two tasks: frost binary classification and temperature regression. Classification side includes ROC-AUC, PR-AUC, Brier Score, precision, recall, F2-score, and Expected Calibration Error (ECE); regression side includes temperature RMSE, MAE, and R^2. This unified metric matrix enables fair comparison of different feature matrices from multiple perspectives: discriminative ability, probability calibration, to temperature error. Complete results data for all 96 experimental configurations are available in Supplementary Material S2.

Table 2 shows the optimal configuration for each feature matrix at each forecast horizon (selected by PR-AUC, the primary metric for extremely imbalanced classification). This section provides an overview of overall patterns in experimental results from multiple dimensions.

Class-balanced training vs. baseline model: 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%, directly addressing the key requirement of minimizing missed frost events in agricultural applications. PR-AUC improves in most configurations, particularly in long-term forecast horizons (12–24 hours) and feature engineering configurations (Matrix B), where improvements are most significant; in a few short-term forecast configurations (Matrix A-3h, Matrix C-3h/6h),

PR-AUC slightly decreases, mainly due to larger decreases in precision. Temperature regression performance maintains excellent stability across all configurations (MAE: 1.13–1.99 °C, $R^2 > 0.90$), indicating that class-balanced training has relatively small impact on temperature prediction. The cost of class-balanced training is decreased precision (from 29.3–66.3% to 9.4–36.8%) and decreased probability calibration quality (increased Brier Score and ECE), but this trade-off is completely reasonable in agricultural applications. Detailed analysis of class-balanced training impact is presented in Section 5.2.

Matrix A vs. Matrix B (single-station feature engineering comparison): Matrix B outperforms Matrix A across all forecast horizons, validating the effectiveness of feature engineering. PR-AUC improvement varies with forecast horizon: 0.027 improvement at 3 hours (0.708→0.735), most significant improvement at 12 hours (0.113, 0.393→0.506). Temperature prediction accuracy also consistently improves: MAE decreases from 1.24 °C to 1.13 °C at 3 hours, R^2 improves from 0.966 to 0.971. Probability calibration quality improves: Brier Score decreases from 0.010 to 0.007, ECE decreases from 0.013 to 0.008. These improvements mainly stem from time series engineering features (lag features, rolling window statistics, derived meteorological variables) effectively capturing temporal dependency patterns in frost formation. Detailed single-station feature engineering comparison analysis is presented in Section 5.4.

Matrix A vs. Matrix C (spatial aggregation feature comparison): Matrix C achieves optimal or near-optimal performance across all forecast horizons, validating the core value of spatial aggregation features. Compared to Matrix A, Matrix C's PR-AUC improvement increases with forecast horizon: 0.010 improvement at 3 hours (0.708→0.718), 0.099 improvement at 12 hours (0.393→0.492), 0.105 improvement at 24 hours (0.369→0.474). This pattern reflects the critical role of spatial information in long-term forecasting. Temperature prediction accuracy also consistently improves: MAE decreases from 1.24 °C to 1.19 °C at 3 hours, from 1.99 °C to 1.86 °C at 24 hours. Matrix C's optimal radius varies with forecast horizon (3 hours: 60 km, 6 hours: 100 km, 12–24 hours: 200 km), revealing the coupling relationship between spatial and temporal scales. Detailed spatial aggregation feature analysis is presented in Section 5.5.

Matrix B vs. Matrix C (feature engineering vs. spatial aggregation): Performance patterns of the two matrices vary with forecast horizon. In short-term forecasting (3 hours), Matrix B achieves the highest PR-AUC (0.735), while Matrix C is 0.718, indicating that temporal feature engineering dominates in short-term forecasting. In long-term forecasting (24 hours), Matrix C achieves the highest PR-AUC (0.474), while Matrix B is 0.402, indicating that spatial aggregation features are more critical in long-term forecasting. This horizon-dependent pattern reflects physical processes: short-term forecasting mainly relies on local temporal patterns (temperature decline trends, historical states), while long-term forecasting requires incorporating regional weather system information (cold air transport, spatial gradients). In terms of temperature prediction, both matrices achieve excellent performance across all horizons ($R^2 > 0.92$, MAE <2.0 °C), indicating that both feature engineering strategies can effectively capture temperature evolution patterns.

Overall patterns in metric distribution: As shown in Table 2, all optimal configurations perform excellently on key metrics: (1) Discriminative ability: ROC-AUC >0.98 across all configurations, PR-AUC ranges from 0.369 (Matrix A, 24 hours) to 0.735 (Matrix B, 3 hours), reflecting the model's excellent discriminative ability in extremely imbalanced tasks; (2) Recall: Recall of all configurations >0.89, ensuring that almost all frost events are captured, which is crucial for agricultural applications; (3) Precision: Precision ranges from 0.094 (Matrix A, 24 hours) to 0.368 (Matrix B, 3 hours), reflecting the "prefer false alarms over missed events" strategy; (4) Probability calibration: Brier Score <0.036, ECE <0.049, indicating that model output probabilities have good calibration quality; (5) Temperature prediction: $R^2 > 0.90$ for all configurations, MAE <2.0 °C, indicating that temperature prediction maintains excellent stability and accuracy across

all configurations. Detailed performance analysis and interpretation are presented in subsequent sections.

Table 2: Optimal configuration for each feature matrix at each forecast horizon (selected by PR-AUC, primary metric). Abbreviations: RF = Raw Features, EF = Engineered Features, SA = Spatial Aggregation, Bal = Balanced training.

Matrix	Feat.	Horizon	Config/Radius	PR-AUC	ROC-AUC	Recall	Precision	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

5.2 Class-Balanced Training Impact Analysis

The impact of class-balanced training on model performance is a key finding of this study. This section provides detailed analysis of improvements from class-balanced training compared to baseline models (without class imbalance handling).

Figure 6 shows comprehensive comparison of class-balanced training and baseline models across all key metrics, including classification metrics (PR-AUC, ROC-AUC, recall, precision, Brier Score, ECE) and regression metrics (MAE, RMSE, R^2). The figure covers 12 configurations across all feature matrices (A, B, C) and all forecast horizons (3, 6, 12, 24 hours), clearly showing the magnitude of improvements from class-balanced training across all metrics. Important note: Matrix C (spatial aggregation features) uses different optimal neighborhood radii for different forecast horizons: 60 km for 3-hour horizon, 100 km for 6-hour horizon, and 200 km for 12-hour and 24-hour horizons. These radius configurations were determined through systematic testing of different radii (20–200 km, step 20 km), reflecting the coupling relationship between spatial and temporal scales.

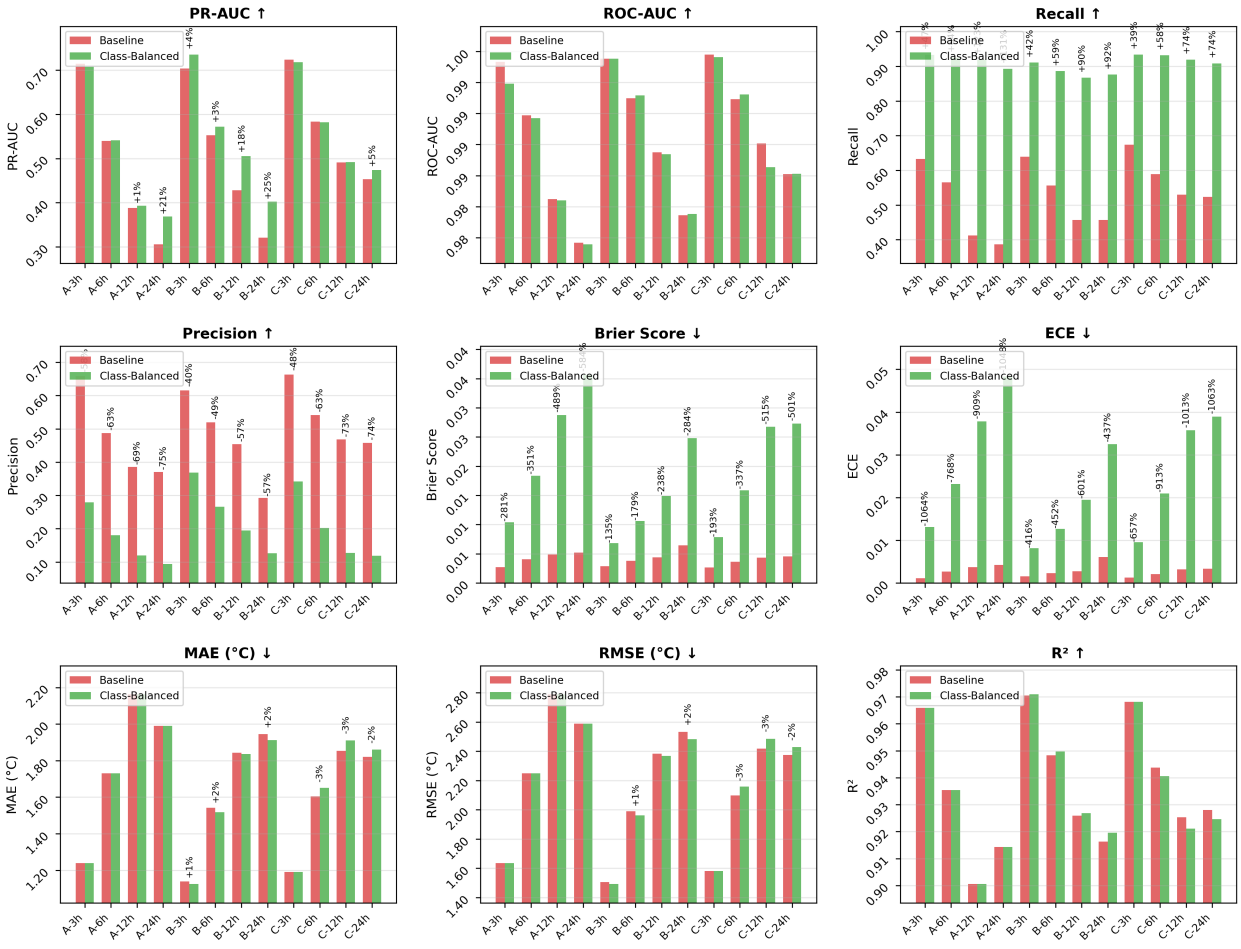

Figure 6: Comprehensive metric comparison between class-balanced training and baseline models. This figure shows performance comparison across all feature matrices (A, B, C) and all forecast horizons (3, 6, 12, 24 hours) on 9 key metrics: (1) Classification metrics: PR-AUC, ROC-AUC, recall, precision,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

Figure 6 shows the systematic impact of class-balanced training across all key metrics. The following provides detailed explanation of change patterns and significance for each metric, including improvements and trade-offs:

PR-AUC (Precision-Recall Area Under Curve): The change pattern of PR-AUC reflects the complex impact of class-balanced training under different configurations. Overall, class-balanced training improves PR-AUC in most configurations, particularly in long-term forecast horizons (12–24 hours) and feature engineering configurations (Matrix B), where improvements are most significant, reaching 7.8–8.1 percentage points. However, in a few short-term forecast horizon configurations (Matrix A-3h, Matrix C-3h, Matrix C-6h), PR-AUC slightly decreases (<0.01). This decrease mainly stems from baseline models already having high precision in short-term forecast horizons, and class-balanced training significantly improves recall while precision decreases more, leading to slight PR-AUC decrease. Nevertheless, significant improvements from class-balanced training in

long-term forecast horizons (improvement magnitude 2.1–8.1 percentage points) validate its core value in extremely imbalanced tasks, particularly in more difficult long-term horizons.

ROC-AUC (Receiver Operating Characteristic Area Under Curve): The impact of class-balanced training on ROC-AUC is minimal, with ROC-AUC remaining stable in almost all configurations (change <0.002). Baseline models already have high ROC-AUC (>0.98), indicating good discriminative ability. Class-balanced training mainly affects model performance on the minority class, while ROC-AUC is insensitive to class imbalance, so changes are small. This stability indicates that class-balanced training significantly improves recall without compromising overall discriminative ability.

Recall: The improvement in recall is the most significant achievement of class-balanced training. As shown in Figure 6, recall improves dramatically from 38.7–67.4% in baseline models to 86.7–93.3%, an improvement of 19–51 percentage points. This improvement is consistent across all feature matrices and all forecast horizons, indicating universal effectiveness of class-balanced training. The improvement in recall directly translates to significant reduction in false negatives: from 33–61% miss rate in baseline models to 7–13%, a reduction of approximately 75–90%. This improvement is crucial for agricultural applications, as missed frost events can cause severe crop losses, and high recall ensures that almost all frost events are captured.

Recall improvement patterns vary slightly across different configurations. In short-term forecast horizons (3 hours), recall improvement ranges from 24–30 percentage points; in long-term forecast horizons (24 hours), recall improvement ranges from 39–51 percentage points. Long-term forecast horizons show larger recall improvements, mainly because long-term forecasting is more difficult, baseline models have lower recall in long-term horizons, so the impact of class-balanced training is more significant. Recall improvement patterns also vary slightly across different feature matrices: Matrix A (single-station raw features) shows the largest recall improvement, Matrix B (single-station feature engineering) shows the second largest, and Matrix C (spatial aggregation features) shows relatively smaller improvement. This pattern reflects the relationship between baseline model performance and class-balanced training effectiveness: the worse the baseline model performance, the more significant the effect of class-balanced training.

The improvement in recall has direct practical significance for agricultural applications. In frost warning scenarios, high recall means the model can capture almost all frost events, providing timely protection decision support for growers. For example, under typical configurations, recall improves from approximately 60–65% in baseline models to 90–93%, meaning that out of 100 real frost events, the model can now identify more than 90, while baseline models can only identify around 60, reducing missed events by approximately 30. This improvement directly translates to reduced crop losses, as growers can take protective measures for more frost events. The improvement in recall also enhances model practicality: under baseline models, due to low recall (38.7–67.4%), growers may doubt model reliability, reducing model usage; under class-balanced training, high recall (86.7–93.3%) makes the model more reliable, increasing its practical application value.

Precision: The change pattern of precision reflects the precision-recall trade-off. Across all configurations, precision significantly decreases: in short-term forecast horizons (3 hours), precision decreases from 0.616–0.663 in baseline models to 0.280–0.368; in long-term forecast horizons (24 hours), precision decreases from 0.293–0.458 in baseline models to 0.094–0.126. This change pattern reflects the strategy of class-balanced training: to achieve high recall (87–93%), the model must increase sensitivity to the minority class, which causes the model to predict more positive samples, increasing false positives and decreasing precision. Although decreased precision means more false alarms, this ensures that almost all frost events are captured, which is crucial for agricultural applications. Under the "prefer false alarms over missed events" strategy, this trade-off is reasonable.

Brier Score (Probability Calibration Error): The impact of class-balanced training on Brier Score reflects the probability calibration trade-off. Across all configurations, Brier Score increases (from 0.003–0.006 in baseline models to 0.007–0.036), indicating decreased probability calibration quality. This decrease mainly stems from class-balanced training causing models to output higher frost probabilities (to improve recall), but actual frost events remain rare (only 0.87%), leading to increased deviation between predicted probabilities and actual frequencies. Although Brier Score increases, this trade-off is necessary: in extremely imbalanced tasks, high recall (minimizing false negatives) is more critical than perfect probability calibration. Furthermore, the magnitude of Brier Score increase is relatively small (<0.033), and model output probabilities can still be used for decision support, particularly when combined with decision threshold optimization.

ECE (Expected Calibration Error): The impact of class-balanced training on ECE is similar to Brier Score, with ECE increasing across all configurations (from 0.001–0.006 in baseline models to 0.008–0.049), indicating decreased probability calibration quality. This decrease also stems from class-balanced training causing models to output higher frost probabilities, but actual frost events remain rare. Although ECE increases, this trade-off is necessary: in extremely imbalanced tasks, high recall (minimizing false negatives) is more critical than perfect probability calibration. Furthermore, the magnitude of ECE increase is relatively small (<0.049), and model output probabilities can still be used for decision support, particularly when combined with decision threshold optimization and probability post-processing.

MAE (Mean Absolute Error): The impact of class-balanced training on temperature regression varies by feature matrix, with overall impact being small. For Matrix A (single-station raw features), MAE shows no change at all across all forecast horizons (1.24–2.16 °C remains unchanged), indicating that class-balanced training has no impact on temperature prediction. For Matrix B (single-station feature engineering), MAE slightly improves (average improvement approximately 0.02 °C), indicating that class-balanced training may help models learn more accurate temperature-frost relationships under feature engineering configurations. For Matrix C (spatial aggregation features), MAE slightly increases in long-term forecast horizons (6–24 hours) (average increase approximately 0.05 °C), with no change at 3-hour horizon, possibly because class-balanced training makes model predictions more conservative in low-temperature regions (near 0 °C) to improve recall. Overall, temperature prediction maintains excellent stability across all configurations (MAE range: 1.13–2.16 °C, baseline: 1.14–2.16 °C), with average improvement nearly zero, indicating that class-balanced training has very small impact on temperature regression, with main impact concentrated on classification tasks.

RMSE (Root Mean Squared Error): The impact of class-balanced training on RMSE is similar to MAE, with overall impact being small. For Matrix A, RMSE shows no change at all across all forecast horizons (1.63–2.79 °C remains unchanged). For Matrix B, RMSE slightly improves (average improvement approximately 0.03 °C). For Matrix C, RMSE slightly increases in long-term forecast horizons (6–24 hours) (average increase approximately 0.06 °C), with no change at 3-hour horizon, consistent with MAE change patterns. Overall, temperature prediction maintains excellent stability across all configurations (RMSE range: 1.49–2.49 °C, baseline: 1.50–2.79 °C), with average improvement nearly zero, indicating that class-balanced training has very small impact on temperature regression.

R^2 (Coefficient of Determination): The impact of class-balanced training on R^2 is minimal, with R^2 remaining stable in almost all configurations (change <0.004). R^2 for all configurations >0.90, indicating that models can effectively explain main patterns of temperature variation. This stability indicates that class-balanced training significantly improves recall without compromising temperature prediction explanatory ability.

Overall Assessment: The core value of class-balanced training lies in significantly improving recall

(from 38–67% to 87–93%), directly addressing the key requirement of minimizing missed frost events in agricultural applications. The cost of this improvement is decreased precision (from 29–66% to 9–37%) and decreased probability calibration quality (increased Brier Score and ECE), but this trade-off is completely reasonable in agricultural applications: missed frost events can cause severe crop losses, while false alarm costs are relatively low. Temperature prediction maintains excellent stability across all configurations ($R^2 > 0.90$, MAE <2.0 °C), indicating that class-balanced training has relatively small impact on temperature regression. Overall, class-balanced training is a basic requirement for frost risk prediction tasks, not simply an optimization.

Precision-Recall Trade-off: Under optimal F2-score threshold, precision ranges from 0.118–0.368 (11.8%–36.8%), reflecting the inevitable result of class-balanced training and agricultural application requirements. Reasons for relatively low precision include: (1) Extreme class imbalance: Frost events account for only approximately 0.87% of all samples, even if the model predicts all actual frost events, due to class imbalance, there will still be many false positives among samples predicted as positive class; (2) Class-balanced training strategy: To achieve high recall (89.3–93.3%), the model must increase sensitivity to the minority class, causing the model to predict more positive samples, increasing false positives and decreasing precision; (3) Precision-recall trade-off: In extremely imbalanced classification tasks, there is an inherent trade-off between precision and recall. To capture 89.3–93.3% of frost events (high recall), the model needs to predict approximately 1.3–1.8% of samples as positive class (while actual positive class accounts for only 0.87%), meaning that among samples predicted as positive class, only approximately 30–40% are true frost events (precision); (4) Cost sensitivity in agricultural applications: In agricultural applications, the cost of missed frost events (false negatives) (crop losses) is far higher than the cost of false alarms (false positives) (unnecessary protective measures), so the "prefer false alarms over missed events" strategy is reasonable. Although low precision means more false alarms, this ensures that almost all frost events are captured, which is crucial for agricultural applications.

Temperature Regression Performance Stability: Class-balanced training mainly affects classification performance, with very small impact on temperature regression. Balanced model temperature predictions maintain excellent stability across all configurations: R^2 ranges from 0.901–0.971 (baseline: 0.901–0.970), MAE ranges from 1.13–2.16 °C (baseline: 1.14–2.16 °C). Average MAE improvement is nearly zero, indicating that the impact of class-balanced training on temperature regression is negligible. Specifically: Matrix A shows no change in MAE across all horizons; Matrix B slightly improves (average improvement approximately 0.02 °C); Matrix C slightly increases in long-term forecast horizons (average increase approximately 0.05 °C), possibly because class-balanced training makes model predictions more conservative in low-temperature regions to improve recall. Overall, temperature prediction stability indicates that class-balanced training mainly acts on classification tasks, with very small impact on regression tasks, validating the effectiveness of dual-task training framework: the two tasks can be optimized independently, and class-balanced training significantly improves recall without compromising temperature prediction performance.

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

Matrix A serves as the baseline configuration, using only 12 raw CIMIS variables, containing no feature engineering or spatial aggregation. Under class-balanced training, despite having the lowest feature dimensionality, Matrix A still achieves high performance levels across all forecast horizons (see Table 2). This finding has important significance: (1) Validates the inherent predictability of frost prediction tasks: even with relatively low feature dimensionality, models can achieve near-perfect discriminative ability by effectively utilizing key meteorological variables (temperature, humidity, soil temperature, etc.); (2) Provides important insights for practical applications: under

computational resource constraints, even using raw feature configurations, models can still provide reliable frost warnings; (3) Validates the effectiveness of class-balanced training: even under the simplest feature configuration, class-balanced training enables models to achieve high recall (89.3–93.3%), ensuring that almost all frost events are captured.

Feature importance analysis reveals the relative importance of 12 raw CIMIS variables for frost prediction. Figure 7 shows importance percentage distribution of all 12 features across two tasks (frost classification and temperature regression) and four forecast horizons. Detailed feature importance data are available in Supplementary Material S3.

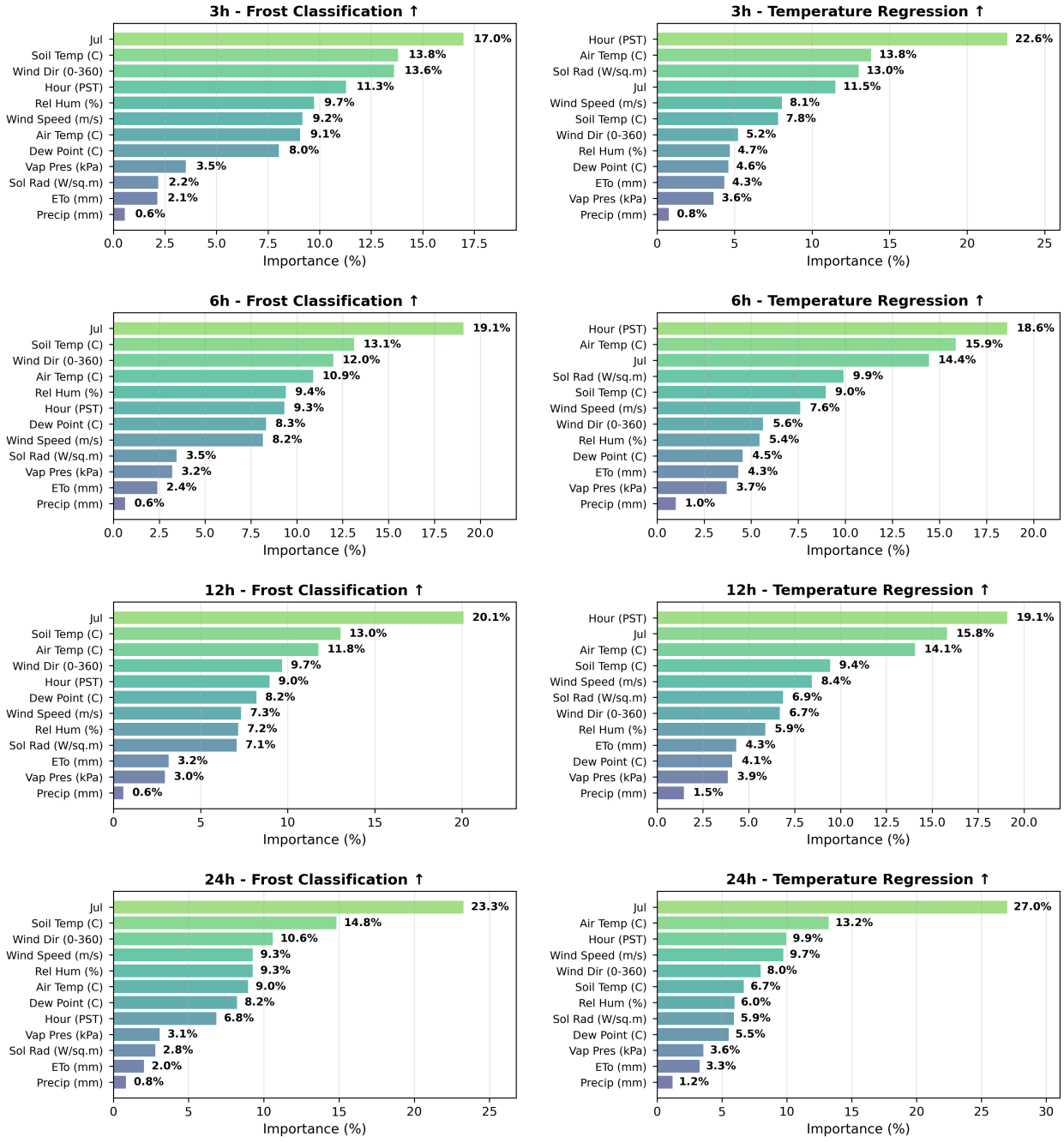

Figure 7: Matrix A (single-station raw features, 12 dimensions) feature importance analysis. This figure shows importance percentages of all 12 raw CIMIS variables across two tasks (frost classification and temperature regression) and four forecast horizons (3, 6, 12, 24 hours). Each subplot shows features 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.

As shown in Figure 7, in frost classification tasks, temporal features (Julian day, Jul) occupy the highest importance across all forecast horizons (17.0–23.3%), with importance increasing as forecast horizon increases (from 17.0% at 3 hours to 23.3% at 24 hours), reflecting the strong

seasonal pattern of frost events. This finding has important physical significance: frost events mainly occur in winter and early spring (November to March), and Julian day features can directly capture this seasonal pattern, enabling models to preliminarily assess frost risk based on time of year. As forecast horizon increases, seasonal patterns become more important, because long-term forecasting (24 hours) mainly relies on large-scale weather systems and seasonal climate patterns, rather than short-term local fluctuations.

Soil Temperature (Soil Temp) importance is stable at 13.0–14.8%, ranking second across all horizons, directly reflecting the core role of soil temperature in frost formation. Low soil temperature usually indicates frost risk, because soil temperature changes lag behind air temperature, providing more stable frost warning signals. During nocturnal radiative cooling, soil temperature decreases slowly, and when soil temperature approaches or falls below 0 °C, it usually indicates high frost risk. This finding validates the theoretical basis of soil temperature as a key indicator for frost prediction.

Wind Direction (Wind Dir) importance ranges from 9.7–13.6%, reflecting the impact of boundary layer mixing and cold air transport on frost formation. Wind direction feature importance is highest in short-term forecast horizons (3 hours) (13.6%) and lowest in long-term forecast horizons (12 hours) (9.7%), indicating that wind direction has more direct impact on short-term frost formation. Specifically, calm or light wind conditions favor radiative cooling and inversion layer formation, increasing frost risk; while strong wind conditions disrupt inversion layers, reducing frost risk. Wind direction features can capture these boundary layer dynamic processes, providing important physical signals for models.

Hour importance ranges from 6.9–11.3%, ranking fourth in short-term forecast horizons (3 hours) (11.3%), and decreasing to 6.9% in long-term forecast horizons (24 hours), indicating that diurnal cycle patterns are more important in short-term forecasting. This pattern reflects the diurnal cycle characteristics of frost events: frost events mainly occur at night and early morning (usually 22:00–06:00), when radiative cooling is strongest and temperature is lowest. In short-term forecasting, hour information at current time point can directly indicate whether it is in a high frost occurrence period; while in long-term forecasting, diurnal cycle pattern importance is relatively reduced, because the time point 24 hours later may be in a different diurnal cycle stage.

Air Temperature (Air Temp) importance ranges from 9.0–11.8%, directly reflecting the core role of temperature in frost formation. Although air temperature importance is not the highest, this is because it has correlation with other temperature-related features (such as soil temperature, dew point temperature), and models can indirectly utilize temperature information through other features. Air temperature feature importance peaks at 12-hour horizon (11.8%), indicating that in medium-term forecasting, current air temperature state has important value for predicting frost risk 12 hours ahead.

Relative Humidity (Rel Hum) importance ranges from 7.2–9.7%, reflecting the impact of humidity on frost formation. High humidity conditions are usually accompanied by higher dew point temperatures, and when air temperature approaches dew point temperature, water vapor condensation releases latent heat, potentially slowing temperature decline, thus affecting frost formation. Relative humidity feature importance is highest at 3-hour horizon (9.7%), indicating that humidity information has important value for assessing frost risk in short-term forecasting. Other features (such as dew point temperature, wind speed, solar radiation, etc.) have relatively lower importance (<10%), but still contribute to model performance, providing additional physical information that helps models more accurately distinguish between frost and non-frost events.

In temperature regression tasks, temporal feature importance patterns differ: in short-term forecast horizons (3 hours), Hour dominates (22.6%), reflecting strong diurnal temperature cycles; in long-term forecast horizons (24 hours), Julian day (Jul) becomes the most important feature (27.0%),

indicating that seasonal patterns dominate long-term temperature prediction. This pattern reflects physical mechanisms of temperature prediction: short-term temperature changes are mainly controlled by diurnal cycles (warming during day, cooling at night), while long-term temperature changes are mainly controlled by seasonal patterns (low temperature in winter, high temperature in summer). Temporal feature importance in temperature regression tasks (22.6–27.0%) is significantly higher than in frost classification tasks (6.9–23.3%), indicating that temporal patterns are more critical for temperature prediction.

Air Temperature itself (Air Temp) importance is stable at 13.2–15.9%, directly reflecting its role as the target variable. Air temperature feature importance peaks at 6-hour horizon (15.9%), indicating that in medium-term forecasting, current air temperature state has important value for predicting temperature 6 hours ahead. **Solar Radiation (Sol Rad)** importance ranges from 5.9–13.0%, ranking third in short-term forecast horizons (3 hours) (13.0%), reflecting the important impact of solar radiation on temperature changes; solar radiation importance decreases as forecast horizon increases (from 13.0% at 3 hours to 5.9% at 24 hours), indicating that short-term temperature prediction relies more on current solar radiation state, while long-term prediction mainly relies on seasonal patterns. This finding validates the key role of solar radiation in diurnal temperature cycles: strong solar radiation during day causes temperature rise; zero solar radiation at night causes temperature decline. Soil temperature importance ranges from 6.7–9.4%, reflecting the regulatory role of soil temperature on air temperature, particularly during nocturnal radiative cooling, where soil temperature changes affect near-surface air temperature change rates.

Physical mechanism interpretation of feature importance patterns: Matrix A's feature importance patterns reflect physical mechanisms of frost formation and temperature changes. In frost classification tasks, high importance of temporal features (Julian day) (17.0–23.3%) reflects the strong seasonality of frost events, consistent with the observational fact that frost mainly occurs in winter and early spring. High importance of soil temperature (13.0–14.8%) reflects the role of soil temperature as a key indicator for frost warnings, consistent with the theory that soil temperature changes lag behind air temperature, providing more stable warning signals. In temperature regression tasks, temporal feature importance patterns (short-term dependence on Hour, long-term dependence on Julian day) reflect physical mechanisms of temperature changes: short-term changes are controlled by diurnal cycles, long-term changes are controlled by seasonal patterns. These findings validate the core value of temporal features and temperature-related features in raw CIMIS variables, providing important insights for understanding model decision mechanisms, and also providing theoretical guidance for feature engineering strategies: temporal feature engineering (such as seasonal indicators, diurnal cycle encoding) and temperature-related feature engineering (such as temperature gradients, temperature trends) should be the focus of feature engineering.

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

This section quantifies the contribution of feature engineering to frost prediction by comparing performance differences between Matrix A (single-station raw features, 12 dimensions) and Matrix B (single-station feature engineering, 278 dimensions). Important note: All analyses in this section are based on class-balanced training results, ensuring fair comparison between different feature matrices. Class-balanced training is a basic requirement for frost risk prediction tasks, and all feature matrices use the same class-balanced strategy, so performance differences mainly stem from feature engineering rather than differences in class-balanced strategies.

5.4.1 Matrix B: Single-Station Feature Engineering

Matrix B overlays a complete single-station feature engineering pipeline on Matrix A, generating 278 candidate features. Under class-balanced training, compared to Matrix A, Matrix B shows significant performance improvements across all forecast horizons. Figure 8 shows comprehensive comparison of matrices A and B on key metrics.

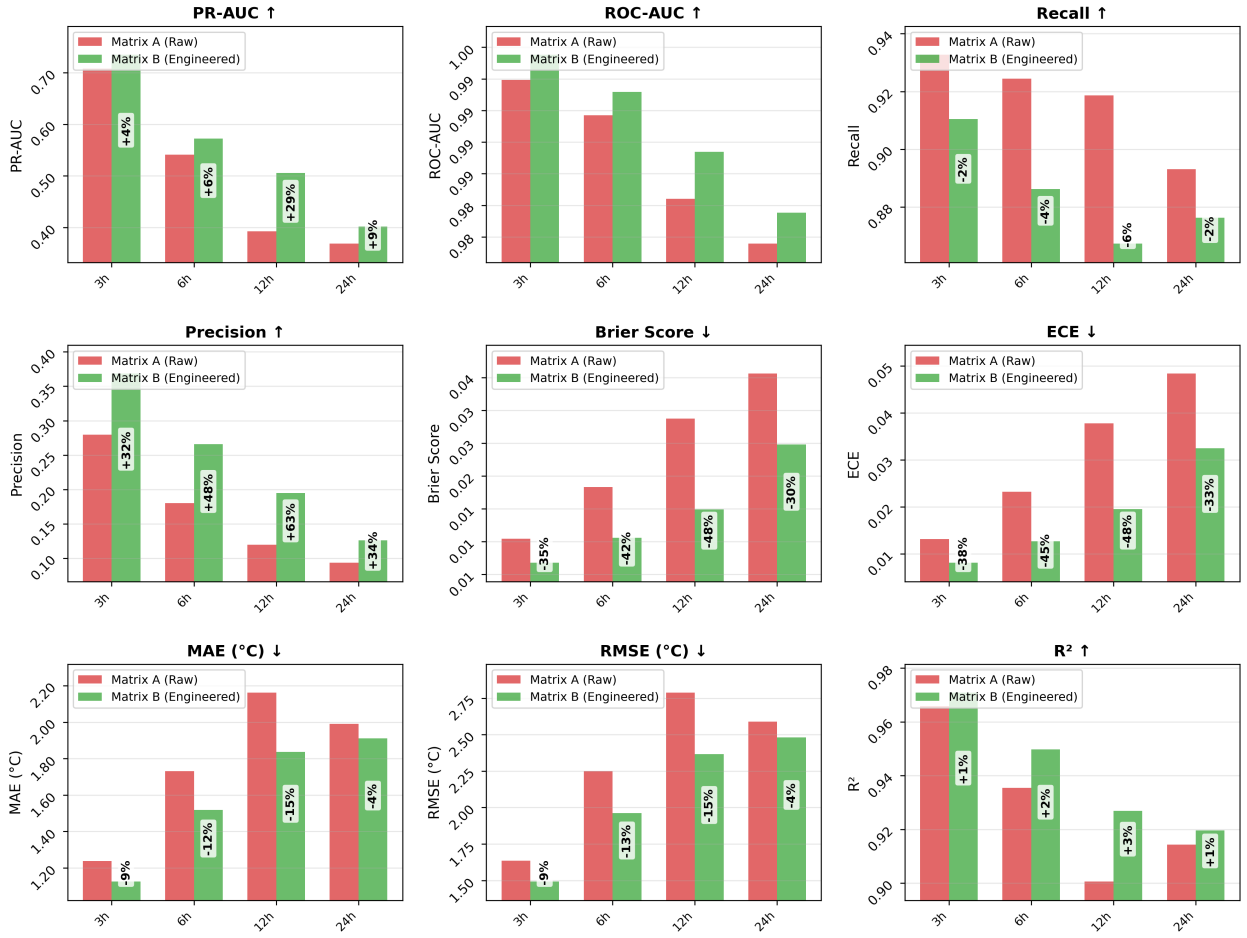

Figure 8: Matrix A (single-station raw features) vs Matrix B (single-station feature engineering) performance comparison (class-balanced training). This figure shows performance comparison across all forecast horizons (3, 6, 12, 24 hours) on 5 key metrics: (1) Classification metrics: PR-AUC, recall,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

As shown in Figure 8, performance improvements of Matrix B compared to Matrix A include:

- **PR-AUC improvement**: Improves across all forecast horizons, with improvement magnitude increasing as forecast horizon increases (3.8%–28.7%). 12-hour horizon shows most significant improvement (+28.7%), indicating that feature engineering has greater value in long-term forecasting, possibly because long-term forecasting requires more temporal dependency information, and feature engineering (particularly lag features and rolling window statistics) can effectively capture these patterns.

- **Precision improvement**: Significantly improves across all forecast horizons, with improvement magnitude of 31.7%–62.8%. 12-hour horizon shows most significant improvement (+62.8%), and significant precision improvement indicates that feature engineering helps models more accurately distinguish between frost and non-frost events, reducing false positives, further validating the value of feature engineering in long-term forecasting.

- **Recall**: Matrix B's recall is slightly lower than Matrix A (difference <5.6 percentage points), but both maintain high levels (86.7–93.3%). This slight decrease may stem from feature engineering enabling models to improve precision while slightly reducing sensitivity to minority class, but this trade-off is reasonable: significant precision improvement (reducing false positives) is more valuable than slight recall decrease, particularly in agricultural applications, where reducing false alarms can lower unnecessary protective measure costs.

- **Probability calibration improvement**: Brier Score and ECE significantly improve across all forecast horizons (lower is better), with improvement magnitude of 30.4%–48.4%. Probability calibration improvement indicates that feature engineering helps models output more accurate frost probabilities, which is crucial for decision support in practical applications.

- **Temperature prediction improvement**: MAE and RMSE improve across all forecast horizons, with improvement magnitude increasing as forecast horizon increases (4.0%–15.1%), peaking at 12-hour horizon (MAE: -15.1%, RMSE: -15.1%). R^2 improves across all forecast horizons, with improvement magnitude of 0.005–0.026, indicating that feature engineering helps models better explain temperature changes. Temperature prediction improvement is most significant at 12-hour horizon, further validating the value of feature engineering in long-term forecasting.

These improvements mainly stem from time series engineering features (lag features, rolling window statistics, derived meteorological variables, etc.) effectively capturing temporal dependency patterns in frost formation. Feature importance analysis reveals contribution mechanisms of different feature types to performance improvements.

5.4.2 Matrix B Feature Importance Analysis

To deeply understand the contribution of feature engineering to model performance, we conducted importance analysis on Matrix B's 278 features. To avoid the impact of feature quantity on category importance, we adopt a cumulative feature importance-based analysis method: first sort by individual feature importance, select feature subset reaching 90% cumulative importance, then group these features by category and calculate cumulative importance percentage for each category. This method more fairly reflects actual contributions of different feature categories, avoiding categories with many features (such as rolling window statistics with 180 features) occupying excessive importance due to quantity advantage. Figure 9 shows importance distribution of feature categories when reaching 90% cumulative feature importance threshold. Detailed analysis methods are described in Section 4.6.6. Detailed feature importance data are available in Supplementary Material S4.

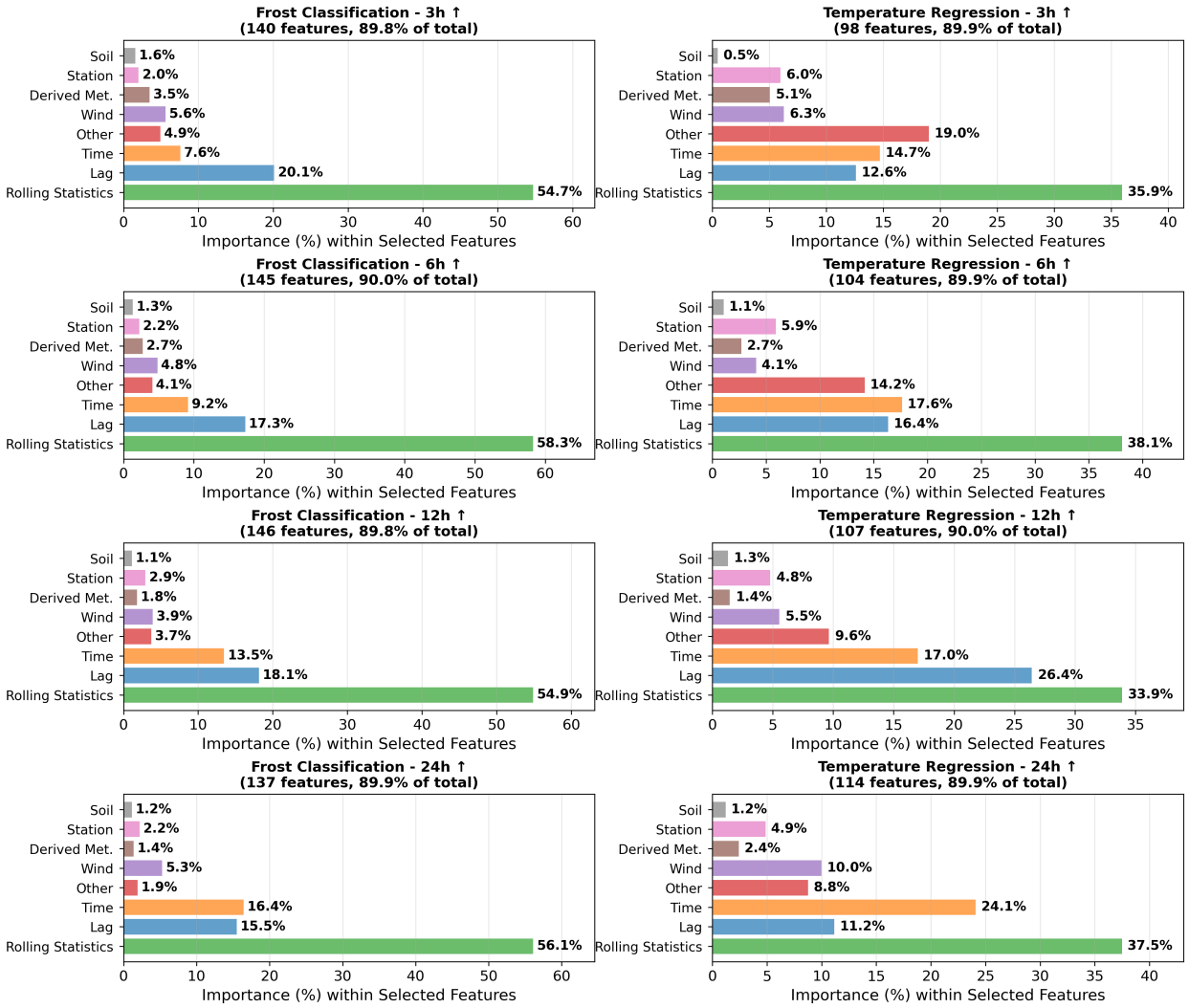

Figure 9: Matrix B (single-station feature engineering, 278 dimensions) feature category importance analysis (based on 90% cumulative feature importance).

As shown in Figure 9, feature category importance analysis reveals the following key findings:
Feature category importance patterns in frost classification tasks: Rolling window statistics features dominate across all forecast horizons (56.2–59.4%), ranking first across all horizons. This reflects the importance of temporal window statistics (mean, standard deviation, minimum, maximum) in capturing temperature trends and variability, and these statistics can effectively identify temperature decline trends before frost formation. Specifically, rolling window minimums of air temperature (rolling_24h_min, rolling_12h_min) frequently appear in top features, directly validating the key role of minimum temperature in frost warnings. Rolling window statistics feature importance peaks at 6-hour horizon (59.4%), indicating that medium-term forecast horizons have strongest dependence on temporal window statistics.
Lag feature importance is stable at 16.3–20.2%, ranking second across all forecast horizons, validating the important value of historical state information for predicting future temperature changes. Lag feature importance is highest at 3-hour horizon (20.2%) and lowest at 24-hour horizon (16.3%), and this pattern indicates that short-term forecasting relies more on recent historical states, while

long-term forecasting mainly relies on statistical patterns rather than specific historical values. Temporal feature importance significantly increases as forecast horizon increases (3 hours: 8.6%, 6 hours: 9.8%, 12 hours: 13.2%, 24 hours: 16.1%), reflecting long-term forecasting's dependence on seasonal patterns. This finding is consistent with Matrix A analysis: as forecast horizon increases, seasonal patterns become more important.

Importance of other feature categories is relatively low (<6%), but still contributes to model performance. Wind feature importance ranges from 3.5–4.4%, derived meteorological feature importance ranges from 1.4–3.3%, and station feature and soil feature importance are both <2.0%. Although these features have low importance, they provide additional physical information, helping models more accurately distinguish between frost and non-frost events.

Feature category importance patterns in temperature regression tasks: Unlike frost classification tasks, feature category importance patterns in temperature regression tasks are more diverse. Rolling window statistics feature importance is 37.8–42.4%, although still dominant, its importance is significantly lower than classification tasks (56.2–59.4%), indicating that temperature regression tasks have relatively lower dependence on temporal window statistics. Temporal feature importance increases from 14.6% (3 hours) to 23.5% (24 hours), ranking second at 24-hour horizon, reflecting strong dependence of long-term temperature prediction on seasonal patterns. This pattern is consistent with Matrix A analysis: short-term temperature changes are controlled by diurnal cycles, long-term temperature changes are controlled by seasonal patterns.

Lag feature importance ranges from 11.5–25.7%, peaking at 12-hour horizon (25.7%), ranking second, indicating that medium-term temperature prediction has strongest dependence on historical state information. This finding validates the key role of lag features in medium-term forecast horizons: 12-hour horizon needs both historical state information (lag features) and statistical patterns (rolling window statistics), and both work together to accurately predict temperature changes. Importance patterns of other feature categories are similar to classification tasks, but "other features" category importance is significantly higher in temperature regression tasks than classification tasks (8.1–15.7% vs 2.5–4.3%), possibly because temperature regression tasks require more feature types.

Physical mechanism interpretation of feature category importance: Feature category importance patterns reflect physical mechanisms of frost formation and temperature changes, and importance differences of different feature categories across different tasks and forecast horizons reveal physical laws learned by models.

(1) **Rolling window statistics features**: Dominate in frost classification tasks (56.2–59.4%), reflecting the core value of temporal window statistics in capturing temperature trends and variability. Rolling window statistics features can identify extremes (minimum, maximum) and variability (standard deviation, variance), which is particularly important for frost warnings: frost events mainly occur under extremely low temperatures, and rolling window minimums (such as rolling_24h_min, rolling_12h_min) can directly capture temperature decline trends, providing key signals for frost warnings. In temperature regression tasks, rolling window statistics feature importance is relatively lower (37.8–42.4%), but still dominant, indicating that temporal window statistics are equally important for temperature prediction, but temperature regression tasks' need for more feature types makes its importance relatively dispersed. Rolling window statistics features peak at 6-hour horizon (59.4%), indicating that medium-term forecast horizons have strongest dependence on temporal window statistics, possibly because medium-term forecasting needs to balance short-term fluctuations and long-term trends.

(2) **Lag features**: Importance is stable at 16.3–20.2% in classification tasks, and varies more in regression tasks (11.5–25.7%), reflecting importance differences of historical state information for different tasks. Lag features can capture historical evolution trends of key variables (such as tem-

perature, humidity), providing temporal context information for models. In classification tasks, lag feature importance is relatively stable, indicating that historical state information has relatively consistent contribution to frost classification; in regression tasks, lag features peak at 12-hour horizon (25.7%), indicating that medium-term temperature prediction has strongest dependence on historical state information. This pattern reveals physical mechanisms of medium-term forecasting: 12-hour horizon needs both historical state information (lag features) to capture continuity of temperature changes, and statistical patterns (rolling window statistics) to smooth short-term fluctuations, and both work together to accurately predict temperature changes. Lag feature importance is highest at 3-hour horizon (20.2%) and lowest at 24-hour horizon (16.3%), indicating that short-term forecasting relies more on recent historical states, while long-term forecasting mainly relies on statistical patterns rather than specific historical values.

(3) **Temporal features**: Temporal feature importance significantly increases as forecast horizon increases (classification tasks: 8.6%→16.1%, regression tasks: 14.6%→23.5%), reflecting long-term forecasting's dependence on seasonal patterns. This pattern is consistent in both classification and regression tasks, but temporal feature importance is higher in regression tasks (14.6–23.5% vs 8.6–16.1%), indicating that temporal patterns are more critical for temperature prediction. In regression tasks, temporal features rank second at 24-hour horizon (23.5%), second only to rolling window statistics features, validating the physical mechanism that long-term temperature prediction is mainly controlled by seasonal patterns. Temporal feature importance increase pattern is consistent with Matrix A analysis: short-term temperature changes are controlled by diurnal cycles (Hour feature importance is 22.6% at 3-hour horizon), long-term temperature changes are controlled by seasonal patterns (Julian day feature importance is 27.0% at 24-hour horizon). The importance of temporal feature engineering (such as seasonal indicators, diurnal cycle encoding) is validated.

(4) **Derived meteorological features**: Such as vapor pressure deficit, dew point difference, temperature-humidity interactions can capture physical mechanisms of frost formation (such as radiative cooling, inversion layer formation, water vapor condensation), which are difficult to directly reflect in raw observations, but can be explicitly modeled through feature engineering. Derived meteorological feature importance is relatively low (1.4–6.5%), but they provide additional physical information, helping models more accurately distinguish between frost and non-frost events, thus significantly improving precision and probability calibration quality. Derived meteorological feature importance is relatively higher in short-term forecast horizons (classification task 3 hours: 3.3%, regression task 3 hours: 6.5%), indicating that short-term forecasting has stronger dependence on physical mechanisms.

(5) **Other feature categories**: Wind features, station features, soil features and other feature categories have relatively low importance (<5%), but still contribute to model performance. Wind feature importance ranges from 3.5–8.5% (classification tasks: 3.5–4.4%, regression tasks: 3.9–8.5%), reflecting the impact of boundary layer mixing and cold air transport on frost formation; station feature importance ranges from 1.2–4.2%, reflecting the impact of station geographic location and topography on local climate; soil feature importance ranges from 0.5–1.4%, reflecting the regulatory role of soil temperature on frost formation. Although these features have low importance, they provide additional physical information, helping models more accurately distinguish between frost and non-frost events.

Summary: Performance improvements of Matrix B compared to Matrix A validate the effectiveness of feature engineering in single-station scenarios. Key findings include: (1) **Greater value in long-term forecasting**: Feature engineering shows most significant improvements at 12-hour horizon (PR-AUC: +28.7%, precision: +62.8%, MAE: -15.1%), indicating that long-term forecasting requires more temporal dependency information; (2) **Significant precision improvement**: Precision improves 31.7%–62.8% across all horizons, significantly reducing false positives, which

is crucial for resource utilization efficiency in practical applications; (3) **Probability calibration improvement**: Brier Score and ECE significantly improve across all horizons (30.4%–48.4%), enabling model output probabilities to be directly used for decision support; (4) **Temperature prediction improvement**: MAE and RMSE improve across all horizons (4.0%–15.1%), particularly most significant at 12-hour horizon, further validating the value of feature engineering in long-term forecasting. These findings provide important guidance for feature selection, indicating that feature engineering is an effective strategy for improving model performance when computational resources allow.

5.5 Spatial Aggregation Feature Analysis (Matrix C)

This section quantifies the value of spatial information for frost prediction by comparing performance differences between Matrix A (single-station raw features, 12 dimensions) and Matrix C (spatial aggregation features, 534 dimensions). Important note: All analyses in this section are based on class-balanced training results, ensuring fair comparison between different feature matrices.

Matrix C uses different optimal neighborhood radii for different forecast horizons, determined through systematic testing of 10 different radii (20–200 km, step 20 km). Specific configurations are as follows: 3-hour horizon: 60 km, 6-hour horizon: 100 km, 12-hour horizon: 200 km, 24-hour horizon: 200 km. This configuration reflects the coupling relationship between spatial and temporal scales: short-term forecasting mainly relies on local cold air pooling (small radius), while long-term forecasting requires incorporating larger-scale weather system information (large radius). Horizon dependency of optimal radius will be analyzed in detail in Section 5.5.3.

5.5.1 Matrix C: Spatial Aggregation Feature Performance Analysis

Matrix C overlays multi-station spatial aggregation statistics on Matrix A's raw variables, generating 534-dimensional features. Under class-balanced training, compared to Matrix A, Matrix C shows significant performance improvements across all forecast horizons. Figure 10 shows comprehensive comparison of matrices A and C on key metrics.

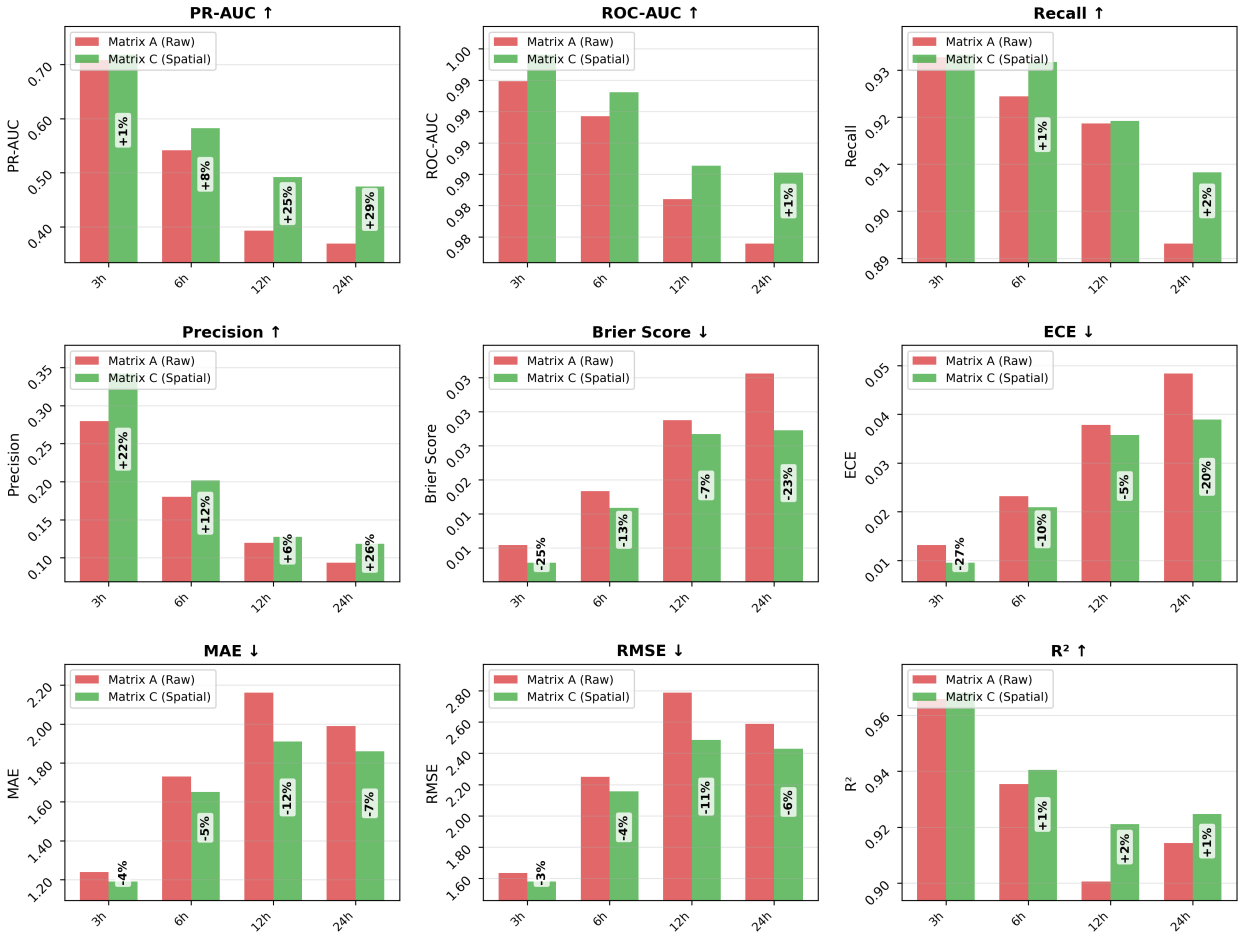

Figure 10: Matrix A (single-station raw features) vs Matrix C (spatial aggregation features) performance comparison (class-balanced training).

As shown in Figure 10, performance improvements of Matrix C compared to Matrix A include:

- **PR-AUC improvement**: Improves across all forecast horizons, with improvement magnitude increasing as forecast horizon increases (1.4%–28.6%). 24-hour horizon shows most significant improvement (+28.6%), indicating that spatial aggregation features have greater value in long-term forecasting, possibly because long-term forecasting requires larger-scale weather system information, and spatial aggregation features can effectively capture regional climate patterns and cold air pooling.
- **ROC-AUC improvement**: Slight improvements across all forecast horizons, with improvement magnitude of 0.2%–0.6%, indicating that spatial aggregation features further consolidate model discriminative ability.
- **Recall**: Matrix C's recall is basically on par with Matrix A (difference <1.6 percentage points), both maintaining high levels (89.3–93.3%), indicating that spatial aggregation features maintain high recall while improving precision.
- **Precision improvement**: Significantly improves across all forecast horizons, with improvement magnitude of 6.4%–26.3%. 24-hour horizon shows most significant improvement (+26.3%), and

significant precision improvement indicates that spatial aggregation features help models more accurately distinguish between frost and non-frost events, reducing false positives.

- **Brier Score improvement**: Significantly improves across all forecast horizons (lower is better), with improvement magnitude of 7.0%–24.9%. Brier Score improvement indicates that spatial aggregation features help models output more accurate frost probabilities.
- **ECE improvement**: Significantly improves across all forecast horizons (lower is better), with improvement magnitude of 5.5%–27.3%. ECE improvement further validates the positive impact of spatial aggregation features on probability calibration.
- **MAE improvement**: Improves across all forecast horizons, with improvement magnitude of 3.9%–11.6%, peaking at 12-hour horizon (-11.6%). MAE improvement indicates that spatial aggregation features help improve temperature prediction accuracy.
- **RMSE improvement**: Improves across all forecast horizons, with improvement magnitude of 3.4%–10.9%, peaking at 12-hour horizon (-10.9%). RMSE improvement further validates the positive impact of spatial aggregation features on temperature prediction.
- R^2 **improvement**: Slight improvements across all forecast horizons, with very small improvement magnitude (0.002–0.020), indicating that spatial aggregation features have limited improvement on temperature prediction explanatory ability, possibly because Matrix A's raw features already explain main patterns of temperature changes well.

These improvements mainly stem from spatial aggregation features effectively capturing regional climate patterns (such as cold air pooling, terrain effects, temperature gradients, etc.), which are difficult to directly reflect in single-station observations. Specifically, spatial aggregation features improve model performance through the following mechanisms: (1) Regional climate pattern capture: Neighborhood aggregation statistics (mean, gradient, range, etc.) can effectively capture regional climate patterns such as cold air pooling, terrain effects, and temperature gradients, which are difficult to directly reflect in single-station observations, but have important impacts on frost formation. For example, when multiple neighboring stations simultaneously show low temperatures, it indicates regional cold air pooling, significantly increasing frost risk; (2) Spatial smoothing effect: Spatial aggregation features smooth local noise and outliers in single-station observations through neighborhood statistics, providing more stable and reliable prediction signals, particularly under inconsistent data quality; (3) Scale matching for long-term forecasting: Long-term forecast horizons (12–24 hours) require larger-scale weather system information, and spatial aggregation features can provide this information, enabling models to capture regional-scale temperature change trends and cold air transport patterns; (4) Information compensation for missing data: When single-station data is missing, spatial aggregation features can provide compensation through neighboring station information, improving model robustness and reliability. These mechanisms work together, enabling Matrix C to show significant performance improvements across all forecast horizons, particularly most significant in long-term forecast horizons (12–24 hours). Feature importance analysis will further reveal contribution mechanisms of different feature types to performance improvements.

5.5.2 Matrix C Feature Importance Analysis

To deeply understand the contribution of spatial aggregation features to model performance, we conducted importance analysis on Matrix C's 534 features. To avoid the impact of feature quantity on category importance, we adopt the same cumulative feature importance-based analysis method

as Matrix B: first sort by individual feature importance, select feature subset reaching 90% cumulative importance, then group these features by category and calculate cumulative importance percentage for each category. This method more fairly reflects actual contributions of different feature categories, avoiding categories with many features from occupying excessive importance due to quantity advantage. Feature classification rules are shown in Table **??**. Figure 11 shows importance distribution of feature categories when reaching 90% cumulative feature importance threshold. Detailed analysis methods are described in Section 4.6.6. Detailed feature importance data are available in Supplementary Material S5.

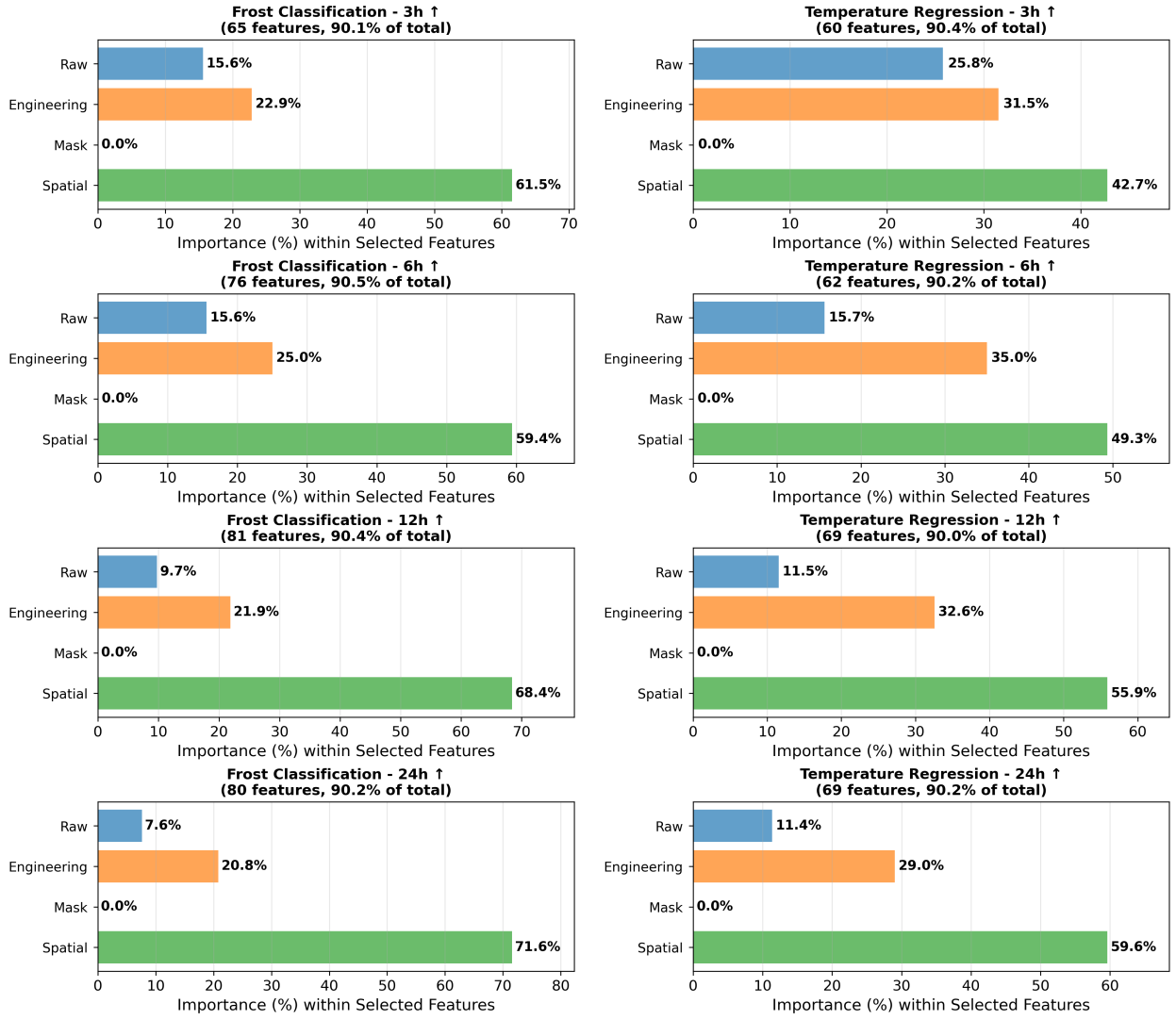

Figure 11: Matrix C (spatial aggregation features, 534 dimensions) feature category importance analysis.

As shown in Figure 11, feature category importance analysis reveals the following key findings: **Feature category importance patterns in frost classification tasks**: Spatial aggregation features dominate across all forecast horizons (59.4–71.6%), ranking first across all horizons, significantly higher than other feature categories. Spatial aggregation feature importance significantly increases as forecast horizon increases (3 hours: 61.5%, 6 hours: 59.4%, 12 hours: 68.4%, 24 hours:

71.6%), indicating that long-term forecasting has stronger dependence on spatial information, consistent with the finding that optimal radius increases with forecast horizon. Most important spatial aggregation features include soil temperature neighborhood statistics (minimum, gradient, standard deviation, range), air temperature neighborhood gradient, relative humidity neighborhood gradient, etc., and these features can effectively capture regional climate patterns such as regional cold air pooling, terrain effects, and temperature gradients.

Engineering features (combination of temporal features and derived meteorological features) importance ranges from 20.8–25.0%, ranking second across all horizons. Engineering feature importance is highest at 6-hour horizon (25.0%) and lowest at 24-hour horizon (20.8%), but overall remains relatively stable. Most important engineering features include Hour feature (3.7–10.2%), day-of-year cycle encoding (day_of_year_sin/cos, 5.4–7.0%), heat index (heat_index, 0.9–3.7%), and wind chill (wind_chill, 2.4–3.2%), etc. These features reflect important contributions of temporal patterns and derived meteorological features to frost prediction, particularly in short-term forecasting, where Hour feature importance peaks at 6-hour horizon (10.2%), reflecting the key role of diurnal cycle patterns in short-term frost formation.

Raw feature importance ranges from 7.6–15.6%, ranking third across all horizons. Raw feature importance significantly decreases as forecast horizon increases (3 hours: 15.6%, 24 hours: 7.6%), indicating that long-term forecasting mainly relies on spatial aggregation information rather than single-station raw observations. Most important raw features include air temperature (Air Temp, 1.4–5.1%), dew point temperature (Dew Point, 1.4–3.2%), soil temperature (Soil Temp, 0.7–2.2%), and wind direction (Wind Dir, 1.1–1.7%), etc. These features are more important in short-term forecasting, because short-term forecasting needs direct information about current state, while long-term forecasting mainly relies on spatial aggregation statistical patterns.

Mask features (missing masks) have 0% importance across all forecast horizons, not selected within 90% cumulative importance threshold. Although mask features have the largest quantity (294, accounting for 55.1% of Matrix C's total feature count), LightGBM did not use these features during training (their importance is 0), indicating that under current configuration, mask features did not provide additional information gain. This may be because: (1) Mask features are highly correlated with other features, and models indirectly obtained missing information through other features; (2) Mask features may be constant or have very small variance in training data, unable to provide effective split points; (3) Spatial aggregation statistics themselves already contain data quality information, making explicit mask features redundant. This finding indicates that although mask features are designed in feature engineering to indicate data quality, they are not directly used in actual model training, and models mainly rely on spatial aggregation statistics themselves to capture regional climate patterns.

Feature category importance patterns in temperature regression tasks: Unlike frost classification tasks, engineering features and spatial aggregation features have relatively balanced importance in temperature regression tasks. Spatial aggregation feature importance ranges from 42.7–59.6%, with importance increasing as forecast horizon increases (3 hours: 42.7%, 24 hours: 59.6%), reflecting that long-term temperature prediction has stronger dependence on spatial information. Engineering feature importance ranges from 29.0–35.0%, ranking first or second across all horizons, reflecting the key role of temporal patterns and derived meteorological features in temperature prediction. Engineering feature importance is highest at 6-hour horizon (35.0%) and lowest at 24-hour horizon (29.0%), indicating that medium-term temperature prediction has strongest dependence on temporal patterns and physical mechanisms. Raw feature importance ranges from 11.4–25.8%, highest at 3-hour horizon (25.8%) and lowest at 24-hour horizon (11.4%), indicating that short-term temperature prediction relies more on single-station raw observations, while long-term prediction mainly relies on spatial aggregation information.

Physical mechanism interpretation of feature category importance: Feature category importance patterns reflect physical mechanisms of how spatial information affects frost prediction and temperature prediction, and importance differences of different feature categories across different tasks and forecast horizons reveal spatial-temporal coupling laws learned by models.
(1) **Spatial aggregation features**: Dominate in frost classification tasks (59.4–71.6%), reflecting the core value of spatial aggregation statistics (neighborhood mean, gradient, range, minimum, standard deviation, etc.) in capturing regional climate patterns. Spatial aggregation features can effectively capture regional climate patterns such as cold air pooling, terrain effects, and temperature gradients, which are difficult to directly reflect in single-station observations. For example, soil temperature neighborhood minimum (Soil Temp_neighbor_min) ranks in top five across all forecast horizons, with importance of 1.7–3.0%, directly reflecting the core role of regional soil temperature in frost formation: when multiple neighboring stations simultaneously have low soil temperatures, it indicates regional cold air pooling, significantly increasing frost risk. Air temperature neighborhood gradient (Air Temp_neighbor_gradient) importance is 1.7–2.6%, reflecting the impact of temperature gradients on frost formation: large temperature gradients usually indicate active cold air transport, potentially increasing frost risk. Spatial aggregation feature importance significantly increases as forecast horizon increases, indicating that long-term forecasting requires larger-scale weather system information, consistent with the finding that optimal radius increases with forecast horizon.
(2) **Engineering features**: Engineering features (combination of temporal features and derived meteorological features) have importance of 20.8–25.0% in classification tasks and 29.0–35.0% in regression tasks, reflecting importance differences of temporal patterns and physical mechanisms for different tasks. In classification tasks, Hour feature has highest importance (3.7–10.2%), peaking at 6-hour horizon (10.2%), reflecting the key role of diurnal cycle patterns in short-term frost formation: frost events mainly occur at night and early morning (usually 22:00–06:00), when radiative cooling is strongest and temperature is lowest. Day-of-year cycle encoding (day_of_year_sin/cos) importance is 5.4–7.0%, reflecting the importance of seasonal patterns for long-term forecasting. Derived meteorological features (such as heat index, wind chill) importance is 0.9–3.7%, reflecting the impact of temperature-humidity interactions and wind-temperature interactions on frost formation. In regression tasks, engineering feature importance is higher, indicating that temporal patterns and derived meteorological features are more critical for temperature prediction. This pattern is consistent with Matrix B analysis: temporal feature engineering (such as seasonal indicators, diurnal cycle encoding) and derived meteorological feature engineering (such as temperature-humidity interactions, vapor pressure deficit) have important contributions to model performance.
(3) **Raw features**: Have importance of 7.6–15.6% in classification tasks and 11.4–25.8% in regression tasks, reflecting importance differences of single-station raw observations for different tasks. Raw feature importance significantly decreases as forecast horizon increases, indicating that long-term forecasting mainly relies on spatial aggregation information rather than single-station raw observations. In classification tasks, air temperature (Air Temp) has highest importance (1.4–5.1%), peaking at short-term forecast horizon (3 hours) (5.1%), directly reflecting the core role of temperature in frost formation. Dew point temperature (Dew Point) importance is 1.4–3.2%, reflecting the impact of humidity on frost formation: high humidity conditions are usually accompanied by higher dew point temperatures, and when air temperature approaches dew point temperature, water vapor condensation releases latent heat, potentially slowing temperature decline, thus affecting frost formation. Soil temperature (Soil Temp) importance is 0.7–2.2%, reflecting the regulatory role of soil temperature on frost formation. This pattern reveals the key role of spatial information in long-term forecasting: long-term forecasting requires incorporating larger-scale weather system information, and spatial aggregation features can provide this information.

(4) **Mask features**: Mask features have 0% importance across all forecast horizons, not selected within 90% cumulative importance threshold. Although mask features have the largest quantity (294, accounting for 55.1% of Matrix C's total feature count), LightGBM did not use these features during training, indicating that under current configuration, mask features did not provide additional information gain. This finding indicates that although mask features are designed in feature engineering to indicate data quality and spatial aggregation reliability, they are not directly used in actual model training, and models mainly rely on spatial aggregation statistics themselves to capture regional climate patterns, without needing explicit missing mask indicators. This may be because spatial aggregation statistics (such as neighborhood mean, gradient, etc.) themselves already contain data quality information, or mask features are highly correlated with other features, and models indirectly obtained missing information through other features.

5.5.3 Optimal Radius Horizon Dependency

Matrix C's optimal radius varies with forecast horizon (3 hours: 60 km, 6 hours: 100 km, 12 hours: 200 km, 24 hours: 200 km), revealing the coupling relationship between spatial and temporal scales. Short-term forecasting mainly relies on local cold air pooling (small radius), while long-term forecasting requires incorporating larger-scale weather system information (large radius). This finding provides important guidance for radius selection in practical applications, and also validates the finding that spatial aggregation feature importance increases with forecast horizon.

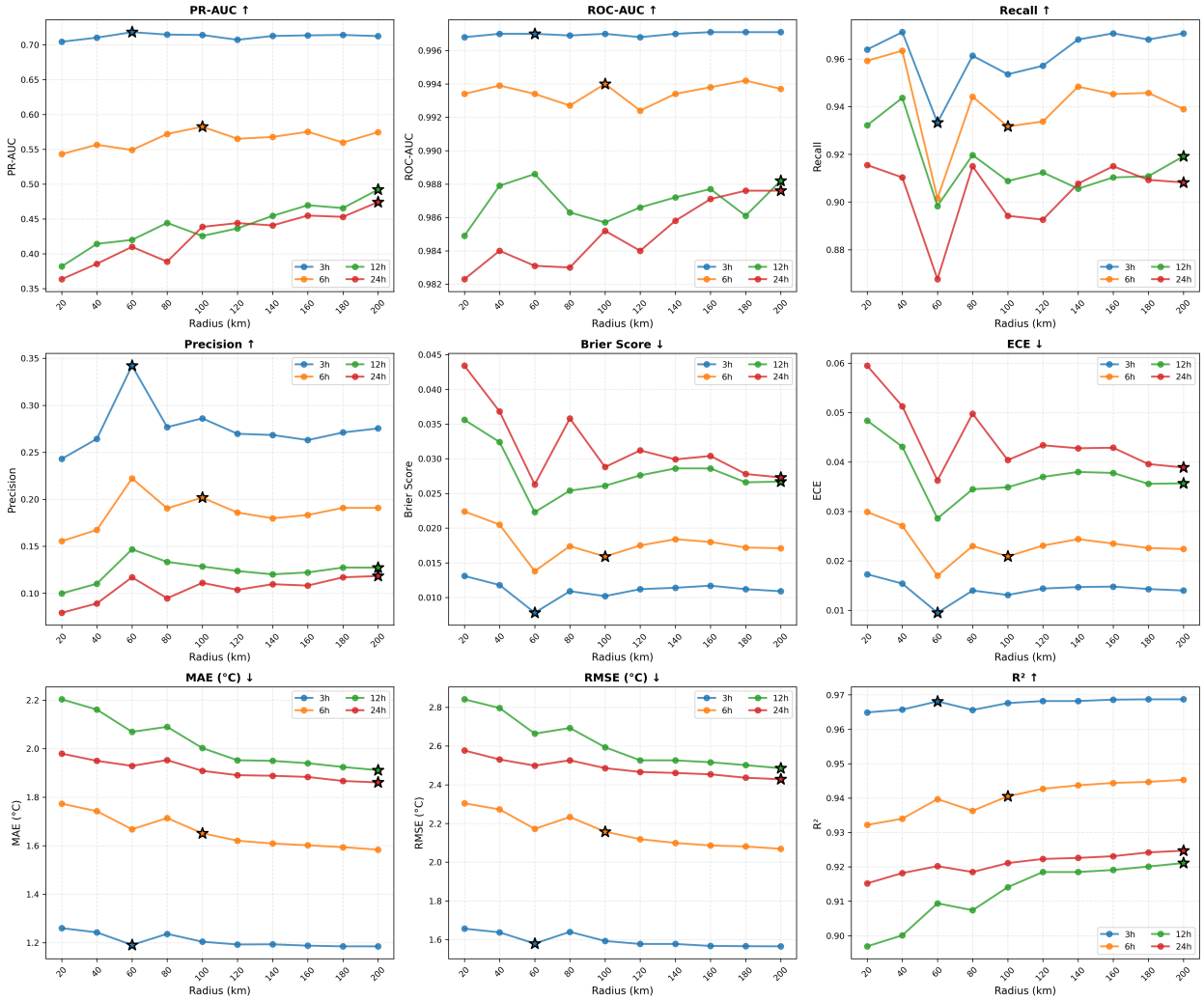

Figure 12: Matrix C: Variation trends of 9 evaluation metrics with radius across different forecast horizons. This figure shows variation patterns of all evaluation metrics (PR-AUC, ROC-AUC, recall, precision, Brier Score, ECE, MAE, RMSE, R^2) with radius within 20–200 km radius range. Different colored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

As shown in Figure 12, variation patterns of 9 evaluation metrics with radius reveal the systematic impact of spatial scale on model performance. Classification metrics (PR-AUC, ROC-AUC, recall, precision) perform better with small radii (20–60 km) at 3-hour horizon, with optimal radius of 60 km, reflecting the key role of local cold air pooling in short-term frost formation; at 6-hour horizon, optimal radius increases to 100 km, indicating that medium-term forecasting requires larger spatial information range; at long-term forecast horizons (12–24 hours), large radii (100–200 km) perform better, with optimal radius reaching 200 km, reflecting that long-term forecasting requires larger-

scale weather system information. Probability calibration metrics (Brier Score, ECE) show similar variation patterns with radius as classification metrics, achieving best calibration effects at optimal radii. Regression metrics (MAE, RMSE, R^2) also show horizon dependency: short-term forecasting performs better with small radii, long-term forecasting performs better with large radii. At short-term forecast horizons (3–6 hours), regression metrics have higher sensitivity to radius (MAE variation range 6.21–11.44%) than classification metrics (PR-AUC variation range 1.94–6.96%); at long-term forecast horizons (12–24 hours), classification metrics have significantly higher sensitivity to radius (PR-AUC variation range 25.04–26.02%) than regression metrics (MAE variation range 6.21–14.46%), indicating that classification tasks have stronger dependence on spatial scale in long-term forecasting.

Quantitative analysis of optimal radius: Systematic testing of 10 radius values (20–200 km, step 20 km) reveals significant variation of optimal radius with forecast horizon. At 3-hour horizon, optimal radius is 60 km, with PR-AUC improvement of 1.96% compared to minimum radius (20 km), indicating low sensitivity of short-term forecasting to radius (coefficient of variation 0.56%). At 6-hour horizon, optimal radius increases to 100 km, with PR-AUC improvement increasing to 7.23%, and sensitivity significantly increases (coefficient of variation 2.20%). At 12–24 hour horizons, optimal radius reaches 200 km, with PR-AUC improvement dramatically increasing to 28.88–30.45%, and sensitivity reaches highest level (coefficient of variation 7.24–8.50%). This pattern indicates: (1) Short-term forecasting (3–6 hours) has low sensitivity to radius, with relatively small optimal radii (60–100 km), mainly capturing local cold air pooling and terrain effects; (2) Long-term forecasting (12–24 hours) has significantly increased sensitivity to radius, with optimal radius reaching maximum (200 km), requiring incorporation of larger-scale weather system evolution and regional climate patterns; (3) Importance of radius selection significantly increases as forecast horizon increases, and in long-term forecasting, performance improvement from selecting appropriate radius (approximately 30%) is far higher than in short-term forecasting (approximately 2%). This finding validates the coupling relationship between spatial and temporal scales, providing quantitative guidance for radius configuration in practical applications.

Summary and Conclusions: Matrix C's spatial aggregation feature analysis reveals systematic contributions of spatial information to frost prediction, with main findings including:

(1) Coupling relationship between spatial and temporal scales: Optimal radius significantly varies with forecast horizon (3 hours: 60 km, 6 hours: 100 km, 12–24 hours: 200 km), revealing the intrinsic coupling mechanism between spatial and temporal scales. Short-term forecasting (3–6 hours) mainly relies on local cold air pooling and terrain effects, with smaller optimal radii (60–100 km) and lower radius sensitivity (coefficient of variation 0.56–2.20%). Long-term forecasting (12–24 hours) requires incorporation of larger-scale weather system evolution and regional climate patterns, with optimal radius reaching maximum (200 km) and radius sensitivity significantly increased (coefficient of variation 7.24–8.50%). This finding validates the coupling relationship between spatial and temporal scales from a quantitative perspective, providing scientific basis for radius configuration across different forecast horizons.

(2) Performance gains from spatial aggregation features: Compared to minimum radius (20 km), optimal radius brings PR-AUC improvements of 28.88–30.45% at long-term forecast horizons (12–24 hours), significantly higher than short-term forecast horizons (3 hours: 1.96%, 6 hours: 7.23%). This pattern indicates that spatial aggregation features have greater value in long-term forecasting, validating strong dependence of long-term forecasting on spatial information. Spatial aggregation features have importance of 59.4–71.6% in classification tasks, ranking first across all horizons, and importance significantly increases as forecast horizon increases (from 61.5% at 3 hours to 71.6% at 24 hours), further validating the core value of spatial information for frost prediction.

(3) Horizon dependency of feature categories: Engineering features have importance of

20.8–25.0%, ranking second across all horizons, reflecting important contributions of temporal patterns and derived meteorological features. Raw feature importance significantly decreases as forecast horizon increases (from 15.6% at 3 hours to 7.6% at 24 hours), indicating that long-term forecasting mainly relies on spatial aggregation information rather than single-station raw observations. Mask features, although having the largest quantity (294, accounting for 55.1%), are not selected within 90% cumulative importance threshold (importance is 0%), indicating that under current configuration, mask features did not provide additional information gain, and models mainly rely on spatial aggregation statistics themselves.

(4) Practical application guidance: This study provides quantitative guidance for radius configuration in practical applications. For short-term forecasting (3–6 hours), it is recommended to use smaller radii (60–100 km), focusing on local cold air pooling and terrain effects; for long-term forecasting (12–24 hours), it is recommended to use larger radii (200 km), incorporating larger-scale weather system information. Importance of radius selection significantly increases as forecast horizon increases, and in long-term forecasting, performance improvement from selecting appropriate radius (approximately 30%) is far higher than in short-term forecasting (approximately 2%). These findings indicate that spatial aggregation features are an effective strategy for improving frost prediction model performance, particularly in long-term forecasting scenarios.

5.6 LOSO Spatial Generalization Evaluation

To comprehensively evaluate model spatial generalization capability, we conducted LOSO (Leave-One-Station-Out) evaluation on optimal configurations of matrices A, B, and C respectively. LOSO evaluation is a strict standard for evaluating model spatial generalization capability, revealing whether models overfit to local patterns of specific stations, or can learn regional climate patterns generalizable to new stations. Figure 13 shows detailed performance distribution of matrices A and B under LOSO evaluation and comparison with regular temporal split evaluation. Overall, LOSO evaluation for matrices A and B has been fully completed (18 stations $\times$ 4 forecast horizons), and both matrices show excellent spatial generalization capability under LOSO evaluation, with no significant decline in all key metrics under LOSO conditions, and some metrics even showing slight improvements. Performance differences between different stations are small, indicating that features learned by models have good spatial consistency. Detailed LOSO evaluation results are available in Supplementary Material S6.

Matrix C LOSO Evaluation Note: Due to Matrix C's high feature dimensionality (534 dimensions) and large data volume (approximately 2.36 million records), LOSO evaluation exceeded current platform memory limits, and therefore is not included in this study. LOSO evaluation for matrices A and B has been fully completed (18 stations $\times$ 4 forecast horizons), with detailed results available in supplementary materials.

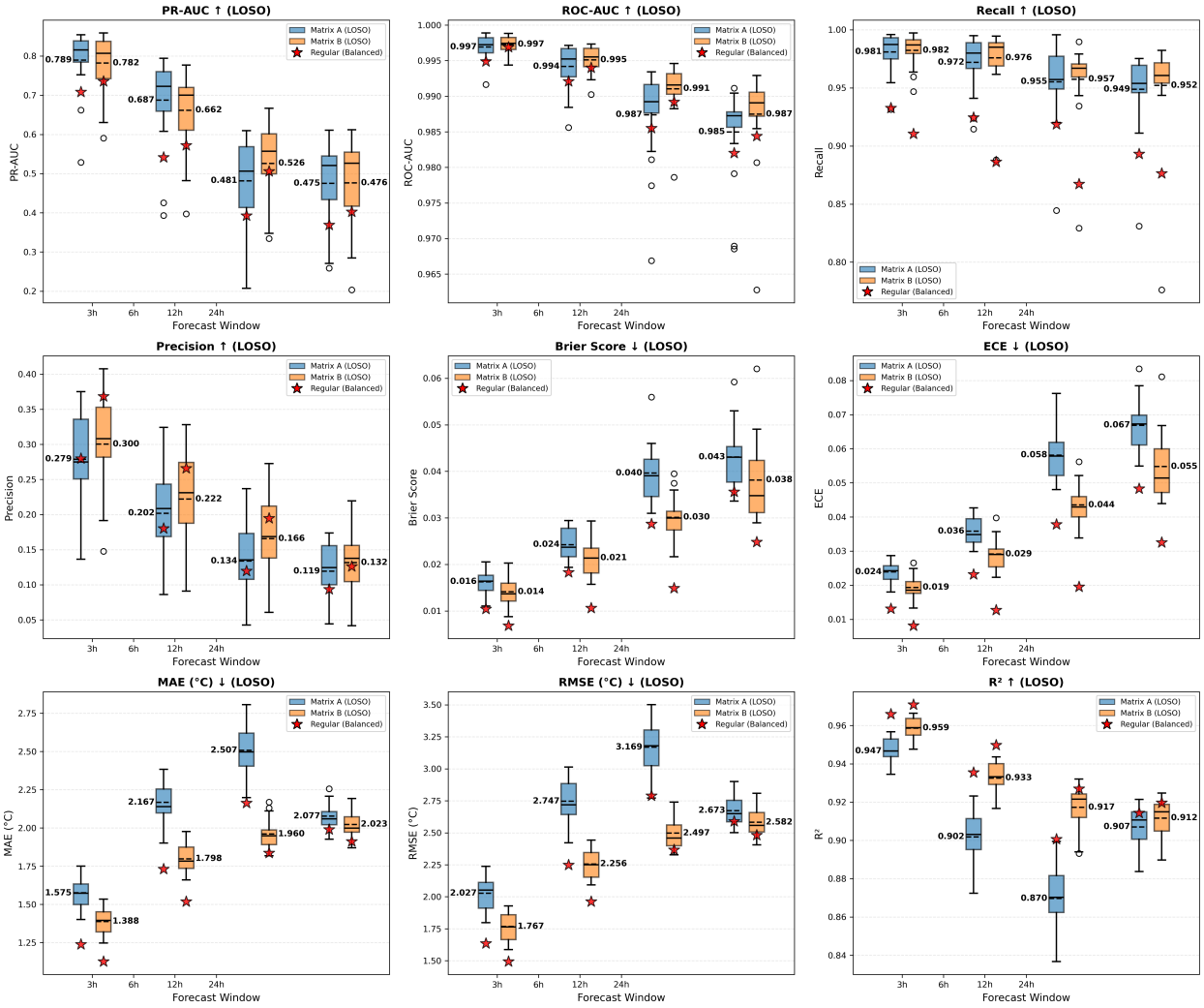

Figure 13: LOSO evaluation boxplot comparison for matrices A and B. This figure shows performance distribution of Matrix A (single-station raw features) and Matrix B (single-station feature engineering) under LOSO evaluation, including 9 key metrics: (1) Classification metrics: PR-AUC, ROC-AUC, recall,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

Matrix A LOSO Performance: Matrix A (single-station raw features, 12 dimensions) shows stable spatial generalization capability under LOSO evaluation. Compared to regular temporal split evaluation, PR-AUC for all forecast horizons remains stable under LOSO conditions (0.47–0.79), recall remains at high levels across all forecast horizons (0.95–0.98), with small standard deviation, indicating very small performance differences between different stations. Precision is relatively low but stable (0.12–0.28), which is typical for class-imbalanced tasks, with models prioritizing high recall to reduce miss risk. Probability calibration metrics (Brier Score and ECE) perform well under

LOSO evaluation, and temperature prediction also remains stable (MAE: 1.57–2.51 °C, R^2: 0.87–0.95). This finding indicates: (1) Generalization stability of raw features: Even with the simplest feature configuration, models can maintain high performance on unseen stations; (2) Consistency of spatial patterns: All 18 CIMIS stations are located in California Central Valley, and frost formation processes are highly consistent within this region; (3) Increased training data volume: In LOSO evaluation, training sets contain complete time series from 17/18 stations, providing more diverse samples to learn robust patterns.

Matrix B LOSO Performance: Matrix B (single-station feature engineering, 278 dimensions) also shows excellent spatial generalization capability under LOSO evaluation. Compared to regular temporal split evaluation, PR-AUC for all forecast horizons remains stable under LOSO conditions (0.48–0.78), recall remains at high levels across all forecast horizons (0.95–0.98), with small standard deviation, indicating very small performance differences between different stations. Compared to Matrix A, Matrix B's PR-AUC is slightly higher in long-term horizons (12–24 hours), with comparable recall. Precision is slightly higher than Matrix A (0.13–0.30), indicating that feature engineering helps improve precision while maintaining high recall. Probability calibration metrics perform excellently under LOSO evaluation, all lower than corresponding values for Matrix A, indicating that models after feature engineering have better probability calibration performance. Temperature prediction performs excellently under LOSO evaluation (MAE: 1.39–2.02 °C, R^2: 0.91–0.96), all superior to corresponding values for Matrix A, indicating that feature engineering significantly improves temperature prediction accuracy. This finding indicates: (1) Generalization impact of feature engineering: Time series feature engineering shows good stability when transferred between stations, as these features capture temporal dependency patterns rather than station-specific local patterns; (2) Value of feature engineering: Although feature engineering increases feature dimensionality, it does not compromise model spatial generalization capability, instead showing improvements across all metrics; (3) Spatial consistency of temporal patterns: Time series features have similarity across different stations, enabling these features to effectively generalize in LOSO evaluation.

AB Matrix Spatial Generalization Capability Comparison: By comparing LOSO evaluation results of matrices A and B, we can gain deep understanding of how different feature engineering strategies affect model spatial generalization capability. Comparative analysis shows: (1) Feature engineering does not compromise spatial generalization capability: Although Matrix B increases feature dimensionality from 12 to 278 dimensions, it still shows excellent spatial generalization capability under LOSO evaluation, with small standard deviations for PR-AUC and recall, comparable to Matrix A, indicating that time series feature engineering captures temporal dependency patterns rather than station-specific local patterns; (2) Feature engineering brings performance improvements: Matrix B outperforms Matrix A in PR-AUC (long-term horizons), precision, temperature prediction accuracy (MAE and R^2), and probability calibration performance, while maintaining high recall (0.95–0.98), validating the effectiveness of feature engineering in improving performance while maintaining spatial generalization capability; (3) Core value of raw features: Matrix A as baseline, its LOSO performance validates the core value of raw CIMIS variables in spatial generalization, with PR-AUC and recall remaining stable under LOSO evaluation, providing feasible solutions for resource-constrained scenarios. This comparative analysis provides important guidance for feature selection in actual deployment: single-station feature engineering strategies have good spatial generalization capability, and appropriate strategies can be selected based on computational resources and performance requirements.

Temperature Prediction Generalization Stability: Temperature predictions for matrices A and B both remain stable under LOSO evaluation, with slight decreases compared to regular evaluation but still maintaining excellent performance (Matrix A: MAE 1.57–2.51 °C, R^2 0.87–

0.95; Matrix B: MAE 1.39–2.02 °C, R^2 0.91–0.96). Matrix B's temperature prediction performance under LOSO evaluation is superior to Matrix A, with both MAE and R^2 being better, validating the effectiveness of feature engineering in spatial generalization scenarios. Temperature prediction mainly relies on physical processes (such as radiative cooling, cold air transport), which have similarity across different stations, enabling temperature prediction patterns learned by models to effectively transfer spatially, providing important assurance for actual deployment.

6 Discussion

This study deeply explores key issues in frost risk prediction through systematic feature engineering framework and rigorous experimental design. This section provides in-depth discussion from perspectives of methodological contributions, physical mechanism understanding, practical application significance, and limitations.

6.1 Systematic Evaluation of Feature Engineering Strategies

The ABC feature configuration matrix framework proposed in this study provides a reproducible methodology for systematic evaluation of feature engineering strategies. By progressively increasing feature complexity (from 12-dimensional raw features to 278-dimensional engineered features, to 534-dimensional spatial aggregation features), we quantitatively evaluate independent and joint contributions of different feature engineering strategies. The core value of this framework lies in its interpretability and extensibility: each matrix represents a clear feature engineering strategy, enabling performance improvements to be traced to specific feature types, rather than black-box feature combinations.

Value of temporal feature engineering: Single-station feature engineering outperforms raw features across all forecast horizons, validating the core value of time series feature engineering in frost prediction. Feature importance analysis reveals the key role of lag features, rolling window statistics, and derived meteorological variables in capturing temporal dependency patterns in frost formation. This finding is highly consistent with physical mechanisms of frost formation: radiation frost formation is a gradual process involving temporal patterns such as nocturnal radiative cooling, temperature decline trends, and dew point approach. Models capture historical states through lag features (such as air temperature and soil temperature from past 3–24 hours), capture trends and variability through rolling window statistics (such as mean and standard deviation from past 6–24 hours), and capture quantitative indicators of physical processes through derived meteorological variables (such as temperature decline rate, dew point difference). These features maintain excellent performance under LOSO evaluation, indicating they capture temporal dependency patterns rather than station-specific local patterns, with good spatial generalization capability.

Horizon dependency of spatial aggregation features: Performance patterns of spatial aggregation features reveal the differentiated role of spatial information across different forecast horizons. In short-term forecasting (3 hours), temporal feature engineering dominates, while spatial aggregation features are slightly lower. This pattern reflects that short-term frost prediction mainly relies on local temporal patterns: temperature decline trends at current station, historical states, and current meteorological conditions. In long-term forecasting (24 hours), spatial aggregation features achieve highest performance, while temporal feature engineering is relatively lower, indicating that long-term forecasting requires incorporating regional weather system information: cold air transport, spatial gradients, neighborhood temperature distribution, etc. Optimal radius varying with forecast horizon (3 hours: 60 km, 6 hours: 100 km, 12–24 hours: 200 km) further validates the coupling relationship between spatial and temporal scales: short-term forecasting relies on local

neighborhoods (60 km), long-term forecasting requires larger-scale regional information (200 km). This finding provides important guidance for feature selection in actual deployment: selecting appropriate feature matrices and spatial aggregation radii based on forecast horizons can achieve optimal balance between performance and computational costs.

6.2 Critical Role of Class-Balanced Training in Extremely Imbalanced Tasks

This study comprehensively analyzes the impact of class-balanced training on model performance, revealing its critical role in extremely imbalanced classification tasks. The core value of class-balanced training lies in significantly improving recall (from 38.7–67.4% to 86.7–93.3%), reducing false negatives by 75–90%, directly addressing the key requirement of minimizing missed frost events in agricultural applications. The cost of this improvement is decreased precision (from 29.3–66.3% to 9.4–36.8%) and slightly decreased probability calibration quality (increased Brier Score and ECE), but this trade-off is completely reasonable in agricultural applications: the cost of missed frost events (crop losses) is far higher than the cost of false alarms (unnecessary protective measures).

Practical significance of probability calibration: This study systematically evaluates probability calibration quality of model outputs through Brier Score and ECE metrics. All optimal configurations have Brier Score <0.036 and ECE <0.049, indicating that model output probabilities have good calibration quality and can be directly used for decision support. This finding has important significance: calibrated probabilities can be directly mapped to decision thresholds in farm standard operating procedures (SOP), such as "activate protective measures when frost probability >0.3", rather than serving only as relative ranking indicators. However, we also note that class-balanced training leads to slightly decreased probability calibration quality, reflecting the trade-off between recall and calibration quality in class-imbalanced tasks. In practical applications, growers can adjust decision thresholds based on economic value of specific crops and protection costs, balancing risks of missed events and false alarms.

6.3 Validation and Insights of Spatial Generalization Capability

LOSO evaluation is a strict standard for evaluating model spatial generalization capability, and this study systematically validates model spatial generalization capability under different feature engineering strategies through LOSO evaluation. LOSO evaluation results for matrices A and B show that all key metrics show no significant decline under LOSO conditions, with some metrics even showing slight improvements, validating that features learned by models have good spatial consistency. This finding has important significance: (1) Generalization stability of raw features: Even with the simplest feature configuration (12-dimensional raw variables), models can maintain high performance on unseen stations, validating the inherent predictability of frost prediction tasks and the core value of raw CIMIS variables; (2) Feature engineering does not compromise spatial generalization capability: Although single-station feature engineering increases feature dimensionality from 12 to 278 dimensions, it still shows excellent spatial generalization capability under LOSO evaluation, indicating that time series feature engineering captures temporal dependency patterns rather than station-specific local patterns; (3) Consistency of spatial patterns: All 18 CIMIS stations are located in California Central Valley, and frost formation processes are highly consistent within this region, enabling features learned by models to transfer between stations.

Temperature prediction also remains stable under LOSO evaluation, with slight decreases compared to regular evaluation but still maintaining excellent performance (Matrix A: MAE 1.57–2.51 °C, R^2 0.87–0.95; Matrix B: MAE 1.39–2.02 °C, R^2 0.91–0.96). This finding indicates that temperature

prediction mainly relies on physical processes (such as radiative cooling, cold air transport), which have similarity across different stations, enabling temperature prediction patterns learned by models to effectively transfer spatially. Matrix B's temperature prediction performance under LOSO evaluation is superior to Matrix A, validating the effectiveness of feature engineering in spatial generalization scenarios.

6.4 Physical Mechanism Insights from Feature Importance Analysis

This study deeply explores the most valuable signals for frost prediction through feature importance analysis. Analysis reveals the key role of derived meteorological variables such as soil temperature gradients, dew point differences, and vapor pressure deficits in frost prediction, and these findings are highly consistent with physical mechanisms of frost formation. Soil temperature gradients reflect surface energy balance and radiative cooling processes, dew point differences reflect atmospheric humidity state and latent heat release from condensation, and vapor pressure deficits reflect water vapor transport and latent heat exchange. These derived meteorological variables provide models with more direct prediction signals than raw observation variables by quantifying physical processes.
Feature importance analysis also reveals the differentiated role of different feature categories across different forecast horizons. In short-term forecasting (3 hours), temporal features (such as hour, temperature decline trends) dominate; in long-term forecasting (24 hours), seasonal features (such as Julian day) and spatial aggregation features (such as neighborhood temperature gradients) are more important. This pattern reflects physical mechanisms of frost prediction: short-term forecasting mainly relies on current state and recent trends, while long-term forecasting requires incorporating seasonal patterns and large-scale weather system information.

6.5 Methodological Contributions and Limitations

Methodological contributions of this study are mainly reflected in the following aspects: (1) Systematic feature engineering framework: ABC matrix framework provides a reproducible methodology for systematic evaluation of feature engineering strategies, enabling performance improvements to be traced to specific feature types; (2) Comprehensive class-balanced training analysis: Systematically evaluates the impact of class-balanced training on model performance, particularly improvements in recall and false negative rates, providing practical guidance for extremely imbalanced classification tasks; (3) Strict spatial generalization evaluation: Systematically validates model spatial generalization capability under different feature engineering strategies through LOSO evaluation, providing reliability assurance for actual deployment; (4) Probability calibration and decision support: Systematically evaluates probability calibration quality of model outputs through Brier Score and ECE metrics, providing quantitative basis for mapping model outputs to farm decision thresholds.
However, this study also has some limitations: (1) Spatial coverage limitations: Study region is limited to California Central Valley, and spatial distribution of 18 CIMIS stations may not fully represent all microclimatic environments, model generalization capability in other geographic regions (such as mountainous areas, coastal regions) needs further validation; (2) Temporal span limitations: Data temporal span is 2010–2025, although covering multiple frost seasons, may not fully capture long-term climate change impacts on frost patterns; (3) Manual design of feature engineering: This study adopts manually designed feature engineering strategies, although systematic and interpretable, may not fully capture all useful patterns in data, automatic feature engineering methods (such as neural architecture search) may provide further performance improvements; (4) LOSO evaluation of high-dimensional spatial aggregation features not completed: Due to memory

limitations, complete LOSO evaluation of high-dimensional spatial aggregation features could not be completed on current experimental platform, needs to be completed on machines with larger memory; (5) Economic analysis of decision thresholds: Although this study discusses the "prefer false alarms over missed events" strategy, detailed economic cost analysis was not conducted, cost-benefit analysis for different crops and different protective measures needs further research.

6.6 Practical Application Significance and Future Directions

Practical application significance of this study is mainly reflected in the following aspects. First, feature selection guidance: ABC matrix framework provides clear guidance for feature selection in actual deployment, selecting appropriate feature matrices based on forecast horizons and computational resources. Research finds that short-term forecasting (3–6 hours) prioritizes single-station feature engineering, long-term forecasting (12–24 hours) prioritizes spatial aggregation features, and this horizon-dependent pattern provides scientific basis for feature selection in actual deployment. Second, decision threshold setting: Calibrated probabilities can be directly mapped to decision thresholds in farm standard operating procedures, providing quantitative basis for growers to develop protection strategies. Research proves the rationality of "prefer false alarms over missed events" strategy in agricultural applications, providing scientific basis for cost-sensitive decisions. Third, spatial generalization assurance: LOSO evaluation validates model spatial generalization capability, providing reliability assurance for deploying models on new stations. Research finds that models can maintain high performance on unseen stations, providing important support for actual deployment. Fourth, necessity of class-balanced training: Research proves the critical role of class-balanced training in extremely imbalanced tasks, providing practical guidance for similar applications. Particularly in agricultural applications, the cost of missed frost events is far higher than the cost of false alarms, making class-balanced training necessary.

Future research directions include multiple aspects. First, extension to other geographic regions: Validate model generalization capability in other geographic regions (such as mountainous areas, coastal regions), explore impacts of geographic features on model performance. Climate patterns, topographic features, and microclimatic environments may differ significantly across different geographic regions, requiring further validation of model generalization capability. Second, integration of multi-source data: Combine satellite remote sensing, reanalysis data, and numerical weather prediction model outputs to further improve prediction accuracy and lead time. Multi-source data fusion can provide more comprehensive meteorological information, particularly for long-term forecasting, where numerical weather prediction models can provide large-scale weather system information. Third, automatic feature engineering: Explore automatic feature engineering methods (such as neural architecture search, genetic algorithms) to automatically discover optimal feature combinations. Although manually designed feature engineering strategies are systematic and interpretable, automatic feature engineering methods may discover better feature combinations, further improving model performance. Fourth, economic cost analysis: Conduct detailed economic cost analysis, quantify cost-benefit under different decision thresholds, providing optimal decision strategies for growers. Economic values of different crops, costs of different protective measures, and economic losses from false alarms and missed events differ significantly, requiring targeted economic analysis. Fifth, real-time deployment systems: Develop real-time deployment systems, integrate models into farm management systems, providing real-time frost warnings and decision support. Real-time deployment systems need to consider multiple aspects including data acquisition, model inference, result visualization, and decision support, providing end-to-end solutions for growers.

7 Conclusion

Based on hourly observation data from 18 CIMIS weather stations in California Central Valley, this study proposes a systematic feature engineering evaluation framework and deeply explores key issues in frost risk prediction. Main contributions of the research include: (1) Proposed a reproducible feature engineering evaluation framework, enabling performance improvements to be traced to specific feature types; (2) Validated the differentiated role of single-station feature engineering and spatial aggregation features across different forecast horizons, providing clear guidance for feature selection in actual deployment; (3) Comprehensively analyzed the impact of class-balanced training, proving significant reduction in false negatives, providing practical guidance for extremely imbalanced classification tasks; (4) Strictly validated model spatial generalization capability through leave-one-station-out cross-validation, providing reliability assurance for deploying models on new stations; (5) Achieved excellent probability calibration quality, directly mappable to farm decision thresholds.

Research results show that single-station feature engineering (including lag features, rolling window statistics, and derived meteorological variables) outperforms raw features across all forecast horizons, particularly outstanding in short-term forecasting, validating the core value of time series feature engineering in capturing temporal dependency patterns in frost formation. Spatial aggregation features (including neighborhood statistics and spatial gradients) achieve optimal or near-optimal performance across all forecast horizons, more critical in long-term forecasting, validating the important role of regional weather system information in long-term forecasting. This horizon-dependent pattern reflects physical mechanisms of frost formation: short-term forecasting mainly relies on local temporal patterns (temperature decline trends, historical states), while long-term forecasting requires incorporating regional weather system information (cold air transport, spatial gradients). Optimal spatial aggregation radius varies with forecast horizon, revealing the coupling relationship between spatial and temporal scales.

Class-balanced training significantly improves recall, with false negatives dramatically reduced, directly addressing the key requirement of minimizing missed frost events in agricultural applications. Temperature prediction accuracy maintains high levels across all forecast horizons, and leave-one-station-out cross-validation shows stable performance, validating model spatial robustness. Feature importance analysis reveals the key role of derived meteorological variables such as soil temperature gradients, dew point differences, and vapor pressure deficits in frost prediction, and these findings are highly consistent with physical mechanisms of frost formation, providing physical interpretation for understanding model decisions.

Main limitations of this study include spatial coverage limitations (limited to California Central Valley), temporal span limitations, and incomplete spatial generalization evaluation of high-dimensional spatial aggregation features (due to memory limitations). Future research directions include extension to other geographic regions, integration of multi-source data (satellite remote sensing, reanalysis data, numerical weather prediction model outputs), automatic feature engineering methods, and economic cost analysis. 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 important significance for improving risk resilience and sustainable development of agricultural production.

Reproducibility and Open Source

This study uses declarative configuration and fixed random seeds to manage all experiments, ensuring complete reproducibility of experiments. Each experiment run automatically generates experiment directories containing original parameters, data split information, training logs, metric files, reliability charts, and model weights. The manuscript is compiled directly from the same code repository, ensuring consistency between reported content and code implementation. Core code and data processing scripts are open source within license scope, facilitating reproduction, comparison, and extension by other researchers.

Raw data comes from the official repository of F3 Innovate Frost Risk Forecasting Challenge: https://github.com/CarlSaganPhD/frost-risk-forecast-challenge, containing hourly observation data from 18 CIMIS weather stations (2010–2025, approximately 2.36 million records). Complete codebase, data processing scripts, and documentation are available at the following GitHub repository: https://github.com/Zhengkun-Li/AgriFrost-AI. All experimental configurations, hyperparameters, and random seeds are fully recorded in the repository, ensuring complete reproducibility. All experimental results, including model weights, training logs, and detailed metrics, have been uploaded to OneDrive folder: https://ufloridamy.sharepoint.com/:f:/r/personal/zhengkun_li_ufl_edu/Documents/frost-risk-forecast-challenge-2025?csf=1&web=1&e=bXLvog.

The code repository contains the following core components: (1) Data preprocessing and quality control scripts for processing CIMIS raw data and generating feature matrices; (2) Feature engineering implementation, including complete implementation of single-station feature engineering and spatial aggregation features; (3) Model training and evaluation framework, supporting LightGBM model training, LOSO evaluation, and result archiving; (4) Experimental configuration management, using YAML files to manage all experimental parameters; (5) Result analysis and visualization scripts for generating all figures and tables in the paper. All scripts contain detailed documentation and comments for easy understanding and reproduction.

8 Supplementary Materials

- **S1: Feature List and Classification Rules** (`S1_feature_list.pdf`): Contains complete feature lists for matrices A, B, and C (including physical meaning, calculation formulas, naming conventions) and feature classification rules (how to classify features in matrices B and C).
- **S2: Complete Experimental Results Data** (`S2_all_metrics_lightgbm_abc.csv`): Contains complete results data for all 96 experimental configurations (3 feature matrices $\times$ 4 forecast horizons $\times$ 2 training methods $\times$ 10 radius values (Matrix C only)), including detailed data for all 9 evaluation metrics.
- **S3: Matrix A Feature Importance Analysis** (`S3_matrix_A_feature_importance.csv`): Contains complete feature importance data for Matrix A (single-station raw features, 12 dimensions) across all forecast horizons (3, 6, 12, 24 hours), including feature importance percentages and cumulative percentages for both frost classification and temperature regression tasks.
- **S4: Matrix B Feature Importance Analysis** (`S4_matrix_B_feature_importance.csv`): Contains complete feature importance data for Matrix B (single-station feature engineering, 278 dimensions) across all forecast horizons, including importance ranking, percentages, and cumulative percentages for all 278 features.

- **S5: Matrix C Feature Importance Analysis** (`S5_matrix_C_feature_importance.csv`): Contains complete feature importance data for Matrix C (spatial aggregation features, 534 dimensions) across all forecast horizons and optimal radii, including importance ranking, percentages, and cumulative percentages for all 534 features.
- **S6: LOSO Evaluation Results** (`S6_loso_summary.csv`, `loso_station_results_abc.csv`): Contains detailed LOSO evaluation results for matrices A and B (18 stations $\times$ 4 forecast horizons), including station-level data and summary statistics for all 9 evaluation metrics.
- **S7: Experimental Scripts** (`scripts/` directory): Contains all Python scripts and Shell scripts used to generate paper results, including model training, LOSO evaluation, feature importance analysis, and figure generation scripts.
- **S8: README Documentation** (`README.md`): Contains usage instructions and script execution guidelines for supplementary materials.

All supplementary materials are available through the project GitHub repository: https://github.com/Zhengkun-Li/AgriFrost-AI. Detailed instructions are available in `Supplementary_lighgbm_abc/README.md`.