Supplementary Material S1: Detailed Feature List for AgriFrost-AI ABCD Feature Matrices

AgriFrost-AI Team

December 6, 2025

Contents

1 Overview

This document provides a detailed list of all features used in the ABCD feature matrices of the AgriFrost-AI system, including feature names, physical meanings, calculation formulas, generation methods, and usage in the four matrices (A, B, C, D). This document aims to provide a complete reference for model reproduction, feature extension, and result interpretation.

2 Matrix A: Single-station + Raw Features (16 dimensions)

Matrix A, as a baseline configuration, only uses raw CIMIS observation variables and basic time encoding, and does not introduce any lag, rolling statistics, or spatial aggregation features.

2.1 Raw CIMIS Variables (12 dimensions)

Table 1 lists the 12 raw CIMIS variables used in Matrix A and their physical meanings.

Table 1: Matrix A: Raw CIMIS Variables List

Variable Name	Physical Meaning and Units	Data Source
`Air Temp (C)`	Near-surface air temperature (Celsius), direct target variable for frost prediction	CIMIS raw observations
`Dew Point (C)`	Dew point temperature (Celsius), reflects air water vapor content and saturation level	CIMIS raw observations
`Rel Hum (%)`	Relative humidity (`Wind Speed (m/s)`	Wind speed (m/s), reflects boundary layer mixing intensity, radiative frost more likely under weak wind conditions
CIMIS raw observations		
`Wind Dir (0-360)`	Wind direction (0–360 degrees), indicates cold air transport pathway	CIMIS raw observations
`Sol Rad (W/sq.m)`	Solar radiation flux (W/m^2), controls daytime surface heat storage	CIMIS raw observations
`Soil Temp (C)`	Shallow soil temperature (Celsius), reflects surface and near-surface heat storage exchange	CIMIS raw observations
`Vap Pres (kPa)`	Vapor pressure (kPa), absolute measure of water vapor content	CIMIS raw observations
`ETo (mm)`	Reference evapotranspiration (mm), comprehensively reflects radiation, temperature, wind speed and humidity conditions	CIMIS calculated values
`Precip (mm)`	Precipitation (mm), affects surface energy balance and soil heat capacity	CIMIS raw observations
`Hour (PST)`	Hour (0–23), used for time feature encoding	Extracted from timestamp
`Jul`	Julian day (1–366), day of year, used for seasonal pattern encoding	CIMIS raw data

2.2 Time Harmonic Encoding (4 dimensions)

To capture diurnal and annual cycle patterns, Matrix A performs cyclic encoding on time variables:

- `hour_sin` = $\sin(2\pi \times \text{hour}/24)$
- `hour_cos` = $\cos(2\pi \times \text{hour}/24)$
- `month_sin` = $\sin(2\pi \times \text{month}/12)$

- `month_cos` = $\cos(2\pi \times \text{month}/12)$

Cyclic encoding avoids time boundary discontinuities (such as the jump between 23:00 and 00:00), enabling the model to learn smooth periodic patterns.

Total dimensions：12 raw CIMIS variables + 4 time harmonic encodings = 16 dimensions

3 Matrix B: Single-station + Engineered Features (278 dimensions)

Matrix B builds upon Matrix A by adding a complete feature engineering pipeline, generating 278 candidate features. This section provides a detailed list of all features by feature category.

3.1 Time Features (15 dimensions)

Table 2 lists the time features of Matrix B.

Table 2: Matrix B: Time Features List

Feature Name	Description and Calculation	Dimensions
`hour`	Hour (0–23), discrete value	1
`day_of_year`	Day of year (1–366)	1
`month`	Month (1–12)	1
`day_of_week`	Day of week (0=Monday, 6=Sunday)	1
`season`	Season (1=Spring, 2=Summer, 3=Fall, 4=Winter)	1
`is_night`	Night indicator: 1 if 18:00–06:00, otherwise 0	1
`hour_sin`, `hour_cos`	Diurnal cycle cyclic encoding	2
`month_sin`, `month_cos`	Annual cycle cyclic encoding	2
`day_of_year_sin`, `day_of_year_cos`	Julian day cyclic encoding	2
`day_progress_sin`, `day_progress_cos`	Day progress cyclic encoding (`day_progress` = `hour` / 24)	2
`frost_season_indicator`	Frost season indicator: 1 if December–April, otherwise 0	1

3.2 Lag Features (50 dimensions)

For 10 core variables, extract 1, 3, 6, 12, 24 hours of historical values．All lag features are computed grouped by station, ensuring no information leakage across stations.

Lag Variable List：

- `Air Temp (C)`：Air temperature
- `Dew Point (C)`：Dew point
- `ETo (mm)`：Reference evapotranspiration
- `Precip (mm)`：Precipitation

- `Rel Hum (%)`：Relative humidity
- `Soil Temp (C)`：Soil temperature
- `Sol Rad (W/sq.m)`：Solar radiation
- `Wind Dir (0-360)`：Wind direction
- `Wind Speed (m/s)`：Wind speed
- `Vap Pres (kPa)`：Vapor pressure

Lag Windows：1h, 3h, 6h, 12h, 24h
Naming Format: `{variable}_lag_{hours}`
Examples：

- `Air Temp (C)_lag_1`：1 hour ago air temperature
- `Dew Point (C)_lag_6`：6 hours ago dew point
- `Soil Temp (C)_lag_24`：24 hours ago soil temperature

Total dimensions：10 variables × 5 lag windows = 50 dimensions

3.3 Rolling Window Statistics (180 dimensions)

For 9 core variables (air temperature, dew point, ETo, precipitation, relative humidity, soil temperature, solar radiation, wind speed, vapor pressure; wind direction does not participate in rolling statistics as it is an angular variable) on 3, 6, 12, 24 hour windows, compute 5 statistics (mean, minimum, maximum, standard deviation, sum).

Rolling Windows: 3h, 6h, 12h, 24h
Statistical Functions：

- `mean`：mean within window
- `min`：minimum within window
- `max`：maximum within window
- `std`：standard deviation within window
- `sum`：sum within window(applicable to precipitation, ETo and other cumulative quantities)

Naming Format: `{variable}_rolling_{window}h_{statistic}`
Examples：

- `Air Temp (C)_rolling_6h_mean`：Recent 6-hour average air temperature
- `Soil Temp (C)_rolling_24h_min`：Recent 24-hour minimum soil temperature
- `Dew Point (C)_rolling_12h_std`：Recent 12-hour dew point standard deviation

Calculation Method：Using pandas `rolling()` function, grouped by station, `min_periods=1` to maximize data utilization.

Total dimensions：9 variables × 4 windows × 5 statistics = 180 dimensions

3.4 Derived Meteorological Features (3 dimensions)

Composite variables calculated based on physical relationships and meteorological principles, capturing interactions between variables:

- `wind_chill`：Wind chill index, quantifies the effect of wind speed on perceived temperature. Calculated when temperature < 10°C, formula is $13.12 + 0.6215T - 11.37V^{0.16} + 0.3965TV^{0.16}$ (where T is temperature in °C, V is wind speed in km/h). In frost scenarios, high wind speed accelerates heat loss, reducing perceived temperature.
- `heat_index`：Heat index, calculated under high temperature and high humidity conditions (temperature > 80°F and relative humidity > 40
- `soil_air_temp_diff`：Soil temperature and air temperature difference (`Soil Temp` − `Air Temp`), reflects surface energy exchange direction. Positive values indicate soil temperature is higher than air temperature (common during daytime), negative values indicate soil temperature is lower than air temperature (common at night). This feature has important value for identifying inversion layers and radiative cooling processes.

Total dimensions：3 dimensions

3.5 Radiation-related Features (4 dimensions)

Solar radiation is the core driver of surface energy balance, directly affecting daytime heating and nighttime cooling processes:

- `sol_rad_change_rate`：Solar radiation change rate (`Sol Rad`(t) − `Sol Rad`$(t-1)$), captures short-term fluctuations in radiation intensity, reflects cloud changes and atmospheric transparency
- `daily_solar_radiation`：Daily cumulative radiation, accumulated from 06:00 to current time. This feature quantifies total daily energy input, high cumulative radiation usually corresponds to stronger daytime heating, may affect nighttime cooling rate
- `nighttime_cooling_rate`：Nighttime cooling rate, calculated only when `is_night=1`. This feature directly captures radiative cooling process, is a key signal for frost forecasting
- `radiation_temp_interaction`：Radiation and temperature interaction term (`Sol Rad` × `Air Temp`), captures nonlinear effect of radiation on temperature

Total dimensions：4 dimensions

3.6 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:

- `wind_dir_sin/cos`：Wind direction cyclic encoding, converts 0–360 degree angles to sin(θ) and cos(θ), avoids angle boundary discontinuity (difference between 359° and 1°)

- `wind_dir_category`：Wind direction categorical encoding, divides 0–360 degrees into 4 quadrants (North, East, South, West), facilitates model learning of impact patterns of different wind directions on frost risk
- `wind_speed_change_rate`：Wind speed change rate (`Wind Speed`$(t) -$ `Wind Speed`$(t-1)$), captures dynamic changes in wind field
- `calm_wind_duration`：Calm wind duration, defined as consecutive duration when wind speed < 1.0 m/s. Calm wind conditions favor radiative cooling, is an important prerequisite for frost formation
- `wind_dir_temp_interaction`：Wind direction and temperature interaction term, captures differential effects of different wind directions on temperature changes (e.g., dry cold wind from inland vs. moist wind from ocean)

Total dimensions：6 dimensions

3.7 Humidity Features (4 dimensions)

Humidity features quantify water vapor content in the atmosphere, for understanding radiative cooling, condensation processes and frost formation mechanisms, which has important significance:

- `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 temperature in °C). This feature reflects maximum water vapor capacity of air at a given temperature
- `dew_point_proximity`：Dew point proximity, calculation formula is `(T - T_dew) / T`, quantifies relative difference between air temperature and dew point. When this value approaches 0, 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
- `vapor_pressure_deficit`：VPD：Vapor pressure deficit, defined as difference between saturation vapor pressure and actual vapor pressure. VPD reflects air "dryness" and cooling potential, high VPD usually corresponds to stronger evaporative cooling effect

Total dimensions：4 dimensions

3.8 Trend Features (1 dimension)

Trend features capture the acceleration of temperature changes, for identifying rapid cooling processes (that may lead to frost), which has important value:

- `cooling_acceleration`：cooling acceleration, based on recent 6 hours of temperature decline rate changes. This feature quantifies the second derivative of cooling rate, positive values indicate cooling acceleration, negative values indicate cooling deceleration. Rapid cooling (high cooling acceleration) is a key signal for frost warning

Total dimensions：1 dimensions

3.9 Station Static Features (4 dimensions)

Geographic attributes merged from CIMIS station metadata, used to characterize spatial heterogeneity:

- `station_id_encoded`：station ID encoding，converts station identifiers to numerical encodings, facilitates model learning of personalized patterns of different stations (such as microclimate, elevation, topography, etc.)
- `region_encoded`：region encoding，classifies and encodes stations by geographic region， captures climate differences at regional scale

Total dimensions：4 dimensions，may vary slightly depending on encoding method。

Total number of features：The above single-station feature engineering pipeline is enabled in Matrix B，actually generates 278 candidate features，including，raw variables (12 dimensions) + time features (15 dimensions) + lag features (50 dimensions) + rolling window statistics (180 dimensions) + derived meteorological features (3 dimensions) + radiation features (4 dimensions) + wind features (6 dimensions) + humidity features (4 dimensions) + trend features (1 dimension) + station static features (4 dimensions).

3.10 Feature Selection Strategy

Based on LightGBM feature importance ranking on Matrix B + 12-hour horizon, adopts a two-stage feature selection:

Stage 1: Train full-feature baseline model, export feature importance files (`frost_feature_importance.csv` and `temp_feature_importance.csv`).

Stage 2：truncate by cumulative importance, select the smallest $k^\star$, such that cumulative importance $\geq 90\%$At 12-hour horizon, $k^\star = 146$

Top-146 feature set：contains approximately 90% cumulative importance, achieves balance between performance and computational cost. Comparative experiments show that Top-146 and full features have ROC-AUC difference < 0.001, while training time is reduced by approximately 35–40%.

4 Matrix C: Neighbor Aggregation + Raw Features (534 dimensions)

Matrix C, on the basis of Matrix A raw variables, adds multi-station spatial aggregation statistics. This configuration aims to evaluate the contribution of spatial information to frost forecasting, especially cold air pooling, inversion layer formation and other spatial patterns. For 27 numerical variables (12 raw CIMIS variables + 15 time features), perform neighbor aggregation at specified radius threshold (systematically tested in experiments 20–200 km, step size 20 km), compute 8 aggregation statistics, generating 216 neighbor aggregation features. Additionally, the system generates missing mask features (293 dimensions) to handle missing data from neighbor stations. Plus raw variables (12 dimensions), time harmonic encoding (2 dimensions) and other features (11 dimensions), total 534 dimensions.

4.1 Neighbor Construction Method

For target station s_0 and radius threshold r (typical values 20–200 km, step size 20 km), calculate Haversine distance based on latitude and 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 radius, ϕ is latitude, λ is longitude.
Neighbor station set $\mathcal{N}(s_0, r) = \{s_i : d(s_0, s_i) \leq r\}$

4.2 Time Alignment

For each neighbor station $s_i \in \mathcal{N}(s_0, r)$, align its time series with target station timestamps (based on `Date` and `Hour` columns), using left join ensures each timestamp of target station has corresponding neighbor data (may be missing).

4.3 Neighbor Aggregation Statistics (216 dimensions)

For 27 numerical variables (12 raw CIMIS variables + 15 time features), perform neighbor aggregation, respectively compute 8 aggregation methods, as shown in Table 3.

27 variables include：

- **12 raw CIMIS variables**：Air temperature、Dew point、Relative humidity、Wind speed、Wind direction、Solar radiation、Soil temperature、Vapor pressure、ETo、Precipitation, Hour, Julian day
- **15 time features**：hour, day_of_year, month, day_of_week, season, is_night, hour_sin, hour_cos, month_sin, month_cos, day_of_year_sin, day_of_year_cos, day_progress_sin, day_progress_cos, frost_season_indicator

Table 3: Matrix C: Neighbor Aggregation Methods List

Aggregation Method	Formula and Physical Meaning	Dimensions
`mean`	$\bar{x} = \frac{1}{\|\mathcal{N}\|} \sum_{s_i \in \mathcal{N}} x_i$，neighbor mean, reflects local average state	27
`max`	$x_{\max} = \max_{s_i \in \mathcal{N}} x_i$，neighbor maximum, identifies extreme values	27
`min`	$x_{\min} = \min_{s_i \in \mathcal{N}} x_i$，neighbor minimum, captures cold air pooling (especially important for temperature/soil temperature)	27
`std`	$\sigma = \sqrt{\frac{1}{\|\mathcal{N}\|-1} \sum_{s_i \in \mathcal{N}} (x_i - \bar{x})^2}$，neighbor standard deviation, reflects spatial variability	27
`median`	neighbor median, more robust to outliers	27
`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, nearby neighbors have larger weights	27
`gradient`	$\nabla x = \bar{x} - x_0$，neighbor mean minus target station value, characterizes spatial gradient (crucial for identifying cold air sinking and inversion layers)	27
`range`	$x_{\max} - x_{\min}$，neighbor maximum minus minimum, reflects range of spatial variability	27

Naming Format: `{variable}_neighbor_{method}`
Examples：

- `Air Temp (C)_neighbor_mean`：neighbor average air temperature
- `Soil Temp (C)_neighbor_min`：neighbor minimum soil temperature，cold air pooling signal，
- `Soil Temp (C)_neighbor_gradient`：neighbor soil temperature gradient，key warning feature，
- `Dew Point (C)_neighbor_std`：neighbor dew point standard deviation

Total dimensions: 27 variables × 8 aggregation methods = 216 dimensions

4.4 Missing Mask Features (293 dimensions)

To handle missing data from neighbor stations，the system generates missing mask features，including：

- **missing masks for neighbor aggregation features**：216 dimensions，，for each aggregation feature, calculate neighbor missing ratio：`{variable}_neighbor_{method}_missing_ratio`，，indicates data quality and spatial coverage of this aggregation feature

- **variable missing ratio**：27 dimensions：for each variable, calculate at timestamp t the proportion of available neighbor data (`{variable}_neighbor_missing_ratio`)
- **missing masks for missing ratio features**：27 dimensions：calculate missing masks for variable missing ratio features themselves
- **missing masks for other features**：23 dimensions：for time features, derived meteorological features and other features, calculate missing masks

Design Motivation：These missing mask features are used to indicate data quality and spatial coverage, help the model understand which features are available/missing, improve model robustness under sparse data conditions.

Total dimensions：216 + 27 + 27 + 23 = 293 dimensions

4.5 Raw Variables (12 dimensions)

Retains all raw CIMIS variables from Matrix A, forms a contrast with neighbor statistics.

4.6 Time Harmonic Encoding (2 dimensions)

only uses `hour_sin/cos` and `month_sin/cos`, total 2 dimensions (different from Matrix A's 4 dimensions, Matrix C uses fewer time encodings).

4.7 Other Features (11 dimensions)

including time discrete features (hour, day_of_year, month, day_of_week, season, is_night), derived meteorological features (such as `has_neighbors`, etc.), total 11 dimensions.

Total dimensions: 216 (neighbor aggregation) + 293 (missing masks) + 12 (raw variables) + 2 (time harmonics) + 11 (other features) = 534 dimensions

5 Matrix D: Neighbor Aggregation + Engineered Features (818 dimensions)

Matrix D combines Matrix B single-station feature engineering (278 dimensions) with Matrix C neighbor aggregation components (216 dimensions neighbor aggregation + 299 dimensions missing masks), overlaying them to form an 818-dimensional high-dimensional feature space. This configuration aims to evaluate the joint effects of time series engineering features and spatial aggregation statistics, exploring performance-complexity trade-offs in complex feature spaces.

5.1 Single-station Engineered Features (278 dimensions)

identical to Matrix B：including：

- Raw variables (12 dimensions)
- Time features (15 dimensions)

- Lag features (50 dimensions)
- Rolling window statistics (180 dimensions)
- derived meteorological features（3 dimensions）
- Radiation features (4 dimensions)
- Wind features (6 dimensions)
- Humidity features (4 dimensions)
- Trend features (1 dimension)
- Station static features (4 dimensions)

Total dimensions: 278 dimensions

5.2 Neighbor Aggregation Features (216 dimensions)

identical to Matrix C, for 27 numerical variables (12 raw CIMIS variables + 15 time features), perform 8 aggregation statistics.

Total dimensions: 216 dimensions

5.3 Missing Mask Features (299 dimensions)

Similar to Matrix C's missing mask features, but with higher dimensions (299 dimensions vs. 293 dimensions), this is because Matrix D generates missing masks for engineered features (such as lag and rolling statistics) also generates missing masks, to improve model robustness under sparse data conditions.

Total dimensions: 299 dimensions

5.4 Other Features (43 dimensions)

including wind features, humidity features, radiation features, trend features, station static features, geographic features and interaction features, etc. It should be noted that feature duplication exists in Matrix D (such as time features in Matrix B and Matrix C's "other features" both appear), in actual implementation these duplicate features are automatically deduplicated, ensuring each feature appears only once in the feature space.

Total dimensions: 43 dimensions

Total dimensions：278 (single-station engineering) + 216 (neighbor aggregation) + 299 (missing masks) + 43 (other features)= 818 dimensions（after deduplication）

5.5 Feature Selection Recommendations

Due to high feature dimensionality in Matrix D（818 dimensions），the following strategies are recommended：

1. **Top-146 Feature Selection**: based on Matrix B feature importance ranking (12 hour horizon), select top 146 single-station features, then overlay Matrix C neighbor features (534 dimensions), total approximately 680 dimensions.