

Supplementary Material S1: Feature List and Classification Rules

1 Overview

This document provides a detailed list of all features used in the ABC feature matrices, including feature names, physical meanings, calculation formulas, and classification rules. This document aims to provide a complete reference for model reproduction, feature extension, and result interpretation.

2 Matrix A: Single-station Raw Features (12 dimensions)

Matrix A serves as a baseline configuration, using only 12 raw CIMIS observation variables without any feature engineering or spatial aggregation.

2.1 Raw CIMIS Variables List

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 (%), together with dew point characterizes air saturation state	CIMIS raw observations
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

Total dimensions: 12 (12 raw CIMIS variables)

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 lists all features by category with detailed calculation formulas.

3.1 Time Features (17 dimensions)

Time features are crucial for frost prediction as frost events exhibit strong diurnal and annual cycle patterns.

Discrete encoding (6 dimensions):

- hour (0-23): Hour of day

- month (1–12): Month of year
- day_of_year (1–366): Day of year
- day_of_week (0–6): Day of week (0=Monday, 6=Sunday)
- season (1–4): Season (1=Spring, 2=Summer, 3=Fall, 4=Winter)
- is_night (binary): 1 if 18:00–06:00, 0 otherwise

Cyclic encoding (8 dimensions): Harmonic encoding using trigonometric functions to avoid boundary discontinuities:

- hour_sin = $\sin\left(\frac{2\pi \cdot \text{hour}}{24}\right)$
- hour_cos = $\cos\left(\frac{2\pi \cdot \text{hour}}{24}\right)$
- month_sin = $\sin\left(\frac{2\pi \cdot \text{month}}{12}\right)$
- month_cos = $\cos\left(\frac{2\pi \cdot \text{month}}{12}\right)$
- day_of_year_sin = $\sin\left(\frac{2\pi \cdot \text{day_of_year}}{365.25}\right)$
- day_of_year_cos = $\cos\left(\frac{2\pi \cdot \text{day_of_year}}{365.25}\right)$
- day_progress_sin = $\sin(2\pi \cdot \text{day_progress})$
- day_progress_cos = $\cos(2\pi \cdot \text{day_progress})$

where day_progress = hour/24 (normalized hour progress, 0–1).

Agricultural indicator (1 dimension):

- frost_season_indicator: Binary indicator marking high frost risk period (December–April) in California

Other time features (2 dimensions):

- day_progress: Normalized hour progress (0–1)
- time_of_day: Time point within the day

3.2 Lag Features (50 dimensions)

Lag features capture historical states of meteorological variables, valuable for predicting future temperature changes.

Calculation formula:

$$x_{\text{lag},h}(t) = x(t - h)$$

where $x(t - h)$ represents the value of variable x at h hours before time t .

Configuration:

- **Variables:** 10 core variables (Air Temp, Dew Point, ETo, Precip, Rel Hum, Soil Temp, Sol Rad, Wind Dir, Wind Speed, Vap Pres)
- **Lag windows:** 1, 3, 6, 12, 24 hours
- **Total:** $10 \times 5 = 50$ dimensions

Naming format: {variable}_lag_{hours}, e.g., Air Temp (C)_lag_12 represents air temperature 12 hours ago.

3.3 Rolling Window Statistics (180 dimensions)

Rolling window statistics capture distribution characteristics of variables within time windows, important for identifying trends, variability, and extremes.

Statistical measures:

- **Mean:** $\bar{x}_w = \frac{1}{n} \sum_{i=1}^n x_i$, reflects average state within window
- **Minimum:** $x_{\min} = \min_i x_i$, identifies extremes, especially important for frost warning
- **Maximum:** $x_{\max} = \max_i x_i$, identifies extremes
- **Standard deviation:** $\sigma_w = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_w)^2}$, quantifies variability
- **Sum:** $\sum_{i=1}^n x_i$, physically meaningful for cumulative quantities (e.g., precipitation, ETo)

Configuration:

- **Variables:** 9 core variables (Air Temp, Dew Point, ETo, Precip, Rel Hum, Soil Temp, Sol Rad, Wind Speed, Vap Pres)
- **Time windows:** 3, 6, 12, 24 hours
- **Total:** $9 \times 4 \times 5 = 180$ dimensions

Naming format: {variable}_rolling_{window}h_{statistic}, e.g., Air Temp (C)_rolling_24h_mean represents mean air temperature over past 24 hours.

3.4 Derived Meteorological Features (7 dimensions)

Derived meteorological features are composite variables calculated based on physical relationships and meteorological principles.

Feature list with formulas:

- **temp_dew_diff** = Air Temp – Dew Point: Quantifies air saturation level. When close to 0, air is near saturation; larger differences indicate drier air, favorable for evaporative cooling.
- **temp_change_rate** = $\text{Air Temp}(t) - \text{Air Temp}(t-1)$: Captures short-term temperature change trend.
- **wind_chill** = $13.12 + 0.6215T - 11.37V^{0.16} + 0.3965TV^{0.16}$ (when $T < 10^\circ\text{C}$, where T is air temperature in $^\circ\text{C}$, V is wind speed in km/h): Quantifies wind speed effect on apparent temperature.
- **heat_index**: Calculated under high temperature and humidity conditions (Air Temp $> 80^\circ\text{F}$ and Rel Hum $> 40\%$). Although typically equals air temperature in frost scenarios, it helps model learn complete temperature-humidity relationships.
- **soil_air_temp_diff** = Soil Temp – Air Temp: Reflects surface energy exchange direction. Positive values indicate soil warmer than air (common during day), negative values indicate soil cooler than air (common at night).
- **temp_decline_rate** = $\frac{\text{Air Temp}(t) - \text{Air Temp}(t-6)}{6}$ ($^\circ\text{C}/\text{hour}$): Quantifies average cooling rate over past 6 hours.
- **vapor_pressure**: Vapor pressure related features

3.5 Wind Features (8 dimensions)

Wind features are crucial for frost prediction as wind speed affects convective mixing intensity and wind direction affects cold air pathways.

Feature list with formulas:

- $\text{wind_dir_sin} = \sin(\theta)$, $\text{wind_dir_cos} = \cos(\theta)$: Cyclic encoding of wind direction (0–360 degrees), avoiding boundary discontinuity (e.g., between 359° and 1°)
- wind_dir_category : Categorical encoding dividing 0–360 degrees into 4 quadrants (North, East, South, West)
- $\text{wind_speed_change_rate} = \text{Wind Speed}(t) - \text{Wind Speed}(t - 1)$: Captures dynamic changes in wind field
- $\text{calm_wind_duration}$: Duration of calm wind conditions (wind speed < 1.0 m/s). Calm conditions favor radiative cooling, important prerequisite for frost formation
- $\text{wind_dir_temp_interaction}$: Interaction term between wind direction and temperature, capturing differential effects of different wind directions on temperature changes

3.6 Soil Features (2 dimensions)

- soil_temp : Soil temperature
- $\text{soil_air_temp_diff} = \text{Soil Temp} - \text{Air Temp}$: Soil-air temperature difference

3.7 Station Features (1 dimension)

- $\text{station_id_encoded}$: Station ID encoded as numerical value
- region_encoded : Region encoded as categorical value

3.8 Other Features (13 dimensions)

Other features include humidity-related variables, radiation-related features, trend features, and all raw CIMIS variables not categorized above.

Total dimensions: 278

4 Matrix C: Spatial Aggregation Features (534 dimensions)

Matrix C adds spatial aggregation features to Matrix A, capturing regional climate patterns by aggregating observations from neighboring stations.

4.1 Raw Variables (12 dimensions)

Matrix C contains the same 12 raw CIMIS variables as Matrix A.

4.2 Neighborhood Aggregation Features (216 dimensions)

Neighborhood aggregation features capture regional climate patterns by aggregating observations from neighboring stations within a specified radius.

Aggregation methods:

- **Mean:** $\bar{x} = \frac{1}{|\mathcal{N}|} \sum_{s_i \in \mathcal{N}} x_i$, where \mathcal{N} is the set of neighboring stations
- **Maximum:** $x_{\max} = \max_{s_i \in \mathcal{N}} x_i$
- **Minimum:** $x_{\min} = \min_{s_i \in \mathcal{N}} x_i$, captures cold air pooling (especially important for temperature/soil temperature)
- **Standard deviation:** $\sigma = \sqrt{\frac{1}{|\mathcal{N}|-1} \sum_{s_i \in \mathcal{N}} (x_i - \bar{x})^2}$, reflects spatial variability
- **Median:** Median of neighborhood values, more robust to outliers
- **Distance-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 $w_i = 1/d_i^2$ (closer stations have higher weights)
- **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)
- **Range:** $x_{\max} - x_{\min}$, neighborhood maximum minus minimum, reflects spatial variability range

Configuration:

- **Variables:** 27 variables (12 raw CIMIS variables + 15 engineered features)
- **Aggregation methods:** 8 methods (mean, max, min, std, median, weighted_mean, gradient, range)
- **Total:** $27 \times 8 = 216$ dimensions

Naming format: {variable}_neighbor_{method}, e.g., Soil Temp (C)_neighbor_min represents minimum soil temperature in neighborhood.

4.3 Missing Mask Features (294 dimensions)

Missing mask features indicate data quality and spatial aggregation reliability.

Feature types:

- (1) **Missing masks for neighborhood aggregation features** (216 dimensions): Binary mask for each aggregation feature:

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

- (2) **Variable missing ratio features** (27 dimensions): Missing ratio for each variable:

$$\text{missing_ratio} = \frac{\text{number of missing neighboring stations}}{\text{total number of neighboring stations}} = \frac{\sum_{s_i \in \mathcal{N}} \mathbf{1}[\text{variable missing at } s_i]}{|\mathcal{N}|}$$

where $\mathbf{1}[\cdot]$ is the indicator function.

- (3) **Missing masks for missing ratio features** (27 dimensions): Missing mask for each missing ratio feature itself
- (4) **Missing masks for other features** (24 dimensions): Missing masks for raw variables and time features

Total dimensions: 294

4.4 Time Harmonic Encoding (2 dimensions)

- $\text{day_of_year_sin} = \sin\left(\frac{2\pi \cdot \text{day_of_year}}{365.25}\right)$
- $\text{day_of_year_cos} = \cos\left(\frac{2\pi \cdot \text{day_of_year}}{365.25}\right)$

4.5 Other Features (10 dimensions)

Other features include time discrete features, derived meteorological features, `has_neighbors` indicator, etc.

Total dimensions: 534 (12 raw + 216 spatial aggregation + 294 masks + 2 time harmonic + 10 other)

5 Feature Classification Rules

This section explains how features are classified in Matrix B and Matrix C for feature importance analysis.

5.1 Matrix B Feature Classification Rules

Table 2 lists the feature classification rules for Matrix B.

Table 2: Matrix B Feature Classification Rules

Feature Category	Count	Classification Rule (based on feature name pattern matching)
Rolling Statistics	180	Features containing <code>rolling</code> , <code>rolling_</code> , or <code>_rolling</code>
Lag Features	50	Features containing <code>lag_</code> or <code>_lag</code>
Time Features	17	Features containing time encoding (hour, date, season, etc.) and time harmonic encoding (sin/cos)
Wind Features	8	Features containing <code>wind_dir</code> , <code>wind_speed</code> , <code>wind_chill</code> , <code>calm_wind</code> and other wind-related keywords
Soil Features	2	Features containing <code>soil_temp</code> , <code>soil_temp_</code> , <code>soil_air_temp</code>
Derived Met. Features	7	Features containing composite variables calculated based on physical relationships (e.g., temperature differences, change rates, vapor pressure)
Station Features	1	Features containing <code>stn_id</code> , <code>station</code> , <code>region</code> , <code>region_encoded</code> and other station encoding features
Other Features	13	Features containing <code>rel_hum</code> , <code>humidity</code> and other humidity-related variables, as well as all raw CIMIS variables not categorized above

5.2 Matrix C Feature Classification Rules

Table 3 lists the feature classification rules for Matrix C.

Table 3: Matrix C Feature Classification Rules

Feature Category	Count	Classification Rule (based on feature name pattern matching)
Spatial Features	216	Features containing neighborhood aggregation statistics, calculated for raw CIMIS variables and engineered features within specified radius using multiple aggregation methods (mean, max, min, std, median, distance-weighted mean, gradient, range), used to capture regional climate patterns and cold air pooling
Mask Features	294	Missing mask features, used to indicate data quality and spatial aggregation reliability, helping model identify reliability of spatial aggregation information
Engineering Features	14	Time features (time encoding and time harmonic encoding) and derived meteorological features (composite variables calculated based on physical relationships), similar to engineering features in Matrix B
Raw Features	10	Raw CIMIS variables (compared to Matrix A, Matrix C contains only partial raw variables), including Air Temp, Dew Point, Rel Hum, Wind Speed, Wind Dir, Sol Rad, Soil Temp, Vap Pres, ETo, etc.

6 Feature Classification Implementation Method

Feature classification is implemented through pattern matching based on feature names. For Matrix B and Matrix C, the system performs feature classification following these steps:

1. **Read feature list:** Extract all feature names from trained models
2. **Pattern matching:** Match each feature against rules in Table 2 and Table 3
3. **Category assignment:** Assign features to corresponding categories
4. **Importance aggregation:** Calculate cumulative importance percentage for each category

Feature classification rules are based on keywords and patterns in feature names, for example:

- Features containing `rolling` are classified as "Rolling Statistics"
- Features containing `lag_` or `_lag` are classified as "Lag Features"
- Features containing `neighbor_` are classified as "Spatial Features"
- Features containing `missing_mask` are classified as "Mask Features"

This pattern-matching based classification method ensures consistency and reproducibility of feature classification.