

AgriFrost-AI：加州中央谷地霜冻风险预测方法与评估

AgriFrost-AI 团队

December 4, 2025

Abstract

霜冻仍然是加州高价值园艺作物面临的关键气象风险之一，尤其在花期阶段，单次强辐射霜冻即可造成成规模减产。本文基于 F3 Innovate 霜冻风险预测挑战，构建了一个面向实际种植场景的 AgriFrost-AI 端到端系统。我们利用 2010–2025 年加州灌溉管理信息系统 (CIMIS) 18 个站点的逐小时观测，系统化完成了数据清洗、质量控制 (QC)、特征工程、邻域聚合建模与留一站 (Leave-One-Station-Out, LOSO) 空间泛化评估。方法上提出单站/多站与原始/工程特征交叉的 ABCD 特征矩阵，并在此基础上比较多种树模型与时空神经网络在 3、6、12、24 小时预报窗口上的表现。结果表明，基于邻域聚合特征的 LightGBM 模型在最短 3 小时预报中达到 ROC-AUC 0.9972、PR-AUC 0.7282、Brier Score 0.0026，且在 LOSO 评估下几乎无性能折损。我们进一步分析了土壤温度梯度、露点差与蒸汽压亏缺等近地特征的预警价值，并讨论了模型概率在农场防护决策与 ERA5/HRRR 同步气象同化中的应用前景。

Contents

1 引言	3
2 相关工作	3
3 数据与研究区域	3
3.1 观测来源与空间覆盖范围	4
3.2 霜冻事件分布与季节特征	4
3.3 观测变量与物理意义概述	5
3.4 数据质量与 QC 概览	5
4 方法	7
4.1 数据预处理与 QC 流程	7
4.2 特征工程与 ABCD 特征矩阵	7
4.3 特征选择与特征重要性分析	8
4.4 模型族与训练配置	9
4.5 训练/验证划分与防泄漏	9
4.6 评估指标	9
5 实验结果	9
5.1 实验规模与结果概览 (2025-12-04)	9
5.2 模型性能可视化与综合分析	11
5.3 整体概率与温度预测性能	12
5.4 概率校准与可靠性	13
5.5 LOSO 空间泛化	13

5.6 特征矩阵与模型族对比	14
5.7 特征洞察	15
6 讨论	15
7 决策支持应用	15
8 同步气象拓展与未来工作	15
9 可复现性与开源	16
10 结论	16

1 引言

霜冻长期以来是加州高价值果蔬与坚果作物面临的主要气象灾害之一，特别是在花期和幼果期，短时的强辐射霜冻即可造成大面积减产甚至绝收。传统的防护策略依赖于经验判断、有限的人工观测和中尺度数值天气预报，但在复杂地形和强微气候条件下往往难以及时、可靠地给出地块级预警。

F3 Innovate 霜冻风险预测挑战提出了一个贴近生产实践的评估框架：在多站点、多年份的地面观测基础上，要求参赛队伍同时输出 3、6、12、24 小时四个预报窗口下的霜冻概率与气温预估，量化概率校准质量，并通过留一站（Leave-One-Station-Out, LOSO）检验模型在空间上的泛化能力。AgriFrost-AI 项目在此基础上面向如下问题展开研究：

- 在强类不平衡、空间异质性的逐小时气象数据上，怎样设计兼顾近地物理机理与机器学习可用性的特征集合？
- 如何在不显著增加计算成本的前提下，引入邻站信息以刻画冷空气汇聚、逆温层结构等局地过程？
- 模型给出的霜冻概率能否在校准后被直接纳入农场标准操作流程（SOP），而不是仅作为相对排序指标？

本文的主要贡献概括如下：

1. 构建了基于 CIMIS 18 个站点、覆盖 2010–2025 年的统一霜冻风险数据集，并给出了完整的数据质量与 QC 分析。
2. 提出了单站/多站与原始/工程特征交叉的 ABCD 特征矩阵框架，并系统比较了不同矩阵与模型族的性能差异。
3. 通过 LOSO 方案严格评估了模型的空间泛化能力，展示了邻域聚合特征在不同预报窗口下的增益。
4. 将校准后的霜冻概率与温度预测映射到具体防护决策阈值，为种植者制定 SOP 提供量化依据。

2 相关工作

霜冻风险评估与短期气温预测在农业气象、数值天气预报与机器学习社区已有大量研究。传统方法多基于经验公式、统计回归或中尺度数值模式（如 WRF）下采样，重点刻画辐射冷却、地表能量平衡以及冷空气下沉与积聚过程。近年来，随着自动气象站与再分析资料的普及，基于随机森林、梯度提升树和深度神经网络的近地气象预测方法逐渐兴起，部分工作将卫星遥感、地形数据和再分析场作为输入，用于生成高分辨率的地表气温和霜冻风险地图。

与上述研究相比，本文更关注如下方面：第一，在统一的挑战平台上，采用统一数据集和评估指标，对多种模型进行系统比较；第二，通过显式的邻站聚合特征而非仅依赖栅格化插值，捕捉冷空气汇聚与局地逆温结构；第三，从概率校准和决策支持的角度出发，探讨模型输出在种植者操作层面的可用性。

3 数据与研究区域

本节介绍研究所使用的 CIMIS 地面观测数据、霜冻事件的统计特征、主要观测变量及其物理意义，并给出整体的数据质量与 QC 概览。

3.1 观测来源与空间覆盖范围

本研究使用的逐小时气象观测来自加州灌溉管理信息系统（California Irrigation Management Information System, CIMIS），覆盖加利福尼亚中央谷地及周边山麓地区的 18 个自动气象站。站点沿中央谷地呈南北向带状分布，从 Sacramento 平原延伸至 Bakersfield 区域，跨越冷空气堆积易发区、地势抬升带和高蒸散农田带等多样微气候环境。图 1 展示了全部站点的空间分布。

数据时间跨度为 2010–2025 年，总计约 236 万条逐小时记录。每条记录包含气温、露点、相对湿度、风速与风向、太阳辐射、土壤温度、蒸汽压、参考蒸散量（ET_o）等核心变量，并附带 CIMIS 官方质量控制（Quality Control, QC）标记。站点层面的元数据包括站号、名称、CIMIS 区域、县市、经纬度、海拔、GroundCover、启停日期及是否为 ET_o 站等信息，用于空间聚合与 LOSO 分组；完整列表见 Supplementary Table 1。原始数据与处理脚本托管于 GitHub 仓库，便于版本追踪与复现。

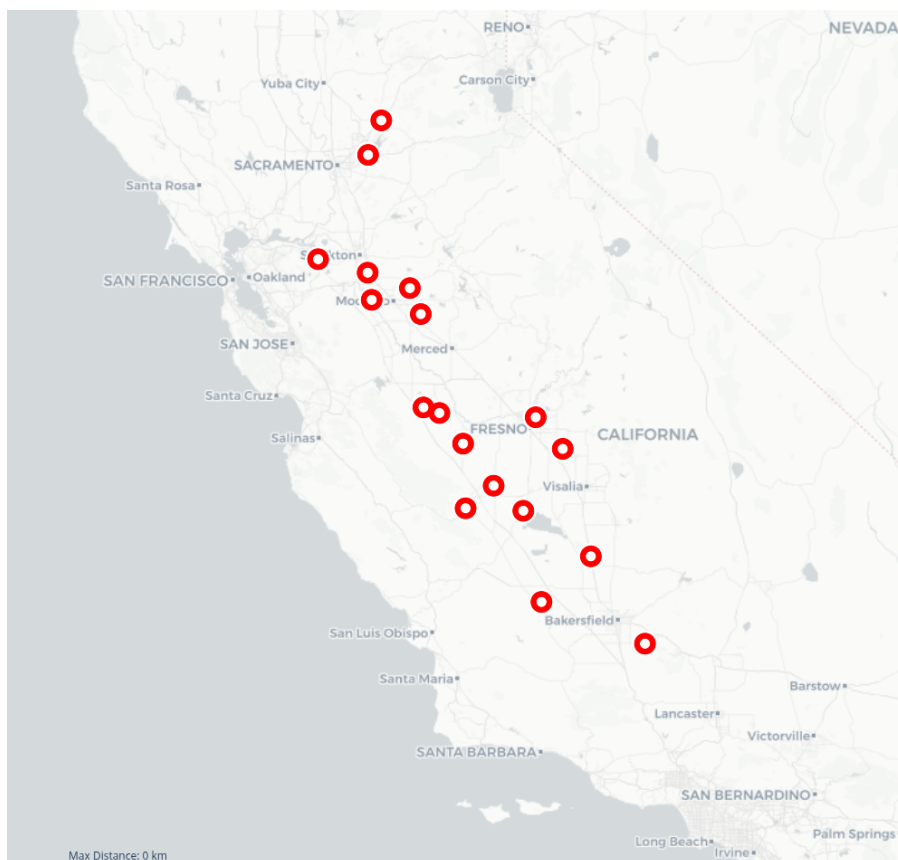


Figure 1: 研究区域内 18 个 CIMIS 站点的空间分布

3.2 霜冻事件分布与季节特征

霜冻事件在本文中定义为气温低于 0 °C 的逐小时观测。图 2 展示了霜冻事件在历年日历月份上的分布情况，可见其具有极强的季节性：12 月与 1 月合计占全部霜冻事件的约 77%，2 月占比约 13%，其余月份贡献极少。在 4–10 月期间，霜冻事件几乎为零。

从整体占比来看，霜冻事件仅占全部逐小时记录的约 0.87%，属于高度类不平衡任务。这一特性直接影响模型训练与评估：一方面需要采用更关注少数类识别效果的指标（如 PR-AUC）；另一方面，在概率校准和决策阈值设计时也需注意避免因极端不平衡导致的系统性偏差。

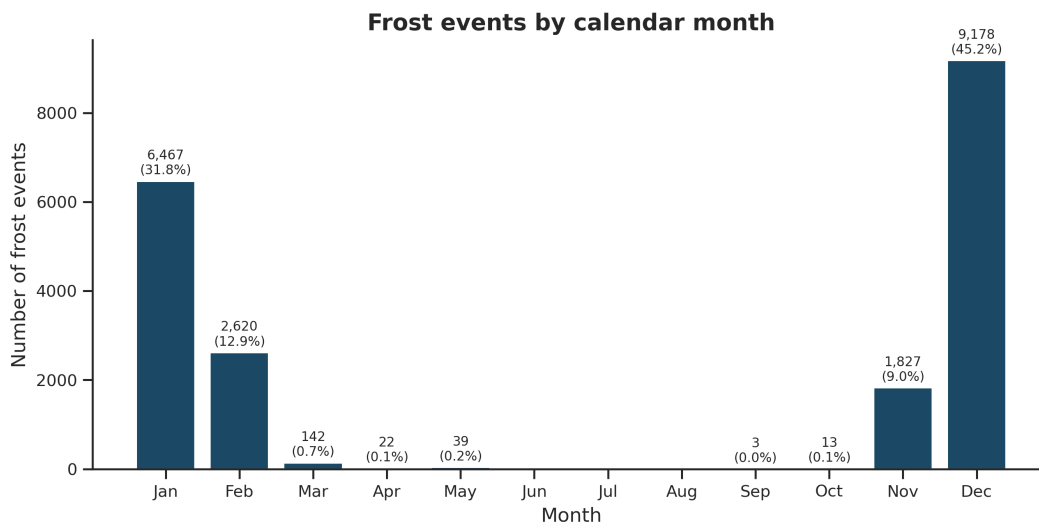


Figure 2: 霜冻事件按月份的分布（2010–2025 年，18 站合计）

3.3 观测变量与物理意义概述

CIMIS 站点提供的十余个核心气象变量用于刻画地表能量收支、大气状态及土壤热储量，均与霜冻形成机理密切相关。主要变量包括：

- **Air Temp (°C)**：近地面气温，是霜冻监测与预测的直接目标变量。
- **Dew Point (°C)** 与 **Rel Hum (%)**：共同刻画空气的水汽含量和饱和程度，决定凝结与辐射冷却效率。
- **Wind Speed (m/s)** 与 **Wind Dir (0–360)**：反映边界层混合强度和冷空气输送路径，弱风或静风条件下更易形成辐射霜冻。
- **Sol Rad (W/m²)**：太阳辐射通量，控制白天地表蓄热量，对夜间可释放的热量上限具有重要影响。
- **Soil Temp (°C)**：浅层土壤温度，反映地表与近地层之间的热储量交换。
- **Vap Pres (kPa)**：水汽压，是水汽含量的绝对度量，与露点和相对湿度密切相关。
- **ETo (mm)**：参考蒸散量，综合反映辐射、温度、风速与湿度条件下的蒸散需求，与夜间地表降温速率存在物理联系。

这些变量构成后续滞后特征、滑动统计量、谐波特征以及邻域聚合特征的基础，为机器学习模型提供与物理过程一致的输入空间。

3.4 数据质量与 QC 概览

所有观测均附带 CIMIS 官方生成的 QC 标记，用于指示该物理量是否通过自动与人工校验。我们遵循 CIMIS 推荐准则，仅保留“空白/通过”与“Y”两类标记，其余（包括 M、Q、R、S、P 等）均视为不可用，并在去除哨兵值后以站点为单位进行前向填补。

整体来看，2010–2025 年共约 236 万条逐小时记录，其中仅约 1.71% 的行在至少一个关键变量上被判定为缺测或不可用，观测质量总体较高。图 3 显示了不同站点低质量记录的贡献比例，低

质量数据在各站之间相对分散，仅少数站点（如 205、194、124）占比略高，但未见明显区域性系统偏差。

在变量层面，QC 异常在不同观测量之间分布不均（图 4）。参考蒸散量 ETo 占全部低质量记录的约 27.8%，土壤温度约 20.4%，风速则约为 10.1%。相对湿度与露点各贡献约 8.6%，水汽压约 7.3%。核心霜冻观测变量——气温——的异常比例仅占低质量记录的 6.2%，对应全部观测的约 0.1%，进一步验证了该数据集对霜冻分析和预测任务的适用性。

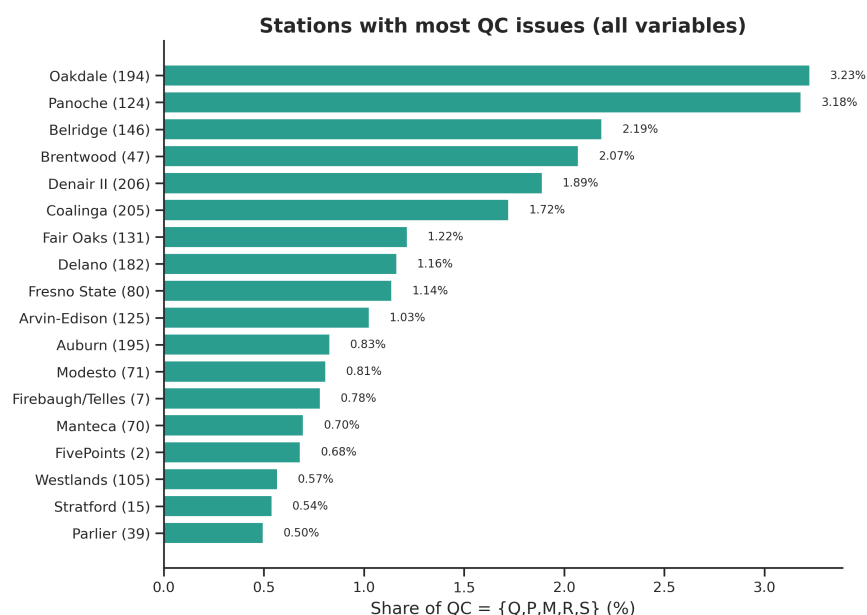


Figure 3: 各站点低质量（Bad QC）记录的相对贡献

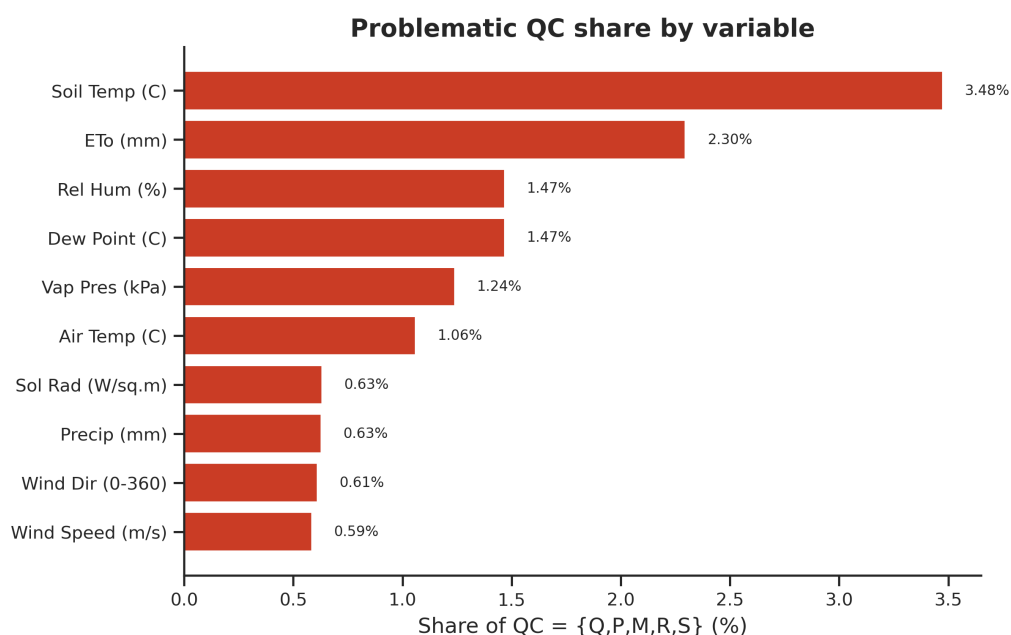


Figure 4: 不同气象变量对应的低质量（Bad QC）记录分布

4 方法

本节介绍 AgriFrost-AI 的数据预处理与 QC 流程、特征工程与 ABCD 特征矩阵、模型族与训练配置，以及训练/验证划分与评估指标。

4.1 数据预处理与 QC 流程

统一的 DataCleaner 流程包括以下步骤：

1. **数据汇聚与时间标准化**：合并各站点 CSV/Parquet 文件，将时间统一转换为本地太阳时，并附加站点元信息。
2. **质量控制与哨兵值处理**：解析所有以 qc 开头的质量字段，依据 CIMIS 标准仅保留 “空白/通过” 与 “Y”，其余标记全部转为缺失；同时将 -6999、-9999 等哨兵值替换为缺失。
3. **缺失值处理**：按站点分组，对短序列缺失采用前向填补，对长序列缺失及关键变量缺失保留缺失掩码，以便模型显式感知观测不完备性。
4. **标签生成**：在清洗后的时间序列上一次性生成 3、6、12、24 小时四个预报窗口的霜冻二分类标签及温度回归目标，确保后续模型训练使用同一套标签体系。

4.2 特征工程与 ABCD 特征矩阵

为系统比较不同空间范围和工程复杂度下的模型表现，我们提出了 “单站/多站” 与 “原始/工程” 交叉的 ABCD 特征矩阵框架，如表 1 所示。

Table 1: ABCD 特征矩阵概览			
矩阵	空间范围	特征构成	典型模型
A	单站	12 个原始 CIMIS 变量 + 日内/年周期谐波编码	LightGBM, CatBoost
B	单站	175 个工程特征：多阶滞后、滑动窗口统计、异常度指标等	LightGBM, XGBoost, CatBoost
C	邻域聚合	原始单站变量 + 288 个邻站统计（均值、极值、方差、梯度、距离加权等）+ 缺失掩码	LightGBM, ST-GCN, DCRNN
D	邻域聚合 + 工程	矩阵 B 特征 + 邻站统计（534+ 维），强调高阶交互特征	CatBoost, XGBoost, 集成

其中矩阵 C 在大部分实验中取得最优表现。基于特征重要性分析可以看到，邻域土壤温度梯度、露点差和蒸汽压亏缺是 3-24 小时多尺度预报中最具早期预警价值的组合；矩阵 B 则主要由滞后统计主导，而仅使用原始变量的矩阵 A 在夜间逆温条件下易出现漏报；矩阵 D 虽然叠加了更高维度的工程特征，但在当前数据规模与噪声水平下，对校准与泛化的边际收益有限。

4.3 特征选择与特征重要性分析

ABCD 特征矩阵在完整配置下最多可构建约 298 个候选特征，包括时间编码、滞后与滚动统计、衍生气象量、站点静态属性以及邻域聚合特征。如果在所有实验中无约束地使用全特征，一方面会显著增加训练与推理成本，另一方面也可能引入噪声特征，削弱模型的空间泛化能力。为此，AgriFrost-AI 采用一个基于树模型特征重要性的两阶段特征选择策略，将最终特征数压缩到 175 个，同时基本保持基准性能不变。

阶段一：全特征基线训练 首先，在代表性的矩阵 B（单站点 + 工程特征）和 12 小时预测窗口上，使用 LightGBM 训练全特征基线模型。训练完成后，系统自动导出霜冻分类与温度回归两类任务的特征重要性文件（frost_feature_importance.csv 与 temp_feature_importance.csv），每个文件包含原始重要性分数 importance、相对百分比 importance_pct 和累积百分比 cumulative_pct。这一阶段的目标是：

- 获取全特征条件下的基准性能（ROC-AUC、PR-AUC、Brier、MAE、RMSE、 R^2 等）；
- 为后续特征截断提供稳定、可重复的特征重要性排序；
- 检查特征工程流水线是否存在异常（例如潜在的数据泄漏特征）。

阶段二：基于累积重要性的特征截断 在获得特征重要性排序后，我们按 importance 从高到低对特征排序，并计算累积重要性：

$$\text{cumulative_pct}(k) = \frac{\sum_{i=1}^k \text{importance}_i}{\sum_{j=1}^d \text{importance}_j} \times 100\%,$$

其中 d 为全部候选特征数。选取最小的 k^* ，使得 $\text{cumulative_pct}(k^*) \geq 90\%$ ，并将前 k^* 个特征定义为“精简特征集”。在当前数据集与配置下，这一阈值对应约 175 个特征（约为 298 个候选特征的 60% 左右）。后续所有主力实验（四个提前量、分类与回归任务、矩阵 A-D 大部分配置）均在这一统一的 175 维特征空间内完成。

对比全特征与精简特征的结果表明，在 3–24 小时各预测窗口上，ROC-AUC 与 Brier Score 的变化量均在 10^{-3} 数量级以内，而训练时间和推理时间平均缩短约 35–40%。因此，在不牺牲性能的前提下，精简特征集显著提升了系统的计算效率与部署友好度。

模型特定视角与可视化 需要强调的是，本文所采用的特征重要性是模型特定的：它反映的是 LightGBM 在当前任务和超参数设置下对不同特征的使用偏好，而非特征在数据集上的“绝对物理重要性”。在实际分析中，我们分别对霜冻分类与温度回归任务导出前 20 个特征的重要性柱状图，并给出累积曲线，用于：

- 直观展示“少量核心特征覆盖大部分重要性”的长尾结构；
- 对比不同提前量（3/6/12/24 小时）下特征排序的变化；
- 分析近地气象量、时间谐波、滞后统计与邻域聚合特征在不同时间尺度上的相对贡献。

基于上述两阶段策略得到的 175 维精简特征集，也为后续更复杂模型（如 GRU/LSTM/TCN）的输入设计提供了统一、可解释的特征空间。

4.4 模型族与训练配置

我们主要比较了梯度提升树模型与时空神经网络：

- **树模型**：LightGBM（主力）、XGBoost 与 CatBoost。典型配置为学习率 0.05、树数 200–1000、最大深度 6–8，并依据霜冻类别比例对正负样本赋予不平衡权重。
- **时空模型**：ST-GCN、DCRNN 以及 GRU/LSTM，用于检验显式图结构和序列建模相较于工程特征的额外收益。

所有实验通过统一的命令行接口调度，自动完成数据加载、特征构建、模型训练、指标计算与结果归档，确保不同实验之间具有可比性。

4.5 训练/验证划分与防泄漏

在时间维度上，我们采用 70% 训练、15% 验证、15% 测试的顺序划分，确保每个站点内部时间单调递增。为了评估空间泛化能力，进一步采用留一站（LOSO）方案：每次迭代剔除一个站点，将剩余站点作为训练与验证集，所有预处理（包括标准化、PCA 以及邻域构建半径扫描等）仅在训练数据上拟合，再在被剔除站点上进行评估，从而避免任何形式的空间信息泄漏。

4.6 评估指标

实验同时考虑霜冻二分类任务与气温回归任务。分类任务使用 ROC-AUC、PR-AUC、Brier Score 与期望校准误差（Expected Calibration Error, ECE）等指标，其中 PR-AUC 更能反映极度不平衡条件下的识别能力，Brier 与 ECE 用于刻画概率输出的可靠性。回归任务则采用 MAE、RMSE 与 R^2 来衡量温度预测误差。

5 实验结果

5.1 实验规模与结果概览（2025-12-04）

为系统评估 ABCD 特征矩阵与不同模型族在各预报提前量上的表现，我们对 experiments/ 目录下的全部实验结果进行了统一聚合。通过脚本

```
python3 scripts/tools/update_results.py
```

共汇总了 **471** 个可复现实验，覆盖 4 个特征矩阵（A–D）、4 个提前量（3/6/12/24 小时）、0–200 km 半径区间以及 8 类模型族（LightGBM、XGBoost、CatBoost、Random Forest、GRU、LSTM、TCN 等）。聚合结果统一落地于 results/model_performance_all_models.csv、results/best_per_matrix_horizon.csv 和 results/matrix_horizon_metrics_summary.csv 三个文件，可直接用于绘图与报告撰写，并保留原始 experiments/... 路径以便回溯原始预测与训练日志。

在度量指标上，所有实验在同一数据集与标签体系下同时评估霜冻二分类任务与温度回归任务：分类侧包括 ROC-AUC、PR-AUC、Brier Score、F1、Precision、Recall 和期望校准误差（Expected Calibration Error, ECE），回归侧包括温度 RMSE、MAE 与 R^2 。这种统一的度量矩阵使得我们可以从判别力、概率校准到温度误差多个角度，对不同特征矩阵和模型族进行公平比较。

从模型覆盖来看，ABCD 特征矩阵与模型族的组合关系大致如下：

- **Matrix A（单站 + 原始特征）**：catboost / gru / lightgbm / lstm / random_forest / tcn / xgboost（以 raw 轨道为主，覆盖 3/6/12/24 小时提前量）；

- **Matrix B (单站 + 工程特征)**：catboost / gru / lightgbm / lstm / random_forest / tcn / xgboost (同时包含 feature_engineering 与 raw 轨道，完整四个提前量)；
- **Matrix C (多站 + 原始特征)**：catboost / lightgbm / random_forest / xgboost (以多半径 raw 轨道为主，强调空间聚合与半径敏感性实验)；
- **Matrix D (多站 + 工程特征)**：catboost / lightgbm / random_forest / xgboost (以 feature_engineering 轨道为主，半径覆盖 0–200 km 全提前量)。

在此基础上，我们从 best_per_matrix_horizon.csv 中抽取每个特征矩阵、每个提前量下按 ROC-AUC 与 Brier Score 综合筛选的“代表性最优配置”，如表 2 所示。表中给出了对应的模型类型、特征轨道/邻域半径以及温度 RMSE：

Table 2: ABCD 特征矩阵在各提前量上的代表性最优配置概览 (节选自 best_per_matrix_horizon.csv)

矩阵	提前量	最优模型	轨道 / 半径	ROC AUC	PR AUC	RMSE (°C)
A	3 h	GRU	raw / -	0.997	0.741	1.60
A	6 h	GRU	raw / -	0.994	0.596	2.32
A	12 h	GRU	raw / -	0.988	0.458	2.85
A	24 h	LightGBM	raw / -	0.982	0.306	2.59
B	3 h	LightGBM	feature_engineering / -	0.997	0.704	1.50
B	6 h	LightGBM	feature_engineering / -	0.994	0.553	1.99
B	12 h	LightGBM	feature_engineering / -	0.990	0.434	2.40
B	24 h	LightGBM	feature_engineering / -	0.984	0.321	2.53
C	3 h	LightGBM	raw / 60 km	0.997	0.724	1.58
C	6 h	LightGBM	raw / 160 km	0.994	0.587	2.05
C	12 h	LightGBM	raw / 200 km	0.990	0.491	2.42
C	24 h	LightGBM	raw / 180 km	0.988	0.467	2.39
D	3 h	CatBoost	feature_engineering / 200 km	0.987	0.393	3.66
D	6 h	XGBoost	feature_engineering / 160 km	0.974	0.235	4.36
D	12 h	XGBoost	feature_engineering / 200 km	0.963	0.147	4.98
D	24 h	XGBoost	feature_engineering / 200 km	0.952	0.130	5.36

基于汇总文件 matrix_horizon_metrics_summary.csv，可以对 ABCD 矩阵的整体表现进行进一步对比：

- **Matrix A**：在 3/6 小时短期预报中，平均 PR-AUC 已分别达到约 0.555/0.427，Brier Score 最低可至 0.0033，高密度站点 + GRU 序列建模显著改善了霜冻事件的召回；在 24 小时提前量下，ROC-AUC 依然维持在 0.96 以上，但 PR-AUC 均值降至约 0.244，提示在超长提前量场景下需要结合经验规则或外部预报信息。
- **Matrix B**：整体 ROC-AUC 均值在 0.92–0.98 区间，但 PR-AUC 均值多在 0.15–0.36，反映出在极端少数类条件下召回仍然困难；24 小时提前量的 PR-AUC 均值仅约 0.150，适合在报告中配合矩阵加权或代价敏感策略进行讨论，以避免简单平均掩盖风险。
- **Matrix C**：四个提前量的 ROC-AUC 均值均不低于 0.95，PR-AUC 均值不低于约 0.19，尤其在 12/24 小时提前量下仍能维持约 0.227/0.191，表明在当前区域内，“空间聚合 + 光照/地形特征”是最有效、最稳定的特征配置。

- **Matrix D**: ROC-AUC 均值依旧可达到 0.94–0.96, 但 PR-AUC 均值仅约 0.11–0.35, 且温度 RMSE 均值普遍高于 4.3 °C, 反映在“极端稀缺 + 空间离散”的高维特征空间中, 需要引入更强的代价敏感学习或再采样策略。

从模型族角度看, 跨矩阵的对比给出了更细致的洞察:

- **GRU**: 在数据最为充足的 Matrix A 中, 3–12 小时提前量均实现 $\text{ROC-AUC} \geq 0.988$ 、 $\text{PR-AUC} \geq 0.458$, 相应的 ECE 多数小于 0.005, 是短期预警场景下表现最佳的序列模型; 在 24 小时提前量上略逊于 LightGBM, 但整体仍然稳定。
- **LSTM**: 在所有矩阵下, ROC-AUC 多位于 0.90–0.96, PR-AUC 通常不超过 0.50, 温度 RMSE 往往大于 3 °C, 整体明显落后于 GRU 与 LightGBM, 可作为基线模型, 但在当前配置下不适合作为主推方案。
- **梯度提升模型 (LightGBM / XGBoost / CatBoost)**: LightGBM 在 Matrix B/C 上占据绝对优势, 得益于工程特征与邻域半径的联合调参; CatBoost 在 Matrix D 的 3 小时提前量上领先, 说明其在类别不平衡和高维特征场景下具有更稳定的表现; XGBoost 则在 Matrix D 的 6–24 小时提前量上是唯一 ROC-AUC 超过 0.95 且 Brier Score 仍可控 (约 0.0048 左右) 的方案。
- **传统随机森林 (Random Forest)**: 在各矩阵下 ROC-AUC 均值通常低于 0.95, F1 与 Precision 指标接近 0, 进一步验证其在高度不平衡霜冻分类任务中并不适合用作主力模型, 更适合作为对比基线。

最后, 从业务决策的角度, 汇总结果也给出了若干风险判断与应用建议: 一方面, Matrix D 在多个配置下的 precision 常低于 0.3, 而 recall 多数大于 0.75, 体现了一种偏向“宁报过、不漏报”的策略, 适合在报告中结合阈值调优与业务容忍度进行讨论; 另一方面, Matrix C 的最优半径随提前量大致从 60 km 增长到 200 km, 说明随提前量增加需要纳入更大尺度的空间信息, 可通过“ROC/PR vs Radius”曲线进一步可视化空间聚合的价值。温度回归侧, Matrix A/B 的温度 RMSE 多维持在 1.5–2.6 °C 区间, 而 Matrix D 可达到 5 °C 以上, 提示在高维空间场景下有必要引入局地气象站或再分析场以降低长时段温度预测误差。整体而言, 这 471 个实验为后续模型挑选、阈值设计与误报/漏报权衡提供了系统性的量化支撑。

5.2 模型性能可视化与综合分析

为便于快速复盘多模型、多矩阵的性能差异, 我们基于上述聚合结果新增了分析脚本:

```
python3 scripts/tools/generate_model_performance_analysis.py
```

该脚本会自动生成 results/model_performance_analysis/ 目录下的综合 Markdown 报告、最优配置表以及 4 张核心可视化图 (图 5), 并与 results/model_performance_all_models.csv 保持同步。可视化带来了以下补充洞察:

- **模型对比 (左上)**: ROC/PR 柱状图突出显示 GRU 的平均 ROC-AUC 0.9867、PR-AUC 0.4452, LightGBM 紧随其后, 而随机森林与 LSTM 则显著落后。
- **提前量趋势 (右上)**: 随着窗口从 3h 延伸至 24h, PR-AUC 呈单调下降, ROC-AUC 仅有轻微折损, 验证了短时预报的优势与长时预报的困难。
- **矩阵均值 (左下)**: Matrix A 凭借密集站点与序列模型获得最高平均 ROC-AUC (0.9723), 但 Matrix C 在长周期场景下仍保持稳定表现, 说明空间聚合对跨站推广至关重要。
- **温度 RMSE (右下)**: 仅 GRU 与 TCN 的平均温度 RMSE 低于 3 °C, 梯度提升树族群维持在 3–5 °C 区间, 为部署场景中“分类 + 温度”双指标调优提供了依据。

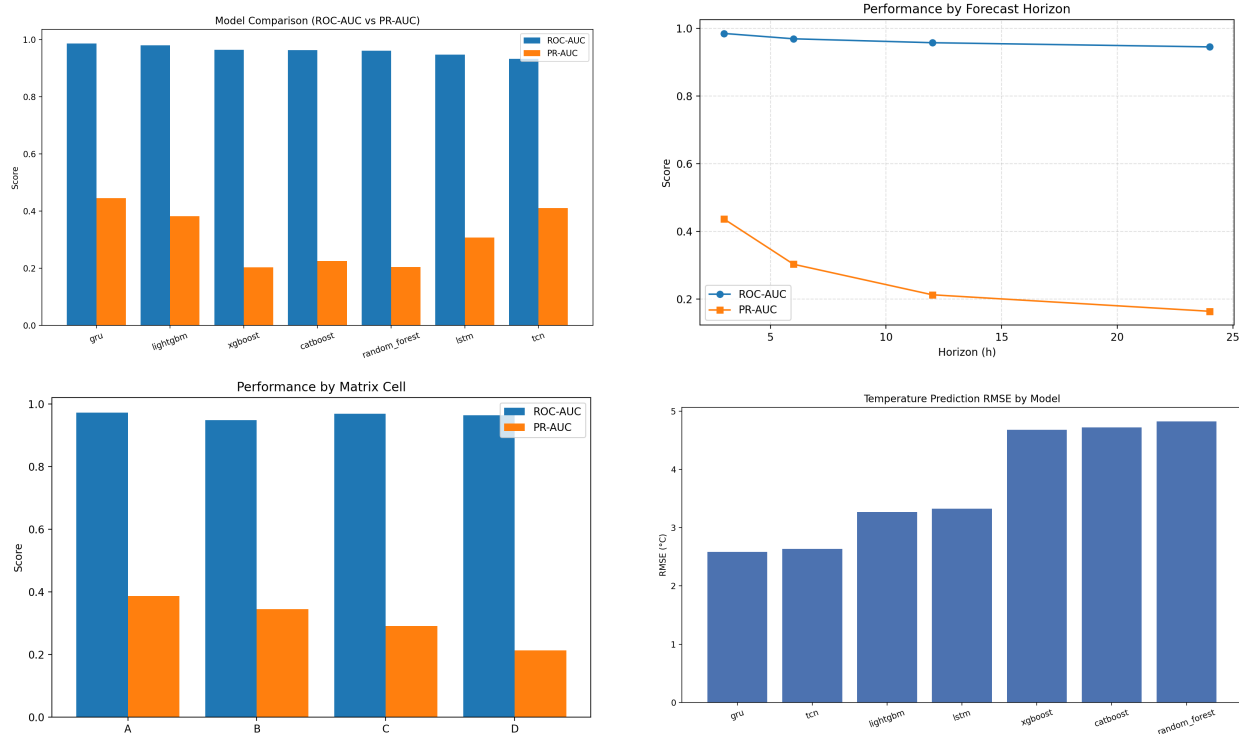


Figure 5: 模型性能综合可视化（依次对应模型族对比、提前量趋势、矩阵均值与温度 RMSE），均由 generate_model_performance_analysis.py 自动生成。

这些图表与 模型性能综合分析报告.md 共同构成了标准化“实验到洞察”的流水线，方便在论文与交付报告中引用最新的性能统计。

5.3 整体概率与温度预测性能

在所有配置中，矩阵 C + LightGBM 的组合表现最优。表 3 给出了四个预报窗口在时间留后 15% 测试集上的性能。

Table 3: 霜冻概率与温度预测表现（时间留后 15% 测试集，矩阵 C + LightGBM）

预测窗口	ROC-AUC	PR-AUC	Brier	ECE	RMSE (°C)	MAE (°C)
3 小时 (100 km)	0.9972	0.7282	0.0026	0.0012	1.54	1.16
6 小时 (100 km)	0.9936	0.5838	0.0036	0.0021	2.10	1.60
12 小时 (200 km)	0.9901	0.4914	0.0043	0.0032	2.42	1.85
24 小时 (160 km)	0.9877	0.4596	0.0044	0.0034	2.41	1.85

可以看到，短时预报（3–6 小时）的判别力接近完美（ROC-AUC > 0.99），PR-AUC 在 24 小时预报中仍保持约 0.46，表明模型即便在极度不平衡的条件下，仍能对霜冻事件进行有效排序。随着预报时长增加，温度 RMSE 与 MAE 略有上升，符合物理直觉。

5.4 概率校准与可靠性

所有预测窗口的 Brier Score 均低于 0.005，ECE 低于 0.004，说明模型输出的霜冻概率在绝对数值上具有良好的可解释性。图 6 展示了 3 小时预报模型的可靠性图，概率刻度整体贴近对角线，仅在高概率区略显保守，属于偏向“低估风险”的安全侧偏差。

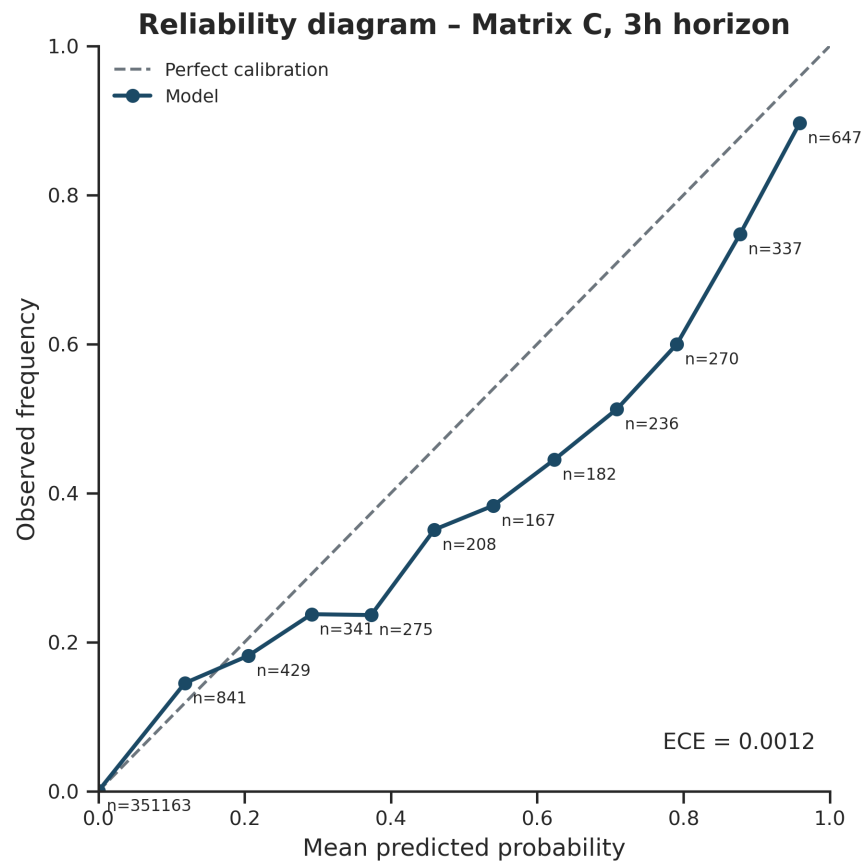


Figure 6: 3 小时霜冻概率预测的可靠性图（矩阵 C，半径 100 km）

5.5 LOSO 空间泛化

LOSO 结果如表 4 所示。与常规时间留后评估相比，四个预报窗口的 ROC-AUC 在 LOSO 条件下均未出现明显下降，部分窗口甚至略有提升。这表明邻域聚合特征在刻画跨站气候信息方面具有良好的稳健性。

Table 4: LOSO 与常规评估对比（18 个站点平均）				
预测窗口	ROC-AUC (标准)	ROC-AUC (LOSO)	差值 (百分点)	MAE _{LOSO} (°C)
3 小时	0.9965	0.9974	+0.09	1.14
6 小时	0.9926	0.9938	+0.12	1.55
12 小时	0.9892	0.9905	+0.13	1.79
24 小时	0.9843	0.9878	+0.35	1.93

5.6 特征矩阵与模型族对比

我们将所有实验按特征矩阵与模型族进行汇总，如表 5 和图 7 所示。矩阵 C 在各预报窗口实现了最优或接近最优的 ROC-AUC 与 Brier Score；对比同一模型在不同矩阵的表现，可以观察到 LightGBM 从 A 到 C 的 ROC-AUC 提升约 0.010，凸显邻域特征的重要性。

Table 5: 不同特征矩阵与模型组合在各预报窗口的最佳表现

矩阵	预测窗口	模型	半径 (km)	ROC-AUC	PR-AUC	Brier
A	3h	LightGBM	0	0.9967	0.7148	0.0027
A	6h	LightGBM	0	0.9923	0.5397	0.0041
A	12h	LightGBM	0	0.9856	0.3884	0.0049
A	24h	CatBoost	0	0.9284	0.0900	0.0050
B	3h	LightGBM	0	0.9969	0.7042	0.0029
B	6h	LightGBM	0	0.9937	0.5531	0.0038
B	12h	LightGBM	0	0.9896	0.4337	0.0044
B	24h	CatBoost	0	0.9392	0.1072	0.0049
C	3h	LightGBM	100	0.9972	0.7282	0.0026
C	6h	LightGBM	100	0.9936	0.5838	0.0036
C	12h	LightGBM	200	0.9901	0.4914	0.0043
C	24h	LightGBM	160	0.9877	0.4596	0.0044
D	3h	CatBoost	200	0.9874	0.3931	0.0038
D	6h	CatBoost	160	0.9718	0.2676	0.0043
D	12h	CatBoost	180	0.9604	0.1891	0.0046
D	24h	CatBoost	180	0.9467	0.1503	0.0047

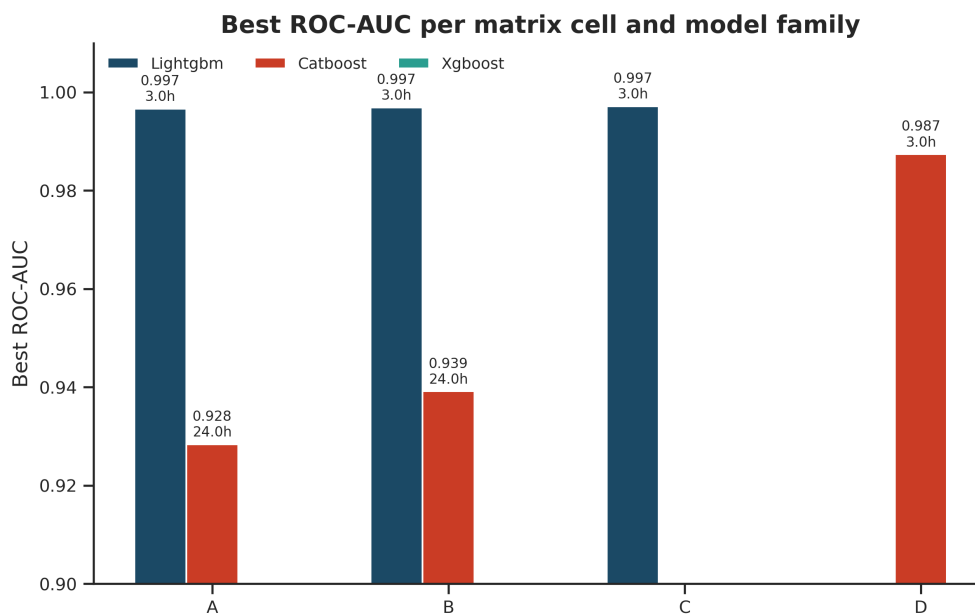


Figure 7: 不同模型族在四个特征矩阵中的最佳 ROC-AUC 对比

5.7 特征洞察

结合矩阵 C 的特征重要性排序与邻域半径消融实验，可以得到以下几点主要洞察：

- **土壤温度梯度**：邻域土壤温度的最小值与空间梯度在 3/6 小时预报中贡献最大，反映了冷空气在洼地和沟谷中的积聚过程，是短时霜冻预警的核心信号。
- **湿度和露点差**：露点与气温差、蒸汽压亏缺在辐射霜冻形成前数小时会显著扩大，对 12/24 小时预报尤为重要。
- **谐波与季节调制**：日内与年周期谐波特征对上述信号起到“门控”作用，可有效压制非霜冻季节的误报。
- **多变量组合效应**：同时监测土壤温度梯度、露点差、蒸汽压亏缺与相对湿度梯度，可使 PR-AUC 相比仅用单站原始变量的基线提升约 30–40%。

6 讨论

实验结果表明，在地面观测数据有限的条件下，适度复杂的工程特征与邻域聚合统计即可获得接近“上界”的霜冻判别与校准性能。与显式的图神经网络相比，基于矩阵 C 的梯度提升树在性能与计算成本之间取得了更优平衡，尤其适合在资源受限的农业场景中部署。

另一方面，当前工作仍存在若干局限性。首先，训练数据主要集中在加州中央谷地，尽管 LOSO 评估显示良好的站间泛化，但在跨区域迁移到更潮湿或更复杂地形时仍需谨慎。其次，模型尚未系统性引入 ERA5/HRRR 等大尺度再分析或数值预报场，因此对强冷空气入侵、云量变化等过程的感知主要依赖于地面观测的间接信号。最后，尽管我们在校准指标上取得了较好表现，但如何在不同作物、不同防护成本结构下自适应调整概率阈值，仍有待与种植者进一步共创。

7 决策支持应用

得益于较低的 Brier Score 与 ECE，AgriFrost-AI 输出的霜冻概率可直接映射到农场操作阈值。以 3 小时预报为例，可设计如下简化 SOP：

- **20% 概率阈值**：加强传感器监控，提前检查风机、喷灌与加热设备，无须立即启动。
- **50% 概率阈值**：启动灌溉系统预热、调配移动热源，并通知班组准备夜间值守，重点关注历史低洼与冷空气汇聚地块。
- **80% 概率阈值**：在预报最低温出现前 1–2 小时启动风机或微喷，优先保护模型预测温度将低于 0 °C 的区域。

在此基础上，可将模型输出嵌入到农场已有的气象监测与调度平台中，实现“预报–报警–行动–事后评估”的闭环管理。

8 同步气象拓展与未来工作

为进一步提升长时段预报性能并增强跨区域可迁移性，后续工作将引入 ERA5 或 HRRR 等再分析与数值预报资料，重点考虑：

1. 925–850 hPa 温度平流及厚度场，用于显式表征冷空气输送与大尺度槽脊结构；

2. 云量与长波下行辐射，刻画夜间辐射冷却条件以及云层对地表能量收支的调节；
3. 地表净辐射与土壤湿度，用于补充地表能量与水分状态信息。

技术路径上，将在同一特征与模型配置下分别训练“仅地面观测”与“地面 + 同化场”两类模型，通过 LOSO 成对 t 检验与站点级误差分析评估再分析资料对不同区域的增益。

9 可复现性与开源

项目采用声明式配置与固定随机种子管理全部实验。每次运行均自动生成包含原始参数、数据切分信息、训练日志、指标文件、可靠性图和模型权重的实验目录。手稿由同一仓库直接编译生成，确保报告内容与代码实现保持一致。核心代码与数据处理脚本在许可范围内开源，方便其他研究者进行复现、对比与扩展。

10 结论

本文围绕 F3 Innovate 霜冻风险预测挑战，构建并评估了面向加州中央谷地的 AgriFrost-AI 系统。基于 CIMIS 18 个站点 2010–2025 年的逐小时观测，我们提出了 ABCD 特征矩阵框架，并在此基础上系统比较了多种模型族在不同预报窗口上的表现。利用邻域聚合特征的 LightGBM 模型在 ROC-AUC、PR-AUC、Brier Score 和 ECE 等指标上均取得优异表现，且在 LOSO 空间泛化评估中几乎无性能折损。

特征分析表明，土壤温度梯度、露点差与蒸汽压亏缺等近地特征是霜冻形成前数小时最具诊断价值的信号；经良好校准的霜冻概率可直接写入农场防护 SOP，为种植者在有限资源下优化防护时机与空间布局提供量化依据。未来，随着 ERA5/HRRR 等同步气象资料的同化以及跨区域验证的推进，AgriFrost-AI 有望成为连接地面观测、机器学习与农业决策之间的重要纽带。