Data Slayers : Frost Risk Forecasting Data Challenge

Austin Tathong
Donovan Butler

December 2025

1 Modeling Pipeline

1.1 Challenge Background

Frost is one of the most damaging weather risks for California agriculture. In the United States, the economic losses due to frost damage exceed all other weather-related phenomena. This challenge required us to build short-term frost risk forecasting models using 15 years of hourly data from 18 CIMIS weather stations across diverse topographies and climates within the Central Valley. Our goal was to evaluate whether machine learning models can be used as a viable alternative or even better predictor than traditional heuristics.

1.2 Overall Objectives

The task was to develop a machine learning model that predicts the probability of a frost event ($T < 0$°C) and the expected temperature (T) and compare it with baseline traditional methods within horizons of 3, 6, 12, and 24 hours.

1.3 Dataset Used

The challenge utilized the `CIMIS Hourly Multi-Station Dataset` [1], comprising 15 years of meteorological observations collected from 18 distinct weather stations across California's Central Valley various different stations, such as `2-FivePoints, 7-Firebaugh, 15-Stratford, 39-Parlier, 47-Brentwood, 70-Manteca, 71-Modesto,` and others, represent a wide range of elevations, microclimates, and agricultural regions.

Each station provided a structured `.csv` file containing station specific meta data (Station ID, Date, Hour, etc.) as well as hourly measurements of key environmental variables (Air Temperature, Humidity, Wind Speed, Solar Radiation, etc).

Note: Optional external datasets (`ERA5/HRRR`) could have been utilized, but due to time constraints we opted to focus on near-surface variables.

1.4 Preprocessing

The initial step in training a machine learning model for data prediction is to ensure the accuracy of the data through preprocessing. Our goal during this phase was to thoroughly clean the datasets, engineer relevant features, and create labels for the model to learn from over time. Although the `CIMIS` dataset was partially cleaned, further preprocessing was necessary to enhance data integrity, prevent temporal leakage, and develop meaningful features.

Note: All preprocessing steps were performed on the complete dataset, ensuring that the outputs were properly labeled for effective model training.

In order to properly clean the data for further analysis and feature engineering, we applied the following rules:

- Removed quality control (QC) flag columns
- Replaced all spaces (" ") in column names to underscores ("_")
    - Example: `"Air Temp (C)"` → `"Air_Temp_(C)"`

This was essential because quality control was not a variable needed during the model training process. Additionally, replacing spaces with underscores helped prevent potential warnings that could arise from the use of spaces.

After cleaning our data, we proceeded to prune and select our features. We opted to include most variables from the datasets, excluding those related to the relevant station metadata (`'Stn_Id'`, `'Stn_Name'`, `'CIMIS_Region'`). Additionally, we decided to exclude the variables for (`'Date'`, `'Hour_(PST)'`, `'Jul'`) because we encoded these variables in sin/cos format, which will be detailed later in the next following process.

Next is feature creation. Based on the variables we chose to include, we also developed additional features derived from the original ones in the dataset. We created the following features:

- **Dewpoint Depression (T_a - T_d)** – `dewpoint_dep`
    - Large depressions → efficient radiative cooling (air dry)
    - Small depressions → frost unlikely until saturation
- **Is Dim (Little Sunlight)** - `is_dim`
    - True → more likely of frost due to little sunlight
    - False → less likely of frost due to sunlight
- **Is Very Calm Winds** - `is_very_calm`
    - True → more likely frost due to weak mixing
    - False → less likely frost due to air mixing warming surfaces

- **Is Winter** - `is_winter`
    - True → frost more climatologically likely
    - False → frost less likely (warmer seasons)
- **Hour Cyclical** - `hour_sin, hour_cos`
    - Certain hour angles → frost most likely near sunrise
    - Daytime angles → frost unlikely
- **Day Cyclical** - `day_sin, day_cos`
    - Winter angle range → frost more common
    - Summer angle range → frost rare
- **Air Temp Rolling Min** - `air_temp_roll_min{H}h`
    - Low values → recently cold → frost more likely
    - Higher values → less recent cold → frost less likely
- **Air Temp Change** - `air_temp_change_{H}h`
    - Large negative → cooling trend → frost more likely
    - Positive/near zero → warming/stable → frost less likely
- **Dew Point Rolling Min** - `dew_point_roll_min{H}h`
    - Low values → dry air → needs more cooling → frost less immediate
    - High values → moist air → saturation close → frost more likely
- **Dew Point Change** - `dew_point_change_{H}h`
    - Positive → moistening → saturation closer → frost more likely
    - Negative → drying → frost less likely
- **Wind Speed Rolling Min** — `wind_spd_roll_min_{H}h`
    - Low values → long calm periods → strong cooling → frost more likely
    - Higher mins → mixing present → frost less likely
- **Wind Speed Change** — `wind_spd_change_{H}h`
    - Negative → winds weakening → frost more likely
    - Positive → winds increasing → frost less likely
- **Solar Radiation Rolling Min** — `sol_rad_roll_min_{H}h`
    - Low values → extended darkness/clouds → frost more likely
    - Higher mins → some sunlight present → frost less likely
- **Solar Radiation Change** — `sol_rad_change_{H}h`
    - Positive → sunrise/clearing → warming → frost less likely
    - Negative → sunset/clouding → cooling → frost more likely

Finally in our preprocessing pipeline is our creating for target values, for our model to fit towards. These were our following targets:

- **Frost Classification Target** — `frost_{H}h`
    - True (1) → Air temperature drops below 0°C within next H hours
    - False (0) → Air temperature stays ≥ 0°C within next H hours
    - Purpose → binary frost prediction (yes/no frost event ahead)

- **Temperature Regression Target** — `temp_{H}h`
    - Value = actual air temperature H hours in the future
    - Purpose → continuous prediction of future temperature trajectory

Note: `H` is a variable representing the forecasting horizon, meaning the model predicts frost likelihood and temperature `H` hours into the future (3, 6, 12, or 24 hours ahead).

Through doing each one of these steps in order it allows us to ensure a safe dataset to work from, extract more meaning through the data through the feature selection, and ensure our model is fitting towards the targets which are to classify for frost event and to regress to a temperature within a {H} hour horizon.

1.5 Models

For the baseline we evaluated three baselines commonly used in environmental and time-series predictions: climatology, dew point and persistence. Climatology is a baseline that predicts frost likelihood based solely on the historical frost frequency for the current month. Dew Point predicts frost likelihood using frost rates conditioned on the current dew point temperature. Persistence simply assumes that if frost is occurring now, it will continue; if not, frost is unlikely in the near future. With these three models, we created a baseline with which to compare the performance of the more advanced machine learning models.

We successfully trained a total of 160 machine learning models. This follows from the need to develop two model types, a classifier for forecasting frost events and a regressor for predicting future temperature, across four forecasting horizons: 3, 6, 12, and 24 hours. Our full framework consisted of 18 teacher models, one student model, and one general model, each trained at every horizon. This structure allowed us to evaluate generalization, station-specific learning, and the benefits of knowledge distillation at multiple temporal scales.

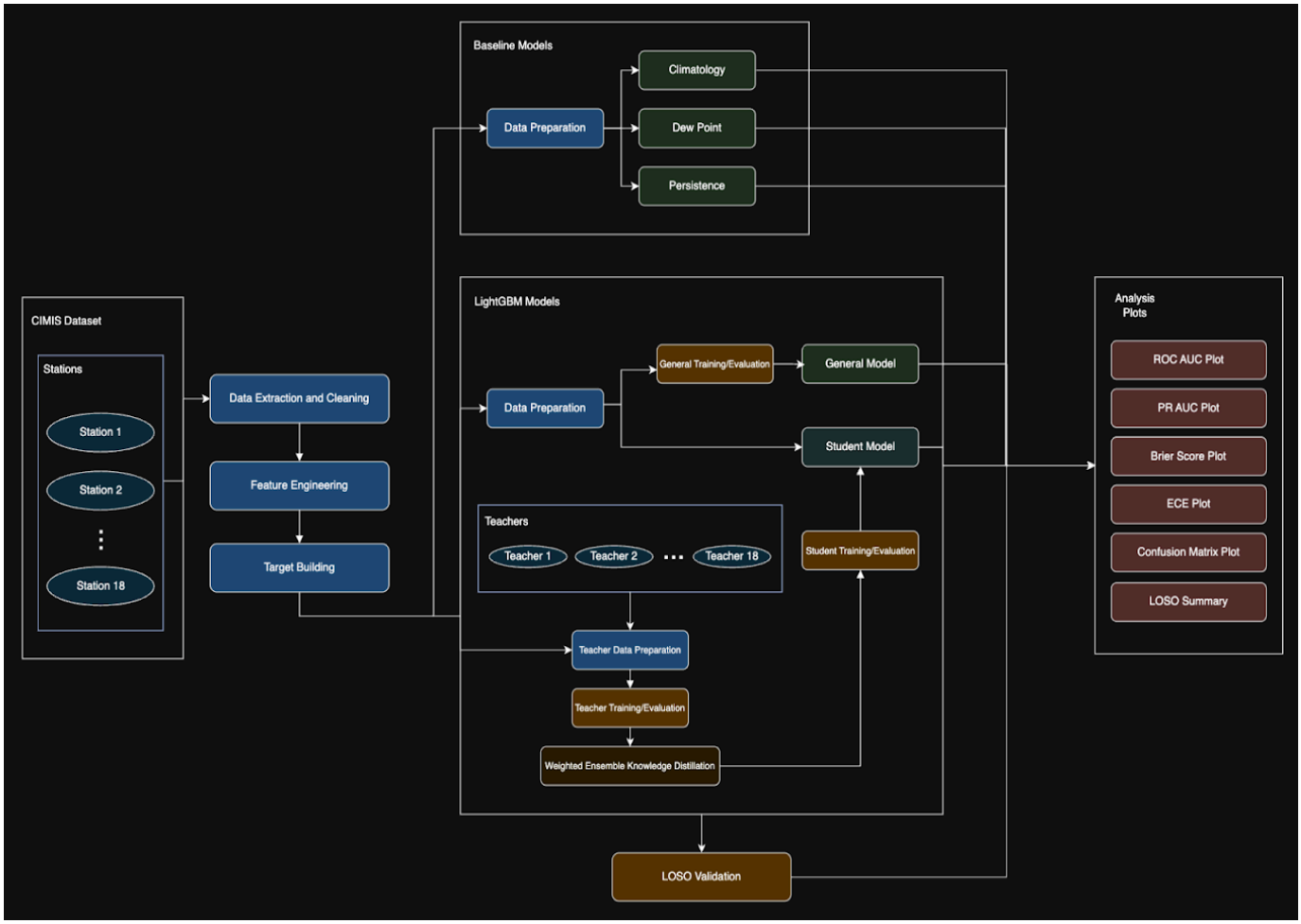
The machine learning model we selected for forecasting frost events and future temperatures was LightGBM (Light Gradient Boosting Machine), a tree-based ensemble method. We chose LightGBM because tree-based models perform exceptionally well on tabular datasets and offer strong modularity for feature engineering. Additionally, LightGBM's boosting framework, where each new tree learns from the errors of the previous ones, allows the model to iteratively refine its predictions by adjusting weights over time. This gives us many of the optimization advantages associated with deep learning models, while retaining the speed, efficiency, and interpretability of a tree-based approach. This was done for future scenarios where the model could be embedded into robotic systems for real-time environmental monitoring.

We trained and evaluated two variations of the LightGBM model. The first was a general model trained on data from all stations in the dataset. The second used an ensemble knowledge-distillation approach

with a custom weighting mechanism. In this setup, multiple Teacher LightGBM models were trained individually on specific stations, and their learned patterns were used to guide and influence the training of a Student LightGBM model that was trained on the full multi-station dataset.

To convert predicted frost probabilities into binary frost-event decisions, we created a hyperparameter called `threshold` where it allows us to choose how confident the model has to be before we classify it. We the variable `threshold` to 0.3. That is, the model predicts a frost event for a given forecasting horizon when its estimated probability exceeds 0.3. This threshold was selected to balance sensitivity and specificity in a way that prioritizes early frost detection, which is more valuable operationally than minimizing false alarms.

1.6 Pipeline Diagram:



This diagram illustrates the full workflow of our system, from data extraction and feature construction, to model deployment across the various architectures, to the generation of performance plots used for analysis.

1.7 Calibration & Reliability

To evaluate the quality of our forecasting with our model, various calibration and reliability metrics were used:

- **Brier Score**
  - The mean squared error between the forecast probability and the observed event, and how far probabilities are from the truth
- **Reliability Diagram**
  - A visual diagnostic showing whether predicted probabilities match real outcomes
- **Expected Calibration Error (ECE)**
  - A summary statistic that measures how well the probabilities are calibrated
- **Area Under the Precision-Recall Curve (PR-AUC)**
  - A discrimination statistic that measures how well the model finds positives when they're rare
- **Area Under the Receiver Operating Characteristic (ROC-AUC)**
  - A summary statistic that measures how well the probabilities are calibrated to the overall data
- **Normalized Confusion Matrix**
  - A matrix that shows the ratio of how many predictive positives were actually positives and how many predicted negatives were actually negative.

1.8 Baseline Models

Before evaluating machine learning models, it is important to establish how well simple, non-ML models perform on the frost risk prediction. These baseline models provide the minimum performance level that any advanced model must surpass to be considered useful.

For the baseline model development we started by concatenating all the stations into one dataframe. From the `Date` and `Hour (PST)` columns we were able to create a new timestamp column called `Datetime` (formatted as the date followed by the hour of day). We handled 24:00 cases and added a `Stn Id` column for grouping. We then narrowed the dataframe features to just `Air Temp (C)` and frost_now, and built frost labels for 3, 6, 12, and 24 hour horizons. With this we were then able to evaluate the different baseline models.

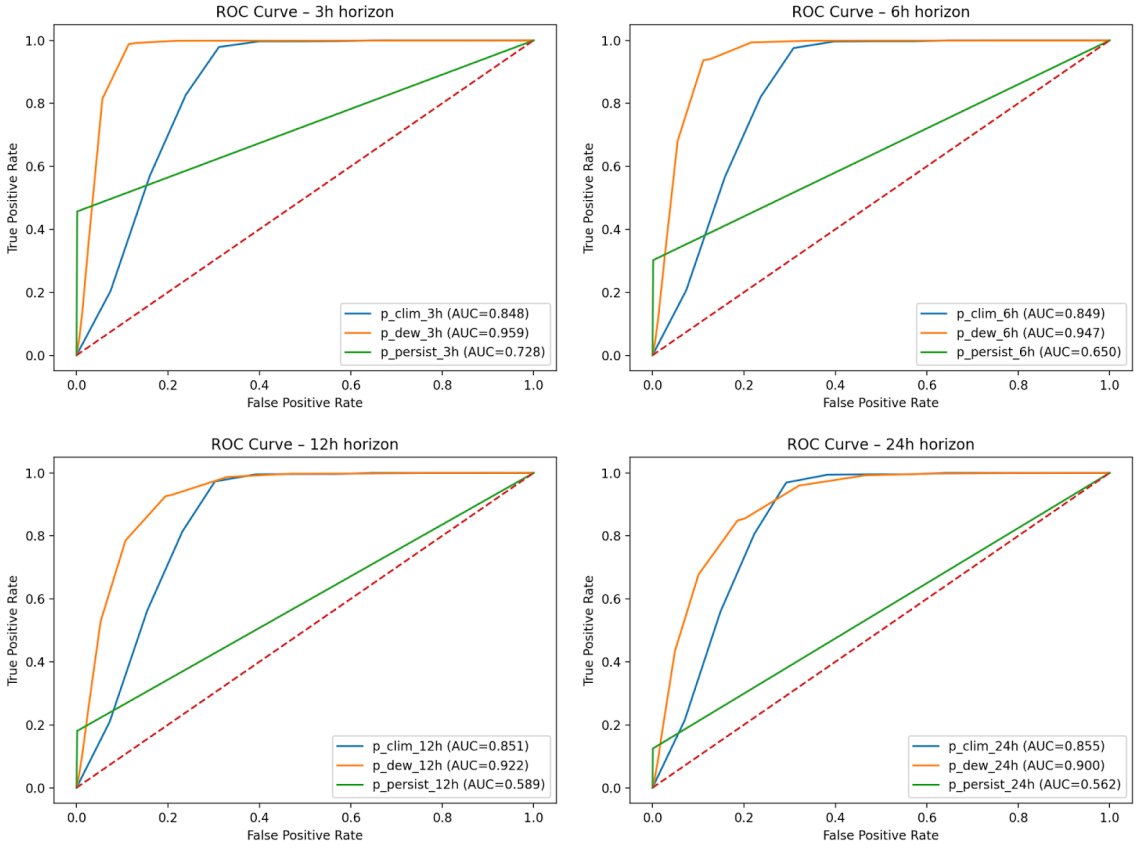
First, we created a climatology baseline which for each horizon computed the probability of frost as the historical frost frequency for that given month.
For the dew-point baseline, we split dew points into bins like -20℃ to -18℃, -18℃ to -16℃ all the way to +18℃ to +20℃. For each bin, it looks at the historical chance of frost when dew point was in that range, essentially computing the frost probability based on similar dew-point conditions in past data.

The persistence baseline was the most simple. Essentially, if there is frost now, then there will be frost soon (in the next hour). It does not consider any temperature trends, dew point, season, time or anything else, it just outputs simple binary probabilities based on the current frost status.

From these models we created a csv containing the probabilities output from each baseline, labeled as `frost_next_{H}h, p_clim_{H}h, p_dew_{H}h, and p_persist_{H}h`. This allowed us to plot ROC curves, PR curves, calibration and reliability curves, and develop other metrics to measure the baseline models' performance. We will now analyze and discuss these results.

1.8.1 ROC Curve Analysis (Discrimination Ability)

The ROC curves illustrate how well each baseline distinguishes frost from non-frost conditions.



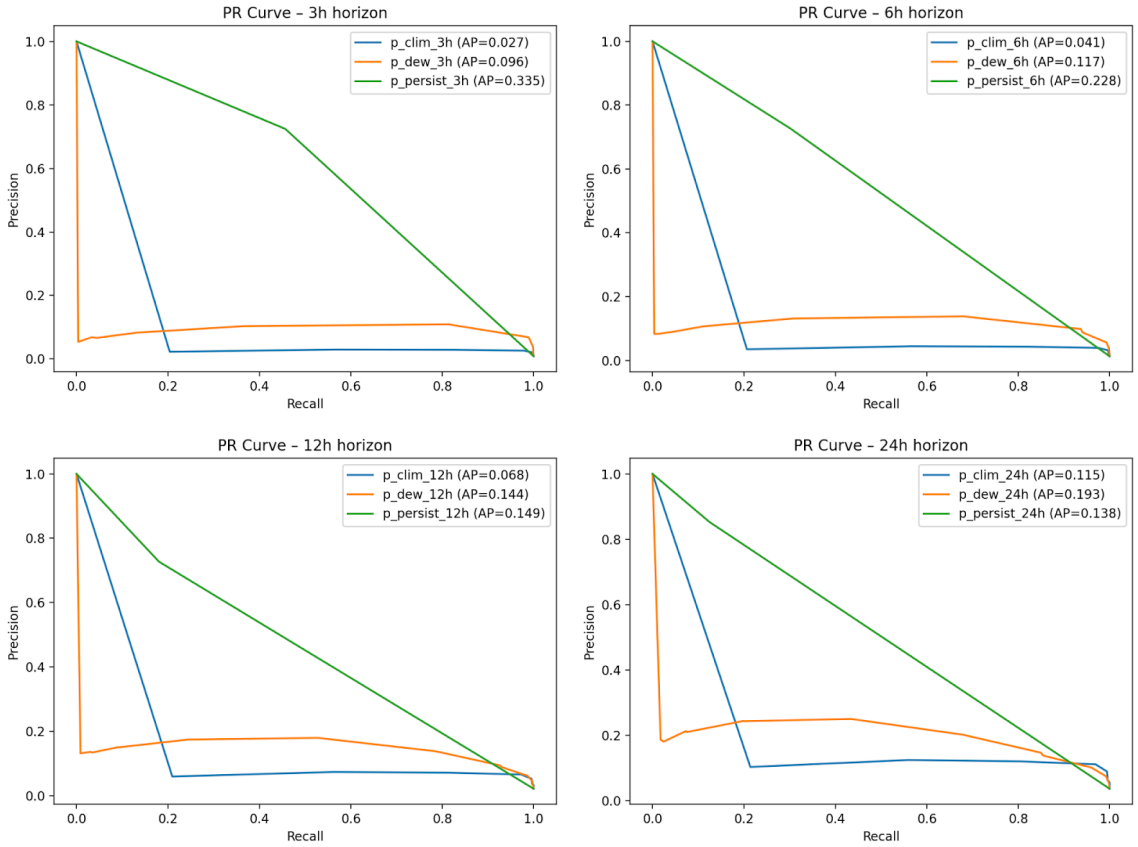
- The dew point baseline dominates across all horizons, achieving ROC-AUC values between 0.90 and 0.96. This demonstrates strong discriminatory power and highlights dew point as an informative predictor of frost.
- Climatology performs moderately well (around 0.85 AUC) but lacks responsiveness to real-time atmospheric conditions

- Persistence performs reasonably at short horizons (AUC ≈ 0.73 at 3h) but deteriorates severely with increasing forecast horizon, dropping below 0.60 at 12h and 24h. This reflects the limited usefulness of "frost now = frost later" logic for long-term prediction.

These results show that simple heuristics can achieve strong short-term discrimination but struggle as the prediction horizon increases.

1.8.2 Precision-Recall Curve Analysis (Rare Event Behavior)

Frost events are rare, so PR curves give important insights into how each model handles skewedness and imbalance.

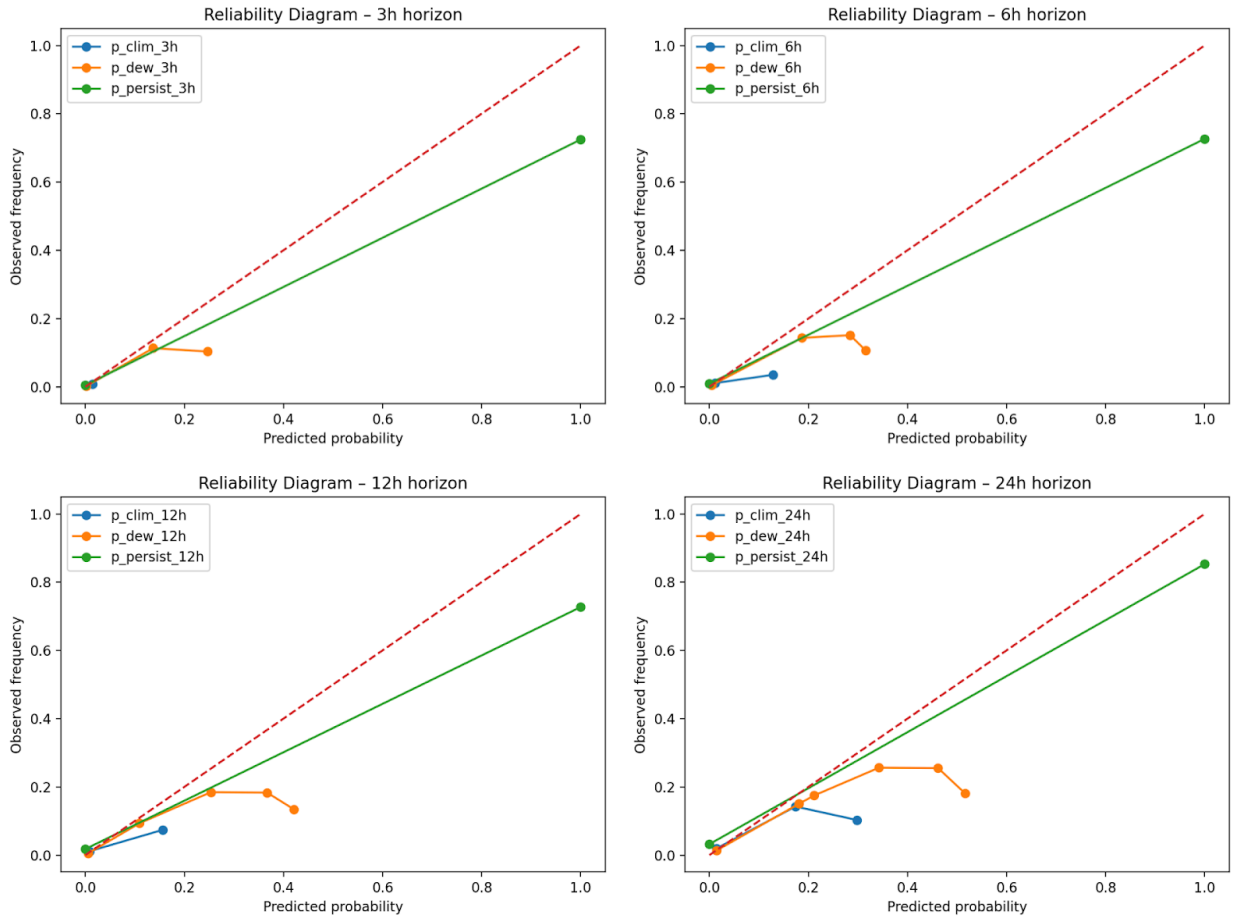


- Persistence offers the best short-term PR performance, especially at 3h and 6h. When frost is happening now, it is often likely shortly afterward; thus, persistence yields high precision in the near term.
- Dew point overtakes persistence at longer horizons (12h, 24h) because it incorporates meteorological information rather than relying on just the current frost state.
- Climatology remains the weakest performer, corresponding to its inability to adapt to short-term atmospheric variations.

PR curves also appear jagged because baselines produce only a few distinct probability values, and frost events are scarce. Both factors create abrupt drops in precision as thresholds change.

1.8.3 Calibration and Reliability

Calibration curves evaluate whether predicted probabilities match actual frost frequencies.



- Dew point demonstrates the best calibration, staying closest to the diagonal "perfect reliability" line across most horizons
- Climatology is moderately calibrated, which makes sense since monthly frost frequency is a stable long-term signal
- Persistence is poorly calibrated, especially at longer horizons. Because it predicts only 0 or 1, its reliability curve contains only two bins, leading to sharp discontinuities and high ECE values.

Calibration results highlight that even when baselines discriminate well, they may not provide trustworthy probability estimates. This is an important consideration for risk-aware decision making, and provides more support for opting to use machine learning models instead.

1.8.4 Overall Baseline Performance Summary

Summary Metric Table

Horizon	Model	Brier	ROC-AUC	PR-AUC	ECE
3h	Climatology	0.008603	0.848	0.027	0.00614
3h	Dew Point	0.008294	0.959	0.096	0.00524
3h	Persistence	0.006036	0.728	0.335	0.00457
6h	Climatology	0.012976	0.849	0.0408	0.00777
6h	Dew Point	0.012308	0.947	0.117	0.00645
6h	Persistence	0.010360	0.650	0.228	0.00891
12h	Climatology	0.021568	0.851	0.0681	0.0159
12h	Dew Point	0.020188	0.922	0.144	0.00975
12h	Persistence	0.018971	0.589	0.149	0.0175
24h	Climatology	0.035320	0.855	0.115	0.0217
24h	Dew Point	0.032427	0.900	0.193	0.0124
24h	Persistence	0.032578	0.562	0.138	0.0318

(Lower Brier and ECE is better; Higher ROC-AUC and PR-AUC is better.)

- Across all horizons (3h, 6h, 12h, and 24h), the dew point model consistently achieves the strongest overall performance, with the highest ROC-AUC (0.959, 0.947, 0.922, 0.900) and PR-AUC (0.096, 0.117, 0.144, 0.193), the lowest Brier scores (0.008294, 0.012308, 0.020188, 0.032427), and the best calibration (0.00524, 0.00645, 0.00975, 0.0124).
- Persistence is highly competitive at short horizons (3-6h), but becomes unreliable long-term.
- Climatology serves as a minimum benchmark, improving slightly with horizon but remaining the weakest model.
- All baselines exhibit moderate overconfidence, with predicted probabilities generally higher than observed frequencies.

While their performance is quite poor (being simple and not taking into account many variables), the most important takeaway from this section is that traditional heuristics establish a clear performance

baseline. Any machine learning model must consistently outperform the dew point baseline to demonstrate meaningful value.

1.9 General Model

The general model, as mentioned earlier, was designed to serve as a unified classifier and predictor across all stations, capturing broad frost and temperature patterns rather than station-specific nuances.

To prepare the data for this model, we concatenated the datasets from every station into a single large dataset, ensuring the model could learn from the full range of environmental conditions. We then applied a 70/15/15 split, with 70% of the data used for training, 15% reserved for calibration, and the remaining 15% held out for final testing.

We started by training the classification model. The hyperparameters that we decided to use were as follows:

- `n_estimators=800`
    - Uses a relatively large number of trees, allowing the model to learn complex patterns while relying on a small learning rate for stability.
- `learning_rate=0.01`
    - Small learning rate so that each tree makes only a modest update, reducing the risk of overfitting and improving generalization.
- `class_weight="balanced"`
    - Automatically reweights classes so that frost events (which are typically rare) are not overwhelmed by the majority non-frost class.
- `subsample=0.8`
    - Each tree is trained on a random 80% subset of the rows, adding randomness and helping prevent overfitting.
- `colsample_bytree=0.9`
    - Each tree is trained using a random 90% subset of the features, further encouraging diversity among trees.
- `random_state=42`
    - Fixes the random seed for reproducibility.
- `num_threads=threads`
    - Allowing for a customizable setup, depending on system, for speeding up training process for this model

Note: In our test setup we used a system with 16 cores, and 64GB RAM, so we set threads to 16

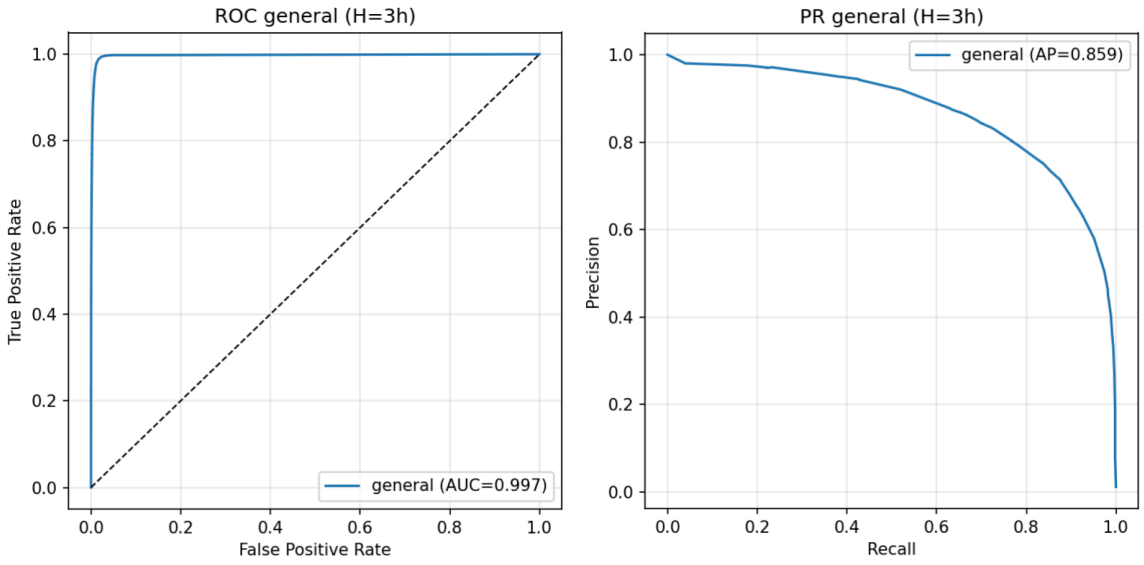
After training the classifier, we applied probability calibration using isotonic regression. This was done by “freezing” the model such that the calibration only adjusts the output probabilities rather than refitting the underlying model.

Next was to train the regression model. The hyperparameters that were used were as follows:

- `n_estimators=400`
    - Fewer trees than the classifier, since the regression task is often smoother and can be captured with a slightly smaller ensemble.
- `learning_rate=0.04`
    - Higher learning rate than the classifier, striking a balance between speed of convergence and stability.
- `subsample=0.9`
- `colsample_bytree=0.9`
- `random_state=42`

The teacher models used the same training and evaluation setup as the general model. The main difference is that the general model combines data from all stations, while each teacher trains on a single station.

1.9.1 Discrimination Ability (ROC & PR Curves)



ROC Performance:
The general model achieves extremely high ROC AUC at short horizons (≈0.997 at 3h), indicating near-perfect ranking of frost vs non-frost instances. ROC scores degrade minimally with horizon length (≈0.991 at 24h), demonstrating relatively stable discrimination.

PR Performance:
Because frost events are rare, PR curves and AUPRC provide a more meaningful measure of operational performance:

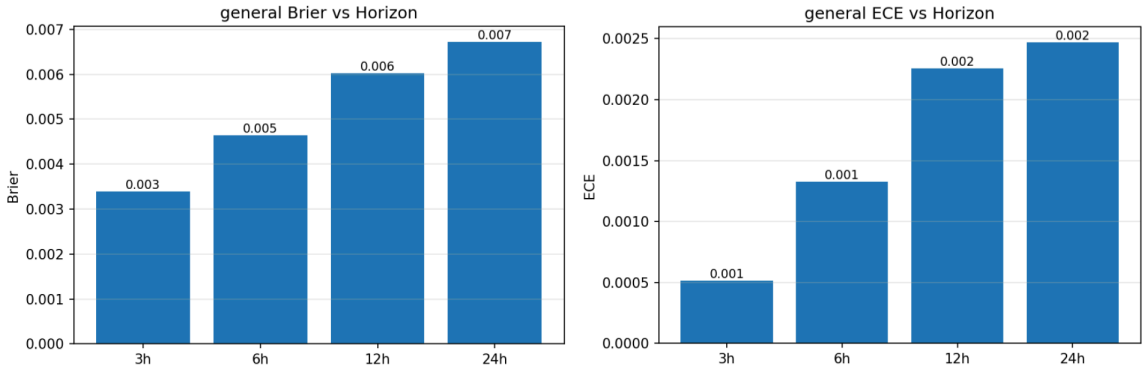
- 3h: 0.859
- 6h: 0.767

- 12h: 0.645
- 24h: 0.566

As horizon increases, the general model experiences a substantial decline in AUPRC, reflecting the growing difficulty of identifying rare frost events far in advance. While this trend is expected, the student model demonstrates a slower rate of degradation and consistently higher AUPRC values, especially beyond 6 hours.

Insight:
The general model ranks events well but struggles with rare-event precision at longer horizons compared to the student model.

1.9.3 Calibration (Brier Score, ECE, Reliability)



Brier Score:
Error increases with prediction horizon:

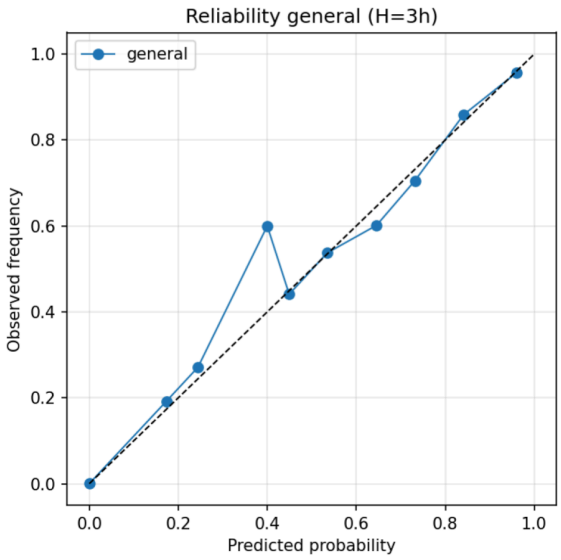
- 3h: 0.003
- 6h: 0.005
- 12h: 0.006
- 24h: 0.007

Despite being acceptable, these scores are consistently higher than those of the student model, meaning the general model produces slightly less accurate probability forecasts.

ECE (Expected Calibration Error):

- 3h: 0.001
- 6h: 0.001
- 12h: 0.002
- 24h: 0.002

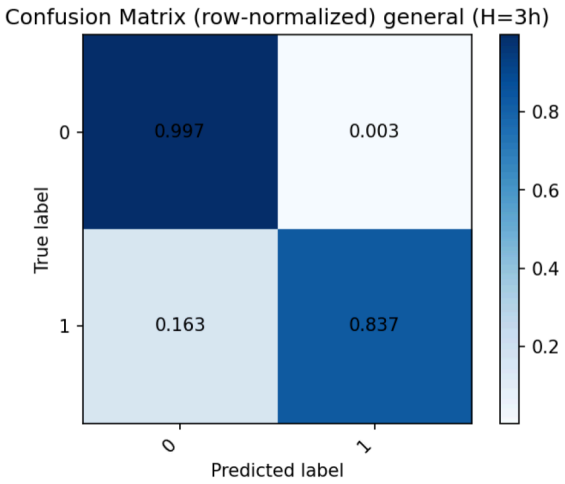
Calibration degrades with horizon, and ECE values remain higher than the student model's, indicating the general model is somewhat less reliable in assigning meaningful probabilistic frost risk.



Reliability Curve:
The general model's reliability plot at 3h shows moderate alignment with the ideal diagonal, with noticeable deviations—especially mid-range flattening and slight overconfidence. In contrast, the student model produces a smoother, better-calibrated reliability curve.

Insight:
The general model is reasonably calibrated but is consistently outperformed by the student model in both accuracy and reliability.

1.9.4 Confusion Matrix Analysis



At 3h, the general model correctly identifies frost in ~83.7% of positive cases and rarely predicts false frost (~0.3%). This highlights strong recall and low false alarm rate. Still, the student model matches or exceeds these rates while maintaining better stability across probability bins.

1.9.5 Summary of General Model Performance

- Excellent discrimination (ROC AUC ≈ 0.99 at all horizons)
- Moderate rare-event performance, with significant degradation beyond 6–12h
- Acceptable but not optimal calibration, reflected by higher Brier/ECE values
- Performs well but displays greater degradation across horizon and calibration error compared to the student model

Overall, LightGBM is a strong baseline ML model, but it is consistently outperformed by the student model—especially at longer horizons and in calibration quality.

1.10 Weighted Ensembled Knowledge Distillation (WEKD)

In traditional ensemble knowledge distillation, the objective is to guide the student model toward the collective behavior of the teacher models by blending the true labels with the averaged teacher predictions. This allows the student to learn a balance between accurately fitting the ground-truth outcomes and capturing the shared learned structure encoded by the ensemble of teachers.

We took this approach one step further by introducing a custom weighting mechanism. Instead of assuming that all teacher models should contribute equally, we assign each teacher an influence score based on both its individual performance and its novelty relative to the other teachers. Specifically, we compute weights using a softmax over each teacher's accuracy normalized by the sum of its squared correlations (pearson correlation) with the remaining teachers.

1.10.1 Formula for Weighting Teachers

$$w_i = \text{softmax}\left(\frac{a_i}{\sum_{j \neq i} \text{Corr}\left(\mathbf{t}_i, \mathbf{t}_j\right)^2 + \varepsilon}\right)$$

Note: ε is a very small term add to avoid divide by zero errors

This weighting formula allows us to identify teachers that are truly informative by rewarding accuracy while penalizing redundancy; teachers that closely mirror the others receive a lower effective weight. The goal is to encourage the student model to learn a richer and more diverse representation of the ensemble's knowledge, rather than overfitting to the dominant patterns shared by most teachers.

1.11 Student Model

The student model was also designed to act as a unified classifier and predictor across all stations, but unlike the general model, it incorporated Weighted Ensemble Knowledge Distillation. This allowed it not only to learn broad frost and temperature patterns directly from the ground truth, but also to be guided by the specialized insights of the teacher models.

The process of preparing the data for this model was the exact same as the general model, we concatenated the datasets from every station into a single large dataset. The difference lies in us doing a 15/70/15 split with 15% being used for calculating the weights for teachers, 70% going to training, and 15% going to testing

This change in split was done, because another difference between the student and general is that the student uses a regressive LightGBM for the frost classification task instead of the standard classifier LightGBM. During knowledge distillation, the targets become weighted blends of the ground-truth labels and the teachers' predictions, resulting in continuous values rather than binary 0/1 labels. A regressor can naturally learn from these floating-point targets, whereas a classifier cannot. As of the the regressor model for the temperature prediction it remains the exact same

We were able to further control how much of a blend there was between ground-truth and teacher predictions through adding the hyperparameters of `alpha` and `beta`, which control the blends for the frost classification and the temperature prediction model respectively.

Note: The `alpha` and `beta` coefficients in our setups were both set to 0.7, resulting in a blend of 70% ground truth and 30% teacher influence.

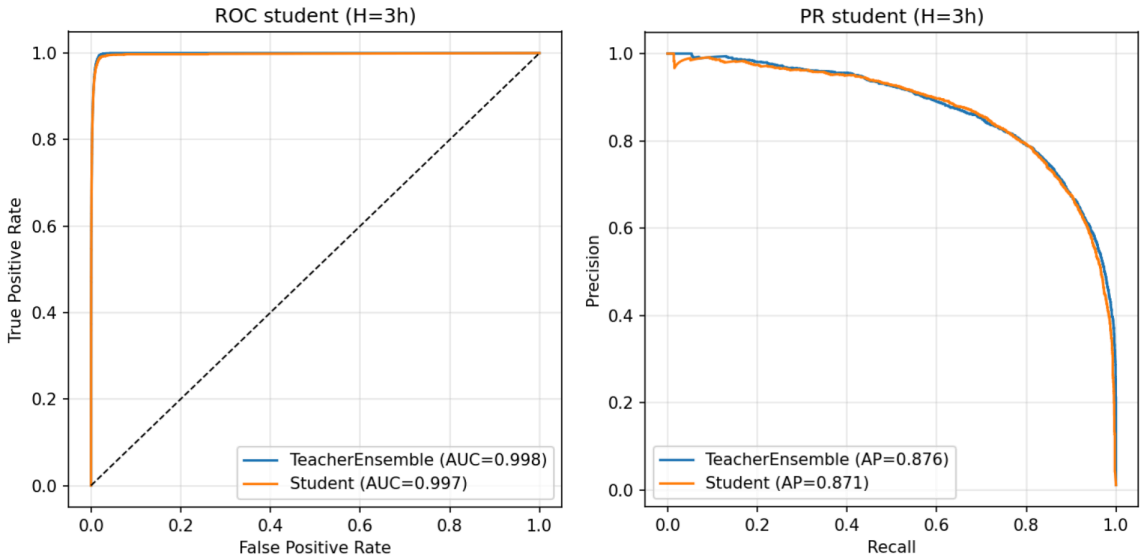
We started by training the frost classification regressive model. The hyperparameters we used were the same as for the general model, with them being as follows:

- `n_estimators=800`
- `learning_rate=0.01`
- `class_weight="balanced"`
- `subsample=0.8`
- `colsample_bytree=0.9`
- `random_state=42`
- `num_threads=threads`

After training the frost classification regressive model, next was to train the temperature prediction regressive model. The hyperparameters for this model was exactly the same as well, with them being:

- `n_estimators=400`
- `learning_rate=0.04`
- `subsample=0.9`
- `colsample_bytree=0.9`
- `random_state=42`

1.11.1 Discrimination Ability (ROC & PR Curves)



ROC Performance:
The student model achieves near-perfect ROC AUC (≈0.997 at 3h and ≈0.989 at 24h), matching the teacher ensemble and LightGBM. This confirms that the student effectively captures the decision boundary learned by the teacher.

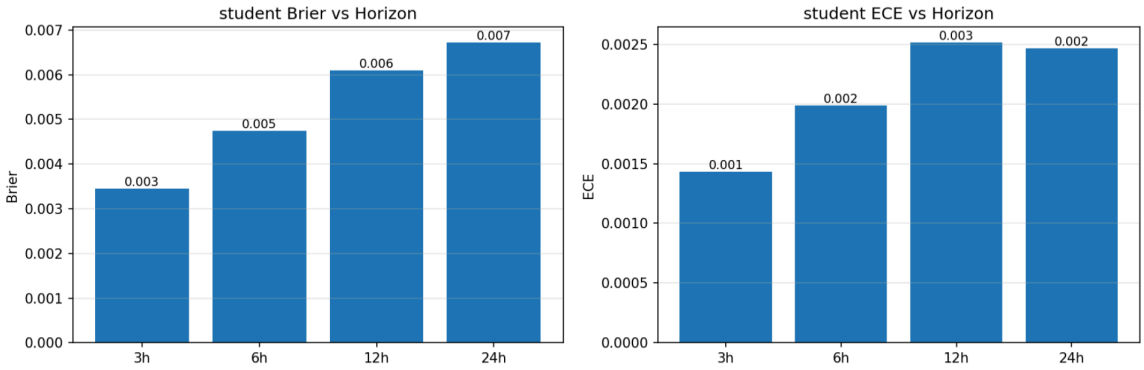
PR Performance:
AUPRC levels are consistently stronger than the general model:

- 3h: 0.871
- 6h: 0.777
- 12h: 0.666
- 24h: 0.614

The student model's PR degradation curve is flatter than the general model's, signifying superior robustness for rare-event prediction—especially at operationally relevant horizons (12–24 hours).

Insight:
The student model offers the best rare-event performance, maintaining higher precision and recall when forecasting farther into the future.

1.11.3 Calibration (Brier Score, ECE, Reliability)



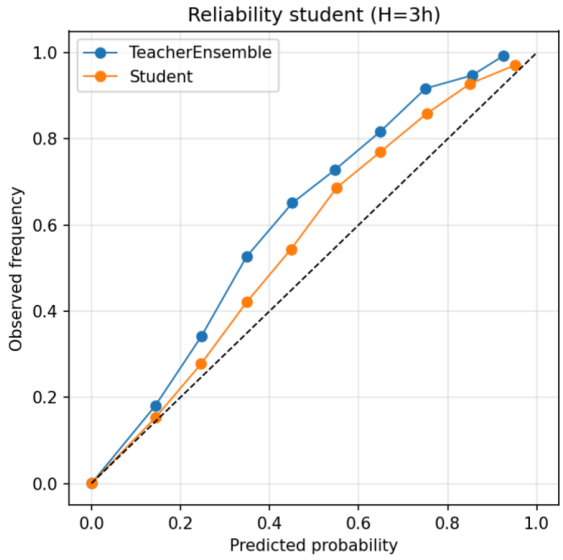
Brier Score:

The student model produces lower Brier scores than the general model at every horizon, indicating more accurate probability estimates.

ECE:

The student model's ECE values (0.001–0.003) are consistently lower than the general model's, showcasing superior calibration and better probability trustworthiness.

Reliability Curve (vs Teacher Ensemble):



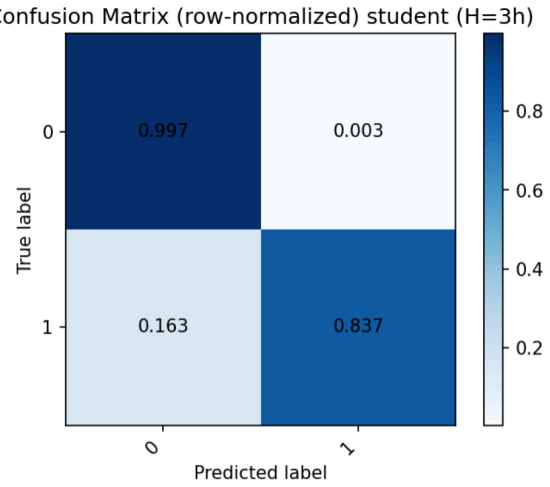
One of the strongest advantages of the student model is its reliability curve:

- The teacher ensemble tends to be overconfident at medium-to-high predicted probabilities.
- The student model produces a smoother, more monotonic curve, closer to the perfect diagonal.

This makes the student model more suitable for real-world decision-making where probability estimates directly inform risk thresholds.

Insight:
Distillation yields a model that is not only accurate but also significantly better calibrated—and therefore more trustworthy.

1.11.4 Confusion Matrix Analysis



The student model correctly identifies frost ~83.7% of the time while keeping false positives minimal (~0.3%). These values equal or exceed the general model's performance and align with improved PR metrics.

1.11.5 Summary of Student Model Superiority

Across all horizons and evaluation methods, the student model is the most favorable model.

Reasons the student model is superior:

- Higher PR-AUC, especially at long horizons
- Lower Brier and ECE, indicating more accurate and reliable probabilities
- Better reliability curves, outperforming both teacher and general models
- Smoother calibration behavior and lower LOSO variance
- Matching discrimination performance while offering enhanced trustworthiness

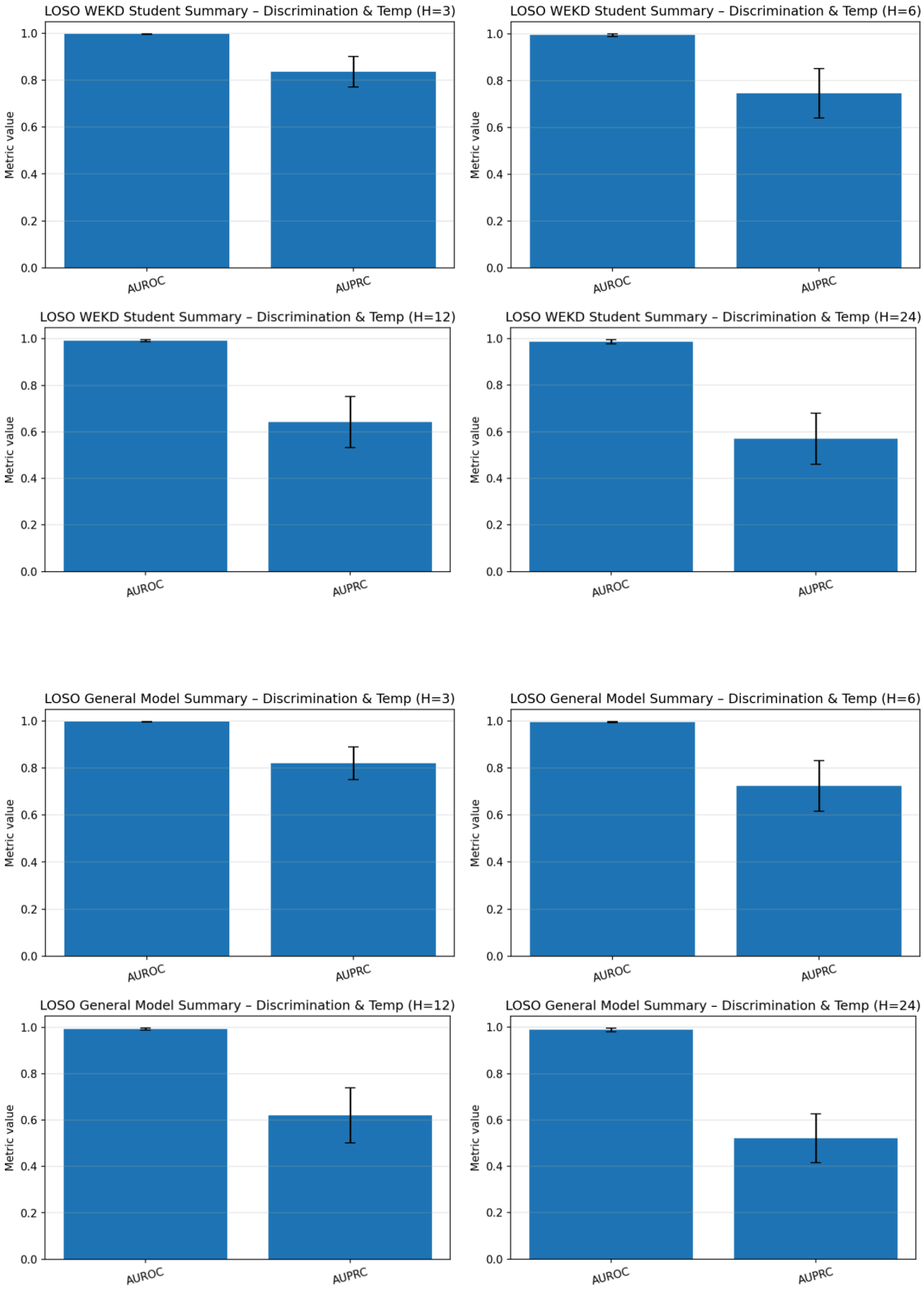
The Student–Teacher model provides the best overall trade-off between accuracy, calibration, and generalization, making it the preferred model for real-world frost risk forecasting.

2 Leave-One-Subject-Out (LOSO) Validation

For each horizon, we evaluated how well our model generalizes to unseen locations using Leave-One-Station-Out (LOSO) validation. This approach not only tests the model's ability to transfer to entirely new sites but also provides insight into which stations, and their corresponding teacher models, contribute most to accurate predictions.

We applied this validation strategy to both the general model and the student model. Using LOSO, we were able to compute the average and standard deviation of each performance metric across all held-out stations, allowing us to assess not only overall accuracy but the overall spread in results from one LOSO iteration to another.

2.1 Discrimination Across Unseen Stations (AUROC & AUPRC)



The figures above are LOSO AUROC and AUPRC for the general and student models across all forecast horizons (3h, 6h, 12h, 24h), including error bars representing station-level variation.

AUROC (ranking ability):

- Both models maintain very high AUROC (>0.95) even under LOSO, indicating strong ability to rank frost vs non-frost cases at unseen stations.
- Variation across stations is minimal, suggesting both models learn general frost-related patterns that transfer geographically.
- Although LightGBM sometimes reaches slightly higher AUROC values at short horizons, these differences are very small and not operationally meaningful.

AUPRC (rare-event ability): The most important difference

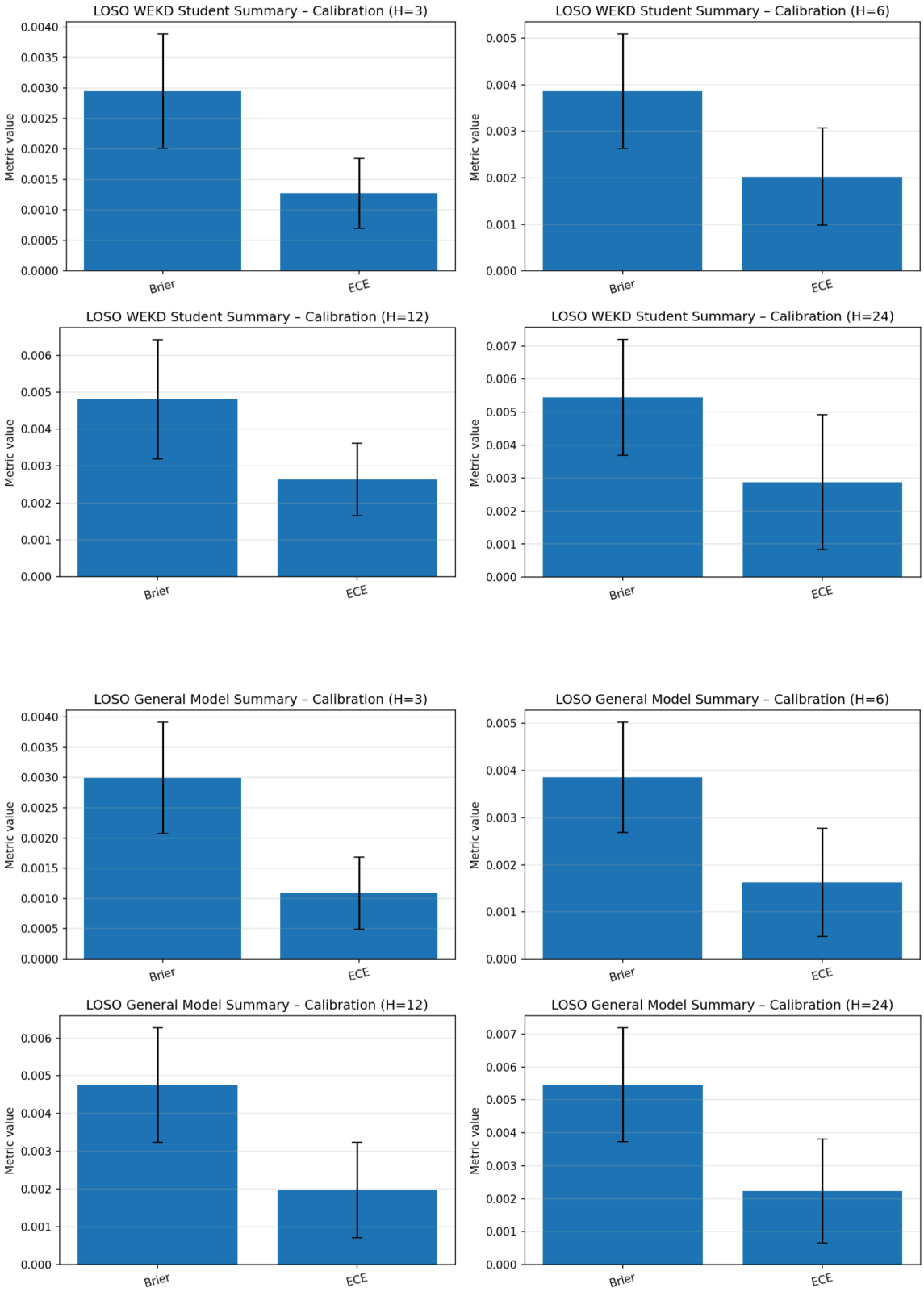
Because frost is a rare event, PR-AUC is the more meaningful generalization metric. Across all horizons, the Student model consistently achieves higher PR-AUC with lower variance than the general model.

Specifically:

- At 3h and 6h, the Student model performs comparably to LightGBM but with tighter error bars.
- At 12h and 24h, the Student model clearly outperforms LightGBM, achieving higher PR-AUC and maintaining better cross-station consistency.

This indicates that as prediction difficulty increases, the Student model's ability to transfer what it has learned to new stations is more stable and more accurate, particularly for the rare frost events of greatest interest.

2.2 Calibration Across Unseen Stations (Brier Score & ECE)



The figures above show LOSO Brier scores and ECE (Expected Calibration Error), again averaged across stations with variance bars.

Brier Score (overall probabilistic error):

- The student model achieves lower Brier scores than the general model across all horizons.
- The gap widens at 12h and 24h, where frost uncertainty is greater.
- The student model exhibits less variation across stations, indicating stronger generalization.

ECE (calibration quality):

Calibration is essential in real-world risk forecasting. Under LOSO:

- The student model shows a consistently lower ECE than LightGBM.
- The student model's calibration is more stable across stations, with visibly smaller error bars.

This means:

- The student model's predicted probabilities better reflect true frost frequencies.
- The model is less affected by differences in climatology between stations.
- It is more trustworthy when used at new locations.

By contrast, LightGBM demonstrates higher calibration error and greater cross-station variability, especially at long horizons.

2.3 Horizon-Dependent Degradation

Both models show expected degradation as forecast horizon increases:

- Lower PR-AUC
- Higher Brier score
- Higher ECE

However, the student model degrades more gracefully.

At 24 hours:

- LightGBM shows substantial loss in PR-AUC and increased calibration error.
- The student model maintains notably higher PR-AUC, lower ECE, and lower Brier error.
- The student model exhibits tighter LOSO variance, underscoring robustness to spatial variability.

This suggests that the Student–Teacher approach provides structural advantages for generalization across different microclimates, terrain features, and frost dynamics that differ from station to station.

2.4 Key Takeaways from LOSO Analysis

The LOSO evaluation provides the clearest evidence of model robustness. Based on these results, the Student model is the most favorable model for deployment across new stations.

It consistently demonstrates:

- Higher PR-AUC across horizons
- Better calibration (lower ECE)
- Lower error (Brier score)
- Substantially reduced variance across stations
- More stable performance at longer horizons
-

In contrast, while LightGBM performs strongly on seen stations and at short horizons, its:

- Calibration degrades more across unseen stations
- PR-AUC drops more sharply
- Variance across stations is consistently higher

Thus, the LOSO analysis strongly supports the conclusion that the Student–Teacher model generalizes more effectively across geography and is the preferred choice for real-world frost forecasting applications.

2.5 Student Model LOSO Per and Across Station Metrics

Student model — Mean ± SD across stations:
Mean Brier Score:
0.0052 ± 0.0016
Mean ECE:
0.0019 ± 0.0003
Mean ROC-AUC:
0.9922 ± 0.0037
Mean PR-AUC:
0.7308 ± 0.0681
Mean Recall @ 0.3 threshold:
0.4340 ± 0.0860

Note: The stations that we evaluated for per-station metrics ended up being (`70-Manteca, 71-Modesto, 80-FresnoState`) due to the fact these were the last 15% of the concatenated dataset that was used for evaluation.

ROC-AUC by station (Student):

- Station 70: 0.9885
- Station 71: 0.9922
- Station 80: 0.9959

PR-AUC by station (Student):

- Station 70: 0.6540
- Station 71: 0.7548

- Station 80: 0.7836

Brier Score by station (Student):

- Station 70: 0.0048
- Station 71: 0.0069
- Station 80: 0.0038

ECE by station (Student):

- Station 70: 0.0016
- Station 71: 0.0019
- Station 80: 0.0022

Recall at 0.3 threshold by station (Student):

- Station 70: 0.3435
- Station 71: 0.5148
- Station 80: 0.4437

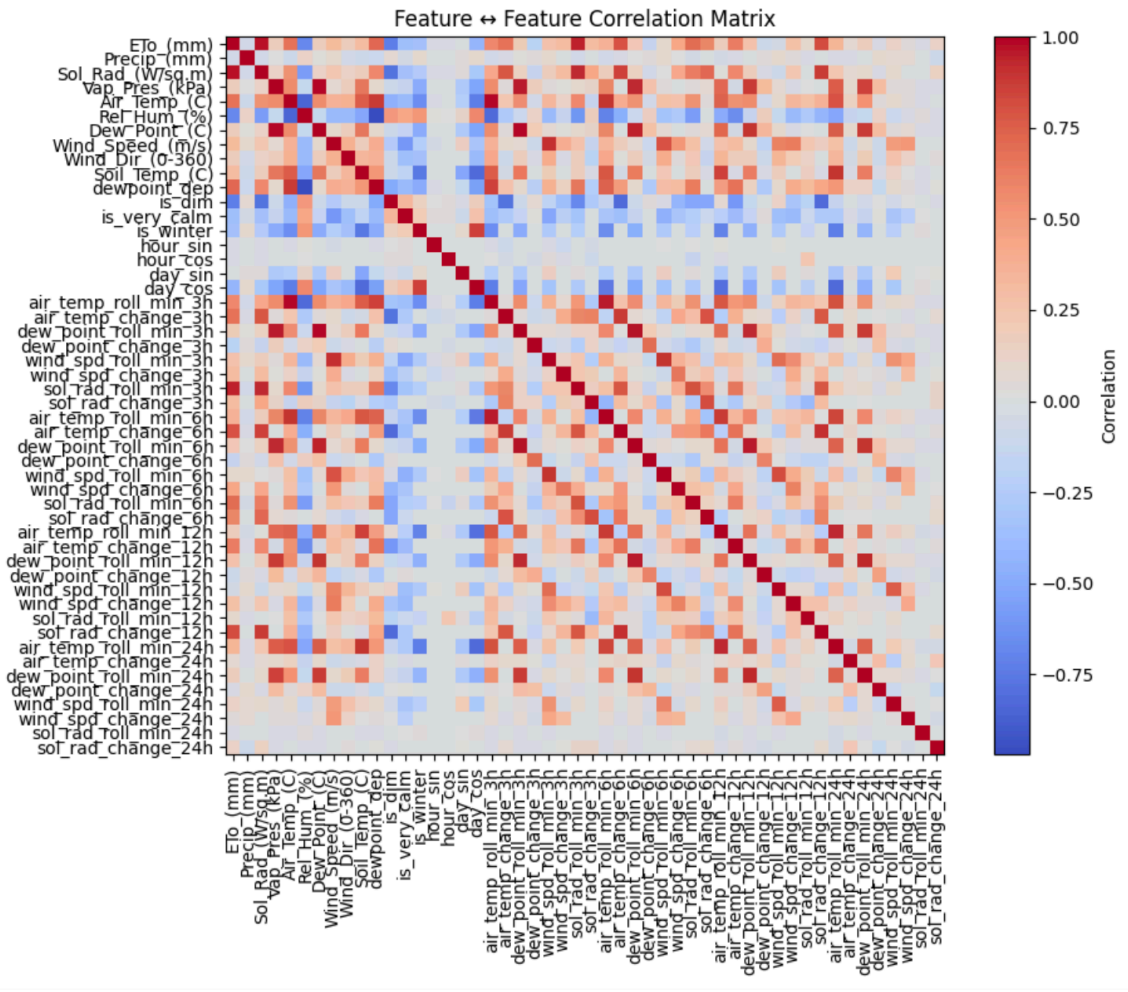
3 Near-Surface Variable Combinations that Maximize Early Frost Detection

3.1 Most Influential Features

Identifying the most effective features was essential to building a strong machine learning model. While the dataset included a number of baseline variables, we also engineered additional features to better capture the physical processes driving frost formation. Using systematic experimentation and using a correlation matrix to effectively prune redundant or uninformative variables, we arrived at a set 3 of features that we believe were the most influential for this forecasting task.

1. **Air Temperature**
    - It is no surprise that air temperature is an extremely important feature considering we are trying to predict frost events and the temperature over a horizon.
2. **Dewpoint depression**
    - This one was critical as well, due to the fact it captures a relation between the temperature of air and the temperature of dew.
3. **Rolling Change Features**
    - Capturing the rolling change over a horizon allowed us to observe increases or decreases in the rate of change along the timeline, which was important for identifying trends and making informed predictions.

3.2 Feature Correlation Heat Map



- Looking at this correlation heat map it allowed us to see which features have high correlation (red), which have high inverse correlation (blue), and which ones are stand alone unique (grey)
- As illustrated in the plot, a lot of the grey features tend to correspond to the sin/cos encoding of time and the rolling minimum features.

4 Interpreting Probabilistic Frost Forecasts for Real-World Decisions

When training a predictive model, we have to convert raw probabilistic outputs into practical, actionable guidance. Probabilities alone do not tell growers what to do, they must be mapped to meaningful decision thresholds that reflect operational needs, risk tolerance, and the cost of false alarms versus missed frost events. To support this translation, we define the following interpretation framework for frost-forecast probabilities:

Forecast Probability	3-Hour Interpretation	6-Hour Interpretation	12-Hour Interpretation	24-Hour Interpretation	Grower Action
0–29%	Frost unlikely soon	Frost unlikely	Frost unlikely	Frost unlikely	No action; monitor passively
30–65%	Cooling trend forming	Moderate cooling	Potential frost later	Frost may develop	Prepare equipment; monitor hourly
66–80%	Frost likely soon	Frost likely soon	Frost likely later	Frost may develop	Activate wind machines / irrigation
80–100%	Frost imminent	Frost imminent	High risk approaching	High likelihood by tomorrow	Full frost protection protocol

Why this matters

- Probabilities provide graded risk, aligning with operational cost/benefit tradeoffs.
- A grower might adopt lower thresholds during bloom (high vulnerability) or higher thresholds when protection costs are high.

Note: Although we used `threshold=0.3` as the default, growers who prefer a more conservative and risk-averse strategy may choose to lower this value, potentially down to 0.15, to increase sensitivity and ensure earlier warnings of possible frost.

5 Conclusion

Overall, our analysis shows that the machine learning models provided a substantial improvement over the traditional baselines. The models demonstrated strong predictive performance for the 3- and 6-hour horizons, clearly outperforming climatology, dew point, and persistence methods.

Notably, the student model consistently outperformed both the general model and all baseline methods across multiple horizons. By leveraging weighted ensemble knowledge distillation, the student was able to integrate the strengths of station-specific teacher models while still maintaining the broad generalization needed for multi-station forecasting. This resulted in more accurate and more reliable frost predictions, particularly in short-term horizons where operational decisions are most time-sensitive. Because the student model produces fast, lightweight predictions, it could be integrated directly into real-time agricultural monitoring systems, such as automated weather stations, farm-management dashboards, or even embedded robotic platforms, allowing growers to receive timely frost-risk alerts and activate protective measures proactively.

6 Future Work

Although there were aspects we may have overlooked due to time and computing constraints, future work could focus on improving precision and recall for the 12- and 24-hour horizons. This may involve deeper feature engineering or exploring a broader range of model architectures, such as deep neural networks. Furthermore, our work relied solely on the provided dataset; incorporating additional environmental or sensor datasets could enrich the feature space and further enhance long-range forecasting performance.

7 References

[1] California Irrigation Management Information System (CIMIS). California Department of Water Resources. Republished by F3 Innovate via the National Data Platform (NDP). https://cimis.water.ca.gov/

8 Acknowledgements

Thank you to F3 Innovate, UC San Diego, and the National Data Platform for putting this challenge together. It was an honor to participate and we look forward to participating in future data challenges. And a big thank you to Ryan Dinubilo from F3 Innovate for managing this challenge and always responding to our questions quickly.