# Frost Risk Forecast Data Challenge

Group members : Aziz Saries, Tyler Roberts

**1.) Describe your full modeling pipeline, including which datasets you used, how you preprocessed them, and how your model was trained and evaluated. Include a flowchart or diagram to visualize your workflow.**

For this project, we used the full CIMIS weather dataset, which contains hourly meteorological observations across 18 stations and 15 years across California. The hourly measurements taken include air temperature, wind speed and direction, solar radiation, humidity, and reference evapotranspiration. We did not incorporate any external datasets aside from this csv file which contained all the available weather information from all 18 stations. To improve time-based interpretability, we added month, year, and day columns to the dataset. We also added a binary frost variable that was synchronized with each hour's meteorological record so that each observation indicated whether a frost event occurred. This allowed the model to learn the atmospheric conditions that typically precede frost.

After constructing the initial dataset, we assessed the extent of missing values. The overall level of missing data was low, with no feature exceeding approximately 2.5% missing values. Importantly, missing values were confined only to the weather measurement variables, while all other predictors such as station ID, date, and quality control flags were fully complete. Because the missing values were feature specific and relatively minimal, we chose to impute missing values rather than remove entire rows, which helped preserve rare frost events and maintain a representative dataset. The imputation was made with the median value from the respective columns as most of our weather variables are skewed, making the median a more robust representation of the typical value of a data as opposed to the mean.

To construct the supervised-learning targets for each forecast horizon (3,6,12 and 24 hours), we generated horizon-specific variables by shifting the frost indicator and temperature series forward by 3,6,12 and 24 hours respectively. This ensures that each row represents the information available at time "*t"* while the target reflects the conditions at time "*t + horizon"*. If the shifted temperature value was missing (due to no measurement existing *"h"* hours ahead), we excluded that row from training and evaluation for that horizon, so that each example has a valid target we can compare our results to. Each horizon was then modeled as an independent prediction task, meaning that preprocessing, feature selection, and cross-validation were executed separately for the 3,6,12 and 24 hour horizons.

To train the models, we first defined which columns could be used as features and which would not be appropriate. We removed station name, region, date, hour, all the quality-control columns, and all the "future" target columns (such as frost_3h_ahead or temp_3h_ahead) to prevent accidental label leakage. Since the dataset did not include meaningful categorical predictors, the feature matrix consisted entirely of numeric variables. These numeric features were passed through a preprocessing pipeline that applied median imputation, which was referenced earlier, followed by standardization. The preprocessing pipeline was implemented using a ColumnTransformer, which was then fit on the training fold and then applied to the corresponding test fold to ensure that no information from the test data influenced the training process.

Regarding the CIMIS quality-control (QC) flags, we did not remove or filter rows based on these indicators, nor did we use them as predictive features. The QC columns were excluded from the feature matrix to avoid introducing metadata that might somehow encode information about downstream sensor conditions or processing steps. This ensured that the models learned only from meteorological variables rather than from any internal CIMIS data-quality annotations.

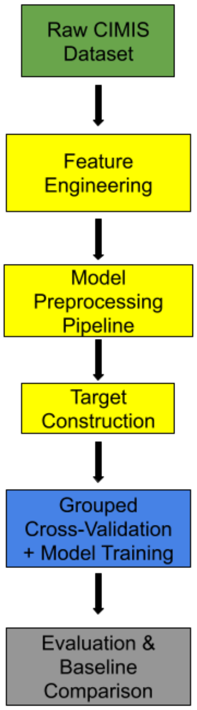
Due to frost hours being significantly more rare than non-frost hours, the classification problem is imbalanced. We did not apply any resampling or class weighting because the model will ultimately be evaluated and used in a real-world setting where frost events are presumably just as rare. Therefore, preserving the natural frequency of the frost events should lead to a more realistic probability estimate. Instead, we relied on probabilistic metrics such as Brier score (measures the accuracy of predicted probabilities), ROC-AUC (evaluates how well the model separates frost from non-frost hours), and average precision (A.K.A. PR-AUC, which measures how well the model identifies these rare frost events without generating excessive false alarms). Together, these metrics also allowed us to assess the calibration of the predicted probabilities. Although we evaluated calibration using reliability diagrams and ECE, we did not apply any post-hoc calibration; all predicted frost probabilities come directly from the model outputs.

Because the challenge explicitly asks for the probability that a frost event (T < 0 °C) occurs at each horizon, we formulated frost prediction as a binary probabilistic classification problem and used tree-based ensembles to model it. The engineered predictors such as rapid cooling rate, dewpoint depression collapse, calm winds, low radiation, and nighttime indicators create a strong, nonlinear separation between frost and non-frost hours. This then led me to apply classification models such as Random Forest and XGBoost, since they can best “learn” threshold-like decision rules and the complex interactions in our weather data. These models were chosen because they produce reliable probability estimates through their ensemble structure, making them well-suited for calibrated frost-risk forecasting.

We did not perform hyperparameter tuning inside the GroupKFold process. Instead, both Random Forest and XGBoost were trained using fixed hyperparameters chosen before cross-validation began. Random Forest was run with its standard defaults (including 100 estimators and unrestricted depth), while XGBoost used a consistent set of manually selected parameters (such as learning rate, max_depth, n_estimators, subsample, and colsample_bytree) that were chosen during preliminary offline experimentation. Keeping hyperparameters fixed ensures that the GroupKFold evaluation measures true spatial generalization across stations rather than simply reflecting horizon-specific tuning or variability introduced by hyperparameter optimization.

We trained both Random Forest and XGBoost models for every horizon and evaluated them under identical GroupKFold conditions. GroupKFold was specifically chosen to ensure generalization across stations, which addresses the challenge requirement of models performing well on unseen locations. While the two models performed similarly in terms of RMSE for temperature regression, XGBoost showed slightly lower error at the shorter horizons. More importantly, the frost-classification calibration metrics favored XGBoost, for example, its reliability diagrams were consistently tighter. However, the Random Forest tended to be mildly overconfident at higher predicted frost probabilities.(As seen in Figure 13/Figure 14) XGBoost also showed lower variance across folds and more stable performance on the rare-event frost classification task, therefore we use XGBoost as our primary model in Questions 2-4 while still reporting Random Forest results for comparison.

**Flow Chart**


Raw CIMIS Dataset
Feature Engineering
Model Preprocessing Pipeline
Target Construction
Grouped Cross-Validation + Model Training
Evaluation & Baseline Comparison

## 2.) How does model performance degrade when applied to unseen stations?

We began our model training and evaluation by first creating a naive predictor to establish a baseline level of performance. For frost prediction at 3, 6, 12, and 24 hours, the naive approach simply assumed that the previous hour's conditions would remain the same into the next forecast period. In other words, if frost occurred in the previous hour, the model would predict frost again, and if it did not, it would assume no frost. For temperature prediction, we followed the same idea by using the current hour's temperature as the prediction for the next hour. This gave us a simple benchmark to compare against more advanced models and helped us confirm that our final model was truly learning meaningful patterns rather than just repeating recent values.

Because we wanted the model to generalize across space, not just time, we used grouped cross-validation based on station ID. Specifically, we used GroupKFold with 18 splits, where each split keeps samples from the same station together in either the training or test set but never both. This setup mimics a “leave-one-stations-out” approach and lets us see how well the model performs on the unseen station. For each forecast horizon (3, 6, 12, and 24 hours ahead), we created two targets, one for frost (a classification problem) and one for temperature (a regression problem). For a given horizon, we masked out rows where the temperature target was missing, then ran the grouped cross-validation loop. In each fold, we fit the preprocessing pipeline on the training data, transformed both train and test, and then trained a Random Forest classifier to predict frost and a Random Forest regressor to predict temperature.

During each fold, we evaluated the Random Forest classifier using the Brier score, which measures the quality of the predicted probabilities, and we saved the predicted probabilities and true labels so we could later make reliability diagrams and other calibration plots. For the temperature model, we evaluated performance using root mean squared error (RMSE). After looping through all folds, we averaged the Brier scores and RMSE values and also computed their standard deviations to summarize performance for that horizon. We then combined all the classifier predictions across all folds to compute additional metrics, including ROC-AUC, precision–recall AUC, and expected calibration error (ECE), and plotted reliability diagrams to see how well the predicted probabilities matched observed frost frequencies.

We repeated the same overall training and evaluation process using XGBoost models: an XGBClassifier for frost and an XGBRegressor for temperature at each horizon. The XGBoost models used a slightly different set of hyperparameters (number of trees, learning rate, max depth, subsampling, and column subsampling) but were still trained within the same grouped cross-validation structure with the same preprocessing, horizons, and metrics. This allowed us to directly compare Random Forest and XGBoost on the same task, using the same folds and evaluation criteria. Comparing these models against our naive predictor showed how much value we gained from using more advanced, nonlinear models that can capture complex weather patterns leading up to frost.

Across all horizons, both models produced strong results, but the differences became clearer as the lead time increased. At 3h and 6h, both Random Forest and XGBoost achieved very high ROC-AUC scores (~0.99) and low Brier scores (~0.004–0.006), with XGBoost showing slightly lower temperature RMSE. As the horizon increased to 12h and 24h, forecast difficulty naturally increased. The Brier scores rose slightly, PR-AUC dropped toward ~0.55–0.60, and temperature RMSE approached ~2.5°C. The main distinction came from XGBoost’s reliability curves staying closer to the ideal diagonal and consistently having lower ECE values (roughly 0.0008 across all horizons), while Random Forest ECE increased at longer horizons (0.0025 at 12hr & 0.0010 at 24hr. These dips in performance are a sign that the model slightly underperformed when applied to unseen locations rather than being robust and having similar performance across all horizons/stations. Overall, XGBoost delivered more stable performance across stations, better probability calibration, and lower variance across folds, which is why it became the preferred model going forward.

**3.) Which combinations of near-surface variables maximize early frost detection skill?**

To improve early frost detection, we performed targeted feature engineering based on physical principles of radiational cooling. We computed dewpoint depression as the difference between air temperature and dewpoint temperature, which serves as a direct measure of how close the air is to saturation. A small dewpoint depression value indicates moist air that cannot hold much more water vapor, which accelerates radiative heat loss after sunset and is strongly associated with overnight frost formation. This measure is also fundamental in agricultural meteorology because the dew point governs many derived variables (e.g., vapor pressure, relative humidity, wet bulb temperature, and vapor pressure deficit) and is widely used to anticipate minimum temperatures critical for crop protection.

We also calculated short-term temperature change by differencing current temperature values with temperatures from several hours earlier, allowing the model to capture rapid cooling rates after sunset, which are a strong precursor to frost. Rapid temperature drops during the early evening are one of the clearest precursors to frost because they signal that radiational cooling has begun efficiently and that the atmosphere is stabilizing. Additionally, we created a calm-wind indicator by flagging periods when wind speed fell below 1 meter per second. Calm or near-calm flow causes dense cold air to remain near the surface of the Earth and therefore doesn’t mix with the warmer air. This prevention of vertical mixing of the hot and cold air substantially increasing frost likelihood.

We also computed vapor pressure deficit (VPD) to quantify atmospheric dryness. Higher VPD values indicate drier air and enhanced radiative cooling, while sudden decreases in VPD often precede frost formation. VPD is derived from saturation vapor pressure (denoted $e_s$) minus actual vapor pressure and provides a compact measure of atmospheric dryness and its role in overnight cooling.

To ensure the model properly learned seasonal and directional patterns, we applied cyclical encoding to both the month and wind direction variables using sine and cosine transformations. Without this transformation, the model would treat December (12) and January (1) as being far apart, even though they are consecutive and belong to the same season. The sine/cosine encoding fixes this by placing all months on a circular scale, preserving their seasonal relationships. The same logic can be applied to the wind direction to deal with the numerical separation between between 359° and 1°. Without this transformation, the model would incorrectly treat these values as being far apart despite their clear physical proximity. Finally, all continuous predictor variables were standardized so that differences in measurement units and numerical magnitudes did not cause any single variable to dominate the training process.

In summary, the near-surface variables that maximize early frost detection are those that jointly describe the physical setup for radiational cooling. The strongest signals come from four combinations:

- Low dewpoint depression together with falling temperatures, indicating near-saturated air that cools rapidly after sunset.

- Rapid evening cooling paired with calm winds, which allows cold, dense air to pool at the surface.
- Vapor pressure deficit behavior combined with temperature trends, reflecting how atmospheric dryness influences radiative heat loss.
- Seasonal phase and wind-direction context, captured through cyclical encodings that help the model recognize when and under what flow patterns frost is most likely.

These combinations mirror the physical conditions under which radiation frosts form, and they form the core of our engineered features that make early frost detection possible.

**4.) How can your model's probabilistic forecasts be interpreted for real-world decisions? For example, how might a grower use your predicted frost probabilities (e.g., 20%, 50%, 80%) to decide when to activate frost protection or monitor conditions more closely?**

Because the model is probabilistic, as opposed to a binary prediction, and well-calibrated, its frost probabilities can be interpreted directly as the likelihood of frost under the current conditions. The GroupKFold results showed that XGBoost produced well-calibrated probabilities across all horizons, meaning that a predicted probability (e.g., 20%, 50%, 80%) corresponds closely to the true frequency of frost events at unseen stations. This makes the outputs suitable for operational decision-making, where growers should balance the cost of activating frost protection with the cost of potential crop damage. The Growers can use these values as operational thresholds:

- ≈20% probability – <u>"Monitor closely"</u>
  This level signals that conditions are trending towards a potential frost. A grower would not typically activate protection yet, but should begin checking temperatures more frequently, confirm equipment readiness, and watch for rapid cooling.
- ≈50% probability – <u>"Prepare to act"</u>
  Frost and non-frost outcomes are equally likely. At this threshold, growers usually stage crews and monitor in real time. Since an unforeseen frost could be debilitating to a farmer's yield, they need to be very closely monitoring in this situation. We recommend, if possible, a monitor that can inform the grower of any potential sharp dips on their mobile device.
- ≈80% probability – <u>"Activate protection"</u>
  This indicates high, near-certain frost risk (consistent with the model's reliability curves). Growers would typically start irrigation, wind machines, or heaters immediately to mitigate crop damage.

Since the model provides probabilities at multiple horizons (3h, 6h, 12h, 24h), growers can use the longer horizons for planning and the shorter horizons for operational decisions, with the 3-hour horizon offering the most trustworthy near-term signal.

# Appendix:

Figure 1:

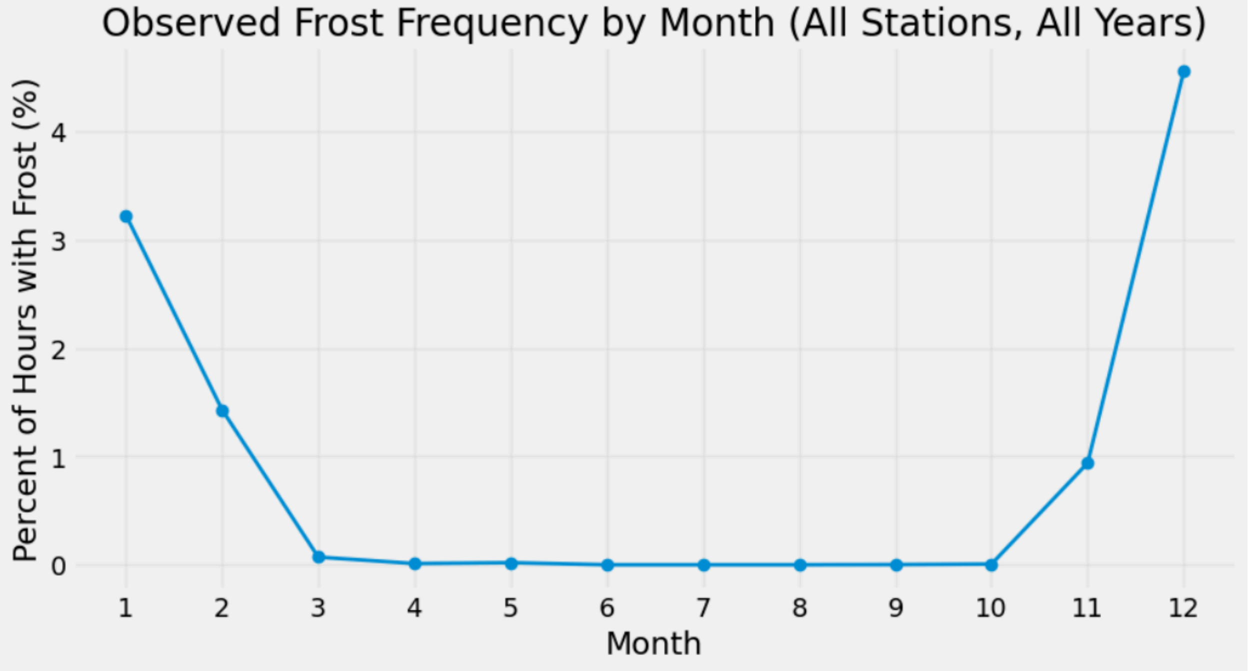


Figure 2:

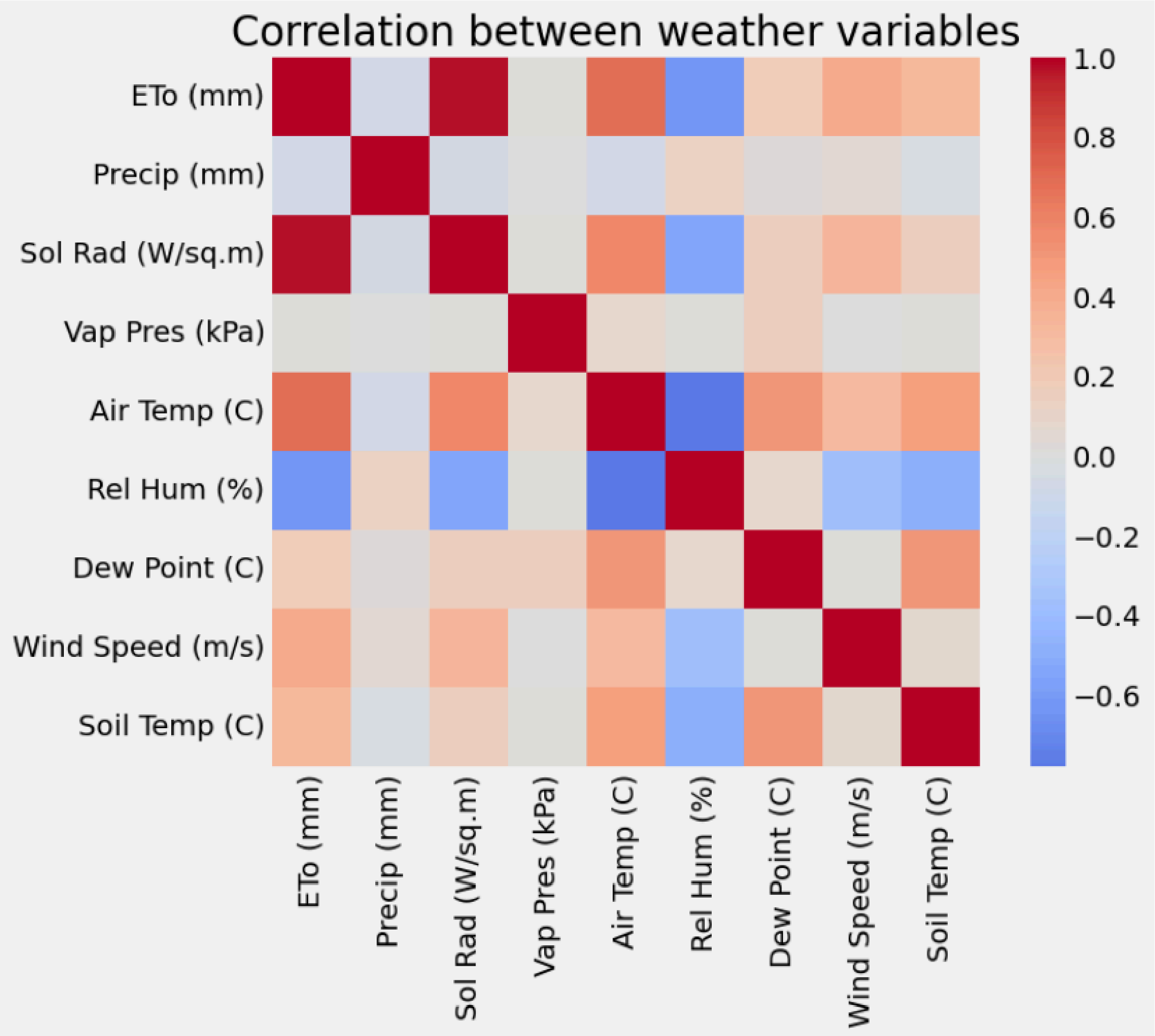
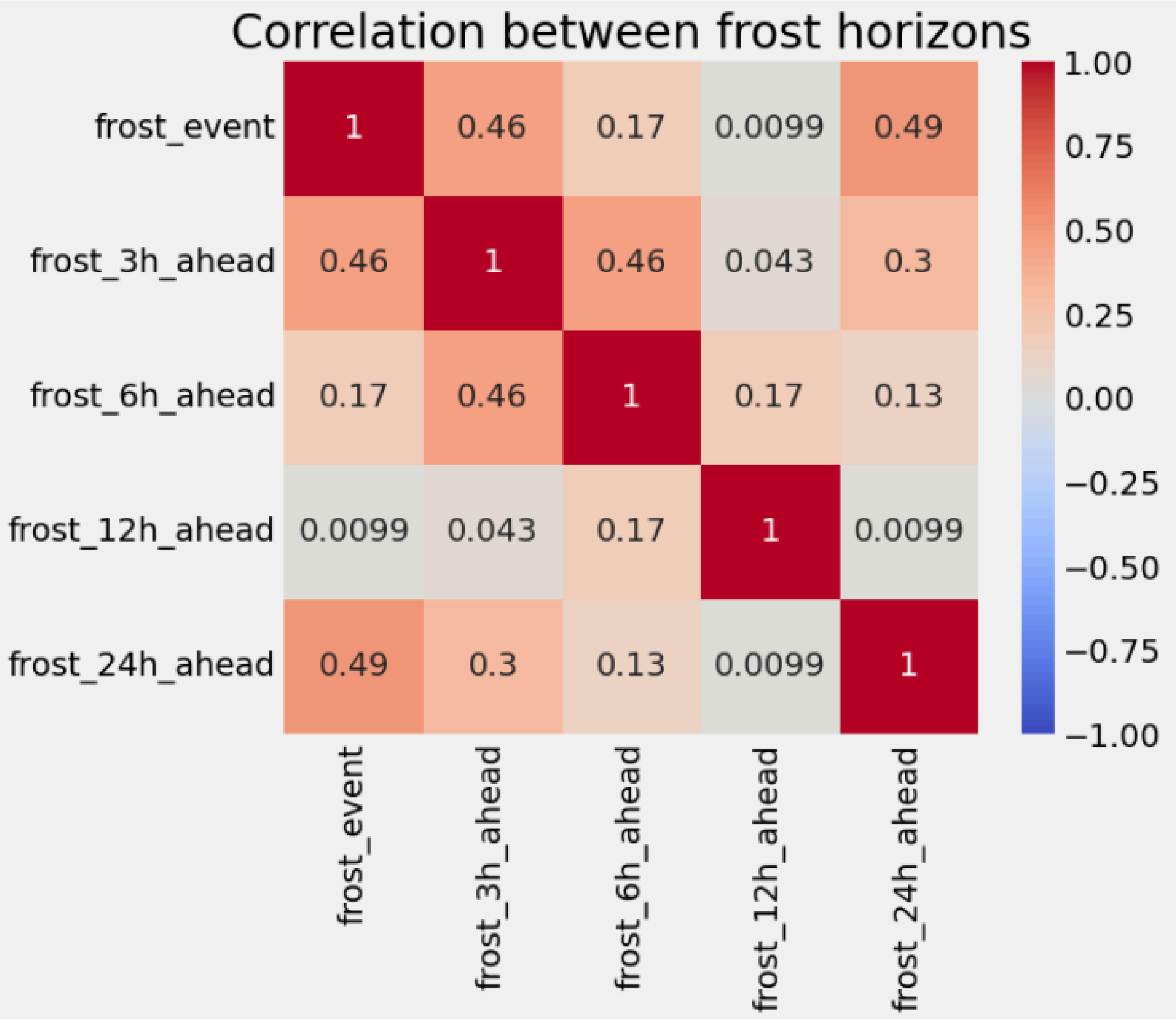

Correlation between weather variables
ETo (mm)
Precip (mm)
Sol Rad (W/sq.m)
Vap Pres (kPa)
Air Temp (C)
Rel Hum (%)
Dew Point (C)
Wind Speed (m/s)
Soil Temp (C)
ETo (mm)
Precip (mm)
Sol Rad (W/sq.m)
Vap Pres (kPa)
Air Temp (C)
Rel Hum (%)
Dew Point (C)
Wind Speed (m/s)
Soil Temp (C)
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Figure 3:



Correlation between frost horizons

|  | frost_event | frost_3h_ahead | frost_6h_ahead | frost_12h_ahead | frost_24h_ahead |
|---|---|---|---|---|---|
| frost_event | 1 | 0.46 | 0.17 | 0.0099 | 0.49 |
| frost_3h_ahead | 0.46 | 1 | 0.46 | 0.043 | 0.3 |
| frost_6h_ahead | 0.17 | 0.46 | 1 | 0.17 | 0.13 |
| frost_12h_ahead | 0.0099 | 0.043 | 0.17 | 1 | 0.0099 |
| frost_24h_ahead | 0.49 | 0.3 | 0.13 | 0.0099 | 1 |

Figure 4

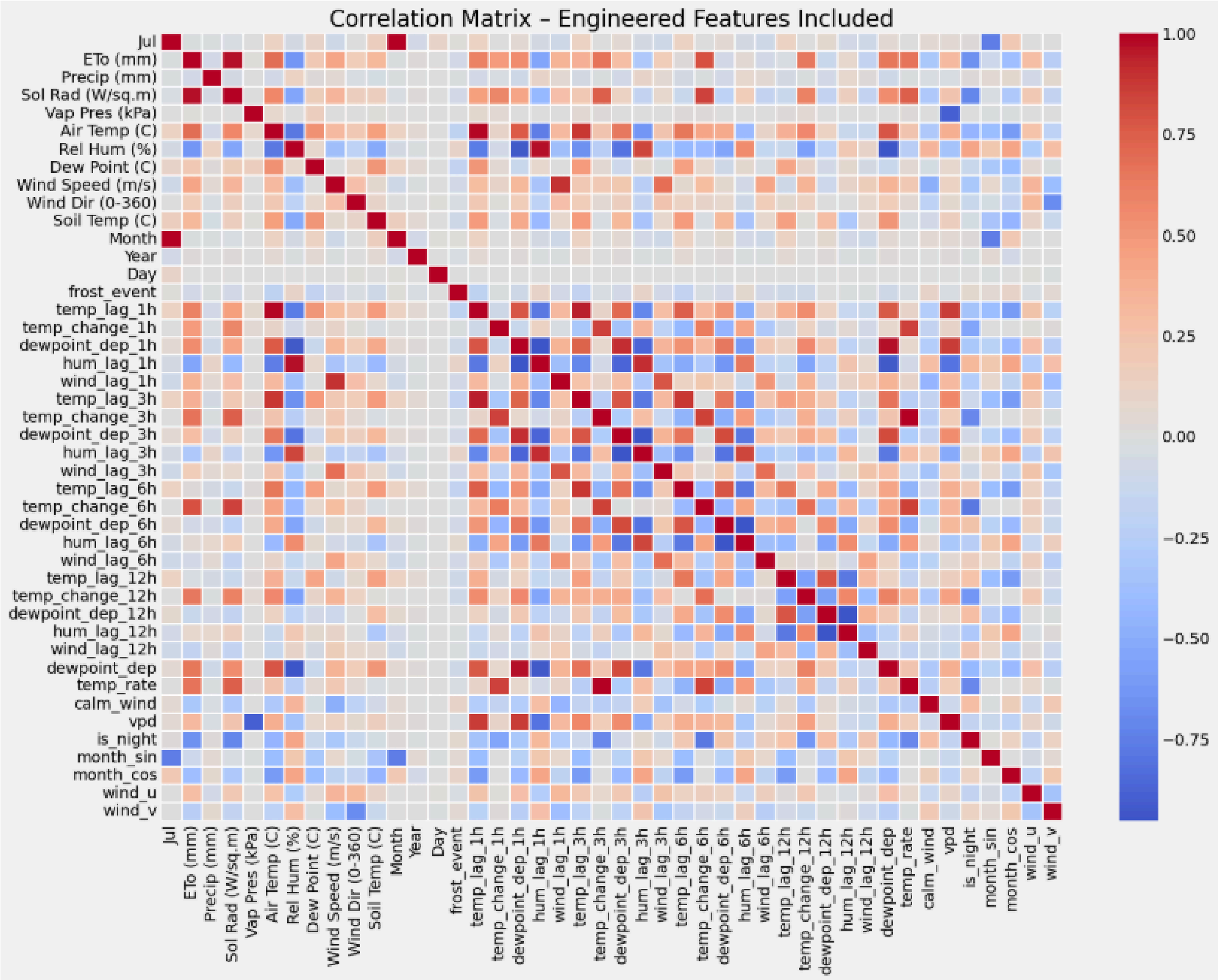


Figure 5:

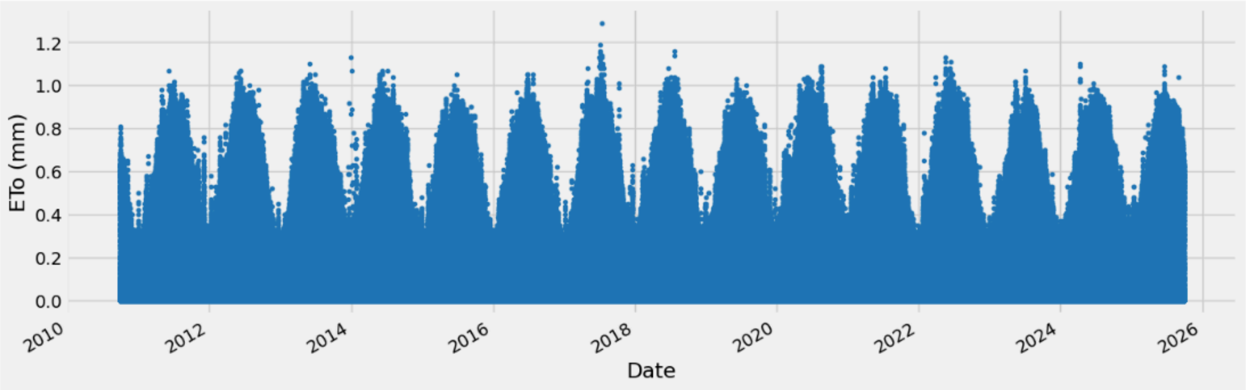


Figure 6:

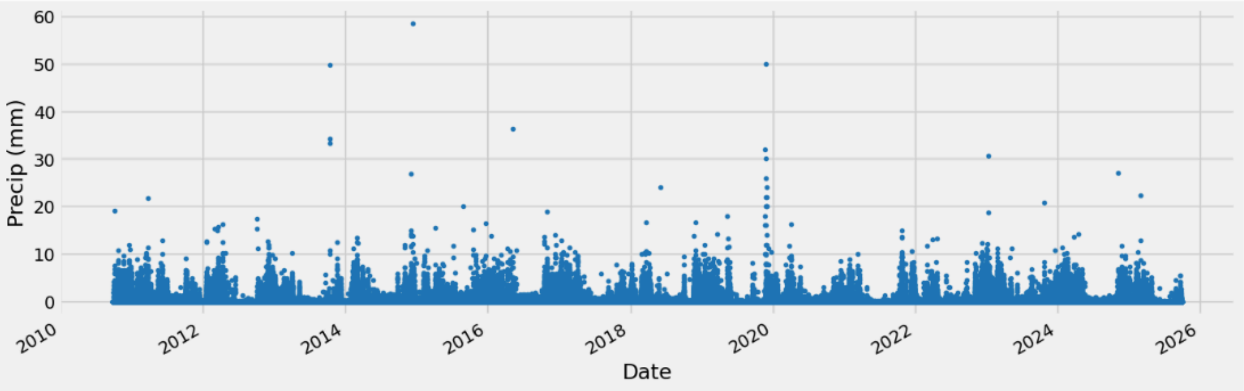


Figure 7:

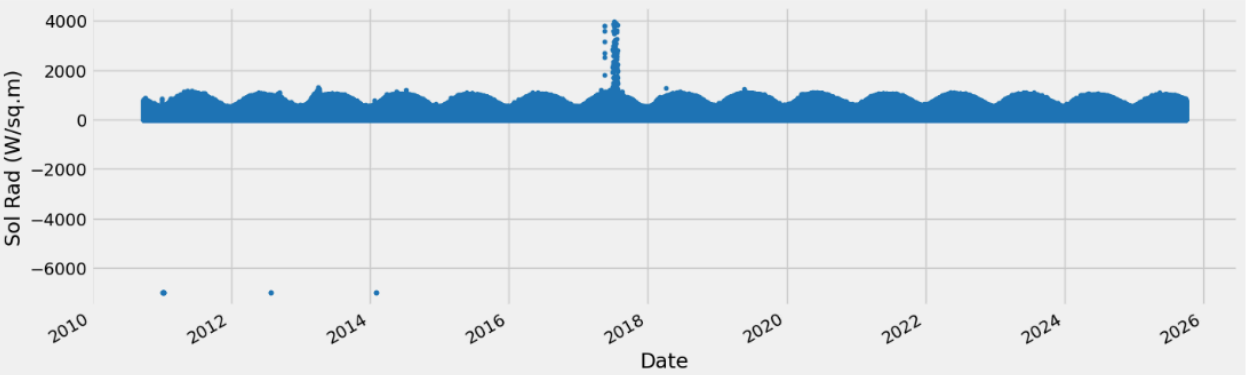


Figure 8:

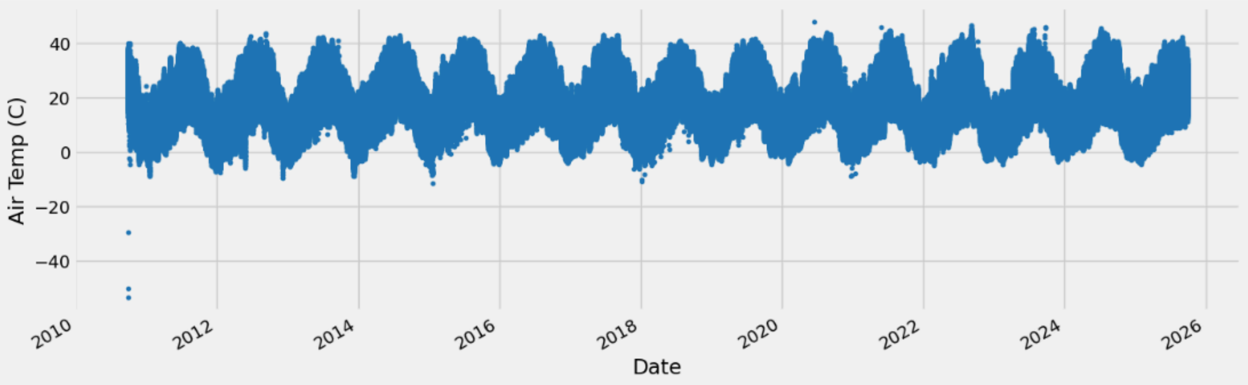


Figure 9:

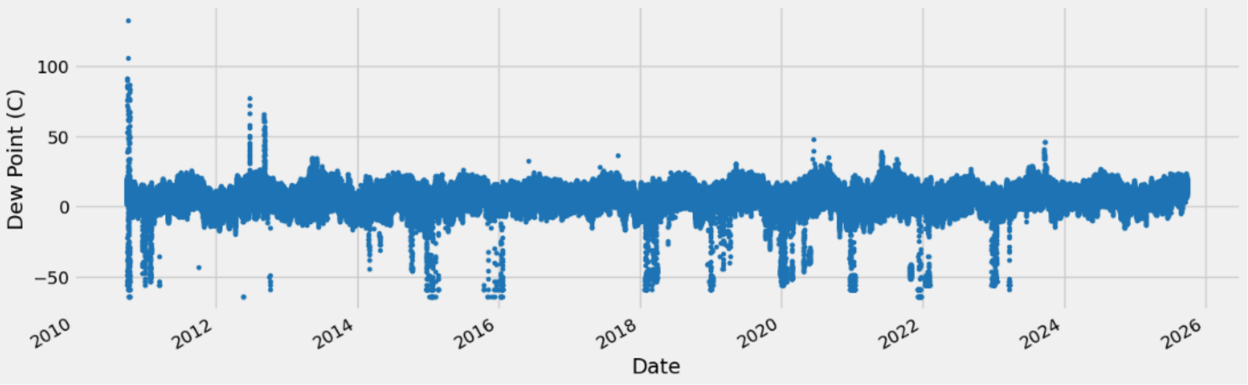


Figure 10:

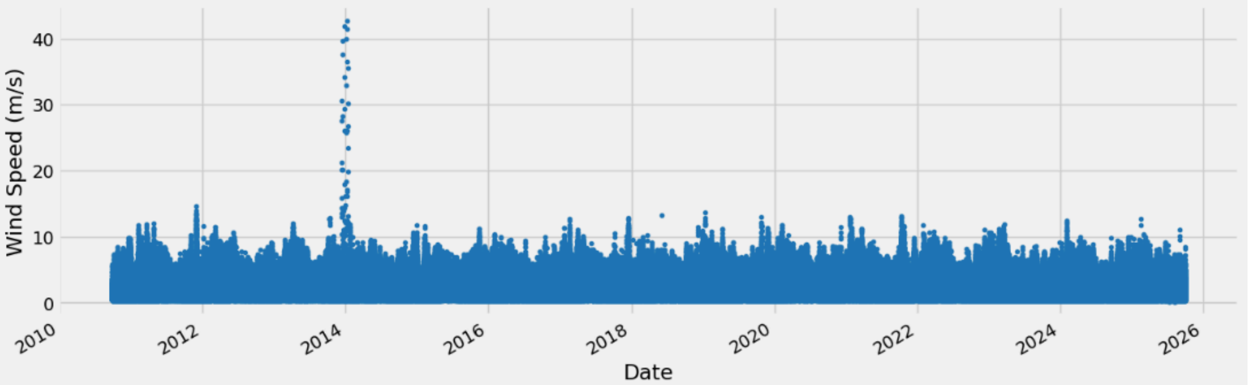
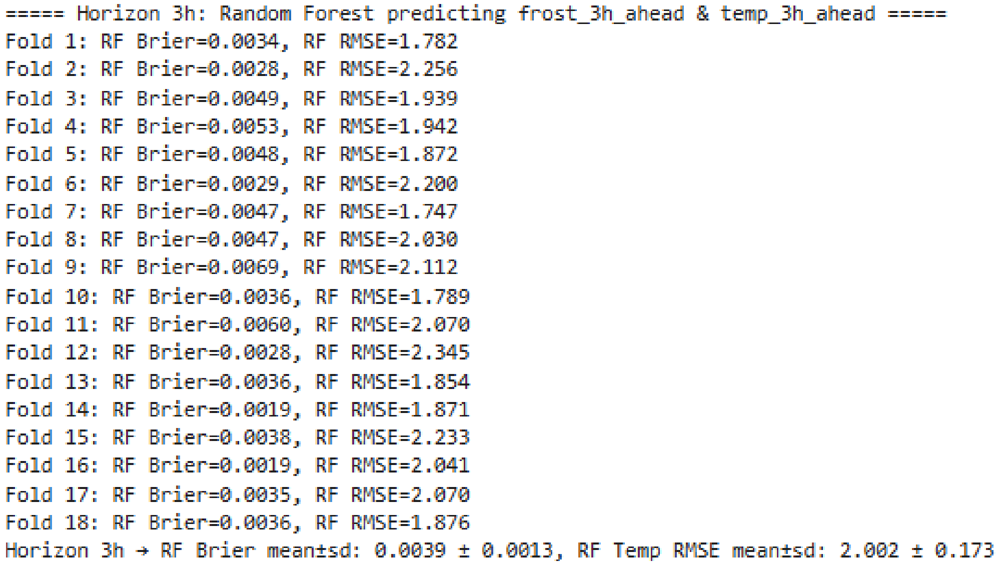
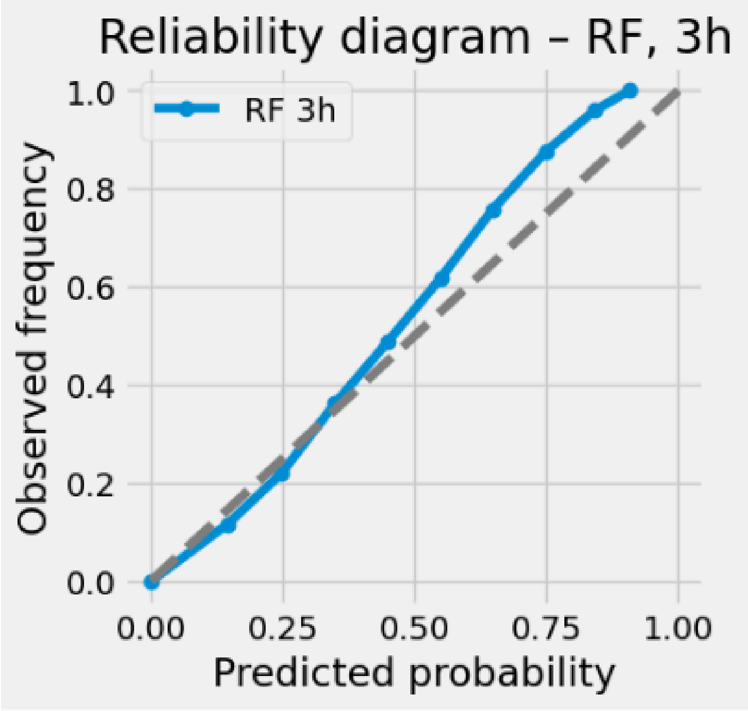


Figure 11:

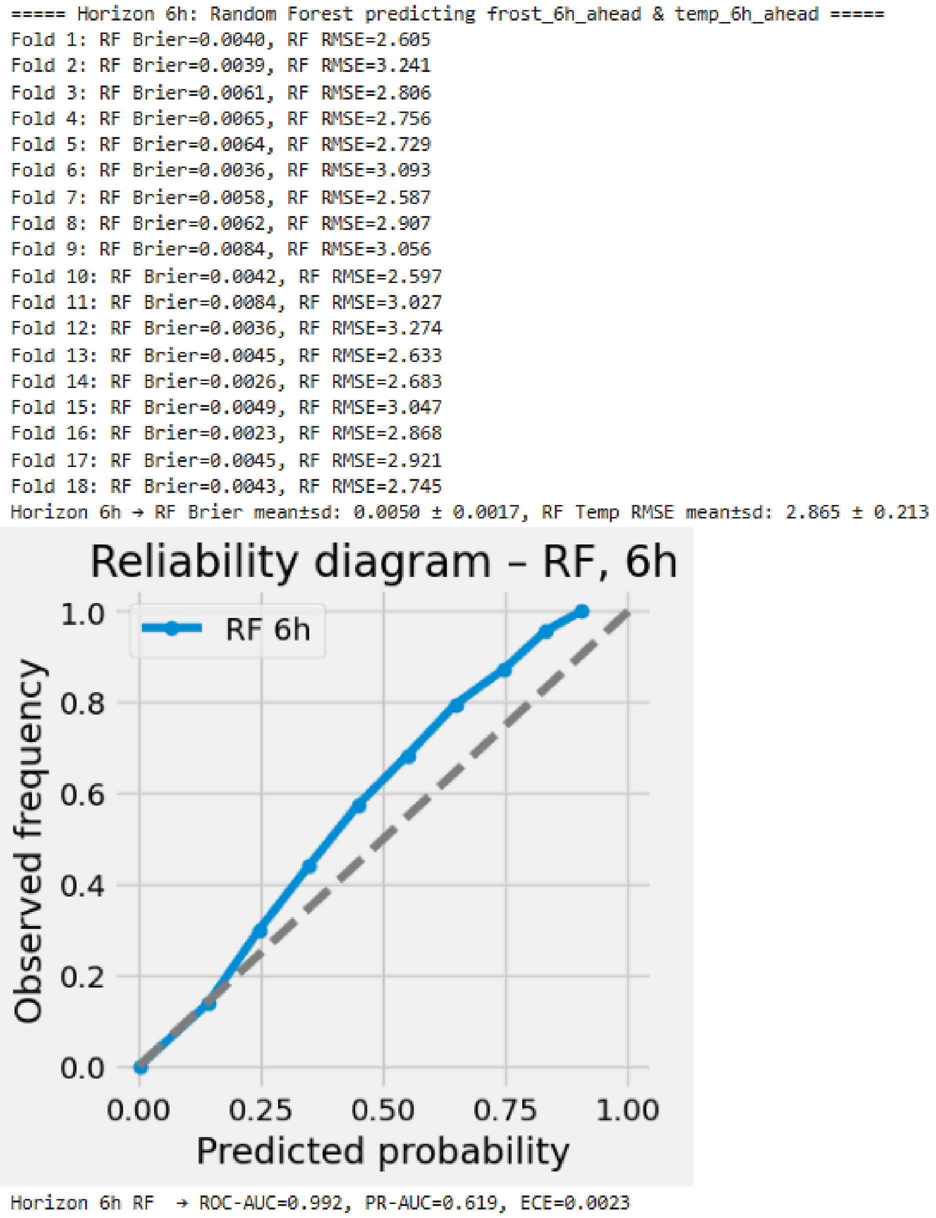


```
===== Horizon 3h: Random Forest predicting frost_3h_ahead & temp_3h_ahead =====
Fold 1: RF Brier=0.0034, RF RMSE=1.782
Fold 2: RF Brier=0.0028, RF RMSE=2.256
Fold 3: RF Brier=0.0049, RF RMSE=1.939
Fold 4: RF Brier=0.0053, RF RMSE=1.942
Fold 5: RF Brier=0.0048, RF RMSE=1.872
Fold 6: RF Brier=0.0029, RF RMSE=2.200
Fold 7: RF Brier=0.0047, RF RMSE=1.747
Fold 8: RF Brier=0.0047, RF RMSE=2.030
Fold 9: RF Brier=0.0069, RF RMSE=2.112
Fold 10: RF Brier=0.0036, RF RMSE=1.789
Fold 11: RF Brier=0.0060, RF RMSE=2.070
Fold 12: RF Brier=0.0028, RF RMSE=2.345
Fold 13: RF Brier=0.0036, RF RMSE=1.854
Fold 14: RF Brier=0.0019, RF RMSE=1.871
Fold 15: RF Brier=0.0038, RF RMSE=2.233
Fold 16: RF Brier=0.0019, RF RMSE=2.041
Fold 17: RF Brier=0.0035, RF RMSE=2.070
Fold 18: RF Brier=0.0036, RF RMSE=1.876
Horizon 3h → RF Brier mean±sd: 0.0039 ± 0.0013, RF Temp RMSE mean±sd: 2.002 ± 0.173
```



```
Horizon 3h RF  → ROC-AUC=0.996, PR-AUC=0.746, ECE=0.0014
```

Figure 12:

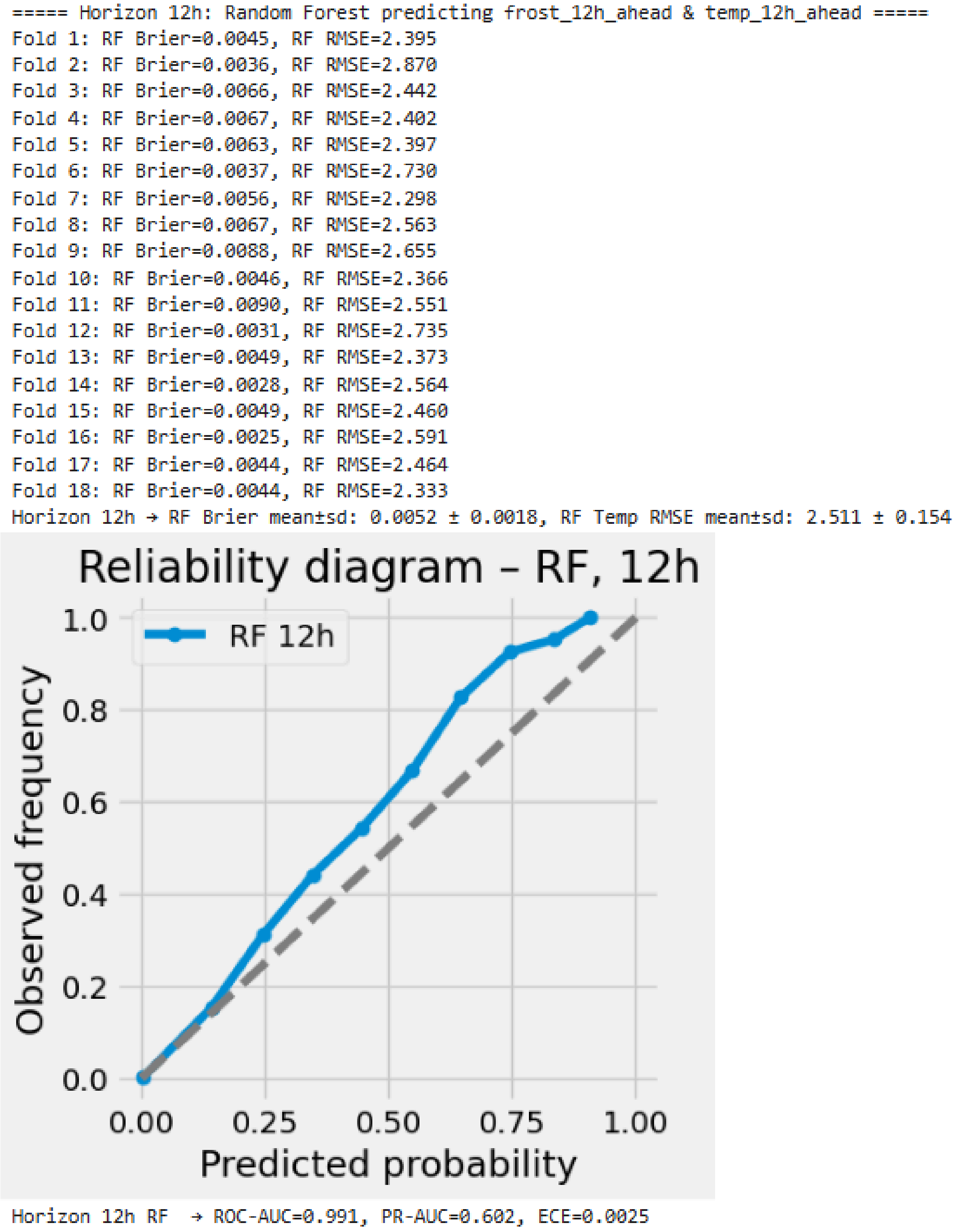


```
===== Horizon 6h: Random Forest predicting frost_6h_ahead & temp_6h_ahead =====
Fold 1: RF Brier=0.0040, RF RMSE=2.605
Fold 2: RF Brier=0.0039, RF RMSE=3.241
Fold 3: RF Brier=0.0061, RF RMSE=2.806
Fold 4: RF Brier=0.0065, RF RMSE=2.756
Fold 5: RF Brier=0.0064, RF RMSE=2.729
Fold 6: RF Brier=0.0036, RF RMSE=3.093
Fold 7: RF Brier=0.0058, RF RMSE=2.587
Fold 8: RF Brier=0.0062, RF RMSE=2.907
Fold 9: RF Brier=0.0084, RF RMSE=3.056
Fold 10: RF Brier=0.0042, RF RMSE=2.597
Fold 11: RF Brier=0.0084, RF RMSE=3.027
Fold 12: RF Brier=0.0036, RF RMSE=3.274
Fold 13: RF Brier=0.0045, RF RMSE=2.633
Fold 14: RF Brier=0.0026, RF RMSE=2.683
Fold 15: RF Brier=0.0049, RF RMSE=3.047
Fold 16: RF Brier=0.0023, RF RMSE=2.868
Fold 17: RF Brier=0.0045, RF RMSE=2.921
Fold 18: RF Brier=0.0043, RF RMSE=2.745
Horizon 6h → RF Brier mean±sd: 0.0050 ± 0.0017, RF Temp RMSE mean±sd: 2.865 ± 0.213
```

Reliability diagram – RF, 6h

```
Horizon 6h RF  → ROC-AUC=0.992, PR-AUC=0.619, ECE=0.0023
```

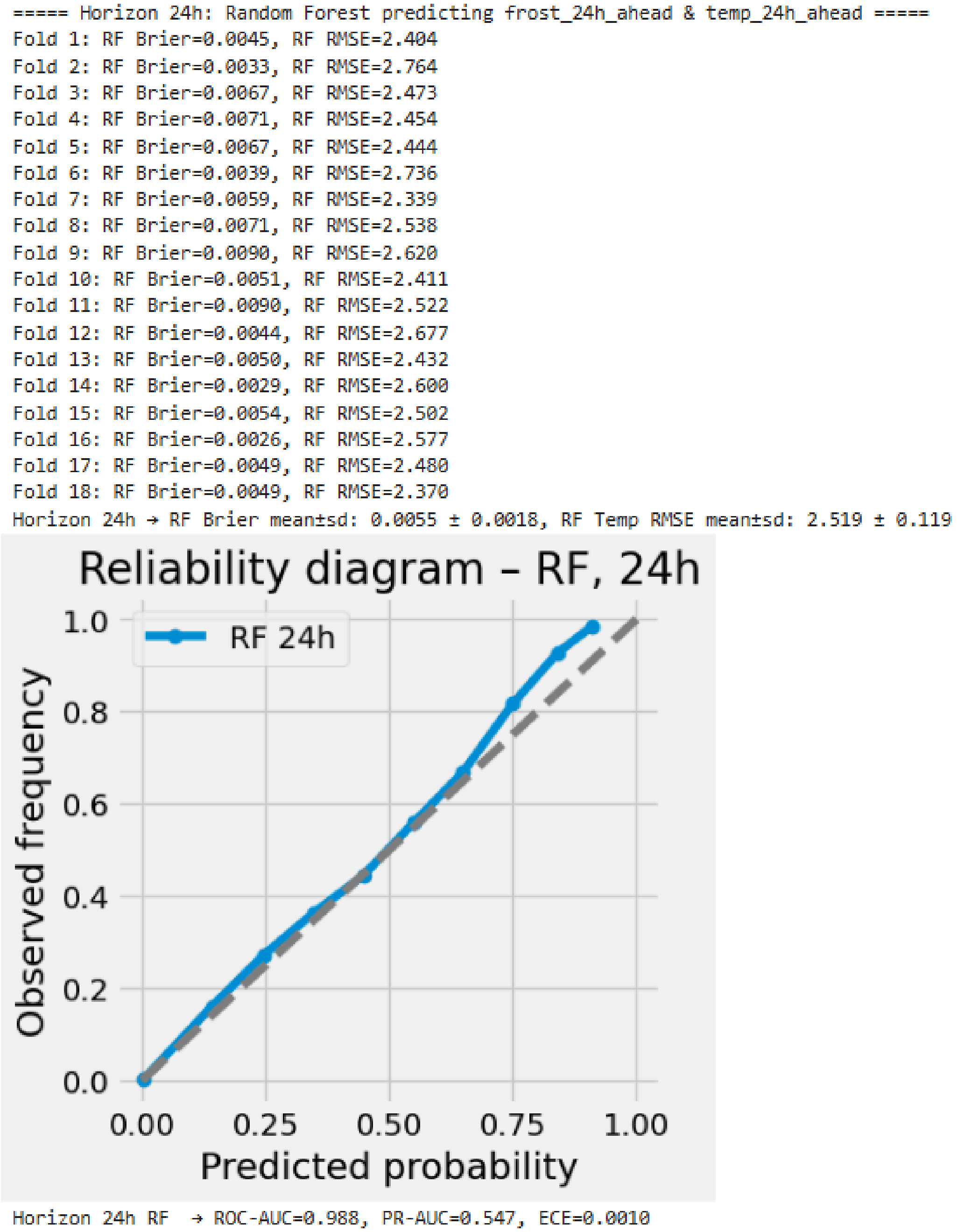
Figure 13:



```
===== Horizon 12h: Random Forest predicting frost_12h_ahead & temp_12h_ahead =====
Fold 1: RF Brier=0.0045, RF RMSE=2.395
Fold 2: RF Brier=0.0036, RF RMSE=2.870
Fold 3: RF Brier=0.0066, RF RMSE=2.442
Fold 4: RF Brier=0.0067, RF RMSE=2.402
Fold 5: RF Brier=0.0063, RF RMSE=2.397
Fold 6: RF Brier=0.0037, RF RMSE=2.730
Fold 7: RF Brier=0.0056, RF RMSE=2.298
Fold 8: RF Brier=0.0067, RF RMSE=2.563
Fold 9: RF Brier=0.0088, RF RMSE=2.655
Fold 10: RF Brier=0.0046, RF RMSE=2.366
Fold 11: RF Brier=0.0090, RF RMSE=2.551
Fold 12: RF Brier=0.0031, RF RMSE=2.735
Fold 13: RF Brier=0.0049, RF RMSE=2.373
Fold 14: RF Brier=0.0028, RF RMSE=2.564
Fold 15: RF Brier=0.0049, RF RMSE=2.460
Fold 16: RF Brier=0.0025, RF RMSE=2.591
Fold 17: RF Brier=0.0044, RF RMSE=2.464
Fold 18: RF Brier=0.0044, RF RMSE=2.333
Horizon 12h → RF Brier mean±sd: 0.0052 ± 0.0018, RF Temp RMSE mean±sd: 2.511 ± 0.154
```

```
Horizon 12h RF  → ROC-AUC=0.991, PR-AUC=0.602, ECE=0.0025
```

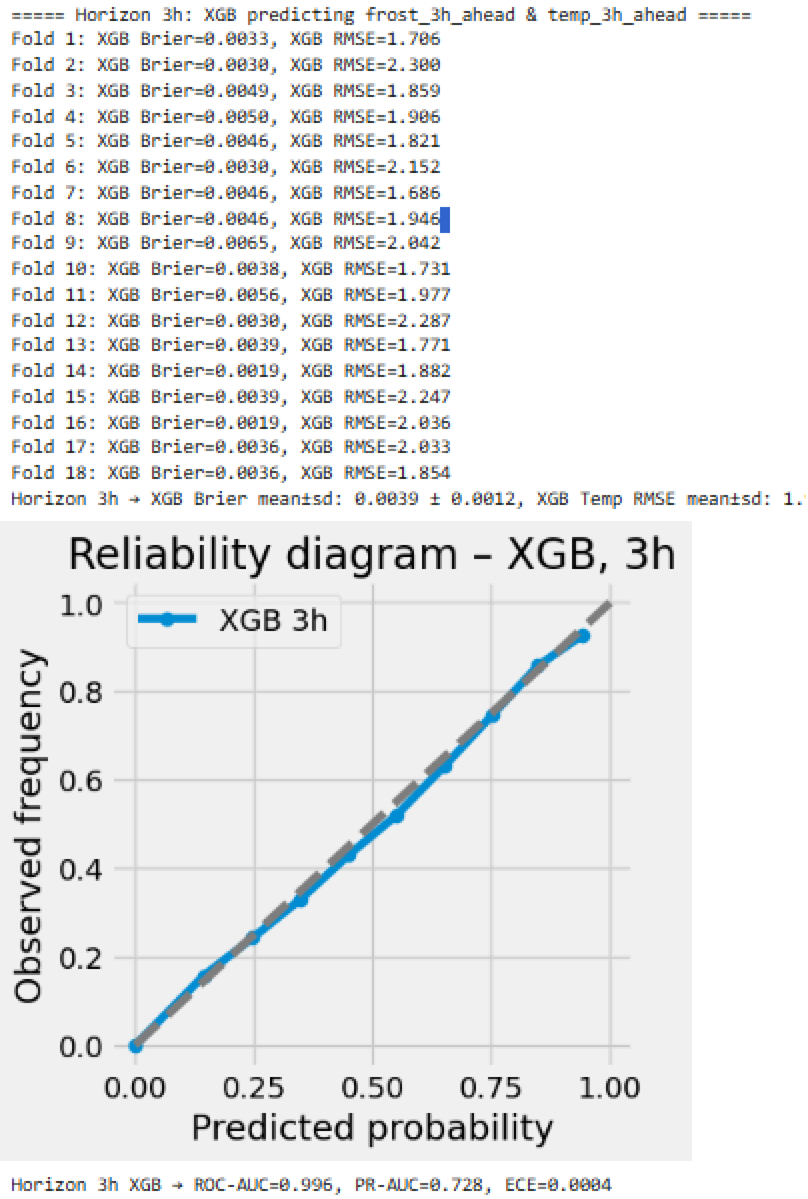
Figure 14:



```
===== Horizon 24h: Random Forest predicting frost_24h_ahead & temp_24h_ahead =====
Fold 1: RF Brier=0.0045, RF RMSE=2.404
Fold 2: RF Brier=0.0033, RF RMSE=2.764
Fold 3: RF Brier=0.0067, RF RMSE=2.473
Fold 4: RF Brier=0.0071, RF RMSE=2.454
Fold 5: RF Brier=0.0067, RF RMSE=2.444
Fold 6: RF Brier=0.0039, RF RMSE=2.736
Fold 7: RF Brier=0.0059, RF RMSE=2.339
Fold 8: RF Brier=0.0071, RF RMSE=2.538
Fold 9: RF Brier=0.0090, RF RMSE=2.620
Fold 10: RF Brier=0.0051, RF RMSE=2.411
Fold 11: RF Brier=0.0090, RF RMSE=2.522
Fold 12: RF Brier=0.0044, RF RMSE=2.677
Fold 13: RF Brier=0.0050, RF RMSE=2.432
Fold 14: RF Brier=0.0029, RF RMSE=2.600
Fold 15: RF Brier=0.0054, RF RMSE=2.502
Fold 16: RF Brier=0.0026, RF RMSE=2.577
Fold 17: RF Brier=0.0049, RF RMSE=2.480
Fold 18: RF Brier=0.0049, RF RMSE=2.370
Horizon 24h → RF Brier mean±sd: 0.0055 ± 0.0018, RF Temp RMSE mean±sd: 2.519 ± 0.119
```

```
Horizon 24h RF  → ROC-AUC=0.988, PR-AUC=0.547, ECE=0.0010
```

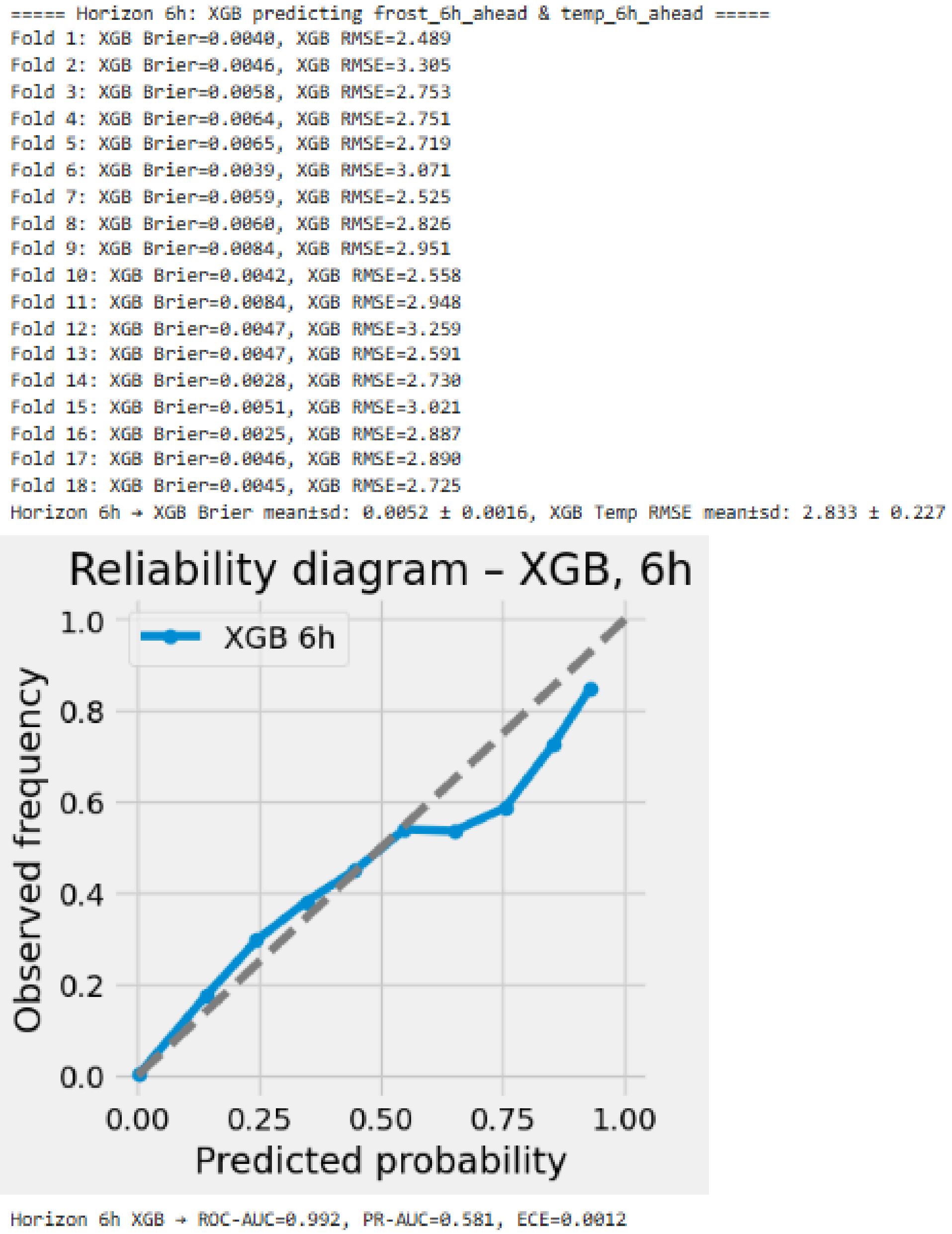
Figure 15:



```
===== Horizon 3h: XGB predicting frost_3h_ahead & temp_3h_ahead =====
Fold 1: XGB Brier=0.0033, XGB RMSE=1.706
Fold 2: XGB Brier=0.0030, XGB RMSE=2.300
Fold 3: XGB Brier=0.0049, XGB RMSE=1.859
Fold 4: XGB Brier=0.0050, XGB RMSE=1.906
Fold 5: XGB Brier=0.0046, XGB RMSE=1.821
Fold 6: XGB Brier=0.0030, XGB RMSE=2.152
Fold 7: XGB Brier=0.0046, XGB RMSE=1.686
Fold 8: XGB Brier=0.0046, XGB RMSE=1.946
Fold 9: XGB Brier=0.0065, XGB RMSE=2.042
Fold 10: XGB Brier=0.0038, XGB RMSE=1.731
Fold 11: XGB Brier=0.0056, XGB RMSE=1.977
Fold 12: XGB Brier=0.0030, XGB RMSE=2.287
Fold 13: XGB Brier=0.0039, XGB RMSE=1.771
Fold 14: XGB Brier=0.0019, XGB RMSE=1.882
Fold 15: XGB Brier=0.0039, XGB RMSE=2.247
Fold 16: XGB Brier=0.0019, XGB RMSE=2.036
Fold 17: XGB Brier=0.0036, XGB RMSE=2.033
Fold 18: XGB Brier=0.0036, XGB RMSE=1.854
Horizon 3h → XGB Brier mean±sd: 0.0039 ± 0.0012, XGB Temp RMSE mean±sd: 1.
```

```
Horizon 3h XGB → ROC-AUC=0.996, PR-AUC=0.728, ECE=0.0004
```

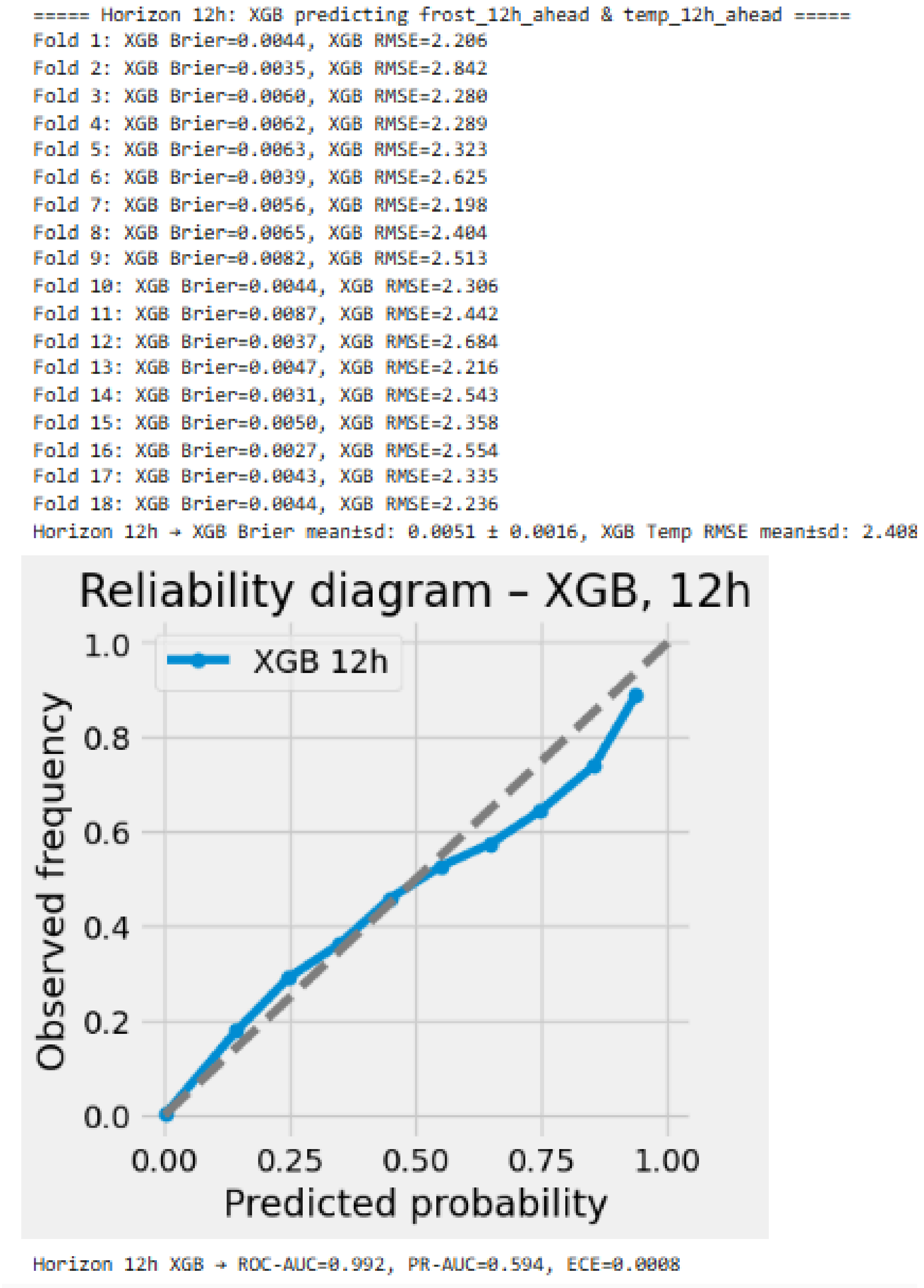
Figure 16:



```
===== Horizon 6h: XGB predicting frost_6h_ahead & temp_6h_ahead =====
Fold 1: XGB Brier=0.0040, XGB RMSE=2.489
Fold 2: XGB Brier=0.0046, XGB RMSE=3.305
Fold 3: XGB Brier=0.0058, XGB RMSE=2.753
Fold 4: XGB Brier=0.0064, XGB RMSE=2.751
Fold 5: XGB Brier=0.0065, XGB RMSE=2.719
Fold 6: XGB Brier=0.0039, XGB RMSE=3.071
Fold 7: XGB Brier=0.0059, XGB RMSE=2.525
Fold 8: XGB Brier=0.0060, XGB RMSE=2.826
Fold 9: XGB Brier=0.0084, XGB RMSE=2.951
Fold 10: XGB Brier=0.0042, XGB RMSE=2.558
Fold 11: XGB Brier=0.0084, XGB RMSE=2.948
Fold 12: XGB Brier=0.0047, XGB RMSE=3.259
Fold 13: XGB Brier=0.0047, XGB RMSE=2.591
Fold 14: XGB Brier=0.0028, XGB RMSE=2.730
Fold 15: XGB Brier=0.0051, XGB RMSE=3.021
Fold 16: XGB Brier=0.0025, XGB RMSE=2.887
Fold 17: XGB Brier=0.0046, XGB RMSE=2.890
Fold 18: XGB Brier=0.0045, XGB RMSE=2.725
Horizon 6h → XGB Brier mean±sd: 0.0052 ± 0.0016, XGB Temp RMSE mean±sd: 2.833 ± 0.227
```

```
Horizon 6h XGB → ROC-AUC=0.992, PR-AUC=0.581, ECE=0.0012
```

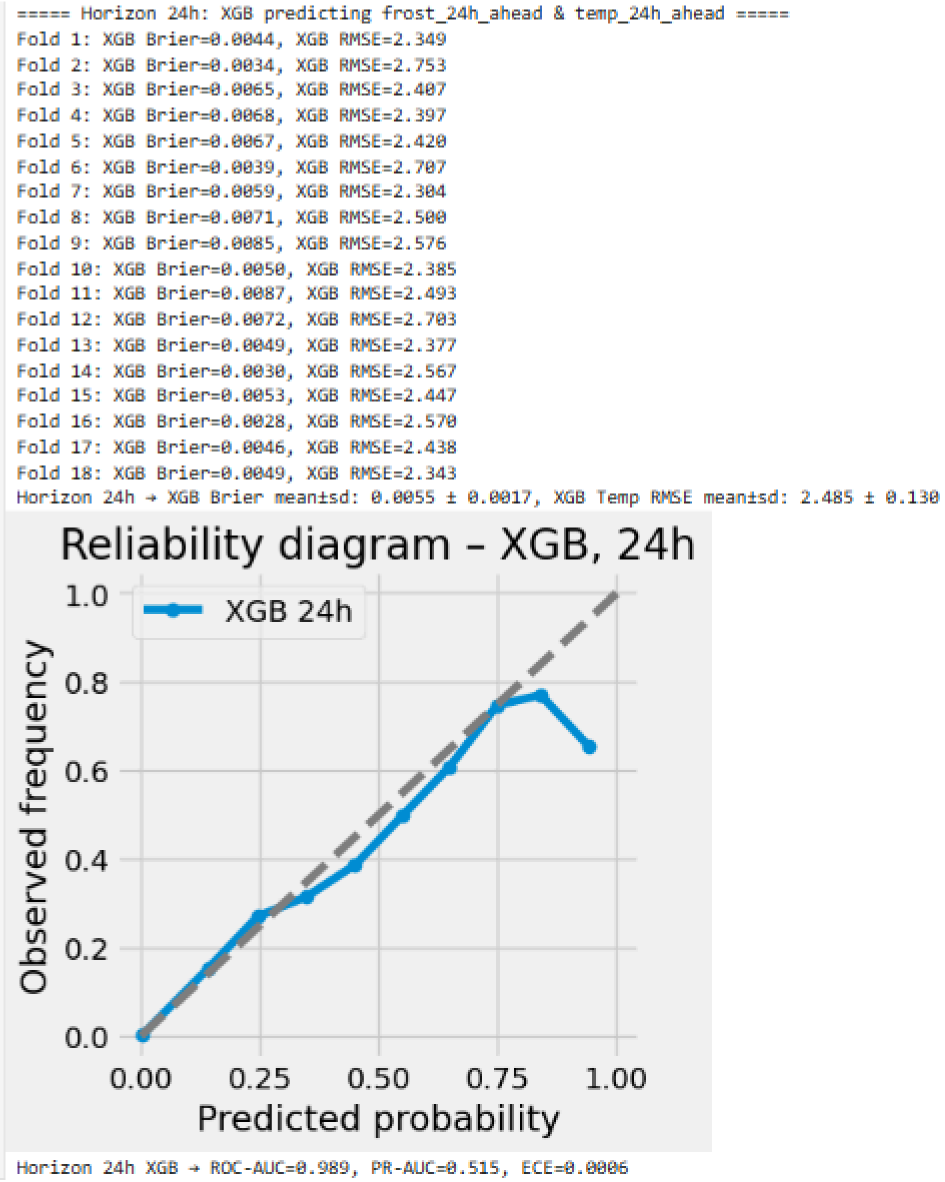
Figure 17:



```
===== Horizon 12h: XGB predicting frost_12h_ahead & temp_12h_ahead =====
Fold 1: XGB Brier=0.0044, XGB RMSE=2.206
Fold 2: XGB Brier=0.0035, XGB RMSE=2.842
Fold 3: XGB Brier=0.0060, XGB RMSE=2.280
Fold 4: XGB Brier=0.0062, XGB RMSE=2.289
Fold 5: XGB Brier=0.0063, XGB RMSE=2.323
Fold 6: XGB Brier=0.0039, XGB RMSE=2.625
Fold 7: XGB Brier=0.0056, XGB RMSE=2.198
Fold 8: XGB Brier=0.0065, XGB RMSE=2.404
Fold 9: XGB Brier=0.0082, XGB RMSE=2.513
Fold 10: XGB Brier=0.0044, XGB RMSE=2.306
Fold 11: XGB Brier=0.0087, XGB RMSE=2.442
Fold 12: XGB Brier=0.0037, XGB RMSE=2.684
Fold 13: XGB Brier=0.0047, XGB RMSE=2.216
Fold 14: XGB Brier=0.0031, XGB RMSE=2.543
Fold 15: XGB Brier=0.0050, XGB RMSE=2.358
Fold 16: XGB Brier=0.0027, XGB RMSE=2.554
Fold 17: XGB Brier=0.0043, XGB RMSE=2.335
Fold 18: XGB Brier=0.0044, XGB RMSE=2.236
Horizon 12h → XGB Brier mean±sd: 0.0051 ± 0.0016, XGB Temp RMSE mean±sd: 2.408
```

```
Horizon 12h XGB → ROC-AUC=0.992, PR-AUC=0.594, ECE=0.0008
```

Figure 18:



```
===== Horizon 24h: XGB predicting frost_24h_ahead & temp_24h_ahead =====
Fold 1: XGB Brier=0.0044, XGB RMSE=2.349
Fold 2: XGB Brier=0.0034, XGB RMSE=2.753
Fold 3: XGB Brier=0.0065, XGB RMSE=2.407
Fold 4: XGB Brier=0.0068, XGB RMSE=2.397
Fold 5: XGB Brier=0.0067, XGB RMSE=2.420
Fold 6: XGB Brier=0.0039, XGB RMSE=2.707
Fold 7: XGB Brier=0.0059, XGB RMSE=2.304
Fold 8: XGB Brier=0.0071, XGB RMSE=2.500
Fold 9: XGB Brier=0.0085, XGB RMSE=2.576
Fold 10: XGB Brier=0.0050, XGB RMSE=2.385
Fold 11: XGB Brier=0.0087, XGB RMSE=2.493
Fold 12: XGB Brier=0.0072, XGB RMSE=2.703
Fold 13: XGB Brier=0.0049, XGB RMSE=2.377
Fold 14: XGB Brier=0.0030, XGB RMSE=2.567
Fold 15: XGB Brier=0.0053, XGB RMSE=2.447
Fold 16: XGB Brier=0.0028, XGB RMSE=2.570
Fold 17: XGB Brier=0.0046, XGB RMSE=2.438
Fold 18: XGB Brier=0.0049, XGB RMSE=2.343
Horizon 24h → XGB Brier mean±sd: 0.0055 ± 0.0017, XGB Temp RMSE mean±sd: 2.485 ± 0.130
```

```
Horizon 24h XGB → ROC-AUC=0.989, PR-AUC=0.515, ECE=0.0006
```