Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

FrostTR: A calibrated and spatially transferable machine-learning framework for 24-hour agricultural frost forecasting across Türkiye

Mehmet Tevfik AĞDAŞ

mtagdas@munzur.edu.tr

Munzur University

Article

Keywords: agricultural frost, machine learning, XGBoost, spatial transferability, probability calibration, early warning

Posted Date: September 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10555645/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

FrostTR: A calibrated and spatially transferable machine-learning framework for 24-hour agricultural frost forecasting across Türkiye

Mehmet Tevfik AĞDAŞ

Department of Computer Technologies, Çemişgezek Vocational School, Munzur University, Tunceli 62600, Türkiye

ORCID: 0000-0002-5608-6240

Corresponding author: Mehmet Tevfik AĞDAŞ; mtagdas@munzur.edu.tr

Article type: Research Article

Abstract

Reliable short-range frost forecasting is an Earth science informatics problem that requires heterogeneous observations, terrain context, calibrated probabilities, and validation that reflects real temporal and spatial deployment. We developed FrostTR, a machine-learning framework for 24-hour agricultural frost forecasting and minimum-temperature estimation across Türkiye. Hourly NOAA Integrated Surface Database observations from 2015-2025 were assembled for 30 stations and combined with Copernicus GLO-30 terrain descriptors. XGBoost models were trained on 2015-2023, tuned only on 2024, and evaluated on 34,011 observation-supported issue times from 2025. The primary classifier achieved F1 0.8908, AUPRC 0.9681, AUROC 0.9802, Brier score 0.0538, and expected calibration error 0.0066; minimum-temperature mean absolute error was 1.586 °C. Full leave-one-station-out validation across 29 modelled stations yielded macro-F1 0.8474, while nine rolling-origin tests supported temporal stability. Event-based verification detected 1,754 of 1,835 observed frost episodes, with probability of detection 0.9559, false-alarm ratio 0.0665, critical success index 0.8949, and mean warning lead time 20.7 h. Results remained stable under stricter station quality control, alternative issue-time strides, multiple random seeds, missingness ablation, and topographic ablation. FrostTR demonstrates how public Earth-system data can be transformed into calibrated, explainable, and spatially transferable early-warning information, while preserving explicit limits arising from observation gaps and station-specific microclimates.

Keywords: agricultural frost; machine learning; XGBoost; spatial transferability; probability calibration; early warning

Introduction

Near-surface frost is a high-impact atmospheric boundary-layer event whose occurrence depends on radiative cooling, moisture, wind, cloud cover, local exposure, and terrain-controlled cold-air drainage. Because the damaging threshold is crossed at local and often sub-daily scales, useful warning systems must convert irregular station observations and static geospatial information into forecasts that are both discriminative and probabilistically reliable. This combination of heterogeneous Earth observations, spatiotemporal feature construction, model validation, and decision-oriented output places agricultural frost forecasting squarely within Earth science informatics.

The spatial structure of frost is strongly conditioned by terrain. Local atmospheric decoupling, slope orientation, relief, and cold-air pooling can create minimum-temperature contrasts that are poorly represented by regional mean conditions (Chung et al. 2006; Daly et al. 2010; Lundquist et al. 2008). Static digital elevation models can therefore provide useful context when meteorological station networks are sparse, although their added value must be quantified rather than assumed.

Machine-learning approaches have increasingly been applied to frost prediction, temperature downscaling, and agricultural early warning (Liakos et al. 2018). Previous studies have used artificial neural networks, random forests, support-vector machines, convolutional models, and numerical-model post-processing at scales ranging from individual farms to national station archives (Diedrichs et al. 2018; Noh et al. 2021; Rozante et al. 2023; Talsma et al. 2023; Pérez Tárraga et al. 2025; Bhatti et al. 2026). These studies demonstrate the promise of data-driven forecasting, but their reported skill is not always directly transferable to operational deployment because forecast horizon, sampling design, spatial holdout strategy, and probability calibration differ substantially.

Five recurring methodological gaps motivate the present study. First, random or weakly separated train-test designs can leak highly similar weather sequences across partitions. Second, single-site validation does not establish transfer to an unseen station. Third, threshold metrics alone do not show whether forecast probabilities are reliable enough for cost-sensitive decisions. Fourth, overlapping issue times and incomplete station reporting can produce overly optimistic uncertainty estimates. Fifth, event-level warning performance and lead time are often omitted even though they determine operational usefulness. An Earth science informatics framework should address these issues jointly and document how data quality, terrain context, and model stochasticity affect the reported conclusions.

Table 1. Selected frost-forecasting studies and the methodological position of FrostTR.

Study	Scope	Forecast target	Model family	Validation design	Remaining limitation
Ghielmi & Eccel (2006)	Multiple Italian sites	Overnight minimum T	ANN vs descriptive models	Site-specific historical tests	No national spatial holdout or probability calibration
Diedrichs et al. (2018)	IoT crop-field network	Short-range frost	Random forest	Local deployment evaluation	Limited geographic scope
Noh et al. (2021)	Korean orchard-oriented stations	Hourly frost	LR, tree, RF, SVM	Held-out records	Orchard-specific design
Talsma et al. (2023)	Point-scale farms, New Mexico	6-48 h	RF, DNN, CNN	Temporal/seasonal evaluation	Single regional setting
Rozante et al. (2023)	Southern Brazil	Frost post-processing	MLP ANN	2016 validation; 2017 test	Dependent on numerical-model inputs

Study	Scope	Forecast target	Model family	Validation design	Remaining limitation
Shammi et al. (2024)	Wheat field thermography	Early frost detection	Four ML classifiers	Field experiments	Detection rather than regional forecasting
Pérez Tárraga et al. (2025)	274 Spanish stations	Next-day frost	KNN, SVM, GBM, RF, Extra Trees	Historical station archive	Daily resolution; spatial transfer not central
Bhakare et al. (2025)	Two Alpine valleys	Daily Tmin downscaling	ANN, RF, CNN	Valley-scale validation	Downscaling rather than event warning
Bhatti et al. (2026)	Cranberry beds, Massachusetts	Spring frost	Random forest	Operational case study	Crop- and region-specific
Present study	30 assembled; 29 modelled stations across Türkiye	24-h frost probability and minimum temperature	Calibrated XGBoost classification and regression	2015-2023 train; 2024 validation; 2025 test; full 29-station LOSO; rolling origin	Observation gaps and station-specific microclimates remain

Table 1 positions FrostTR within the frost-forecasting literature. The contribution is not a new boosting algorithm; it is an integrated, reproducible evaluation architecture that combines a national station network, leakage-safe temporal separation, calibrated probability assessment, full leave-one-station-out validation, rolling-origin testing, event-level verification, and explicit sensitivity analyses.

The objectives were to: (i) develop a calibrated 24-hour frost-probability and minimum-temperature framework from public meteorological and terrain data; (ii) preserve strict chronological separation between training, threshold selection, and evaluation; (iii) quantify transfer to unseen stations and stability across test years; (iv) test sensitivity to station quality, missingness descriptors, terrain variables, issue-time overlap, and random initialization; and (v) translate window-level predictions into event detection and warning lead time. The resulting framework is intended as an Earth science information product for agricultural early warning rather than a direct model of crop injury.

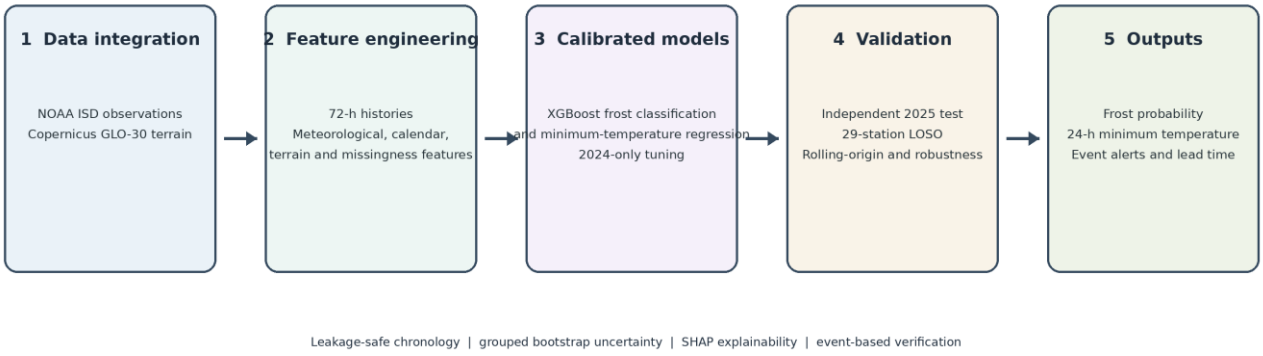


Fig. 1 Conceptual architecture and validation scope of the FrostTR framework

Fig. 1 summarizes the information flow from public Earth-observation archives through spatiotemporal feature engineering and calibrated ensemble learning to independent temporal, spatial, robustness, explainability, and event-based verification. The outputs are meteorological frost probability, 24-hour minimum temperature, and event-warning lead time.

Methods

Study design and data sources

Hourly surface observations were obtained from the NOAA Integrated Surface Database (ISD), a harmonized global archive of synoptic and hourly meteorological reports (Smith et al. 2011). The workflow identified 126 stations within Türkiye, assembled 30 stations for the analytical network, and retained 29 stations in the observation-supported primary evaluation. Static terrain context was derived from the Copernicus GLO-30 digital elevation model. All reported predictors therefore originated from public Earth-system observations or reproducible transformations of those observations; no numerical weather-prediction or ERA5 variables were used in the fitted models.

The analysis was restricted to October-May issue times. This choice concentrated the modelling task on meteorologically plausible frost periods while retaining autumn, winter, and spring transitions. All

timestamps were stored in UTC; cyclical local-hour and day-of-year variables were derived after conversion to Europe/Istanbul time.

Table 2. Earth-system data sources, feature groups, and analytical roles.

Source/group	Resolution	Variables	Period	Role
NOAA Integrated Surface Database (ISD)	Hourly surface observations	Temperature, dew point, pressure, wind, sky condition, precipitation	2015-2025	Primary meteorological source
Copernicus GLO-30	Static terrain	Elevation, local mean and SD, relief, TPI, mean/max slope, aspect sine/cosine	Static	Topographic context
Derived temporal variables	Calendar encoding	Hour and day-of-year sine/cosine	Hourly	Seasonality and diurnal phase
Missingness descriptors	Observation availability	Missing and imputed flags/source codes	Hourly	Sensitivity to reporting patterns
Target construction	Future observed/completed temperature	24-h minimum temperature, binary frost, four-level severity	6, 12, 24, 48 h generated; 24 h analysed	Primary and secondary endpoints

Table 2 documents the provenance, resolution, and analytical role of each data component. NOAA observations provide the dynamic atmospheric state, Copernicus terrain descriptors provide static geospatial context, calendar variables encode diurnal and seasonal phase, and missingness descriptors allow explicit sensitivity analysis of reporting patterns.

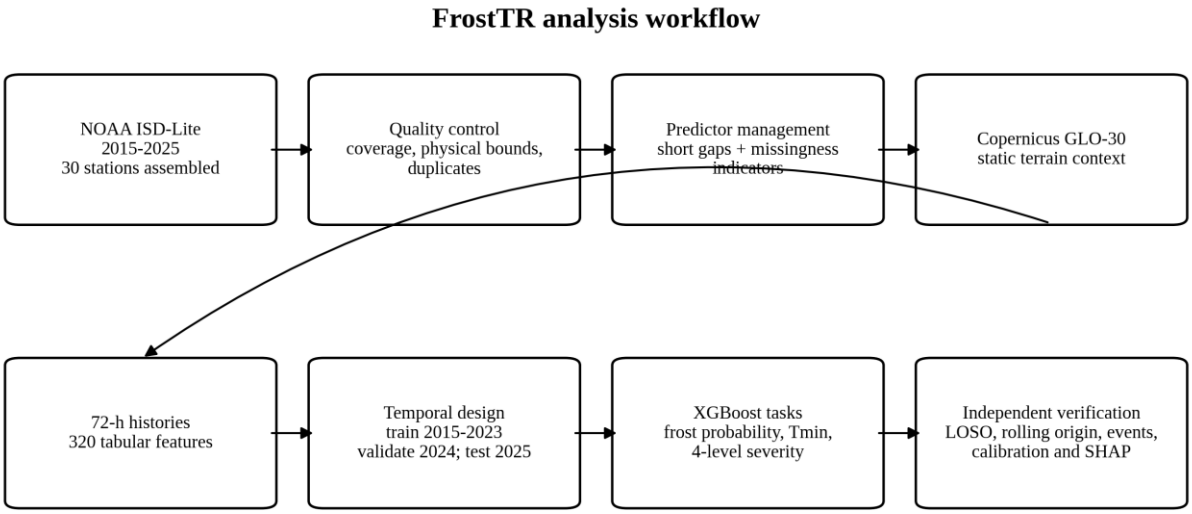


Fig. 2 End-to-end Earth science informatics workflow from public data acquisition to independent validation

Fig. 2 shows the separation between data preparation, temporal modelling, and independent verification procedures.

Station selection and quality control

Station quality was evaluated using overall hourly coverage, cold-season coverage, the number of frost days, and physical plausibility. Pre-specified preferred thresholds were overall coverage ≥0.65, cold-season coverage ≥0.78, and at least 10 frost days. Twelve stations met the complete strict-QC criterion. To represent the national range of coastal, inland, plateau, and high-elevation climates, the extended network also retained lower-coverage stations. The strict-QC and extended networks were subsequently evaluated as separate experiments rather than treated as interchangeable datasets.

Raw values were converted to physical units and screened using temperature -60 to 60 °C, dew point -80 to 50 °C, sea-level pressure 800-1100 hPa, wind speed 0-80 m s−1, non-negative precipitation, and a dew-point constraint that removed values exceeding air temperature by more than 2 °C. Duplicate reports within an hour were collapsed. Original observations were retained in *_obs variables, and every model-side completion step was flagged.

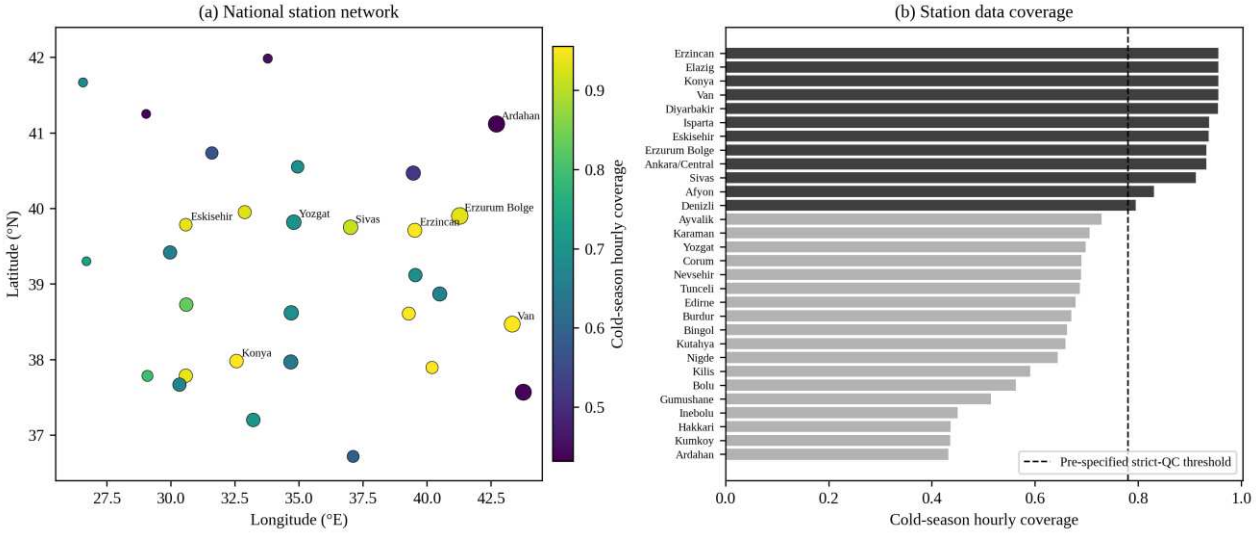


Fig. 3 Spatial distribution and cold-season coverage of the 30 assembled stations; 29 stations contributed observation-supported primary endpoints. Marker size is proportional to elevation, and the dashed line denotes the pre-specified 0.78 strict-QC threshold

Fig. 3 demonstrates broad geographic and elevational coverage but also the pronounced difference between the 12 strict-QC stations and the extended network.

Calendar completion, residual missingness, and target integrity

A complete hourly calendar index was created for each station for 2025, while direct observation availability varied by station and timestamp. Calendar indexing was not treated as equivalent to complete observation. Across contributing stations, the observed-temperature ratio was generally about 0.61–0.64. This level of support can arise from irregular or omitted hourly reports, station-specific reporting protocols, maintenance interruptions, and aggregation of source reports to a regular hourly grid; it should not be interpreted as a uniform sensor-failure rate. The primary endpoint therefore retained only issue times for which at least 75% of the subsequent 24 hourly temperatures were directly observed. Short internal predictor gaps of up to three hours were eligible for time interpolation; no reanalysis predictor was included in the reported model fits. Remaining ancillary gaps were represented by explicit missingness descriptors and handled through XGBoost's native missing-value routing. Raw observations were preserved in separate variables and were never overwritten.

This distinction is central to endpoint validity. The primary 24-h regression target was the minimum original observed temperature in the strictly future interval t+1,...,t+24, and a window was retained only when at least 75% of those future hourly observations were available. Binary frost was defined as minimum temperature ≤0 °C. Severity was coded as no frost (>0 °C), light (-2<T≤0 °C), moderate (-4<T≤-2 °C), and severe (≤-4 °C). A secondary calendar-completed target was generated for sensitivity analysis, but it was never presented as equivalent to the observation-supported primary endpoint.

A simultaneous “all core raw channels complete” indicator was structurally uninformative because at least one ancillary raw channel was absent at every timestamp in every station. It was therefore removed from Online Resource 1, Table S2 and was not used for station selection, model fitting, or inference. Data support is instead reported through observed-temperature availability, mean model-side imputation, observation-supported target availability, and completed-target availability. The analysis did not assume that missingness was completely at random; removal of missingness descriptors tests whether the model exploited reporting patterns, but it does not by itself exclude selection bias in the eligible endpoint windows.

Window representation and feature sets

For each issue time, the preceding 72 contiguous hours formed the predictor history. The main analysis used a three-hour issue stride. Dynamic meteorological and missingness channels were summarized by the final value, mean, standard deviation, minimum, maximum, and endpoint-to-endpoint trend. Additional predictors included the preceding 24-h minimum and mean temperature. Static topographic variables were appended once. This process yielded 92 meteorology-only variables, 308 meteorology-plus-missingness variables, 104 meteorology-plus-topography variables, and 320 variables in the full configuration.

Temporal partitioning and sampling

The primary chronological design trained models on 2015–2023, used 2024 exclusively for early stopping and decision-threshold selection, and reserved 2025 for independent observation-supported evaluation. The primary test comprised 34,011 eligible windows from 29 stations and retained its natural frost prevalence of 0.3605. The corresponding validation set contained 30,619 windows, and the pre-2024 training pool contained 236,771 windows. Binary-model fitting used a reproducible stratified cap of 180,000 training-only windows with approximately one positive example for every two negative examples; validation and test data were neither balanced nor downsampled. Fold-specific LOSO and rolling-origin models used a 60,000-window training cap for computational feasibility.

Table 3. Temporal partitions and validation designs.

Experiment	Training period	Validation	Test	Training sample	Validation n	Test n/folds	Purpose
Primary temporal test	2015-2023	2024	2025	236,771 pool; 180,000 fit	30,619	34,011	Observation-supported endpoint
Completed-target sensitivity	2015-2023	2024	2025	180,000 fit	57,840	57,360	Secondary endpoint; 40.6% mean target imputation
Full LOSO	All non-holdout stations before 2024	Non-holdout 2024	Held-out station in 2025	60,000/fold	Varied	29 completed folds	Unseen-station transfer
Rolling origin	All years before validation year	Test year - 1	2017-2025	60,000/fold	Varied	Nine test years	Temporal stability
Overlap sensitivity	2015-2023	2024	2025	60,000	Full	3-, 24-, 72-h strides	Checks dependence on overlapping windows

Table 3 defines the chronological and resampling structure that underpins the reported estimates. The main model was trained on 2015–2023, tuned only on 2024, and evaluated on 34,011 observation-supported issue times from the 2025 calendar. Rolling-origin and full leave-one-station-out experiments extended this separation across years and stations. Consequently, threshold selection, early stopping, and model fitting were isolated from the final 2025 endpoint.

Models and baselines

The principal learner was XGBoost, selected because it supports nonlinear interactions, regularization, scalable histogram tree construction, and native routing of missing values (Chen and Guestrin 2016). The main models used 450 estimators, maximum depth 6, learning rate 0.04, minimum child weight 2, subsampling 0.85, column sampling 0.85, L1 regularization 0.05, L2 regularization 1.0, histogram tree construction, and 40-round validation-based early stopping. Binary frost, continuous minimum temperature, and four-class severity were modelled separately. Inverse-frequency sample weights were applied to the severity task only.

The final configuration was fixed in the executable analysis before scoring the 2025 endpoint and was not optimized on test data. Early stopping used only 2024 validation loss, and the decision threshold was selected only from 2024 predictions. The same depth, learning rate, regularization, subsampling, and column-sampling values were retained across the primary, strict-QC, missingness, LOSO, and rolling-origin analyses; only the estimator and training-sample caps were reduced for repeated fold-level experiments. Online Resource 1, Table S6 lists the complete configuration and test-isolation rules.

Three operationally interpretable baselines were used. Persistence treated current temperature as the predicted future minimum and transformed distance below 0 °C with $p=1/[1+\exp(T/1.8)]$. The previous-24-h baseline used the minimum temperature during the preceding day. The climatology baseline used station-month-hour means estimated only from the training period. The binary decision threshold for XGBoost was selected by maximizing F1 on 2024 and then fixed at 0.4942 for the primary 2025 test.

Binary XGBoost used no class-weight multiplier. The training cap was applied only to the pre-2024 pool, whereas validation and test sets retained their natural prevalence. Raw model probabilities were evaluated without post-hoc recalibration; the validation-derived threshold affected only binary decisions, not AUROC, AUPRC, Brier score, ECE, calibration intercept, or calibration slope. The reported calibration analysis therefore tests the probability output directly rather than assuming reliability from discrimination alone.

Validation metrics and statistical analysis

Classification was evaluated with accuracy, balanced accuracy, precision, recall, F1, Matthews correlation coefficient, AUROC, AUPRC, Brier score, and ECE. AUPRC was emphasized because it is more informative than

AUROC when event prevalence is unequal (Davis and Goadrich 2006; Saito and Rehmsmeier 2015). Regression performance was evaluated with MAE, RMSE, and R^2. Calibration was characterized by the reliability curve, calibration intercept and slope, Brier skill relative to climatology, and ECE computed with 10, 15, and 20 bins (Brier 1950; Murphy 1973; Gneiting and Katzfuss 2014; Guo et al. 2017; Van Calster et al. 2019).

Uncertainty was estimated by 2,000 station-day grouped bootstrap resamples, which preserved within-day dependence among overlapping issue times (Efron and Tibshirani 1993). The topographic ablation contained three pre-specified paired inferential contrasts (F1, AUPRC, and MAE); their raw p-values were adjusted jointly using the Benjamini-Hochberg false-discovery-rate procedure. Strict-QC, missingness, issue-stride, multi-seed, rolling-origin, and LOSO analyses were treated as validation or sensitivity summaries rather than as a family of null-hypothesis tests, and no multiplicity-dependent significance claims were made from those summaries. Five independent random seeds assessed model stability, and their 95% intervals used the t distribution with four degrees of freedom. Three-, 24-, and 72-h issue strides quantified sensitivity to overlapping predictor windows. Rolling-origin tests used consecutive validation-test year pairs from 2016/2017 through 2024/2025. Statistical significance was interpreted together with effect size and interval width (Wilks 2016).

For LOSO folds, positive-window counts and Wilson 95% intervals for observed frost prevalence were reported to make rare-event support explicit. Wilson intervals were not applied to F1 because F1 is a nonlinear function of true positives, false positives, and false negatives. In folds with fewer than 50 positive windows, station-level F1 was therefore labelled descriptive and interpreted cautiously rather than presented with a pseudo-precise interval.

Event-based verification

An observed frost episode was defined as one or more consecutive hourly observations ≤0 °C, with a new event starting after an interruption of more than one hour. Positive issue times separated by no more than six hours were grouped into an alert episode. Each observed event was evaluated independently and was counted as detected when at least one positive issue time occurred during the preceding 24 h; the earliest qualifying issue time defined lead time. False-alert episodes were counted when no observed event began within the alert interval or the following 24 h. This event-centric matching is intended as screening-oriented verification rather than a global one-to-one assignment. POD, FAR, CSI, missed-event count, and warning lead time were reported following standard forecast-verification principles (Jolliffe and Stephenson 2012; Wilks 2019).

Explainability

TreeSHAP was applied to a reproducible random sample of 3,000 observation-supported 2025 test windows. Global mean absolute SHAP values quantified average contribution magnitude, the beeswarm plot retained feature-value direction, and a dependence plot described the nonlinear response of the leading recent-temperature feature (Lundberg and Lee 2017). Interpretation was limited to the completed global and dependence analyses shown in the manuscript; no causal interpretation was assigned to SHAP values.

Computational environment

Analyses were executed in a Python 3 Google Compute Engine GPU runtime. The available environment provided 51.0 GB system RAM, 15.0 GB GPU memory, and 235.7 GB disk. During the strict-QC and extended-network sensitivity analysis, the supplied runtime monitor showed 37.7 GB of active system memory, consistent with approximately 40 GB RAM use. This screenshot is reported as a workload snapshot rather than a standardized peak-memory benchmark. The captured GPU allocation was 0.2 GB at that moment, although XGBoost reported CUDA execution during model fitting. Online Resource 1, Fig. S1 documents the environment.

Results

Station network and endpoint coverage

The assembled network covered longitudes 26.7-43.8°E, latitudes 36.7-42.0°N, and station elevations from 4 to 1,869 m. Cold-season coverage ranged from 0.432 to 0.955. Twelve stations passed the strict-QC criterion, whereas the extended analysis retained 29 stations with usable 2025 primary targets. The primary test included 34,011 windows. The secondary completed-target test included 57,360 windows; its mean imputed fraction within the future target interval was 0.406, and only 46.5% of targets were fully observed. These quantities justify treating the completed endpoint as sensitivity evidence only.

Observation-supported 2025 performance

Table 4. Independent 2025 performance of XGBoost and operational baselines.

Model	Accuracy	Precision	Recall	F1	AUROC	AUPRC	MAE (°C)	RMSE (°C)	R^2
XGBoost	0.9220	0.8991	0.8827	0.8908	0.9802	0.9681	1.586	2.105	0.9184
Previous 24-h minimum	0.9002	0.8569	0.8682	0.8625	0.9614	0.9310	2.168	2.931	0.8418
Persistence	0.8306	0.9884	0.5364	0.6954	0.9394	0.9170	5.213	6.675	0.1796
Station-month-hour climatology	0.7770	0.7577	0.5609	0.6446	0.8490	0.7610	3.839	5.207	0.5008

Table 4 is the principal performance table. Relative to persistence, the previous-day minimum, and station–month–hour climatology, FrostTR produced the highest F1, AUPRC, and AUROC and the lowest continuous-temperature error. The concordant gains across discrimination, threshold-dependent classification, and regression indicate that the model improved both frost-warning decisions and minimum-temperature estimation rather than optimizing one objective at the expense of another.

XGBoost correctly classified 31,358 of 34,011 test windows. The confusion counts were 10,824 true positives, 1,438 false negatives, 1,215 false positives, and 20,534 true negatives, corresponding to precision 0.8991, recall 0.8827, and balanced accuracy 0.9134. Its 1.586 °C MAE was 0.581 °C lower than the previous-24-h minimum baseline and 3.627 °C lower than persistence. The model's performance therefore reflected more than short-term temperature autocorrelation.

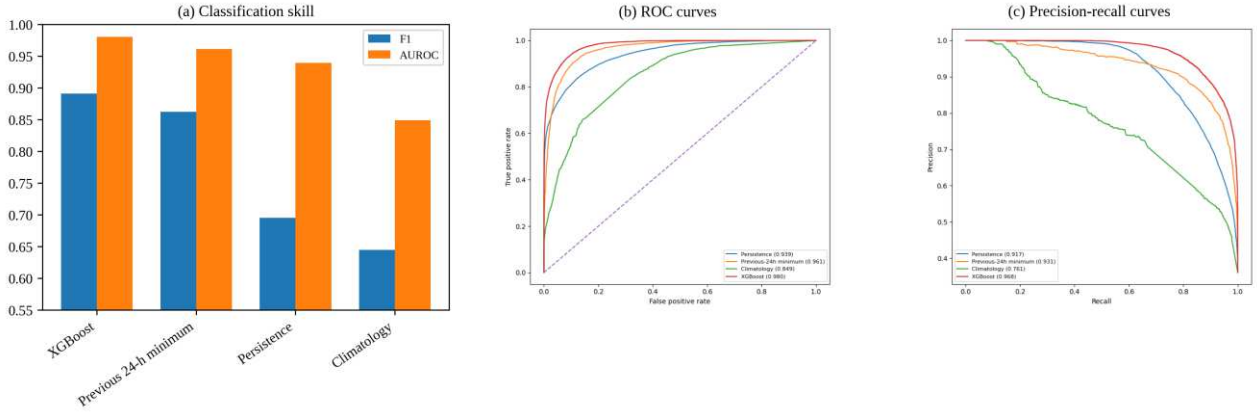


Fig. 4 Main independent 2025 classification results: (a) F1 and AUROC comparison; (b) receiver operating characteristic curves; and (c) precision-recall curves

Fig. 4 shows that XGBoost retained the strongest discrimination across both ROC and precision-recall spaces, while the previous-day minimum remained a competitive but clearly weaker reference.

Four-level frost severity

Table 5. Class-specific performance of the four-level frost-severity model.

Class	Precision	Recall	F1	Support
No frost	0.9762	0.8706	0.9204	21,749
Light	0.3694	0.6883	0.4807	3,391
Moderate	0.4446	0.4022	0.4224	2,996
Severe	0.8460	0.8044	0.8247	5,875
Macro average	0.6591	0.6914	0.6620	34,011
Weighted average	0.8464	0.7997	0.8161	34,011

Table 5 extends the evaluation from binary occurrence to four ordered severity levels, but it also identifies a material weakness of the severity model. Precision was only 0.3694 for light frost and 0.4446 for moderate frost, meaning that 63.1% and 55.5%, respectively, of predictions assigned to those two classes were false positives. Their class-specific F1 values were 0.4807 and 0.4224, substantially below the no-frost and severe-frost values. The quadratic weighted kappa of 0.891 indicates that many errors remained ordinal and close to the correct category, but it does not make the light-versus-moderate distinction operationally reliable. Accordingly, the four-level output should be regarded as exploratory; binary frost probability and continuous minimum-temperature prediction remain the more defensible decision variables.

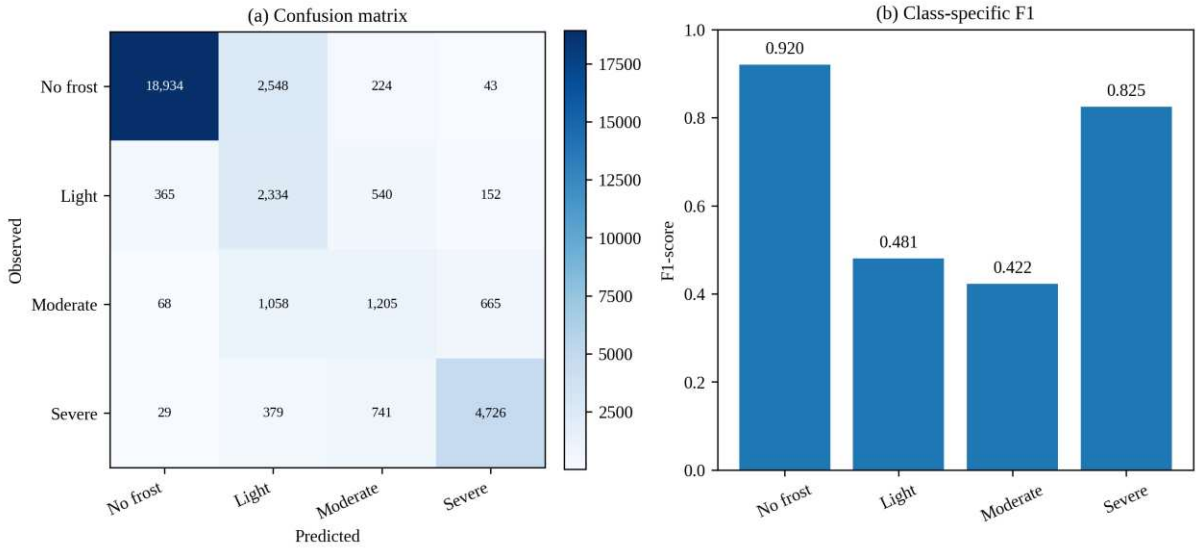


Fig. 5 Four-level frost-severity results: (a) confusion matrix and (b) class-specific F1-scores

Fig. 5 indicates that most errors occurred between adjacent intensity categories rather than between no frost and severe frost.

Quality-control and missingness sensitivity

Table 6. Strict-QC, extended-network, and missingness-indicator sensitivity.

Scenario	Stations	Features	Test n	F1	AUPRC	Brier	ECE	MAE (°C)
Extended full	29	320	34,011	0.8908	0.9681	0.0538	0.0066	1.586
Strict-QC full	12	320	14,270	0.9019	0.9748	0.0559	0.0224	1.562
Extended without missingness	29	104	34,011	0.8922	0.9680	0.0540	0.0099	1.599
Strict-QC without missingness	12	104	14,270	0.9010	0.9741	0.0566	0.0223	1.571

Table 6 tests sensitivity to station-quality restriction and explicit missingness descriptors. The strict-QC network produced slightly higher classification scores, whereas the 29-station extended network retained comparable discrimination and lower calibration error. Removing missingness indicators changed F1 only marginally, which argues against a model that derived its performance primarily from station-specific reporting patterns.

The strict-QC full model achieved F1 0.9019 and AUPRC 0.9748 on 14,270 windows, compared with F1 0.8908 and AUPRC 0.9681 for the 29-station extended network. In the extended network, removing missingness indicators slightly increased F1 from 0.8908 to 0.8922 but increased ECE from 0.0066 to 0.0099 and MAE from 1.586 to 1.599 °C. This pattern argues against a model that primarily memorized station-specific data gaps.

Contribution of topographic context

Table 7. Topographic ablation with raw and Benjamini-Hochberg FDR-adjusted paired-bootstrap probabilities.

Configuration/contrast	F1 or Δ	AUPRC/CI	Raw p; BH-FDR q	ECE	MAE (°C)
Meteorology + missingness	0.8904	0.9675	—	0.0076	1.622
Meteorology + missingness + topography	0.8908	0.9681	—	0.0066	1.586
Paired bootstrap: F1	Δ=0.0005	95% CI -0.0028 to 0.0038	p = .791; q = .791	—	—
Paired bootstrap: AUPRC	Δ=0.0006	95% CI -0.0002 to 0.0015	p = .136; q = .204	—	—
Paired bootstrap: MAE	Δ=0.0361	95% CI 0.0249 to 0.0471	p < .001; q < .001	—	—

Table 7 isolates the contribution of topographic context. After Benjamini-Hochberg adjustment across the three pre-specified paired contrasts, topography retained a small but statistically supported reduction in minimum-temperature MAE ($q < .001$), whereas the paired intervals for F1 and AUPRC included zero and their adjusted probabilities

remained non-significant (q = .791 and q = .204). This pattern supports a focused interpretation: terrain information sharpened continuous temperature estimation, but recent meteorology already captured most of the binary frost signal. Multiplicity correction therefore does not change the substantive conclusion.

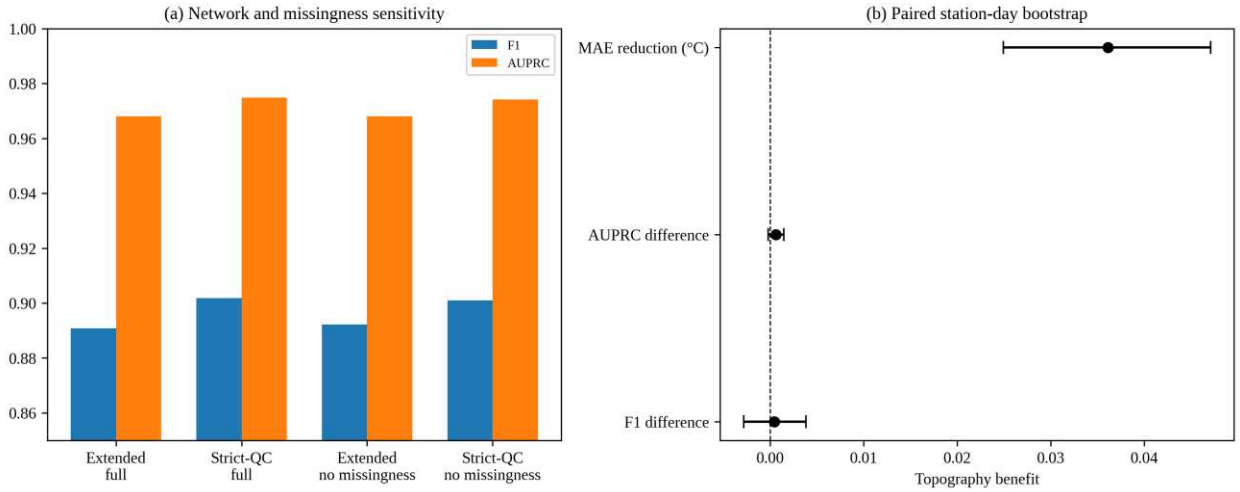


Fig. 6 Sensitivity analyses for network quality, missingness indicators, and topographic context

Fig. 6 confirms that the main classification result was stable across data-quality and missingness specifications, while topography contributed most clearly to continuous temperature estimation.

Overlap and random-seed robustness

Table 8. Window-overlap sensitivity and five-seed stability.

Issue stride	Test n	F1	AUPRC	Brier	ECE	MAE (°C)	RMSE (°C)
3 h	34,011	0.8918	0.9677	0.0540	0.0078	1.608	2.134
24 h	4,266	0.9235	0.9809	0.0420	0.0113	1.204	1.691
72 h	1,393	0.9067	0.9778	0.0452	0.0211	1.254	1.772

The first part of Table 8 evaluates sensitivity to issue-time overlap. F1 and AUPRC remained high when the stride increased from 3 h to 24 h and 72 h, despite the substantial reduction in evaluated windows. The 3-h sensitivity model used the fold-level 60,000-window training cap, whereas the principal model used 180,000 training windows; the small difference between the corresponding 3-h estimates is therefore expected and does not represent a discrepancy in the test set.

Table 8 continued. Five-seed mean, standard deviation, and 95% t interval (df = 4).

Metric	Mean	SD	95% CI lower	95% CI upper
F1	0.892132	0.001063	0.890812	0.893452
AUPRC	0.968325	0.000161	0.968125	0.968525
AUROC	0.980306	0.000092	0.980192	0.980420
BRIER	0.053615	0.000191	0.053378	0.053852
ECE	0.006185	0.000419	0.005665	0.006705
MAE	1.586158	0.002057	1.583604	1.588712
RMSE	2.103338	0.002829	2.099825	2.106851

The second part of Table 8 reports five independent model fits. F1, AUPRC, AUROC, Brier score, ECE, MAE, and RMSE varied only minimally across seeds. The intervals were calculated with the t distribution because only five runs were available. Together with the overlap analysis, these results indicate that the principal estimates were not dependent on a favourable random initialization or densely overlapping issue times.

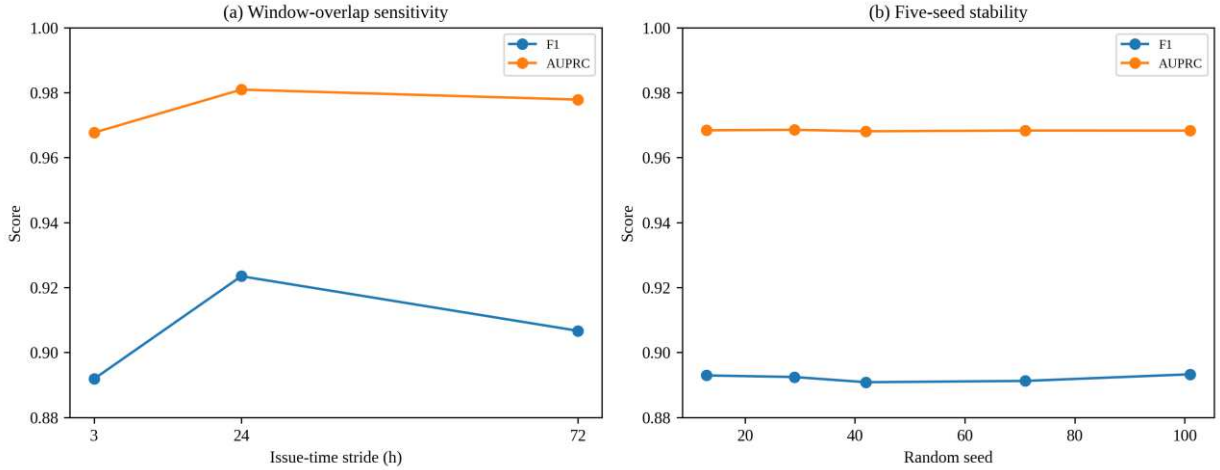


Fig. 7 Robustness to issue-time overlap and random initialization

Fig. 7 indicates that the three-hour main stride did not inflate the result; 24- and 72-h strides produced equal or higher F1 and AUPRC on smaller independent samples.

Spatial transfer to unseen stations

Table 9. Summary of full leave-one-station-out validation.

Measure	Result
Completed stations	29
Macro mean F1	0.8474
Median F1	0.8542
Interquartile range	0.1096
Minimum-Maximum F1	0.5248-0.9732
Macro mean AUPRC	0.9471
Macro mean MAE	1.719 °C

Table 9 summarizes full leave-one-station-out validation. Macro-F1 was 0.8474 across 29 holdouts, demonstrating useful transfer to unseen stations, but this average should not be read as uniform precision across the network. Online Resource 1, Table S3 now reports positive-window counts and Wilson intervals for prevalence. Ayvalık and Kilis contained only 26 and 40 positive windows, respectively, so their F1 values are explicitly treated as descriptive and potentially unstable. The broad station-level range and the low-support flags temper any claim of homogeneous national performance.

LOSO F1 ranged from 0.5248 at Inebolu to 0.9732 at Sivas. The lowest values occurred predominantly at low-prevalence coastal or transitional stations, where a small number of positive windows makes threshold-dependent F1 especially sensitive to a few classification changes. Ayvalık had 26 positive windows and Kilis had 40; Inebolu had 55. These support counts and Wilson intervals for prevalence are shown in Online Resource 1, Table S3. Because prediction-level block-bootstrap intervals were not reconstructed from aggregate fold summaries, the low-support F1 values are presented as descriptive point estimates rather than as precise station-level parameters.

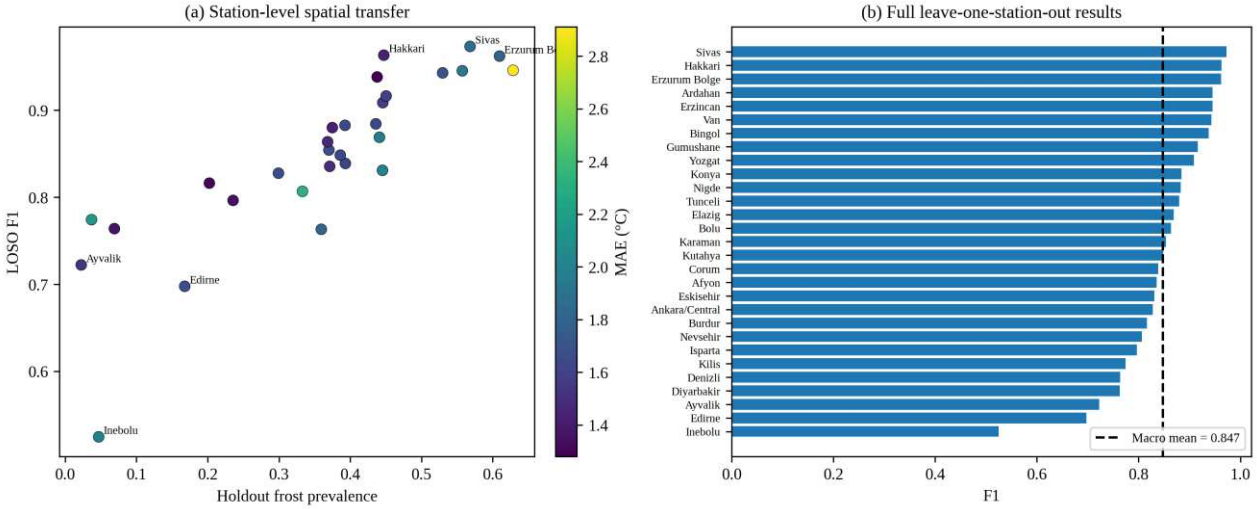


Fig. 8 Full leave-one-station-out validation across 29 stations: (a) F1 versus holdout frost prevalence and (b) station-specific F1 distribution

Fig. 8 shows that station-level F1 was partly associated with event prevalence and that the national mean conceals meaningful coastal and microclimatic variation.

Rolling-origin temporal validation

Table 10. Rolling-origin temporal validation from 2017 to 2025.

Validation year	Test year	Test n	Prevalence	F1	AUPRC	Brier	MAE (°C)
2016	2017	23,712	0.4178	0.8642	0.9525	0.0781	1.861
2017	2018	23,762	0.2912	0.8069	0.9061	0.0780	1.758
2018	2019	26,313	0.3491	0.8541	0.9431	0.0699	1.655
2019	2020	25,783	0.3796	0.8711	0.9543	0.0684	1.713
2020	2021	24,656	0.3636	0.8593	0.9450	0.0723	1.792
2021	2022	24,184	0.3939	0.8778	0.9610	0.0661	1.681
2022	2023	25,881	0.3053	0.8781	0.9564	0.0534	1.549
2023	2024	30,619	0.3462	0.8535	0.9394	0.0729	1.614
2024	2025	34,011	0.3605	0.8918	0.9677	0.0540	1.608

Table 10 evaluates temporal stability across nine rolling-origin test years. Performance varied with annual frost prevalence but showed no progressive deterioration, and 2025 remained within the established range rather than standing as an isolated favourable result. The analysis therefore supports temporal robustness under changing seasonal conditions.

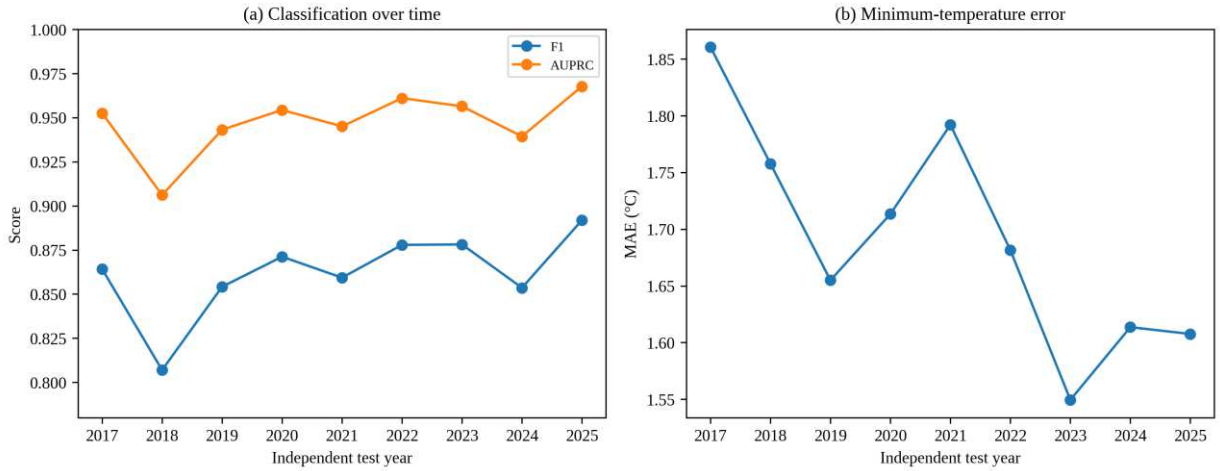


Fig. 9 Rolling-origin temporal performance: (a) F1 and AUPRC and (b) minimum-temperature MAE

Fig. 9 demonstrates year-to-year variation but no progressive degradation, supporting temporal stability across changing frost prevalences.

Event-based operational verification

Table 11. Event-based frost verification for the primary and secondary endpoints.

Endpoint	Events	Hits	Misses	False alerts	POD	FAR	CSI	Mean lead (h)	Median lead (h)
Observation-supported primary endpoint	1,835	1,754	81	125	0.9559	0.0665	0.8949	20.72	22.0
Completed-target sensitivity*	2,690	2,638	52	153	0.9807	0.0548	0.9279	21.50	22.0

Table 11 translates window-level predictions into event-level warning performance. The primary endpoint detected 1,754 of 1,835 observed episodes, with POD 0.9559, FAR 0.0665, CSI 0.8949, and a mean lead time of 20.7 h. Wilson 95% intervals were 0.9455–0.9643 for POD, 0.0561–0.0787 for FAR, and 0.8805–0.9077 for CSI. The completed-target endpoint produced numerically higher event scores, but 40.6% of its future target interval was completed on average; the reduction in target missingness and smoothing of the endpoint can make this secondary analysis appear easier than the observation-supported task.

Important interpretive restriction: completed-target values are sensitivity results only and must not be used to claim superior forecast performance or to replace the observation-supported primary endpoint.

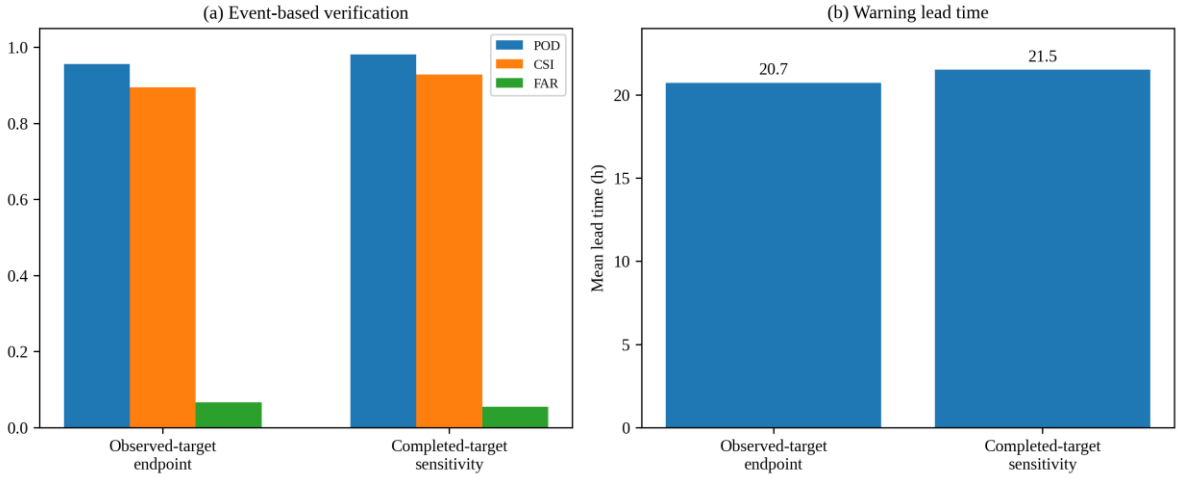


Fig. 10 Event-based POD, CSI, FAR, and warning lead time for observation-supported and completed-target endpoints

Fig. 10 demonstrates that strong window-level discrimination translated into distinct-event detection with operationally useful advance warning.

Probability calibration

Table 12. Calibration measures and station-day bootstrap intervals.

Measure	Estimate	Interpretation/interval
Calibration intercept	0.0824	Ideal = 0
Calibration slope	1.0149	Ideal = 1
Brier score	0.0538	Lower is better
Climatology Brier score	0.1501	Reference
Brier skill score	0.6415	Relative improvement
ECE, 10 bins	0.0065	Sensitivity check
ECE, 15 bins	0.0066	Primary ECE
ECE, 20 bins	0.0069	Sensitivity check
Bootstrap F1	0.8909	95% CI 0.8824-0.8994
Bootstrap AUPRC	0.9682	95% CI 0.9638-0.9721
Bootstrap Brier	0.0538	95% CI 0.0502-0.0574

Measure	Estimate	Interpretation/interval
Bootstrap ECE15	0.0085	95% CI 0.0056-0.0119

Table 12 evaluates probability reliability in addition to ranking and threshold performance. The calibration slope of 1.0149, intercept of 0.0824, low ECE across alternative bin counts, and Brier skill score of 0.6415 indicate close agreement between predicted probabilities and observed frequencies. These measures are important for cost-sensitive decisions because a calibrated probability can support different action thresholds without retraining the model (Gneiting and Katzfuss 2014; Gneiting and Raftery 2007).

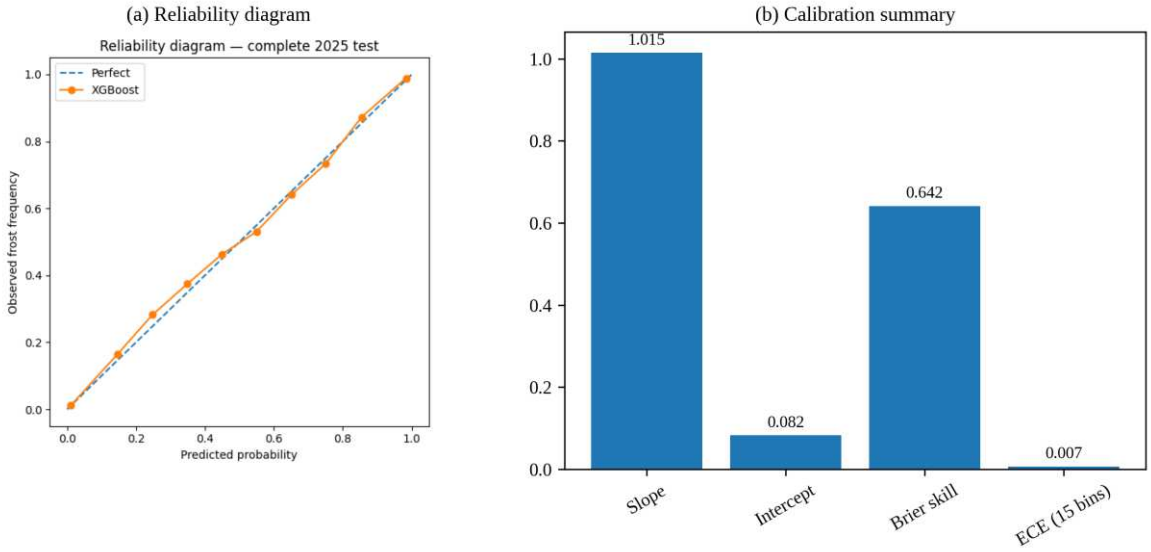


Fig. 11 Reliability and summary calibration statistics for the observation-supported 2025 test

Fig. 11 shows close agreement between predicted frost probabilities and observed event frequencies over most of the probability range.

Model interpretation

The leading global predictor was the preceding 24-h mean temperature, followed by current temperature, current dew point, the 72-h minimum temperature, seasonal phase, pressure, the preceding 24-h minimum, sky condition, time-of-day variables, station elevation, TPI, and latitude. The SHAP beeswarm showed that low recent temperatures and low dew-point values increased the predicted frost probability, while warmer histories reduced it. The dependence curve displayed a nonlinear transition across the recent-temperature range rather than a single hard threshold.

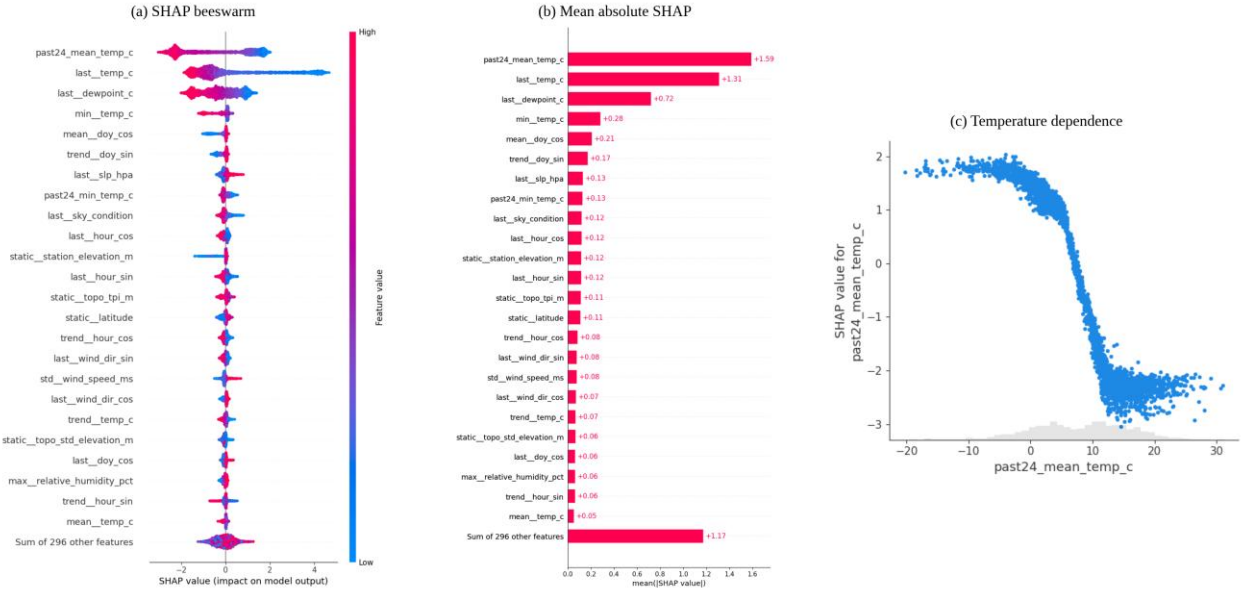


Fig. 12 Global SHAP interpretation: (a) beeswarm distribution, (b) mean absolute SHAP importance, and (c) dependence of model contribution on the leading recent-temperature feature

Fig. 12 shows that the model's dominant signals were physically coherent thermal and moisture variables, with topography and seasonal timing providing secondary context.

Discussion

Principal findings

FrostTR showed consistent discrimination, calibration, minimum-temperature accuracy, and event detection under a chronology that prevented future years from influencing model selection. The observation-supported 2025 endpoint yielded F1 0.8908, AUPRC 0.9681, and MAE 1.586 °C. Five-seed F1 varied by only 0.0011 SD, and 24-h and 72-h issue strides retained F1 values above 0.90. Event analysis detected 1,754 observed frost episodes with a mean warning lead time of 20.7 h. The agreement across these analyses supports the framework as a reliable meteorological exposure-warning layer for plant cold-stress management, while not establishing crop-specific biological injury prediction.

The results compare favourably with prior frost-learning studies while addressing a different validation problem. Talsma et al. (2023) reported useful 6-48 h point-scale minimum-temperature prediction from farm observations, and Noh et al. (2021) demonstrated effective hourly orchard classification. Rozante et al. (2023) improved numerical frost forecasts through neural-network post-processing, whereas Pérez Tárraga et al. (2025) reported high daily AUC across a large Spanish archive. FrostTR adds a national multi-station hourly setting, a fixed future-year threshold, complete LOSO, direct probability verification, and distinct-event metrics. Because datasets, prevalence, horizons, and validation designs differ, the numerical scores should not be interpreted as a ranking of studies.

Spatial generalization and terrain

The 29-station LOSO experiment is the strongest evidence for spatial transfer in this study. Macro-F1 0.8474 indicates that a model trained without any historical examples from the target station retained substantial skill. However, the range from 0.5248 to 0.9732 is equally important. Poorer performance at Inebolu, Ayvalık, and Edirne suggests that rare events, maritime moderation, local exposure, and threshold instability can limit a national model. Operational deployment should therefore use station-specific monitoring and, where possible, local recalibration rather than assuming homogeneous national performance.

Topography improved minimum-temperature MAE by 0.036 °C with a paired 95% confidence interval that excluded zero, but the classification changes were not statistically supported. This nuanced result is physically plausible. Terrain descriptors help resolve the magnitude of nocturnal cooling and cold-air pooling (Chung et al. 2006; Kimura et al. 2023; Lundquist et al. 2008), yet the binary threshold may already be well separated by recent temperature and dew point at many stations. Accordingly, the manuscript treats topography as a modest regression improvement rather than a major classification innovation.

Missing data, severity, and endpoint interpretation

The strict-QC analysis produced higher F1 and AUPRC than the extended network, as expected from more complete observation histories. Nevertheless, two findings support cautious use of the broader network for the primary national analysis. First, F1 remained 0.8908 across 29 stations. Second, removing 216 missingness-related variables did not reduce F1. This shows that explicit reporting-pattern indicators were not necessary for high classification skill. It does not, however, establish that target availability was missing at random. If omitted observations were more common at particular hours, seasons, stations, or extreme cold conditions, the 75% eligibility rule could still induce selection bias. The primary estimates should therefore be interpreted as conditional on adequate future observation support.

The 2025 evaluation requires a clear distinction between the analytical calendar and direct endpoint support. The primary endpoint included only issue times satisfying the pre-specified requirement that at least 75% of the following 24 hourly temperatures were observed, yielding 34,011 test windows. The completed-target analysis contained more windows and higher apparent skill, but 40.6% of its future target interval was completed on average. Its higher POD and CSI should not be interpreted as evidence that the model performed better in real observations; a completed target is smoother, more complete, and partly model-derived. The completed-target analysis is therefore retained only as a secondary stress test of endpoint definition and is not used for comparative performance claims.

The severity model requires a similarly cautious interpretation. Although quadratic weighted kappa was high, precision for light and moderate frost was only 0.3694 and 0.4446. These are the categories closest to operational protection thresholds, and their false-positive burden could lead to unnecessary or mistimed interventions if the four-class label were used directly. Until externally validated or reformulated as an ordinal/probabilistic severity model, the four-level output should be used for exploratory stratification only; operational decisions should rely primarily on frost probability and predicted minimum temperature.

Calibration and operational relevance

Discrimination alone is insufficient for risk-based agricultural decisions. The calibration slope of 1.015, intercept of 0.082, ECE near 0.0066, and Brier skill score of 0.642 indicate that FrostTR probabilities were close to observed frequencies and substantially more informative than climatology. This matters because protection decisions depend on the relative cost of false alarms and missed frost, not on a universal 0.5 threshold (Gneiting and Raftery 2007; Van Calster et al. 2019). The stored probabilities permit farms or warning services to select a threshold that reflects crop vulnerability, water availability, and protection cost.

The primary event-level POD of 0.956 and FAR of 0.067 are especially relevant. Window-based false positives can overstate the burden when several warnings belong to one episode; grouping issue times into alert episodes produced a direct operational measure. The median 22-h lead time is compatible with planning irrigation or other protective action, although crop-specific protection feasibility was not tested in this study.

Implications for Earth science informatics and agricultural early warning

FrostTR illustrates how Earth science informatics can convert heterogeneous station archives and terrain data into operationally interpretable information. Calibrated probabilities permit users to select action thresholds according to the relative costs of false alarms and missed events, while the accompanying minimum-temperature estimate conveys expected event magnitude. These outputs can be combined downstream with locally defined crop thresholds, phenological information, and protection constraints, but those downstream biological and economic decisions are outside the present model and require independent validation.

Explainability

The SHAP results were meteorologically coherent: recent mean temperature, current temperature, dew point, and recent minima dominated the model. This alignment with frost physics reduces concern that performance arose from nonphysical station identifiers or missingness patterns. The secondary importance of elevation, TPI, latitude, pressure, cloud condition, and seasonal phase is consistent with the modulation of nocturnal radiative cooling and cold-air accumulation. SHAP values explain the fitted model rather than causal mechanisms, so they should be read as model-attribution evidence, not as experimental proof of physical causality.

Computational implications

The full comprehensive workflow was memory intensive. The 320-feature tabular representation contained 1,900,440 rows before temporal filtering, and simultaneous strict-QC and extended-network analyses brought active system memory to 37.7 GB. The result is relevant for reproducibility: the final workflow is feasible in a high-memory

cloud notebook, but it may exceed standard 12-16 GB notebook limits. Fold-level caching, Parquet partitioning, and staged LOSO execution are therefore part of the practical design rather than incidental implementation details.

Limitations

Six limitations remain. First, direct station-observation availability was heterogeneous across the 2025 analytical calendar, with observed-temperature ratios generally near 0.61–0.64 among contributing stations. The causes may include irregular source reporting, station-specific protocols, and maintenance gaps, but the present analysis cannot establish that missingness was random; selection bias under missing-not-at-random conditions therefore remains possible. Second, the four-level severity model was not sufficiently precise for light and moderate frost, the decision-relevant boundary region, and should be considered exploratory. Third, the completed-target endpoint contained 40.6% imputed future values on average and produced higher apparent scores; these values must not be interpreted as superior real-world performance. Fourth, LOSO F1 is unstable at rare-event stations: Ayvalık and Kilis contained only 26 and 40 positive windows, and prediction-level block-bootstrap F1 intervals were not available from the retained aggregate fold outputs. Fifth, the study predicts meteorological exposure rather than plant injury. Biological damage depends on tissue temperature, phenological stage, cultivar, acclimation state, humidity, exposure duration, and post-freeze conditions; no physiological, phenological, injury, or yield observations were included. Sixth, LOSO evaluates transfer among existing stations, not farms with different sensor heights, land cover, canopy structure, or cold-air drainage. Independent validation against crop-specific biological outcomes and national meteorological-service observations remains the appropriate next step.

Conclusions

FrostTR provides a calibrated and spatially evaluated Earth science informatics framework for 24-hour agricultural frost forecasting across the evaluated station network in Türkiye. In the observation-supported 2025 test, XGBoost combined F1 0.8908 and AUPRC 0.9681 with ECE 0.0066 and 1.586 °C minimum-temperature MAE. Performance remained stable under stricter station-quality restriction, removal of missingness indicators, wider issue-time strides, five random seeds, nine rolling-origin years, and 29 complete leave-one-station-out folds. Event-level verification further showed high detection with a mean warning lead time of 20.7 h. The main contribution is an integrated validation architecture that treats chronology, spatial transfer, calibration, uncertainty, data quality, and event utility as parts of the same forecasting problem. External testing with denser local networks and field-specific observations remains necessary before operational deployment at farm scale.

List of abbreviations

Abbreviation	Definition
AUPRC	area under the precision-recall curve
AUROC	area under the receiver operating characteristic curve
CSI	critical success index
DEM	digital elevation model
ECE	expected calibration error
FAR	false alarm ratio
ISD	Integrated Surface Database
LOSO	leave-one-station-out
MAE	mean absolute error
POD	probability of detection
QC	quality control
SHAP	SHapley Additive exPlanations
TPI	topographic position index

Statements and Declarations

Data availability

NOAA Integrated Surface Database observations are publicly available through NOAA National Centers for Environmental Information services, and Copernicus GLO-30 terrain data are publicly available through the Copernicus Data Space Ecosystem. The station-level audit tables and fixed supplementary results are provided as Online Resource 1. The executable analysis notebook and usage guide are provided as Online Resource 2. No proprietary data were used.

Code availability

The complete executable notebook and a concise usage guide are supplied as Online Resource 2. The notebook records the chronological partitions, fixed model settings, random seeds, feature construction, validation procedures, and output generation used in this study.

Competing interests

The author declares that he has no competing interests.

Funding

No external funding was received for this study.

Author contributions

Mehmet Tevfik AĞDAŞ conceived the study, developed the methodology and software, curated and analysed the data, performed validation and visualization, and wrote and approved the manuscript.

Acknowledgements

The author acknowledges NOAA NCEI for the Integrated Surface Database and the European Union/ESA Copernicus programme for the GLO-30 digital elevation model.

References

Anandhi, A., Perumal, S., Gowda, P. H., Knapp, M., Hutchinson, S., Harrington, J. J., Murray, L., Kirkham, M. B., & Rice, C. W. (2013). Long-term spatial and temporal trends in frost indices in Kansas, USA. Climatic Change, 120, 169-181. https://doi.org/10.1007/s10584-013-0794-4

Bhakare, S., Matiu, M., Crespi, A., & Zardi, D. (2025). Spatial downscaling of daily temperature minima using machine learning methods and application to frost forecasting in two Alpine valleys. Atmosphere, 16(1), 38. https://doi.org/10.3390/atmos16010038

Bhatti, S., Jeranyama, P., Kennedy, C. D., Buda, A., Millar, D. J., Wiegman, A. R. H., & Zalapa, J. (2026). Forecasting spring frost events in agriculture using machine learning: A case study from southeastern Massachusetts, United States. Smart Agricultural Technology, 13, 101720. https://doi.org/10.1016/j.atech.2025.101720

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939785

Chung, U., Seo, H. H., Hwang, K. H., Hwang, B. S., Choi, J., Lee, J. T., & Yun, J. I. (2006). Minimum temperature mapping over complex terrain by estimating cold air accumulation potential. Agricultural and Forest Meteorology, 137(1-2), 15-24. https://doi.org/10.1016/j.agrformet.2005.12.011

European Space Agency. (2022). Copernicus DEM [Data set]. https://doi.org/10.5270/ESA-c5d3d65

Daly, C., Conklin, D. R., & Unsworth, M. H. (2010). Local atmospheric decoupling in complex topography alters climate change impacts. International Journal of Climatology, 30(12), 1857-1864. https://doi.org/10.1002/joc.2007

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In Proceedings of the 23rd International Conference on Machine Learning (pp. 233-240). Association for Computing Machinery. https://doi.org/10.1145/1143844.1143874

Diedrichs, A. L., Bromberg, F., Dujovne, D., Brun-Laguna, K., & Watteyne, T. (2018). Prediction of frost events using machine learning and IoT sensing devices. IEEE Internet of Things Journal, 5(6), 4589-4597. https://doi.org/10.1109/JIOT.2018.2867333

Efron, B., & Tibshirani, R. J. (1993). An introduction to the bootstrap. Chapman & Hall/CRC.

Erdoğan, Ü., and Erdoğan, Y. (2026). Operational assessment of spring late and autumn early frost risk across Türkiye (1901-2022): Implications for fruit production and apiculture. Journal of Anatolian Environmental and Animal Sciences 2026. https://doi.org/10.35229/jaes.1897127

Ghielmi, L., & Eccel, E. (2006). Descriptive models and artificial neural networks for spring frost prediction in an agricultural mountain area. Computers and Electronics in Agriculture, 54(2), 101-114. https://doi.org/10.1016/j.compag.2006.09.001

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. Annual Review of Statistics and Its Application, 1, 125-151. https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359-378. https://doi.org/10.1198/016214506000001437

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (pp. 1321-1330). PMLR.

Jolliffe, I. T., & Stephenson, D. B. (Eds.). (2012). Forecast verification: A practitioner's guide in atmospheric science (2nd ed.). Wiley.

Kimura, K., Maruyama, A., Sasaki, K., Kudo, K., Tanaka, E., Fushimi, E., & Nakagawa, H. (2023). Fine-scale mapping of daily minimum temperature in a cropland with complex terrains through the combination of a cold flow accumulation model with inversion strength. Agricultural and Forest Meteorology, 329, 109247. https://doi.org/10.1016/j.agrformet.2022.109247

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. Sensors, 18(8), 2674. https://doi.org/10.3390/s18082674

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765-4774).

Lundquist, J. D., Pepin, N., & Rochford, C. (2008). Automated algorithm for mapping regions of cold-air pooling in complex terrain. Journal of Geophysical Research: Atmospheres, 113, D22107. https://doi.org/10.1029/2008JD009879

Murphy, A. H. (1973). A new vector partition of the probability score. Journal of Applied Meteorology, 12(4), 595-600. https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2

Noh, I., Doh, H.-W., Kim, S.-O., Kim, S.-H., Shin, S., & Lee, S.-J. (2021). Machine learning-based hourly frost-prediction system optimized for orchards using automatic weather station and digital camera image data. Atmosphere, 12(7), 846. https://doi.org/10.3390/atmos12070846

Omazić, B., Anić, M., Telišman Prtenjak, M., Kvakić, M., & Blašković, L. (2024). Analysis of different existing measurement-based methods and a new approach for frost probability detection. Agricultural and Forest Meteorology, 347, 109898. https://doi.org/10.1016/j.agrformet.2024.109898

Pérez Tárraga, J., Castillo-Cara, M., Arias-Antúnez, E., & Dujovne, D. (2025). Frost forecasting through machine learning algorithms. Earth Science Informatics, 18, Article 183. https://doi.org/10.1007/s12145-025-01710-6

Pfleiderer, P., Schleussner, C.-F., Kornhuber, K., & Coumou, D. (2019). Increasing risks of apple tree frost damage under climate change. Climatic Change, 157, 515-525. https://doi.org/10.1007/s10584-019-02570-y

Rozante, J. R., Ramirez, E., Ramirez, D., & Rozante, G. (2023). Improved frost forecast using machine learning methods. Artificial Intelligence in Geosciences, 4, 164-181. https://doi.org/10.1016/j.aiig.2023.10.001

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Shammi, S., Diepeveen, D., Zander, S., Jones, M. G. K., & Sohel, F. (2024). Early frost detection in wheat using machine learning from vertical temperature distributions. Computers and Electronics in Agriculture, 221, 108950. https://doi.org/10.1016/j.compag.2024.108950

Smith, A., Lott, N., & Vose, R. (2011). The Integrated Surface Database: Recent developments and partnerships. Bulletin of the American Meteorological Society, 92(6), 704-708. https://doi.org/10.1175/2011BAMS3015.1

Snyder, R. L., & de Melo-Abreu, J. P. (2005). Frost protection: Fundamentals, practice, and economics. Food and Agriculture Organization of the United Nations.

Talsma, C. J., Solander, K. C., Mudunuru, M. K., Crawford, B., & Powell, M. R. (2023). Frost prediction using machine learning and deep neural network models. Frontiers in Artificial Intelligence, 5, 963781. https://doi.org/10.3389/frai.2022.963781

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., Steyerberg, E. W., & Topic Group Evaluating Diagnostic Tests and Prediction Models of the STRATOS Initiative. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, 230. https://doi.org/10.1186/s12916-019-1466-7

Wilks, D. S. (2016). The stippling shows statistically significant grid points: How research results are routinely overstated and overinterpreted, and what to do about it. Bulletin of the American Meteorological Society, 97(12), 2263-2273. https://doi.org/10.1175/BAMS-D-15-00267.1

Wilks, D. S. (2019). Statistical methods in the atmospheric sciences (4th ed.). Elsevier.

Yeşilırmak, E. (2022). Analysis of frost probabilities in Aydın, Türkiye. Turkish Journal of Agriculture - Food Science and Technology, 10(8), 1434-1440. https://doi.org/10.24925/turjaf.v10i8.1434-1440.5177

Zohner, C. M., Mo, L., Renner, S. S., Svenning, J.-C., Vitasse, Y., Benito, B. M., Ordonez, A., Baumgarten, F., Bastin, J.-F., Sebald, V., Reich, P. B., Liang, J., Nabuurs, G.-J., de-Miguel, S., Alberti, G., Antón-Fernández, C., Balazy, R., Brändli,

U.-B., Chen, H. Y. H., ... Crowther, T. W. (2020). Late-spring frost risk between 1959 and 2017 decreased in North America but increased in Europe and Asia. Proceedings of the National Academy of Sciences of the United States of America, 117(22), 12192-12200. https://doi.org/10.1073/pnas.1920816117

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- OnlineResource1FrostTRSupplementaryMaterial.pdf