PROVINCE-LEVEL WHEAT YIELD PREDICTION IN TÜRKİYE: A COMPARISON OF STATIC, ADAPTIVE, ONLINE, AND HYBRID MACHINE LEARNING MODELS

Nuri Berk URAL[1]
[1]Department of Statistics, Çukurova University, Adana, Türkiye
nberkural@gmail.com

Burak ÖZTORNACI[2]
[2]Department of Agricultural Economics, Çukurova University, Adana, Türkiye
burakoztornaci@gmail.com

ABSTRACT

This study compares static, adaptive, and online machine learning models for province-level wheat yield prediction in Türkiye. Crop production statistics covering 79 provinces with sufficient and consistently matched data between 2004 and 2025 were combined with temperature, precipitation, solar radiation, soil moisture, and potential evaporation variables derived from ERA5-Land. The models were trained using data from 2004–2013 and evaluated for 2014–2025 through chronological walk-forward validation. Static Ridge, support vector regression, Random Forest, Extra Trees, and Histogram Gradient Boosting models were compared with sliding-window, expanding-window, recursive least squares, and a Random Forest–recursive least squares hybrid. The eight-year sliding-window Random Forest achieved the best overall performance, with a mean absolute error of 35.07 kg/da, root mean squared error of 47.63 kg/da, coefficient of determination of 0.679, and mean absolute percentage error of 15.83%. Province-clustered bootstrap analysis indicated an average root mean squared error reduction of 2.91 kg/da compared with static Random Forest. However, the difference between sliding-window and expanding-window Random Forest was not statistically conclusive. Neither online recursive least squares nor the hybrid model provided additional predictive improvement. The findings demonstrate that adaptive tree-based models are more effective than static and linear online models for province-level wheat yield prediction.

Keywords: Wheat yield, ERA5-Land, yield prediction, adaptive learning, online learning.

INTRODUCTION

Reliable crop-yield information supports food-supply assessment, agricultural planning, market decisions, and climate-risk management. Wheat is especially important because production outcomes depend on interacting weather, soil, genotype, technology, and management factors. These relationships are nonlinear, spatially heterogeneous, and subject to change over time. As a result, models that perform well under random data partitioning may not retain the same accuracy when they are asked to predict a genuinely later harvest year.

Machine-learning methods have become common in crop-yield research because they can combine heterogeneous predictors and represent nonlinear interactions without requiring a fully specified crop-growth mechanism. A systematic review by van Klompenburg et al. (2020) found that temperature, rainfall, and soil properties were among the most frequently used predictors, while neural networks, tree ensembles, and deep-learning methods were widely

applied. At regional scale, Paudel et al. (2021, 2022) showed that machine learning can support large-scale and multi-level crop-yield forecasting, but also emphasized the importance of reusable workflows, operational baselines, forecast timing, and performance in extreme years. For winter wheat in Germany, Srivastava et al. (2022) demonstrated that nonlinear models could exploit weekly weather, soil, and phenological inputs under temporally ordered evaluation.

Recent research in Türkiye has also moved toward data-driven wheat-yield estimation. Tuğaç et al. (2022) combined MODIS and Landsat vegetation indices with machine-learning models at field scale and showed that crop-development timing affects predictive usefulness. Güngüneş et al. (2025) integrated provincial yield statistics with phenology-grouped climate and soil variables for Central Anatolia. At national scale, Kaleli (2025) reported that adding structural economic inputs such as fertilizer, capital, and labour improved crop-yield forecasts beyond climate-only models. These studies demonstrate the value of environmental and structural information, but their spatial domains, predictors, and validation designs differ substantially.

A central methodological issue is that prediction accuracy depends on how training and test observations are separated. Random splits can mix provinces and years from the same underlying period, allowing persistent spatial or temporal structure to appear as genuine future-year prediction skill. Petrovic et al. (2026), for example, found a large decline in wheat-prediction performance when a Czech crop dataset was evaluated temporally instead of randomly. This problem is especially relevant for administrative panels, where province effects, technology trends, and adjacent-year yield persistence are strong. Strong naive baselines are therefore essential: if a complex model cannot improve on a transparent recent-history rule, its practical value is limited even when its absolute error appears acceptable.

Temporal change creates a second challenge. A model fitted once assumes that the predictor–yield relationship remains stable, although cultivar adoption, irrigation, input use, policy, climate, and data-generation processes may evolve. In data-stream terminology, this is related to concept drift (Gama et al., 2014). Adaptation can be implemented through periodic refitting with an expanding history, refitting with a sliding window that discards older observations, or online coefficient updates. However, online updating is not automatically advantageous: a linear recursive update may be poorly matched to nonlinear, spatially varying yield relationships.

This study addresses these gaps using a nationwide province-level wheat panel for Türkiye. The analysis compares naive, static, adaptive, online, and hybrid models under one common walk-forward protocol. It asks four questions:

1. Do machine-learning models outperform strong recent-yield baselines when predicting later harvest years?
2. Does annual model refitting improve performance relative to a fixed static model?
3. Do online RLS updates or an RF–RLS residual-correction hybrid add value?
4. Do phenology-aligned ERA5-Land predictors improve accuracy beyond recent yield history?

The study's contribution is not merely the identification of a numerically best model. It provides a leakage-controlled comparison that distinguishes three different claims: superiority over a weak one-year persistence rule, superiority over a strong three-year-mean baseline, and incremental predictive value from climate data. These claims are evaluated separately using province-clustered uncertainty estimates.

MATERIALS AND METHODS

Study design and data sources

The unit of analysis was province × harvest year. Crop statistics were obtained from the Turkish Statistical Institute (TurkStat) MEDAS Crop Production Statistics database (Turkish Statistical Institute, 2026). The selected series covered *durum* wheat and wheat excluding *durum*, with sown area, harvested area, production, and reported yield. Monthly land-surface climate data were obtained from the Copernicus Climate Change Service ERA5-Land monthly-averaged reanalysis product (Copernicus Climate Change Service, 2026). ERA5-Land provides spatially complete and physically consistent land-surface variables at approximately 0.1° resolution (Muñoz-Sabater et al., 2021).

Province boundaries were represented using the 2021 first-order administrative layer from geoBoundaries, an open global administrative-boundary database (Runfola et al., 2020). Climate data covered January 2003–December 2025 so that autumn conditions preceding the 2004 harvest could be constructed. The crop panel covered harvest years 2004–2025.

The complete analysis dataset contained 1,738 observations: 79 provinces × 22 harvest years. Rize and Trabzon were excluded because the selected TurkStat wheat series did not contain a usable target panel for these provinces. This exclusion arose from the crop statistics, not from ERA5-Land; the climate extraction contained all 81 provinces.

Crop outcome and historical-yield predictors

Durum and non-*durum* records were aggregated before calculating total wheat yield. For province *i* and harvest year *t*, the target was:

Yield (kg da^{-1}) = [Production (tonnes) × 1,000] / Harvested area(decares).

The calculated target was used because it preserves consistency after combining wheat types. Its maximum absolute difference from TurkStat's integer-reported yield was 0.5 kg da^{-1}, consistent with rounding. No record had harvested area greater than sown area, and no province-year key was duplicated.

Two leakage-safe historical predictors were calculated within province. The first was the previous year's yield. The second was the arithmetic mean of up to the three immediately preceding yields, requiring at least two prior observations. For every evaluated year, both variables were calculated without using the target-year outcome.

ERA5-Land processing and province aggregation

Seven ERA5-Land variables were extracted: 2 m air temperature, soil temperature in layer 1 (0–7 cm), volumetric soil water in layers 1 (0–7 cm) and 2 (7–28 cm), surface solar radiation downwards, total precipitation, and potential evaporation. Temperature was converted from kelvin to degrees Celsius. Surface solar radiation was converted to W m^{-2}. Accumulated precipitation and potential evaporation were converted to monthly millimetres; the ERA5 sign convention for potential evaporation was reversed so that greater positive values represented greater evaporative demand. Climatic water balance was then calculated as precipitation minus potential evaporation.

ERA5 grid-cell centres were assigned to province polygons. When several cells occurred within a province, values were averaged using the cosine of grid-cell latitude as an area weight.

Province-month representations used between 7 and 418 grid centres, depending on province size and geometry. This approach produces administrative-area climate summaries rather than wheat-area-weighted summaries, a limitation considered in Section 4.7.

Phenology-aligned climate features

Monthly variables were aggregated into five fixed periods chosen to approximate the dominant winter-wheat production calendar while retaining nationwide consistency. Temperature, soil moisture, soil temperature, and radiation were averaged using days per month as weights. Precipitation and potential evaporation were summed. Climatic water balance was derived after aggregation. The five periods and eight variables produced 40 climate predictors (Table 1).

Table 1. Predictor groups and phenology-aligned periods.

Predictor group	Definition	Variables or count
Spatial and temporal	Province indicator and harvest year	2
Historical yield	Previous-year yield; mean of previous three years	2
Autumn	October–December before the harvest year	8
Winter	January–February of the harvest year	8
Early spring	March–April	8
Grain filling	May–June	8
Complete growing season	Previous October through June	8

Climate variables in each period: 2 m temperature, layer-1 soil temperature, layer-1 soil water, layer-2 soil water, downward solar radiation, total precipitation, potential evaporation, and climatic water balance.

The periods are phenology-aware but not province-specific phenological observations. Their purpose was to preserve broad stage timing without introducing unavailable local sowing, flowering, or harvest dates. Because current-year information through June was included, the resulting task is a late-season yield-estimation problem. It is not a pre-season forecast.

Preprocessing and feature sets

The combined model used 44 predictors: province, year, two historical-yield predictors, and 40 climate variables. Province was one-hot encoded. Missing numeric values, which were restricted mainly to early historical lags, were replaced by the training-set median. Numeric predictors were standardized for Ridge, support vector regression (SVR), RLS, and the RLS component of the hybrid. Tree-based models used unstandardized numeric values. Preprocessing was fitted only on the training data available to each model at the relevant evaluation step.

For the ablation analysis, the eight-year sliding Random Forest was re-estimated with three feature sets: (i) combined historical yield and climate, (ii) historical yield without climate, and (iii) climate without historical-yield lags. Province and year remained in all three configurations so that the comparison isolated the incremental contribution of the two principal predictor groups.

Model families and update mechanisms

The benchmark included two naive rules and eleven machine-learning configurations (Table 2). Random Forest was included as a nonlinear bagged-tree learner (Breiman, 2001), Extra Trees as a more strongly randomized tree ensemble (Geurts et al., 2006), histogram gradient boosting as a sequential tree-boosting approach related to gradient boosting (Friedman, 2001), and RBF-SVR as a kernel method (Smola and Schölkopf, 2004). Ridge provided a regularized linear benchmark. RLS represented online linear adaptation (Sayed, 2003).

Table 2. Models, temporal update rules, and fixed configurations.

Model	Update rule	Main fixed settings
Previous-year yield	None	Predict the most recent observed yield
Previous-three-year mean	None	Mean of the three most recent observed yields
Static Ridge	Fit once through 2013	$\alpha = 10$
Static SVR	Fit once through 2013	RBF kernel; $C = 100$; $\varepsilon = 0.1$; γ = scale
Static Random Forest	Fit once through 2013	500 trees; min leaf = 2; max features = 0.5
Static Extra Trees	Fit once through 2013	500 trees; min leaf = 2; max features = 0.7
Static HistGradientBoosting	Fit once through 2013	400 iterations; learning rate = 0.05; max leaf nodes = 15; min leaf = 20; L2 = 1
Expanding Ridge / Random Forest	Refit annually	All earlier years; same model settings as static versions
Sliding-8y Ridge / Random Forest	Refit annually	Most recent eight years; same model settings as static versions
Online RLS	Update after each completed year	Ridge initialization $\alpha = 10$; forgetting factor $\lambda = 0.98$
Hybrid static RF + online RLS	Update RLS residual correction annually	Static RF base; residual RLS $\alpha = 20$ and $\lambda = 0.98$

For the hybrid, the base Random Forest was fitted through 2013 and then held fixed. Initial residual-correction targets were generated from temporally out-of-fold Random Forest predictions for 2009–2013, with each year's residuals obtained from a model trained only on earlier years. RLS was fitted to those residuals and subsequently updated once per completed harvest year. Hybrid predictions were the non-negative sum of the static Random Forest prediction and the online residual correction.

Configurations were held constant across evaluation years. No nested temporal hyperparameter-selection loop was conducted. Consequently, the experiment is a controlled comparison of specified model/update strategies, not proof that each model family or memory length was optimally tuned.

Walk-forward evaluation design

The initial training period was 2004–2013. Evaluation proceeded sequentially from 2014 through 2025. In each test year, predictions were generated for all 79 provinces before any observed yield from that year was used for updating. This prevented within-year leakage.

Static models were trained once on 2004–2013. Expanding models were refitted for test year t using all years earlier than t. Sliding models were refitted using years $t - 8$ through $t - 1$. RLS and the hybrid retained their current states while predicting all provinces in year t, then received the full year's labels in one batch. This design reflects annual administrative data availability and makes static, periodically adaptive, online, and hybrid approaches directly comparable.

Performance metrics and uncertainty

The primary metric was root mean squared error (RMSE), expressed in kg da^{-1}. Secondary metrics were mean absolute error (MAE), coefficient of determination (R^2), mean absolute percentage error (MAPE), and mean bias, defined as prediction minus observation. Pooled metrics were calculated over all 948 out-of-time predictions (79 provinces × 12 evaluation years), and yearly metrics were also examined.

Uncertainty in paired model differences was estimated with 2,000 province-clustered bootstrap repetitions. Provinces, rather than individual province-year rows, were sampled with replacement, retaining each selected province's complete 2014–2025 prediction history. For each repetition, improvement was calculated as the reference-model error minus the focal-model error; positive values therefore favoured the focal model. Percentile 95% confidence intervals were obtained from the 2.5th and 97.5th percentiles.

Software and reproducibility

Data processing and modelling were conducted in Python using pandas, NumPy, netCDF4, Shapely, and scikit-learn (Pedregosa et al., 2011). A random seed of 42 was used for stochastic estimators and bootstrap sampling. The analysis-ready panel, scripts, model-level predictions, metrics, and bootstrap summaries were retained as reproducibility files.

RESULTS

Descriptive characteristics

The final dataset contained 1,738 province-year observations from 79 provinces and 22 harvest years. Province-level wheat yield averaged 248.66 kg da^{-1}, with a standard deviation of 82.58 kg da^{-1}. The median was 244.34 kg da^{-1}, and the interquartile range extended from 191.63 to 295.78 kg da^{-1}. Observed yields ranged from 42.43 to 542.81 kg da^{-1} (Table 3).

Table 3. Descriptive statistics of province-level wheat yield, 2004–2025.

Statistic	Value
Province-year observations	1,738
Provinces	79
Harvest years	22
Mean (kg da^{-1})	248.66
Standard deviation (kg da^{-1})	82.58
Minimum (kg da^{-1})	42.43
25th percentile (kg da^{-1})	191.63
Median (kg da^{-1})	244.34
75th percentile (kg da^{-1})	295.78
Maximum (kg da^{-1})	542.81

The annual provincial mean was lowest in 2007 (217.71 kg da^{-1}) and highest in 2023 (291.99 kg da^{-1}). The broad long-term increase was not monotonic. Mean yield fell to 277.32 kg da^{-1} in

2024 and 249.62 kg da^{-1} in 2025. Between-province dispersion remained substantial throughout the period (Figure 1).

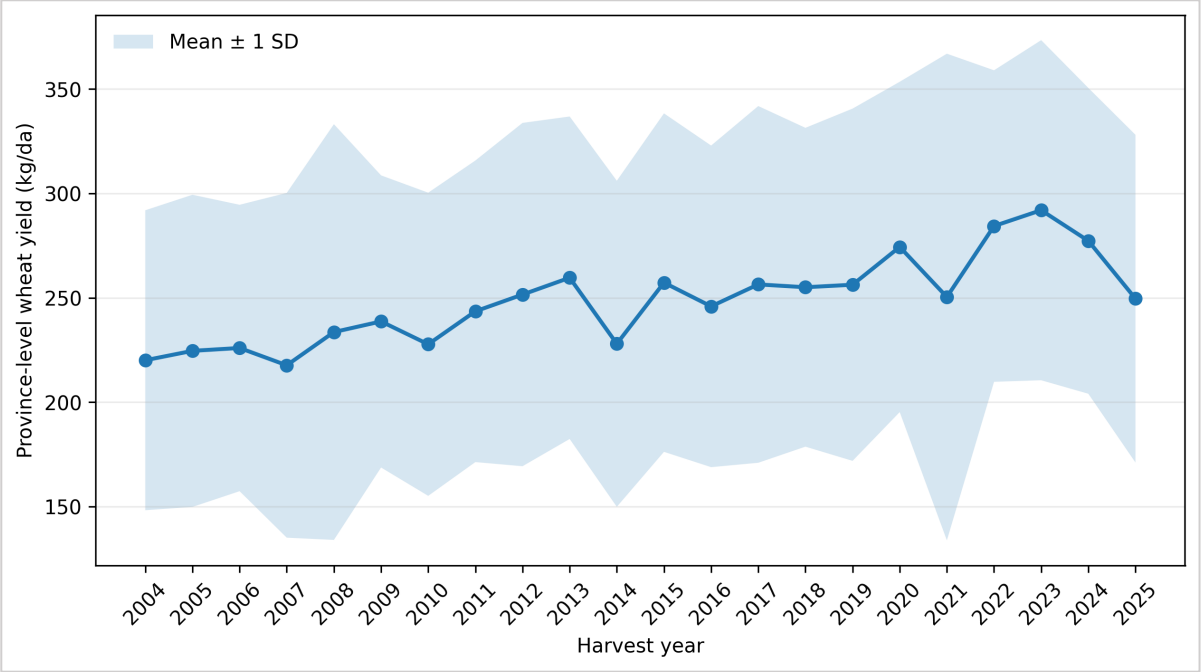

Figure 1. Annual mean and between-province standard deviation of wheat yield, 2004–2025.

Overall walk-forward performance

The previous-three-year mean achieved the lowest pooled prediction error, with MAE of 34.26 kg da^{-1}, RMSE of 47.00 kg da^{-1}, R^2 of 0.688, and MAPE of 15.10% (Table 4; Figure 2). The eight-year sliding Random Forest was the best machine-learning configuration, with MAE of 35.07 kg da^{-1}, RMSE of 47.63 kg da^{-1}, R^2 of 0.679, and MAPE of 15.83%. Expanding Random Forest followed closely with RMSE of 48.06 kg da^{-1}, while static Extra Trees achieved 48.91 kg da^{-1}.

Table 4. Overall walk-forward performance, 2014–2025.

Rank	Model	Type	MAE	RMSE	R^2	MAPE (%)	Bias
1	*Previous-three-year mean*	Naive	**34.256**	**46.995**	**0.688**	**15.096**	-1.939
2	*Sliding-8y Random Forest*	Adaptive	35.066	47.633	0.679	15.834	2.254
3	*Expanding Random Forest*	Adaptive	35.247	48.060	0.674	15.847	0.151
4	*Static Extra Trees*	Static	36.400	48.914	0.662	16.375	0.338
5	*Static Random Forest*	Static	37.342	50.553	0.639	16.793	2.738
6	*Hybrid static RF + online RLS*	Hybrid	38.329	50.831	0.635	17.485	7.301
7	*Static HistGradientBoosting*	Static	37.813	51.068	0.632	17.099	6.510
8	*Expanding Ridge*	Adaptive	38.980	52.412	0.612	18.567	12.547
9	*Sliding-8y Ridge*	Adaptive	41.108	55.013	0.572	19.081	10.344
10	*Previous-year yield*	Naive	39.645	55.511	0.565	17.204	0.836
11	*Static SVR*	Static	44.636	60.299	0.486	20.912	4.881
12	*Static Ridge*	Static	48.277	62.329	0.451	23.485	32.058
13	*RLS (λ = 0.98)*	Online	54.236	73.074	0.246	24.412	-2.303

The difference between the three-year mean and sliding Random Forest was not statistically conclusive. Sliding Random Forest's estimated RMSE improvement relative to the three-year mean was −0.63 kg da⁻¹ (95% CI: −2.04 to 0.81), meaning the simple baseline was numerically better and the interval crossed zero. In contrast, sliding Random Forest improved RMSE over previous-year persistence by 7.90 kg da⁻¹ (95% CI: 5.02–11.13).

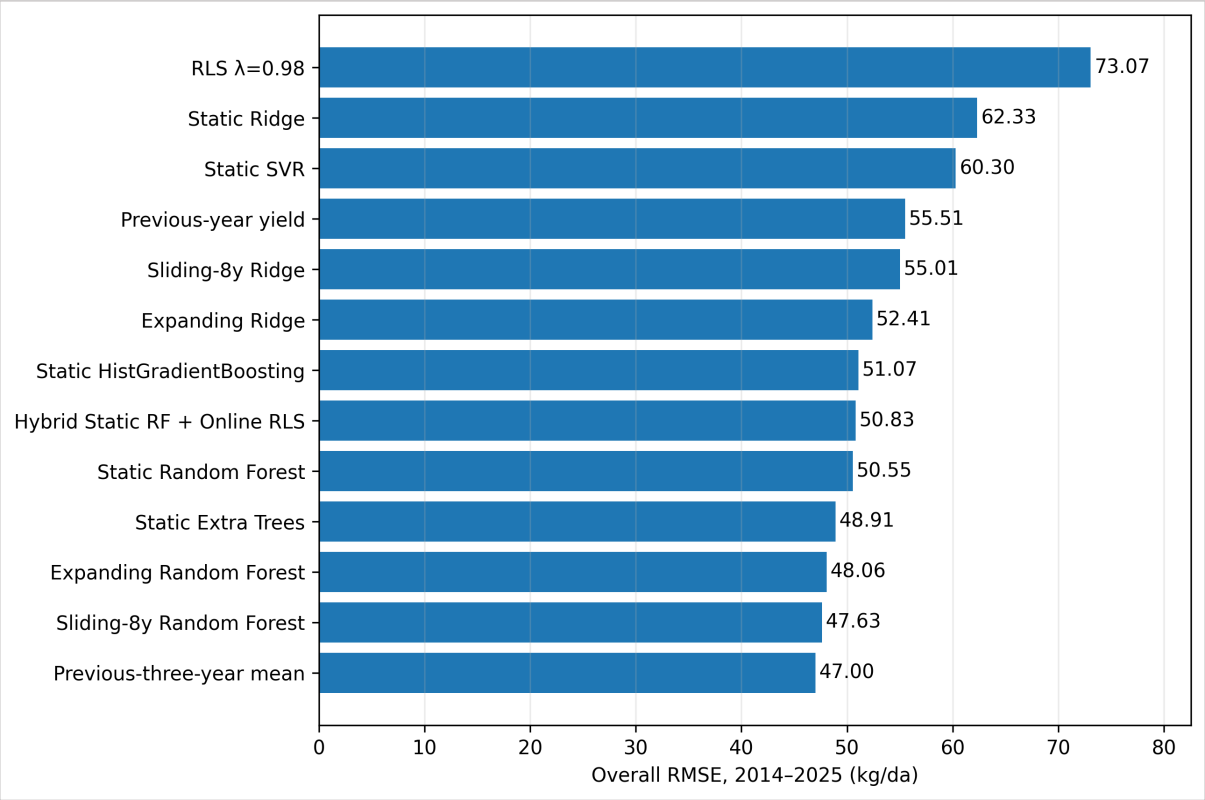

Figure 2. Overall RMSE of all evaluated models and naive baselines, 2014–2025. Lower values indicate better performance.

Year-specific performance

Accuracy varied materially across years, and no model ranked first in every year (Figure 3). The three-year mean had the lowest annual RMSE in 2014, 2015, 2016, 2019, and 2021. Expanding Ridge ranked first in 2017 and 2020; sliding Random Forest ranked first in 2018; static SVR, static HistGradientBoosting, sliding Ridge, and expanding Random Forest ranked first in 2022, 2023, 2024, and 2025, respectively.

Errors increased sharply in 2021. The three-year mean had RMSE of 68.93 kg da⁻¹, compared with 72.89 for sliding Random Forest and 72.97 for expanding Random Forest. In 2024, sliding Ridge achieved the lowest RMSE (37.34 kg da⁻¹), while sliding and expanding Random Forest produced nearly identical values of 42.20 kg da⁻¹. In 2025, expanding Random Forest ranked first at 61.04 kg da⁻¹, compared with 63.72 for the three-year mean and 65.38 for sliding Random Forest. Thus, discarding older observations was not advantageous in every year.

Standalone RLS was unstable: its annual RMSE exceeded 120 kg da⁻¹ in 2015 and 2023 and reached 94.30 kg da⁻¹ in 2025. The hybrid reduced the largest RLS failures but did not consistently improve on static or adaptive tree ensembles.

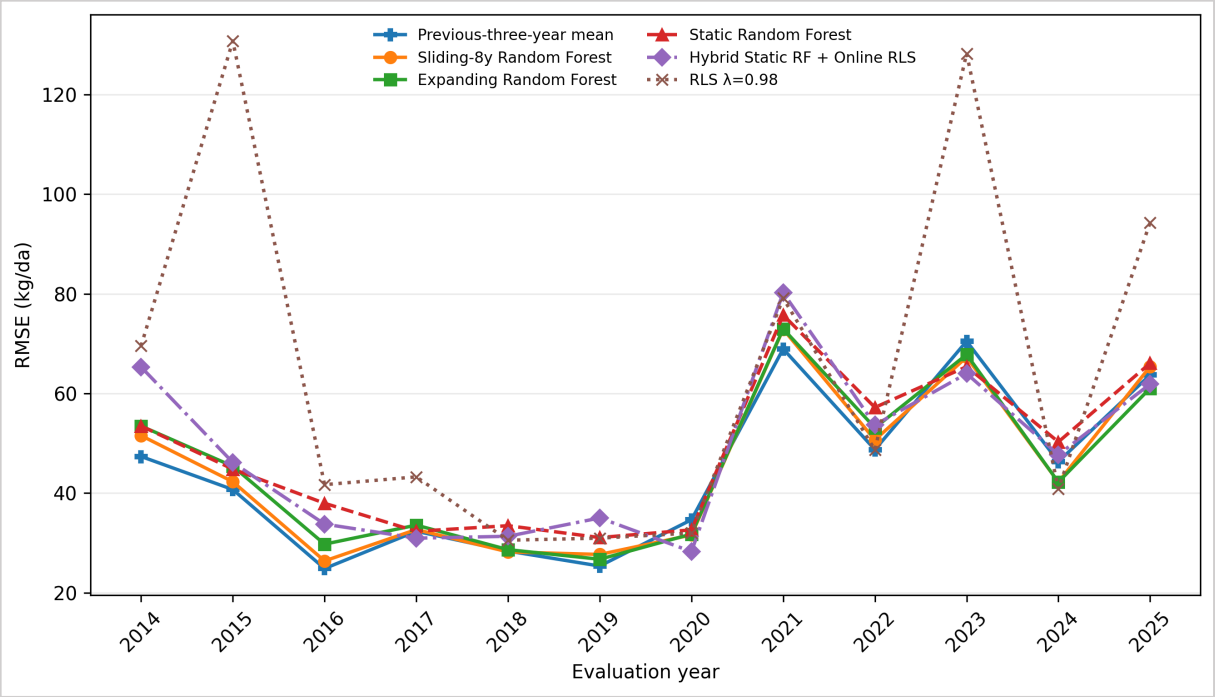

Figure 3. Annual RMSE of selected naive, static, adaptive, online, and hybrid approaches during walk-forward evaluation.

Paired comparisons among machine-learning models

Province-clustered bootstrap comparisons are reported in Table 5 and Figure 4. Sliding Random Forest reduced RMSE by 2.91 kg da^{-1} relative to static Random Forest (95% CI: 1.27–4.65) and by 3.19 kg da^{-1} relative to the hybrid (95% CI: 1.95–4.39). Its advantage over standalone RLS was 25.47 kg da^{-1} (95% CI: 22.84–27.95).

Table 5. Province-clustered bootstrap improvements of Sliding-8y Random Forest relative to reference models.

Reference model	MAE improvement	95% CI	RMSE improvement	95% CI
Previous-three-year mean	-0.795	-1.873 to 0.264	-0.630	-2.045 to 0.806
Previous-year yield	4.617	2.717 to 6.500	7.900	5.022 to 11.131
Expanding Random Forest	0.197	-0.386 to 0.778	0.447	-0.355 to 1.256
Static Extra Trees	1.352	0.145 to 2.690	1.298	-0.088 to 2.774
Static Random Forest	2.268	0.940 to 3.581	2.910	1.266 to 4.653
Hybrid static RF + online RLS	3.252	2.338 to 4.194	3.188	1.952 to 4.393
RLS (λ = 0.98)	19.185	17.117 to 21.235	25.473	22.841 to 27.948

Positive values favour Sliding-8y Random Forest. Confidence intervals are percentile intervals from 2,000 province-clustered bootstrap repetitions.

The estimated RMSE difference between sliding and expanding Random Forest was only 0.45 kg da^{-1} and its confidence interval crossed zero. The interval for static Extra Trees also crossed zero. The evidence therefore supports annual Random Forest refitting relative to a fixed Random Forest, but it does not establish an optimal advantage for the eight-year memory over an expanding history.

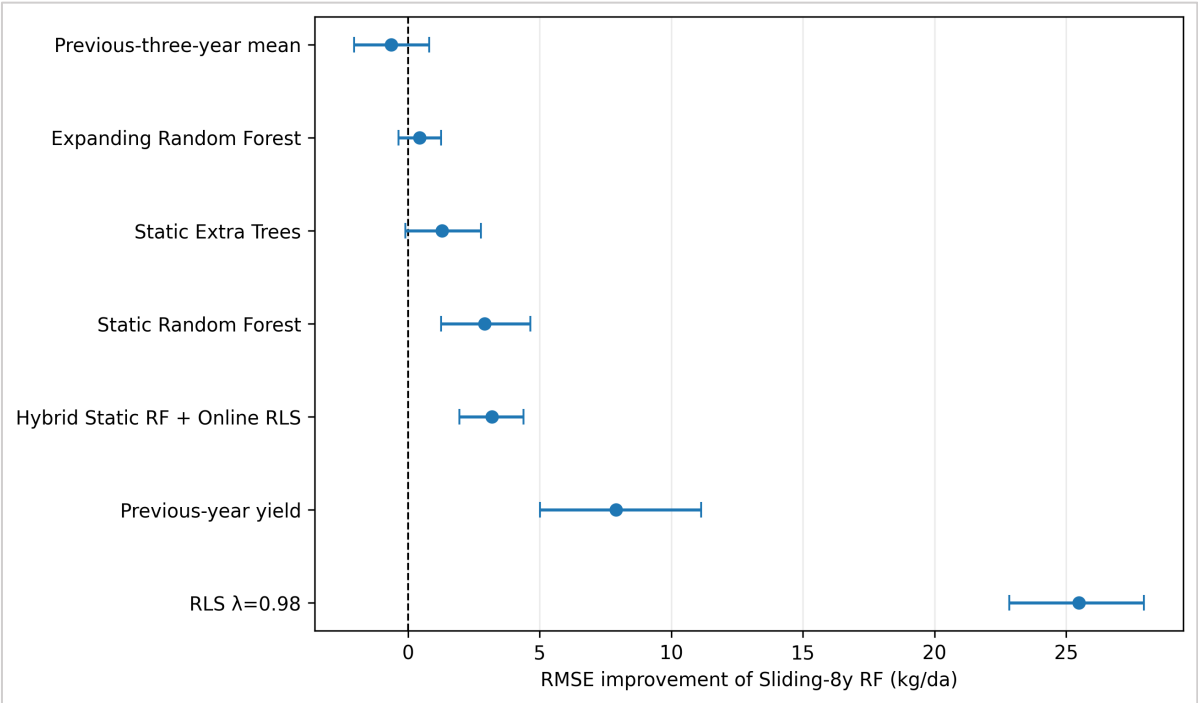

Figure 4. Province-clustered bootstrap estimates of RMSE improvement. Positive values favour Sliding-8y Random Forest; horizontal lines show 95% confidence intervals.

Incremental contribution of climate predictors

Under the same eight-year sliding Random Forest, the combined historical-yield and climate model achieved RMSE of 47.63 kg da^{-1}, compared with 49.82 for the historical-only model and 59.35 for the climate-only model (Table 6; Figure 5). Adding the 40 ERA5-Land predictors to historical yield reduced RMSE by 2.20 kg da^{-1} (95% CI: 1.01–3.34) and MAE by 2.19 kg da^{-1} (95% CI: 1.22–3.14). The combined model improved RMSE over climate alone by 11.66 kg da^{-1} (95% CI: 8.17–15.29).

Table 6. Walk-forward ablation results for Sliding-8y Random Forest.

Feature set	MAE	RMSE	R^2	MAPE (%)	Bias
Historical yield + climate	**35.066**	**47.633**	**0.679**	**15.834**	2.254
Historical yield only	37.244	49.819	0.649	16.818	1.253
Climate only	45.787	59.347	0.502	20.789	-1.957

These results separate incremental climate value from overall superiority over a naive rule. ERA5-Land variables improved Random Forest relative to the same algorithm without climate, but the combined model did not significantly outperform the three-year-mean baseline over the complete evaluation period.

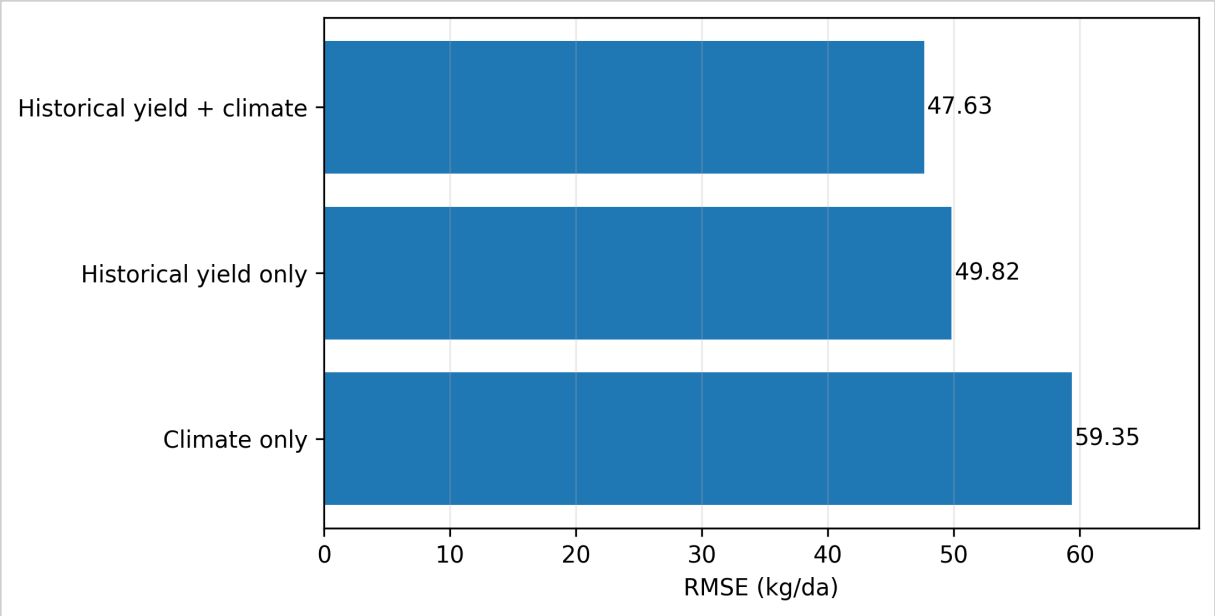

Figure 5. Ablation RMSE under a common Sliding-8y Random Forest and walk-forward protocol.

Annual comparison with the strongest naive baseline

Sliding Random Forest had lower annual RMSE than the three-year mean in four of twelve test years: 2018, 2020, 2023, and 2024. The baseline performed better in the other eight years (Figure 6). The largest Random Forest advantages occurred in 2024 and 2023, whereas the largest disadvantages occurred in 2014 and 2021. This reversal across years explains why the pooled bootstrap interval included zero despite the model's clear improvement over one-year persistence.

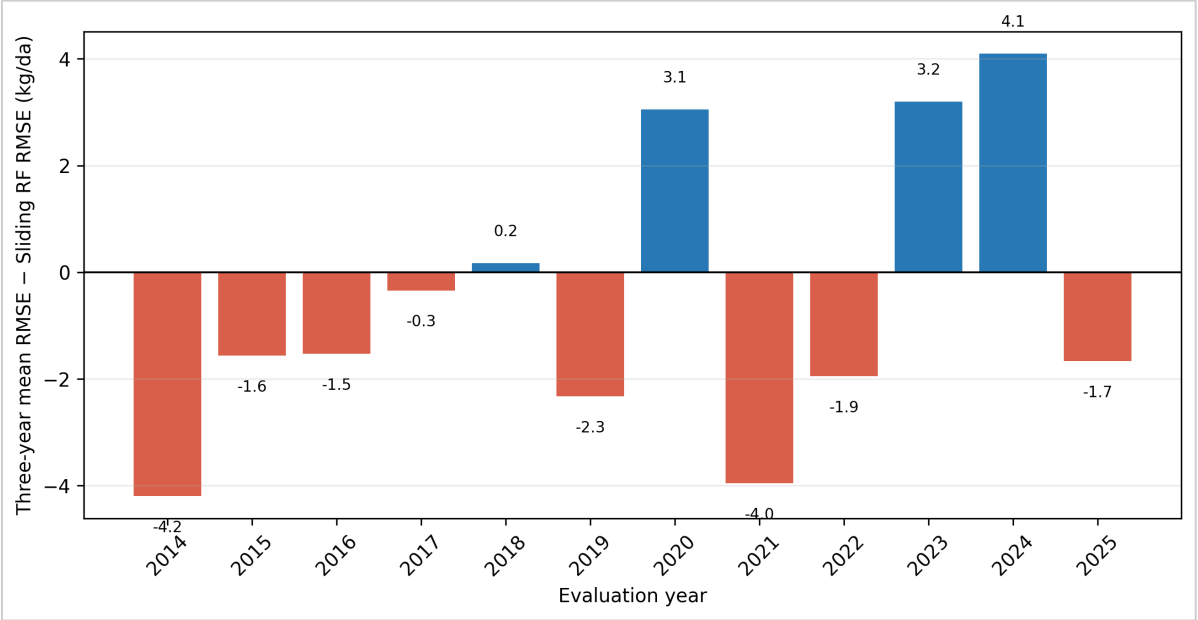

Figure 6. Annual RMSE difference between the previous-three-year mean and Sliding-8y Random Forest. Positive values indicate lower RMSE for Random Forest.

DISCUSSION

Principal findings

Four findings define the study. First, the previous-three-year mean was the strongest numerical predictor overall, and its difference from the best machine-learning model was statistically inconclusive. Second, periodically refitted Random Forest models were the strongest machine-learning group; sliding Random Forest significantly improved on a fixed Random Forest, but not conclusively on expanding Random Forest or static Extra Trees. Third, ERA5-Land climate predictors added a modest and statistically supported improvement beyond recent yield history within the same Random Forest design. Fourth, the tested online RLS and RF–RLS hybrid did not improve predictive accuracy.

These results narrow the claim that can be made. The study does not show that machine learning universally outperforms simple yield-history rules. It shows that temporal updating improves Random Forest relative to leaving it fixed, that climate information contributes incremental signal, and that the benefit of a specific adaptation mechanism depends on its fit to the problem.

Why the three-year mean was difficult to beat

Recent realized yield is a compact proxy for persistent province-level influences that were not directly observed: technology, cultivar mix, irrigation, soil constraints, input intensity, and management. A three-year average also attenuates one-year measurement noise and isolated weather shocks. This explains why it substantially outperformed one-year persistence and remained competitive with models containing 44 predictors.

The result is not evidence that climate is unimportant. Rather, it indicates that recent yield history already embeds both structural conditions and part of the recent environmental signal. Strong naive baselines are scientifically necessary because comparisons only among complex algorithms can exaggerate progress. Petrovic et al. (2026) found that random partitioning produced much higher wheat-prediction R^2 than temporal validation in a Czech regional dataset.

The present walk-forward protocol is deliberately harder than random splitting, but it better approximates use on an unseen harvest year.

The annual reversals are operationally important. Sliding Random Forest improved on the three-year mean in only one third of the evaluation years. A monitoring system should therefore report the machine-learning estimate together with the transparent baseline rather than treating the more complex model as automatically superior.

Climate information added value, but was not dominant

The ablation analysis provides the cleanest evidence for the contribution of ERA5-Land. Adding climate variables to historical-yield predictors reduced RMSE from 49.82 to 47.63 kg da^{-1} under the same algorithm, memory window, and evaluation design. The province-clustered confidence interval excluded zero. Climate information therefore contained predictive content that was not fully represented by recent yields.

The gain was nevertheless modest. Climate-only RMSE was 59.35 kg da^{-1}, substantially worse than historical-only RMSE. Current weather affects yield formation, but realized provincial yield also reflects management, soil, irrigation, cultivar, pests, technology, and economic conditions. Kaleli (2025) similarly found that structural economic inputs improved national crop-yield prediction beyond climate-only specifications in Türkiye. Future models should therefore treat reanalysis climate as one component of a broader information system rather than as a complete explanation of yield variability.

The stage-aligned design is supported by recent research. Tuğaç et al. (2022) found that the usefulness of satellite vegetation information varied by wheat-development period, and Güngüneş et al. (2025) grouped environmental variables by phenological stage in Central Anatolia. Srivastava et al. (2022) used weekly weather, soil, and phenology to capture time-dependent wheat responses across German counties. The present study extends that logic to a nationwide Turkish panel, while showing that fixed monthly stage summaries yield incremental rather than dominant gains once recent yield is included.

Static, expanding, and sliding adaptation

Sliding Random Forest significantly reduced RMSE relative to static Random Forest. This supports updating the model as new annual observations become available. Crop-yield relationships may evolve with changes in technology, cultivar adoption, irrigation, climate, or reporting. Such evolution is compatible with the broader concept-drift framework described by Gama et al. (2014).

However, the comparison between sliding and expanding Random Forest was inconclusive. The eight-year window had the lowest pooled machine-learning RMSE, but its estimated advantage over expanding Random Forest was only 0.45 kg da^{-1}, with a confidence interval spanning zero. In 2025, expanding Random Forest was clearly better than sliding Random Forest. Removing older information may reduce the influence of obsolete relationships, but it also reduces sample size and can discard rare historical conditions that remain relevant. The supported conclusion is therefore that annual refitting helped; the data do not prove that eight years is the optimal memory length.

This result agrees with the broader crop-forecasting literature's emphasis on temporal design. Paudel et al. (2021, 2022) demonstrated the value of operationally structured regional forecasting workflows and noted that performance can deteriorate in extreme years. The present

findings likewise show that model ranking depends on the evaluation year and that pooled averages should not be interpreted as universal dominance.

Why online RLS and the hybrid failed

Standalone RLS had the weakest overall RMSE and pronounced yearly instability. RLS updates a linear coefficient vector efficiently, but the province–climate–yield relationship is nonlinear and spatially heterogeneous. Correlated seasonal predictors and discontinuous year-to-year shocks also challenge a fixed linear forgetting mechanism. The results do not imply that all online learning is unsuitable for agriculture; they show that this particular online learner was poorly matched to the data structure.

The hybrid residual correction was more stable than standalone RLS but remained worse than static and adaptive Random Forest. Its positive mean bias suggests that online corrections were not consistently calibrated across provinces and years. Periodic batch refitting of the nonlinear model was more effective than leaving the nonlinear base fixed and adapting only a linear residual layer.

Future hybrid systems should be more flexible. Possible directions include regional residual models, dynamically weighted ensembles, explicit drift detection, bounded correction terms, and uncertainty-aware updates. Each should still be compared with expanding and sliding refits and with the three-year-mean baseline. Complexity is justified only when it produces a practically meaningful and statistically supported out-of-time gain.

Relationship to recent studies

Direct numerical comparison with published wheat-yield studies is inappropriate because spatial scales, targets, input resolution, forecast dates, and validation schemes differ. Srivastava et al. (2022) used richer weekly soil, weather, and phenology data for German counties. Tuğaç et al. (2022) used high-resolution vegetation indices at field scale. Güngüneş et al. (2025) focused on Central Anatolia, whereas the present study covers nearly the entire country and twelve later test years. Higher R^2 values in narrower or randomly partitioned studies do not necessarily imply better deployable forecasting.

The present contribution is complementary. It provides a balanced 79-province panel, a common feature set, explicit update strategies, strong naive baselines, a 12-year walk-forward evaluation, ablation analysis, and province-clustered uncertainty. This design sacrifices some headline accuracy but produces a more defensible estimate of performance on future administrative data.

Recent evidence from Türkiye's seafood markets likewise shows that forecast rankings depend strongly on the underlying temporal structure of the data. In the presence of a long-run cointegrating relationship, VECM-based approaches outperformed both univariate benchmarks and XGBoost models (Ural & Öztornacı, 2026). Consistent with this context-dependent evidence, the strong performance of the historical-yield baseline in the present study indicates that increasing model complexity alone does not necessarily improve out-of-sample wheat-yield prediction.

Practical implications and future research

The three-year mean should be retained as a mandatory benchmark and fallback in any operational system. Among machine-learning approaches, expanding and sliding Random

Forest are the most defensible candidates, but recent out-of-time metrics and baseline predictions should accompany their outputs.

Future work should define explicit issuance dates, such as pre-season, end of winter, end of April, and end of May, and rebuild predictors using only information available by each date. Wheat-area masks, satellite vegetation indices, local phenology, soil properties, irrigation, management, and economic variables should be added through the same walk-forward ablation design. Prediction intervals and drift-monitoring diagnostics are also necessary before policy deployment. Finally, model and window selection should be conducted through nested temporal validation, with the outer test years reserved strictly for final evaluation.

CONCLUSIONS

This study compared naive, static, adaptive, online, and hybrid models for province-level wheat-yield estimation across 79 provinces of Türkiye. Under a 2014–2025 walk-forward evaluation, the previous-three-year mean achieved the lowest numerical RMSE, while the eight-year sliding Random Forest was the strongest machine-learning approach. Their performance difference was not statistically conclusive. Annual Random Forest refitting improved on a fixed Random Forest, but the evidence did not establish that an eight-year window was superior to an expanding history. Standalone RLS and the RF–RLS hybrid did not improve accuracy.

ERA5-Land variables made a modest but statistically supported contribution beyond historical yield. Thus, climate information was useful, but recent yield history remained the dominant predictive source at province scale. The central scientific conclusion is consequently restrained: temporal adaptation and climate features can improve a specified machine-learning model, yet added complexity does not automatically beat a strong transparent baseline. The current system should be described as late-season provincial yield estimation. Earlier operational forecasting will require forecast-date-specific validation and richer crop, management, and remote-sensing data.

REFERENCES

Breiman, L. (2001). Random forests. *Machine Learning, 45*, 5–32. https://doi.org/10.1023/A:1010933404324

Copernicus Climate Change Service. (2026). *ERA5-Land monthly averaged data from 1950 to present*. Climate Data Store. Dataset accessed 30 July 2026. https://cds.climate.copernicus.eu/datasets/reanalysis-era5-land-monthly-means

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics, 29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), Article 44, 1–37. https://doi.org/10.1145/2523813

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine Learning, 63*, 3–42. https://doi.org/10.1007/s10994-006-6226-1

Güngüneş, R., Ateş, V., Erol, T., & Özek, R. (2025). Phenology-aware machine learning for wheat yield prediction under climate variability: Central Anatolia, Türkiye. *Türkiye Tarımsal Araştırmalar Dergisi, 12*(3), 287–295. https://doi.org/10.19159/tutad.1740059

Kaleli, B. A. (2025). Forecasting crop yields with climate and economic variables: A machine learning approach for Türkiye. *Turkish Journal of Agricultural Economics, 31*(2), 271–283. https://doi.org/10.24181/tarekoder.1699618

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J.-N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data, 13*, 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., & Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. *Agricultural Systems, 187*, 103016. https://doi.org/10.1016/j.agsy.2020.103016

Paudel, D., Boogaard, H., de Wit, A., van der Velde, M., Claverie, M., Nisini, L., Janssen, S., Osinga, S., & Athanasiadis, I. N. (2022). Machine learning for regional crop yield forecasting in Europe. *Field Crops Research, 276*, 108377. https://doi.org/10.1016/j.fcr.2021.108377

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://jmlr.org/papers/v12/pedregosa11a.html

Petrovic, B., Paluba, D., Nofrizal, A. Y., & Kucera, A. (2026). Analyzing meteorological effects on crop yield and yield prediction through machine learning with different data partitioning approaches: A case study from Czech Republic. *Journal of Agriculture and Food Research, 26*, 102662. https://doi.org/10.1016/j.jafr.2026.102662

Runfola, D., Anderson, A., Baier, H., Crittenden, M., Dowker, E., Fuhrig, S., Goodman, S., Grimsley, G., Layko, R., & Melville, G. (2020). geoBoundaries: A global database of political administrative boundaries. *PLOS ONE, 15*(4), e0231866. https://doi.org/10.1371/journal.pone.0231866

Sayed, A. H. (2003). *Fundamentals of adaptive filtering*. Wiley-IEEE Press.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*, 199–222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Srivastava, A. K., Safaei, N., Khaki, S., Lopez, G., Zeng, W., Ewert, F., Gaiser, T., & Rahimi, J. (2022). Winter wheat yield prediction using convolutional neural networks from environmental and phenological data. *Scientific Reports, 12*, 3215. https://doi.org/10.1038/s41598-022-06249-w

Tuğaç, M. G., Özbayoğlu, A. M., Torunlar, H., & Karakurt, E. (2022). Wheat yield prediction with machine learning based on MODIS and Landsat NDVI data at field scale. *International Journal of Environment and Geoinformatics, 9*(4), 172–184. https://doi.org/10.30897/ijegeo.1128985

Turkish Statistical Institute. (2026). *Crop production statistics: MEDAS database*. Dataset accessed 30 July 2026. https://biruni.tuik.gov.tr/medas/?kn=92&locale=en

Ural, N. B., & Öztornacı, B. (2026). Cointegration-based forecasting of sea bass and sea bream prices in Türkiye: Evidence from monthly wholesale markets. Aquaculture Economics & Management, 30(1), 1–17. https://doi.org/10.1080/13657305.2026.2656604

Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture, 177*, 105709. https://doi.org/10.1016/j.compag.2020.105709