remote sensing

MDPI

Article

Plot-Level Yield Modelling of Winter Triticale from Multi-Temporal UAV Multispectral Indices: Validation Across Plots and Seasons

Gergely Dezső [1], Balázs Gáti [1], Judit Csabai [1], Edit Kosztyuné Krajnyák [1], Béla Szabó [1], Zoltán Kovács [1], Shukhrat Shokirov [2], Zokhid Mamatkulov [2], Ilhom Abdurahmanov [2,3] and Zoltán Túri [4,*]

1 Institute of Engineering and Agricultural Sciences, University of Nyíregyháza, Sóstói Street 31/b, 4400 Nyíregyháza, Hungary; dezso.gergely@nye.hu (G.D.); gati.balazs@nye.hu (B.G.); csabai.judit@nye.hu (J.C.); krajnyak.edit@nye.hu (E.K.K.); szabo.bela@nye.hu (B.S.); zoltan.kovacs@nye.hu (Z.K.)
2 Department of Geodesy and Geoinformatics, Tashkent Institute of Irrigation and Agricultural Mechanization Engineers (TIIAME)—National Research University, Kary-Niyaziy 39, Tashkent 100000, Uzbekistan; shokirov@tiiame.uz (S.S.); z.mamatqulov@tiiame.uz (Z.M.); i.abduraxmonov@tiiame.uz or ilhom.isakovich@nstu.uz (I.A.)
3 Namangan State Technical University, Namangan 160800, Uzbekistan
4 Department of Physical Geography and Geoinformatics, University of Debrecen, Egyetem tér 1, 4032 Debrecen, Hungary
* Correspondence: turi.zoltan@science.unideb.hu

Highlights

What are the main findings?

- Pooled UAV index–yield fits were strong but largely reflected the seasonal contrast. Spectral models beat the season-mean reference within 2024, but not within 2025.
- Leave-one-year-out RMSE stayed near that of the training-year mean (≈2.5 t ha^{-1}).

What is the implication of the main finding?

- UAV indices can rank plot conditions within a season, not forecast unseen seasons.
- Plot- and year-blocked validation over several seasons is needed for reliable models.

Abstract

Reported accuracy of unmanned aerial vehicle (UAV)-based yield models depends on whether validation respects repeated plots and seasonal variation. We tested whether temporal descriptors of vegetation indices explain plot-level grain yield of winter triticale and whether it transfers between years. Here, 7 multispectral UAV flights within 190–245 days after sowing were analysed for 20 fixed plots in 2024 and 2025. For normalised difference vegetation index (NDVI), green normalised difference vegetation index (GNDVI), normalised difference red-edge (NDRE), optimised soil-adjusted vegetation index (OSAVI), and wide dynamic range vegetation index (WDRVI), the temporal mean, median, standard deviation, and normalised area under the curve gave 20 candidate predictors. Single-index and multi-index models, with predictors selected by Elastic Net stability selection or a genetic algorithm, were evaluated by nested plot-grouped cross-validation and leave-one-year-out (LOYO) validation against season-mean references. Pooled R^2 reached 0.89 (RMSE = 433 kg ha^{-1}; WDRVI–KNN), but a season-mean reference alone gave 0.83. Within 2024, the best models reduced root mean square error (RMSE) from 571 to about 290–380 kg ha^{-1}; within 2025, none improved on the reference. In leave-one-year-out validation, the lowest RMSE (about 2.0 t ha^{-1}) was only 20% below that of the training-season mean (2.5 t ha^{-1}), and all R^2 values were negative. Yield differences between control and manured plots reversed direction between seasons (−328 and +300 kg ha^{-1}, both 95%

Check for updates

Academic Editor: Wenjiang Huang

Received: 10 September 2026
Revised: 13 November 2026
Accepted: 17 November 2026
Published: 20 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

confidence intervals included zero) and are confounded with plot position. The descriptors can rank plot conditions within a season, but two seasons do not support yield forecasting for unseen years.

Keywords: UAV multispectral imagery; winter triticale; vegetation indices; yield prediction; nested cross-validation

1. Introduction

Reliable estimates of crop yield support field-scale decision making, input allocation, and the evaluation of crop responses to environmental variability [1,2]. Contemporary yield models integrate diverse data sources. For instance, satellite imagery offers broad coverage and has been correlated with grain yield and yield components in cereals, including winter wheat and triticale [3,4]. However, its spatial resolution limits its utility in small experimental plots and for resolving within-field heterogeneity. Unmanned aerial vehicles (UAVs) provide centimetre-scale imagery with flexible acquisition timing, and UAV multispectral data have been used for biomass estimation, growth monitoring, yield estimation, and phenotyping in cereals [5–8]. Systematic reviews summarise the progress in UAV-based yield prediction for grain crops [9]. Proximal sensors remain an important reference for these approaches [10,11], and meteorological or climate variables are frequently added to remote-sensing predictors [12–14].

Spectral vegetation indices condense multispectral reflectance into variables associated with canopy greenness, chlorophyll content, and green biomass. Although the normalised difference vegetation index (NDVI) [15] is the most widely used index in cereals and relates strongly to biomass and nitrogen status [10,11,16], it tends to saturate in dense canopies and exhibits sensitivity to soil background. Consequently, alternative indices have been developed. For instance, the green normalised difference vegetation index (GNDVI) is more sensitive to chlorophyll concentration [17], the optimised soil-adjusted vegetation index (OSAVI) minimizes soil reflectance interference [18], and the wide dynamic range vegetation index (WDRVI) preserves sensitivity under high-biomass conditions [19]. Red-edge indices, such as the normalised difference red-edge (NDRE), have been reported to be informative on nitrogen status, biomass, and grain yield in wheat, particularly at advanced canopy stages [7,8,20]. In a UAV study of winter wheat, NDVI alone was a weaker predictor of above-ground biomass than crop height, and combining spectral and structural information improved the models [5]. The relative merit of individual indices depends on the crop, the sensor, and the growth stage and should, therefore, be evaluated rather than assumed.

Spectral index–yield relationships evolve dynamically throughout the growing season. Previous studies indicate that NDVI–yield correlations strengthen during advanced growth stages [3]. Furthermore, UAV-derived NDVI has successfully explained yield variation across several grain-filling stages in wheat [16]. Multi-temporal UAV data also enable yield assessment prior to flowering [8], while early canopy traits serve as strong yield predictors in oats [21]. Because a single acquisition represents only one developmental stage, repeated flights are commonly condensed into seasonal descriptors of canopy level, variability, and cumulative activity, such as the mean, median, standard deviation, and area under the curve (AUC) of an index time series. These descriptors keep the predictor set compact and interpretable while retaining information on both the magnitude and the duration of canopy greenness. They depend, however, on the acquisition window, the number and timing of flights, and the radiometric consistency among dates [22], which, therefore, need to be documented and, where possible, tied to crop development.

Triticale (×*Triticosecale* Wittmack) is grown for grain and fodder, particularly on marginal, low-fertility soils, but it has received less attention in remote-sensing yield studies than wheat, maize, or rice. Existing triticale studies have related NDVI to grain yield among genotypes [23], described the sensitivity of the spectral response of triticale to water stress [24], followed seasonal NDVI dynamics with satellite and UAV imagery [25,26], estimated dry matter yield in plot trials [27], compared vegetation dynamics among varieties using multi-temporal UAV indices [28], and related Sentinel-2 indices to grain yield and spike number [3]. These studies mostly describe NDVI behaviour or rely on a single index or a few indices. Evidence is particularly needed for organically managed triticale on acidic sandy soils, where low nutrient availability and variable water supply can produce strong seasonal yield contrasts and where organic amendments, such as fermented poultry manure, affect crop performance [29]. Nutrient supply is a major source of canopy and yield variation that spectral indices can register: NDVI from satellite and UAV platforms has been related to early nitrogen status and grain yield in wheat [11], UAV-derived leaf area index (LAI) has supported yield prediction under different nitrogen fertilizer types and rates [30], and UAV multispectral indices have explained variation in nitrogen- and water-use efficiency traits [31]. Organic amendments, such as poultry manure, create contrasts in nutrient supply and thereby widen the range of canopy conditions available for UAV-based assessment; in unreplicated field designs, however, such contrasts are confounded with location, which limits what can be inferred about the treatment itself.

Machine learning algorithms are increasingly deployed to convert remote-sensing predictors into yield estimates, particularly when relationships are nonlinear or involve multiple interacting factors.

Random forest (RF) represents one of the most widely applied methods due to its capacity to handle nonlinearity, robustness against noise, and built-in variable importance measures [32–35]. For example, RF has been combined with vegetation indices for soil nutrient monitoring in winter wheat [36], and ensemble tree methods have been applied to short NDVI time series for early-season wheat yield prediction [12].

Alternative model families—such as regularised linear regression (Elastic Net) [37], partial least squares regression (PLSR), support vector regression (SVR), k-nearest neighbour (KNN), and gradient-boosting frameworks like XGBoost—are also frequently utilised in UAV and satellite time-series studies [4,38,39]. The choice of algorithm must be evaluated alongside dataset size and validation design [40,41].

Although deep learning approaches are increasingly applied to large remote-sensing datasets [41,42], they generally require far more training data than plot-scale experiments provide. Consequently, lightweight alternatives have been reported to generalise better across agricultural domains under data-limited conditions [43].

More complex learners do not necessarily improve grain-yield prediction. For instance, in winter wheat, simple vegetation index regression predicted grain yield with root mean square error (RMSE) of 972 kg ha^{-1}, and advanced machine learning approaches did not significantly improve accuracy [44]. Accuracy figures remain difficult to compare across studies because crop types, yield ranges, sensors, predictor sets, and validation designs differ substantially [45–47].

Because temporal descriptors derived from several indices are strongly collinear, predictor selection and interpretation require particular care. Stability selection, which repeatedly applies a sparse learner to resampled data and retains predictors that are selected consistently [48], combined with Elastic Net regularisation, which copes with groups of correlated predictors [37], provides a principled way to obtain a compact predictor set. Wrapper approaches, such as genetic algorithms, offer a search-based alternative [49]. Impurity-based and permutation importance, Shapley Additive exPlanations (SHAP) [50], and recursive

feature elimination (RFE) [51] provide complementary views of predictor relevance, but importance values of correlated predictors can be unstable [52,53]; agreement among several methods is, therefore, more informative than any single ranking.

The reported accuracy of yield models depends strongly on the validation design. Random k-fold cross-validation assumes independent observations, whereas agricultural data are structured by plots, fields, and seasons; ignoring this structure leads to optimistic estimates, and blocking by group, space, or time is recommended [54,55]. Optimism also arises when predictor selection or hyperparameter tuning uses the data on which performance is evaluated, so these steps need to be nested within the training part of each validation split [56–58]. The consequences are documented: a district-level wheat yield model for Kazakhstan reached $R^2 = 0.646$ under spatial group cross-validation but only 0.413 under temporal validation [14], and in wheat breeding, five-fold cross-validation accuracies of 0.50–0.51 fell to 0.18–0.23 when predicting across years [59]. Limited transferability across years, cultivars, sensors, or regions has also been reported for rice, pea, wheat–barley, and rapeseed models, and has often been attributed to domain shifts in climate, management, and crop growth [13,60–62]. For winter wheat, cross-year and cross-location prediction from UAV data lost substantial accuracy, whereas fine-tuning with target-domain observations improved it [63,64], and transfer performance has likewise depended on source–target differences in soybean and oat [65,66]. In experiments with only a few seasons and repeated plots, observations are not independent, and withholding an entire season is an extrapolation that tests data coverage as much as the algorithm, particularly for tree-based learners, which do not predict outside the range of the training responses. Group-blocked schemes should, therefore, be reported together with within-season performance and a comparison with simple reference predictions, such as season means.

Against this background, the present study examines whether multi-temporal UAV multispectral indices can explain plot-level grain yield of organically grown winter triticale (×*Triticosecale* Wittmack), and how robust this relationship is when the plot and season structure of the data is respected in validation. The specific objectives were as follows: (1) derive compact seasonal descriptors from the time series of area-based mean values of five vegetation indices (NDVI, GNDVI, NDRE, OSAVI, and WDRVI), namely, the temporal mean, median, standard deviation, and AUC (denoted _t_mean, _t_median, _t_std, and _t_auc), which yields 20 candidate predictors; (2) compare grain yield and the UAV-derived descriptors between the control and the poultry-manure-treated management areas in each season; (3) compare the five indices by means of single-index models that use the four descriptors of one index; (4) select a compact set of eight predictors within each training set by Elastic Net bootstrap stability selection and compare this procedure with a consensus genetic algorithm; (5) compare seven model families (Elastic Net, SVR, PLSR, RF, gradient boosting, KNN, and XGBoost), tuned by grid search, using nested, plot-grouped repeated cross-validation and leave-one-year-out validation, with season-mean predictions as a reference, and report pooled as well as within-season performance; (6) interpret the retained predictors using permutation importance, impurity-based importance, SHAP, and RFE, taking predictor collinearity into account.

The comparison of the management areas is descriptive, because the experiment contains one control and one manured area, so that treatment is confounded with location, and the same 20 plots are observed in both seasons. Manure treatment is, therefore, not used as a predictor. The modelling is restricted to UAV-derived spectral predictors from two contrasting growing seasons and 20 fixed plots, which ensures methodological consistency between predictor generation, model training, and validation; weather data are used only to characterise the two seasons.

2. The Study Area and Research Data

2.1. Study Site, Experimental Design, and Weather Conditions

The experimental field is located in the Löszös-Nyírség microregion of the Nyírség region, within the administrative area of Nyírtelek settlement, in Szabolcs-Szatmár-Bereg County, Eastern Hungary (48°00′43.5″N, 21°35′35.5″E; Figure 1). The site belongs to a plain landscape type on unconsolidated deposits, often in low floodplain position [67]. The gently undulating surface of the study area lies at an altitude of about 110 m above sea level and is covered by acidic sandy soils with low humus content and poor nutrient availability.

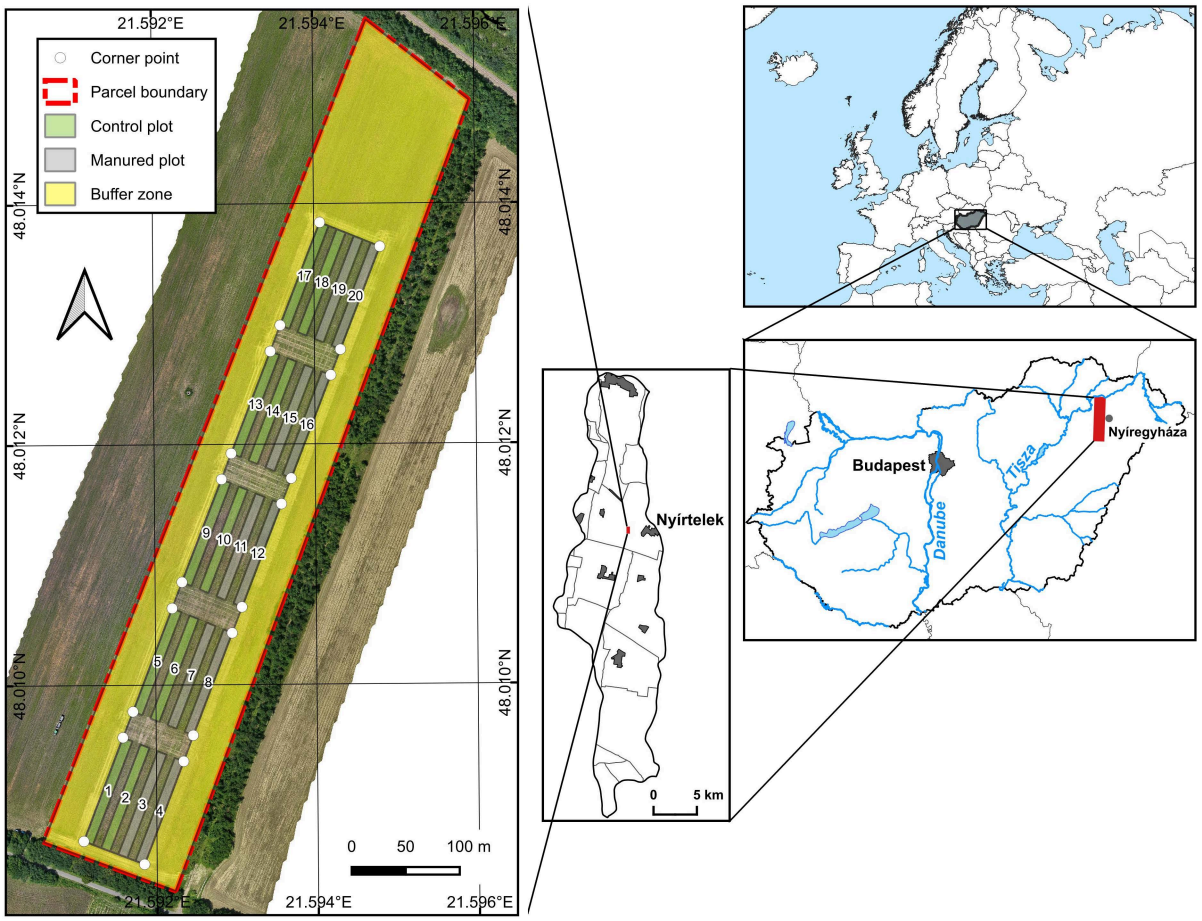

Figure 1. Geographical location of the experimental site.

The study covered two consecutive growing seasons (2023/2024 and 2024/2025). The experimental site has been managed under organic farming practices since 2015. Winter triticale (× *Triticosecale* Wittmack) cv. GK Maros was grown in both seasons. The experiment was conducted on a 10 ha field divided into two 5 ha management areas: an untreated control area and an area treated with poultry manure [29,68] (Figure 1). The purpose of including the two management regimes was to broaden the range of canopy conditions and grain yields available for spectral modelling, rather than to quantify the effect of poultry manure application. Each regime occupied a single management area; therefore, the ten monitored plots within each area represented spatial subsamples, not independent treatment replicates.

Fermented poultry manure (Bio-Fer Natur™, Baromfi Coop Ltd., Nyírjákó, Hungary) was applied to the treated area at 1 t ha^{-1} in autumn before sowing and was incorporated immediately; the treatment was repeated in both seasons, and the control area received no manure. The two regimes were established to generate differences in crop nutritional status

and canopy development [29]. Each area contained 10 fixed plots of 9 × 100 m (900 m^2), which yielded 20 plots monitored in both seasons. Because the experiment comprises one control and one treated area, the treatment is not replicated and is confounded with location. Triticale was sown on 17 October 2023 and on 26 September 2024 at a seeding rate of 200 kg ha^{-1} and was harvested on 8 July 2024 and 30 June 2025, respectively.

Mean temperatures over the August–July reporting periods were 13.2 °C in 2023–2024 and 11.5 °C in 2024–2025, exceeding the 1991–2020 mean of 10.7 °C by 2.5 and 0.8 °C, respectively [69,70] (Table 1). Precipitation totals were also above the corresponding long-term value, but their monthly distribution differed markedly (Table 2). February–March precipitation totalled 23.8 mm in 2024 and 121.2 mm in 2025, compared with the 1991–2020 value of 61.1 mm [69,70]. In contrast, June precipitation reached 182.4 mm in 2024 but only 17.7 mm in 2025, compared with 69.7 mm for the reference period. These observations demonstrate contrasting precipitation patterns. The weather data were used only to characterise the two seasons and were not included in the models.

Table 1. Monthly mean temperatures (°C) for the August–July periods of 2023–2024 and 2024–2025 and the corresponding long-term averages for 1870–2002 and 1991–2020 in Nyíregyháza [69,70].

Month	2023–2024	2024–2025	Multi-Year Average (1991–2020)	Multi-Year Average (1870–2002)
August	22.7	23.9	21.3	19.8
September	19.1	18.3	16.0	15.5
October	13.4	10.6	10.4	9.9
November	5.6	3.0	5.1	4.2
December	2.5	2.2	0.1	−0.4
January	1.0	2.6	−1.1	−2.4
February	7.8	0.0	0.6	−0.1
March	9.4	7.8	5.7	4.6
April	13.7	12.8	11.8	10.7
May	17.5	14.1	16.6	15.9
June	21.2	22.1	20.0	19.0
July	24.1	21.1	21.6	20.6
Average	13.2	11.5	10.7	9.7

Table 2. Monthly precipitation totals (mm) for the August–July periods of 2023–2024 and 2024–2025 and the corresponding long-term averages for 1870–2002 and 1991–2020 in Nyíregyháza [69,70].

Month	2023–2024	2024–2025	Multi-Year Average (1991–2020)	Multi-Year Average (1870–2002)
August	5.0	16.0	46.1	65.0
September	6.4	149.0	54.0	43.0
October	30.8	102.6	48.8	44.0
November	76.0	50.2	41.2	46.5
December	113.6	18.2	45.6	40.5
January	45.0	38.2	34.2	29.5
February	10.4	26.0	32.1	30.0
March	13.4	95.2	29.0	30.0
April	49.8	13.9	39.7	39.5
May	65.8	66.9	61.6	54.0
June	182.4	17.7	69.7	76.0
July	24.6	101.0	72.4	66.5
Total	623.2	694.9	574.4	564.5

2.2. Grain Yield Determination

In both growing seasons, each 900 m^2 experimental plot was harvested individually using a combine harvester. In the experimental area, the margins outside the experimental plots and the intermediate areas between the experimental plots were harvested first. In the fermented poultry-manure-treated and control plots, the weight of the harvested crop was measured using an axle scale placed under the trailer's wheels. On the day of harvest, the moisture content of the harvested grain was determined, and in all cases, it was below the level required for safe storage, so no artificial drying was used. Grain yields were corrected to a standard moisture content of 14% and expressed in kg ha^{-1} based on the harvested plot area.

While harvesting the 900 m^2 plots, samples were taken from the crop for laboratory analysis. During emptying onto the trailer, 2–3 kg grain samples were taken at each unloading and subsequently analysed at the Plant Production Laboratory of the University of Nyíregyháza. A Mininfra SmarT SW-type infrared grain analyser with transmission mode and hectolitre weight module was used to determine grain moisture content. Approximately 300 cm^3 of grain was used per sample for the tests.

Thus, 20 plot-level yield observations were obtained in each growing season, resulting in a total of 40 plot-year observations used as the reference response variable for the remote-sensing and machine learning analyses.

2.3. UAV Data Acquisition and Orthomosaic Generation

Multispectral images were acquired with a DJI Mavic 3 Multispectral (Mavic 3M) UAV (DJI Technology Co., Ltd., Shenzhen, China). The camera records four bands: green (560 ± 16 nm), red (650 ± 16 nm), red-edge (730 ± 16 nm), and near-infrared (860 ± 26 nm) wavelengths. Flights were performed at 55–60 m above ground level, which resulted in a ground sampling distance of approximately 2.5 cm, with 80% forward and lateral overlaps. Flights were carried out in direct sunlight or under broken cloud and in calm or moderate wind, at approximately weekly intervals during the two seasons. Positioning was corrected in real time with an Emlid Reach RS2+ real-time kinematic (RTK) global navigation satellite system (GNSS) receiver (Emlid, Budapest, Hungary) connected to a local base station of the Centipede network. The built-in precision estimate indicated a positional accuracy of a few centimetres. Images were georeferenced to UTM zone 34N (EPSG:32634) and stitched into orthomosaics using Pix4Dmapper 4.8.0 and Pix4Dfields 2.6.2 software (Pix4D SA, Lausanne, Switzerland). Radiometric calibration relied on the DJI Mavic 3M onboard sun sensor, which records real-time solar irradiance data into the metadata of each captured frame. The software automatically integrated these irradiance values with the camera's factory calibration profiles to convert raw measurements into absolute surface reflectance. The analysis used the four-band orthomosaics; the RGB mosaics were not used.

The radiometric consistency among dates affects vegetation indices and temporal descriptors, and the principles of radiometric correction of UAV snapshot-sensor data are reviewed in [22,71].

The subsequent workflow assumes that the orthomosaics represent surface reflectance on a 0–1 scale, and this assumption was checked with a plausibility test (Section 2.4). Only the acquisitions within the temporal window defined in Section 3.2 were used in the analysis.

2.4. Spatial Extraction, Geometric Normalisation, and Quality Control

Prior to harvesting, the corners of the 20 experimental plots were staked out using a Reach RS2+ Multi-Band RTK GNSS receiver in UTM zone 34N (EPSG:32634) with centimetre-level spatial accuracy. These field coordinates were directly applied to the orthomosaics to define the ROIs for NDVI extraction. Buffer zones between the plots were

utilised for field traffic to prevent trampling damage. Non-evaluated areas were cleared first by the combine harvester, followed by the individual harvesting and separate weighing of each plot, ensuring a precise geometric match between the NDVI extraction zones and the harvested areas.

Spectral data were extracted from the harvested strips of the 20 plots. A vector layer with the 20 combine-harvest track polygons defined the extraction area, so the area from which spectral data were taken corresponded to the area from which grain yield was measured; no additional buffer or edge mask was applied. Plot identifiers followed the order of the polygons in the layer (1–20), which is the order of the yield records. The polygons were transformed to the coordinate reference system of each raster when necessary. For each plot and acquisition date, the four bands were clipped to the polygon, and a pixel was assigned to a plot when its centre fell inside the polygon.

To align the plot rasters with the image axes, a rotation angle was derived from the geometry of the first polygon (the azimuth between its leftmost and uppermost vertices) and the same angle was applied to all 20 plots, which assumes parallel harvest strips. The rotation used an affine transformation with nearest-neighbour resampling, so reflectance values were not interpolated; the zero-filled margins created by the rotation were treated as invalid, and each raster was cropped to the bounding box of its valid pixels.

A joint validity mask was applied: a pixel was retained only if it lay within the plot polygon and had a finite value, not equal to 0, −10,000, or the raster no-data value, in all four bands simultaneously. All indices and band statistics of a plot and date were computed from this same set of pixels. No cloud, shadow, or soil-background masking (for example, by an NDVI threshold) was applied; the pixel selection, therefore, removes only missing or invalid values. For each plot and date, the valid pixel fraction was calculated as the ratio of jointly valid pixels to the number of valid near-infrared pixels in the polygon before rotation. A plausibility check of the radiometric scale was also recorded: it was passed when the median of each of the four bands was within the range 0–1.5. This check can reveal rasters that are stored as digital numbers (values far above 1), but it cannot detect values scaled far below the expected reflectance, and it does not constitute a radiometric validation. The processing log recorded all plot–date combinations that failed, and the number of records was compared with the expected number (flights × 20 plots). The quality flags were diagnostic and did not lead to automatic exclusion of flights.

3. Research Workflow and Methods

3.1. The Overall Research Workflow

The workflow consisted of four stages (Figure 2): extraction of plot-level spectral statistics from the orthomosaics, construction of temporal predictors and correlation analysis, single-index models, and the multi-index modelling with predictor selection and validation (Sections 3.4 and 3.5). The last two stages use the same predictor table and the same outer partitions, so that their results can be compared directly. The comparison of the management areas (Section 3.3) was carried out in parallel on the same data.

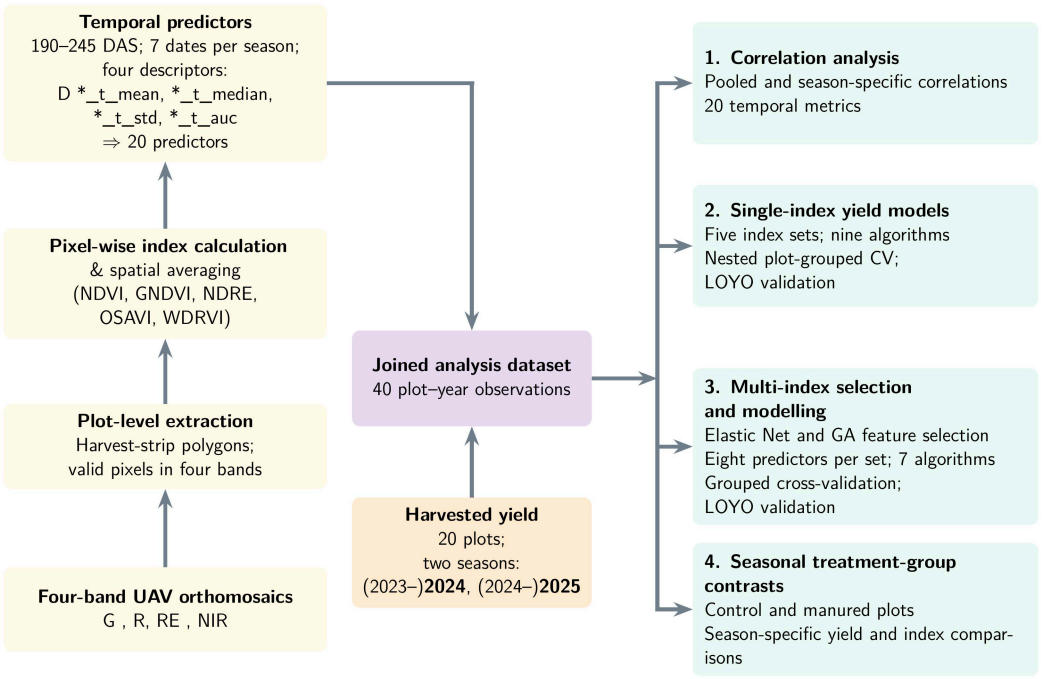

Figure 2. Overview of the research workflow: from UAV orthomosaics to plot-level yield modelling and validation.

3.2. Vegetation Indices, Temporal Window, and Predictors

Ten vegetation indices were calculated at pixel level from the valid reflectance values: NDVI, GNDVI, NDRE, chlorophyll index red-edge (CIre), chlorophyll index green (CIgreen), OSAVI, enhanced vegetation index 2 (EVI2), WDRVI, simple ratio (SR), and modified chlorophyll absorption in reflectance index 2 (MCARI2). Five indices were retained a priori for modelling to represent the main spectral families: NDVI [15], GNDVI [17], NDRE [72], OSAVI [18], and WDRVI [19] (Table 3). The red-edge index and WDRVI were included because NDVI saturates at high leaf area index. The other five indices were computed but not used in the modelling. Divisions with a zero or near-zero denominator were set to missing values without adding a correction term.

Table 3. Vegetation indices retained for modelling (reflectance on a 0–1 scale; G, R, RE, and NIR denote the green, red, red-edge, and near-infrared reflectance).

Index	Formula	Spectral Family	Reference
NDVI	$(NIR - R)/(NIR + R)$	Red/NIR normalised difference	[15]
GNDVI	$(NIR - G)/(NIR + G)$	Green-based normalised difference	[17]
NDRE	$(NIR - RE)/(NIR + RE)$	Red-edge normalised difference	[72]
OSAVI	$1.16\,(NIR - R)/(NIR + R + 0.16)$	Soil-adjusted	[18]
WDRVI	$(0.2\,NIR - R)/(0.2\,NIR + R)$	Reduced saturation at high biomass	[19]

For each index, plot, and acquisition date, the following statistics were calculated from the vector of valid pixel values: number of pixels, mean, median, standard deviation (population form), coefficient of variation, 10th, 25th, 75th, and 90th percentiles, minimum, and maximum. For each band, the mean, median, and standard deviation were calculated. Only the area-based mean of each index (denoted I_a_mean, for example, NDVI_a_mean) was used to construct the predictors, so that each plot is represented by one time series of area-based means per index—the other statistics were archived for diagnostic purposes.

The time axis was expressed as days after sowing (DAS), calculated from the sowing dates in Section 2.1. Before calculating the seasonal predictors, the time series were restricted

to a common window of 190–245 DAS, comprising 7 UAV acquisitions in each growing season (Table 4). This window was fixed in advance and was not optimised statistically. It corresponds approximately to the period from stem elongation to heading and the beginning of flowering of winter cereals, a period in which relationships between canopy indices and grain yield have been reported to be informative [3,8,16].

Table 4. Acquisition dates and days after sowing (DAS) of the UAV flights used in the analysis (window 190–245 DAS). Sowing: 17 October 2023 (2023/2024) and 26 September 2024 (2024/2025).

Season	Acquisition Date	DAS
2023/2024	25 April 2024	191
	1 May 2024	197
	9 May 2024	205
	16 May 2024	212
	22 May 2024	218
	6 June 2024	233
	13 June 2024	240
2024/2025	10 April 2025	196
	17 April 2025	203
	29 April 2025	215
	8 May 2025	224
	14 May 2025	230
	20 May 2025	236
	28 May 2025	244

In 2024, the selected acquisitions were conducted on 25 April (BBCH 31–37), 1 May (BBCH 32–37), 9 May (BBCH 37–39), 16 May (BBCH 39–45), 22 May (BBCH 45–49), 6 June (BBCH 45–49), and 13 June (BBCH 51–61), spanning 49 days. In 2025, the selected acquisitions were conducted on 10 April (BBCH 31–32), 17 April (BBCH 32–39), 29 April (BBCH 32–39), 8 May (BBCH 39–41), 14 May (BBCH 41–49), 20 May (BBCH 49–51), and 28 May (BBCH 51–61), spanning 48 days. Because the sowing dates differ, the same DAS window covers different calendar periods in the two seasons, and the window began about two weeks earlier in 2025 than in 2024. The same DAS window does not necessarily represent the same phenological stages in the two seasons.

For each plot, season, and index, four descriptors of the time series of area-based means within the window were calculated. Summary statistics of short NDVI time series have been used as inputs to tree-based yield models [12]. The temporal mean (I_t_mean) and median (I_t_median) describe the typical level of the index, and the temporal standard deviation (I_t_std; sample standard deviation with $n - 1$ degrees of freedom) describes its variability among dates. The area under the curve was calculated by trapezoidal integration over the DAS axis and divided by the time between the first and last acquisition in the window (49 days in 2024 and 48 days in 2025):

$$\mathrm{I_t_auc} = \frac{\left[\Sigma \mathrm{i}\,{}^{1}\!/_{2}(\mathrm{xi} + \mathrm{xi}+1)(\mathrm{ti}+1 - \mathrm{ti})\right]}{(\mathrm{tn} - \mathrm{t1})}, \quad (1)$$

where xi is the area-based mean of the index at acquisition i and ti is the corresponding DAS. This normalisation accounts for the uneven spacing of acquisitions and expresses the descriptor in the units of the index; it is, therefore, closely related to the temporal mean. Missing values were omitted. The combination of 5 indices and 4 descriptors resulted in 20 candidate predictors for 40 plot-year observations. The predictor table was merged with the yield data using the season and plot identifier as the key.

Pearson correlation coefficients and their p-values were calculated among the 20 predictors and grain yield, for the pooled data ($n = 40$) and separately for each season ($n = 20$), and

the matrices were displayed as lower-triangle heat maps. Pooled correlations are strongly influenced by the difference in yield and index level between the seasons, whereas the seasonal correlations describe the association within a season. The *p*-values were adjusted by the Benjamini–Hochberg (BH) procedure [73] across all 210 unique off-diagonal tests of each correlation matrix (20 predictors and yield), and both unadjusted *p*-values and adjusted q-values are reported—they are used as descriptive measures. The strong correlations among predictors were considered in the choice of the selection and interpretation methods.

3.3. Interpretation of the Management-Area Contrasts

Plots were coded as control (0) or manured (1) according to the treatment table. The assignment was unchanged between the seasons and followed a fixed pattern of plot identifiers: plots 1–2, 5–6, 9–10, 13–14, and 17–18 were controls, and plots 3–4, 7–8, 11–12, 15–16, and 19–20 were manured. Because the same plots are observed in both seasons and the seasons differ strongly in yield and in index level, the comparison was made separately for each season (10 plots per group). Mean, standard deviation, median, minimum, and maximum of grain yield were calculated for each season and group. Differences between the groups (manured minus control) in grain yield and in each of the 20 predictors were tested with Welch's *t* test [74] (two-sided; SciPy (V1.8.0.) [75]) and are reported with the 95% confidence interval of the difference and Hedges' g [76]. For the 20 predictors, the *p*-values were adjusted separately within each season (family of 20 tests) by the Holm [77] and the Benjamini–Hochberg [73] procedures. Because the assignment was systematic and not randomised, the differences between the groups were also summarised within each group of four consecutive plots (two control and two manured plots; five contrasts per season) and tested with a paired *t* test and the Wilcoxon signed-rank test [78] as a descriptive sensitivity analysis. The distributions were presented as box plots and violin plots for each season and group and as a treatment × season interaction plot of mean yield with standard error. Because the experiment is not replicated, the same plots appear in both seasons, and treatment is confounded with plot position, the comparison is descriptive, and the differences cannot be attributed to the manure itself. Manure treatment was not used as a predictor in any model.

3.4. Single-Index Models and Validation

To compare the five indices, regression models were built separately for each index, using its four descriptors (I_t_mean, I_t_median, I_t_std and I_t_auc) as predictors and no predictor selection. Nine model families were evaluated for each index: multiple linear regression (MLR), ridge regression [79], Elastic Net [37], SVR, PLSR, RF, GB, KNN, and XGBoost. The models, pipelines, and hyperparameter grids were those described in Section 3.5. MLR has no hyperparameters, the ridge penalty was tuned (alpha $\in$ {0.01, 0.1, 1, 10, 100}), and the number of PLSR components was limited to the number of predictors (four). The single-index models were evaluated in the same framework as the multi-index models (Sections 3.4 and 3.5): the outer partitions were generated with the same seeds, the same fold structure, and the same plot ordering, so that the test folds are identical and the predictions can be paired fold by fold, hyperparameters were tuned in the same inner grouped cross-validation, and LOYO validation, the reference predictions, and the performance metrics were the same. Paired plot–bootstrap differences in RMSE were calculated between models within an index, between indices within a model, and between each model and Ref_year_mean (pooled and within each season). For the linear models, standardised regression coefficients (change in yield per standard deviation of a predictor) were also recorded, and the importance measures of Section 3.5 were calculated for the tree-based models.

Because each plot is observed in both seasons, the observations were grouped by plot, and the two records of a plot were always kept together in training or test sets. Block-based validation of this kind is recommended for data with a hierarchical or temporal structure [54]. The evaluation used nested cross-validation to avoid the optimistic bias that arises when predictor selection or hyperparameter tuning uses the data on which performance is assessed [56–58]. In the outer loop, the 20 plot identifiers were divided into 5 folds, so that each outer training set contained 16 plots (32 records) and each test set 4 plots (8 records). The division was repeated with 5 random seeds (2026–2030), which gave 25 outer folds. Standardisation, predictor selection (where applied), model selection, and hyperparameter tuning were performed exclusively within the outer training set—the outer test set was used only to obtain predictions from the final model. Hyperparameters were tuned in an inner loop with three-fold grouped cross-validation (GroupKFold, groups = plots) using the negative mean squared error as a criterion.

To assess transfer between seasons, leave-one-year-out (LOYO) validation was also performed: the models, including predictor selection and tuning, were trained on one season and tested on the other, in both directions. Withholding a whole season from a two-season dataset is an extrapolation and tests the coverage of the training data as well as the algorithm. Tree-based models do not predict outside the range of the training responses. The random assignment of plots to folds does not remove the spatial autocorrelation among neighbouring plots, and the management areas form two spatial blocks.

In LOYO validation, the natural reference is the mean yield of the training season (Ref_year_mean; hereafter RYM), which is the benchmark prediction available without spectral information. Because the test season is absent from the training data, Ref_year_mean and Ref_global_mean coincide, and a model was considered to add information only if its LOYO RMSE was lower than that of the RYM. When predicting an unseen target season with a mean shift Δ relative to the training season, the coefficient of determination (R^2), calculated against the mean of the held-out season, approaches $-\left(\frac{\Delta}{\sigma}\right)^2$, where σ is the within-season standard deviation of the target season. Large inter-seasonal yield contrasts ($\Delta \gg \sigma$) thus inherently produce strongly negative LOYO R^2 values. Consequently, LOYO R^2 primarily reflects the magnitude of the seasonal domain shift rather than model accuracy, making relative RMSE against the RYM the more informative evaluation metric. LOYO results represent single estimates for each held-out season and have no repeat-based variability. For each model, the LOYO RMSE was also expressed relative to that of the RYM, (RMSE_model − RMSE_RYM)/RMSE_RYM, in %.

3.5. Predictor Selection, Regression Models, Performance Metrics, and Predictor Importance

A set of 8 predictors was selected from the 20 candidates within each training set. Two procedures were applied. The first was Elastic Net [37] stability selection [48], in which 300 bootstrap samples of the training plots (resampling plots, so that both records of a plot are drawn together) were generated. In each sample, the predictors and yield were standardised, and Elastic Net regularisation paths were computed for l1 ratios of 0.3, 0.5, 0.7, and 0.9 (60 values of the penalty over 3 orders of magnitude). The combination of l1 ratio and penalty at which the number of non-zero coefficients was closest to eight was taken as the sample's selected set. Predictors were ranked by their selection frequency across the bootstrap samples (ties were broken by the mean absolute standardised coefficient), and the eight most frequently selected predictors were retained. In contrast to the original stability selection, no probability threshold was used, and the number of selected predictors was fixed.

The second procedure was a genetic algorithm (GA) [49] that searched for subsets of eight predictors. The population contained 96 subsets and evolved for a maximum of 50

generations, with an elite of 8 subsets, tournament selection (size 3), 10 random immigrants per generation, and an adaptive mutation rate (0.15, increased to 0.45 when the diversity of the population was low). The search stopped after 10 generations without improvement. The fitness of a subset was the lowest inner cross-validated RMSE among three models (random forest with 100 trees, Elastic Net, and PLSR), estimated by grouped four-fold cross-validation repeated twice, plus 0.5 times the standard deviation of that RMSE. The final predictor set was the 8 predictors that occurred most frequently among the 50 best subsets (consensus selection). For both procedures, the frequency with which each predictor was selected across the 25 outer folds was recorded as a measure of selection stability.

Seven regression models were compared using the selected predictors (the single-index analysis in Section 3.4 additionally included MLR and ridge regression [79]): Elastic Net [37], support vector regression with a radial basis function kernel (SVR) [80], partial least squares regression (PLSR) [81], random forest (RF) [32], gradient boosting (GB) [82], k-nearest neighbour (KNN) [83], and XGBoost [84]. Each model was implemented as a pipeline with standardisation of the predictors, and the response was also standardised for Elastic Net and SVR. Hyperparameters were tuned by exhaustive grid search in the inner loop (Table 5), and the model was refitted on the whole outer training set with the best parameters. The models were implemented in scikit-learn [85] and XGBoost [84].

Table 5. Hyperparameter grids searched in the inner cross-validation.

Model	Tuned Hyperparameters	Reference
Elastic Net	alpha {0.001, 0.01, 0.1, 0.5, 1.0}; l1_ratio {0.1, 0.3, 0.5, 0.7, 0.9}	[37]
SVR (RBF)	C {0.1, 1, 10, 100}; gamma {scale, auto, 0.01, 0.1}	[80]
PLSR	n_components {1, 2, 3, 4, 5}	[81]
Random forest	n_estimators {50, 100, 200}; max_depth {none, 5, 10}; min_samples_split {2, 5}	[32]
Gradient boosting	n_estimators {50, 100, 150}; learning_rate {0.01, 0.05, 0.1}; max_depth {3, 5}	[82]
KNN	n_neighbors {3, 5, 7, 9}; weights {uniform, distance}	[83]
XGBoost	n_estimators {50, 100, 150}; learning_rate {0.01, 0.05, 0.1}; max_depth {3, 5}	[84]

Predictions from the outer test folds were pooled for each repeat, and the coefficient of determination (R^2), root mean square error (RMSE), normalised RMSE (NRMSE, RMSE divided by the mean observed yield, in %), mean absolute error (MAE), and bias (mean of predicted minus observed values) were calculated for all observations and separately for each season. R^2 was calculated as $1 - SSE/SST$ from the pooled predictions and can be negative when a model performs worse than the mean of the observations. Metrics are reported as the mean and standard deviation over the five repeats; fold-wise metrics were also recorded. Uncertainty was quantified by a plot-level bootstrap [86]: plots were resampled with replacement ($B = 2000$), both records of a plot being drawn together, and 95% percentile intervals were calculated for the metrics and for paired differences in RMSE between models. The paired differences were calculated between models, between predictor sets, and between each model and the season-mean reference, and a difference was considered distinguishable from zero when its interval excluded zero. The bootstrap reflects the variability of the evaluation sample and does not include the variability from retraining the models. Within-season R^2 and NRMSE were interpreted together with the reference predictions.

The importance of the predictors was determined for the random forest model in each outer fold by four methods: permutation importance [32,52], calculated as the increase in the mean squared error on the outer test fold after permuting a predictor (10 repetitions), impurity-based (Gini, variance reduction) importance, the mean absolute SHAP value [50] computed with the tree explainer on the test fold, and the rank from recursive feature elimination (RFE) [51]. For the single-index models, permutation importance was calculated for

all models, standardised coefficients for the linear models, and SHAP values for gradient boosting and XGBoost. Values were summarised as the mean and standard deviation over the folds in which the predictor was selected. Because the predictors are strongly correlated (Section 3.2), importance is shared among correlated predictors and the values of single predictors can be unstable [52,53]. Importance was, therefore, interpreted at the level of groups of predictors and by agreement among the methods, not as a causal effect.

3.6. Software and Reproducibility

Analyses were implemented in Python 3.11.9 using, among others, rasterio 1.4.4, GeoPandas 1.1.2, Shapely 2.1.2, pyproj 3.7.2, OpenCV 4.13.0.92, NumPy 2.3.5, pandas 2.3.3, SciPy [75], scikit-learn 1.8.0 [85], XGBoost 3.2.0 [84], and SHAP [50]. Random seeds were fixed (Section 3.4), and the intermediate tables and the data underlying each figure were saved as separate files. During manuscript preparation and revision, Claude (Anthropic, Claude Sonnet 5.5) and ChatGPT (OpenAI, GPT-6.1 Sol) were used to assist with English-language editing, improve textual clarity, and refine sentence structure.

4. Results

4.1. Temporal Vegetation Index Metrics and Their Associations with Yield

The 190–245 days-after-sowing (DAS) window included 7 UAV acquisitions in each growing season (Table 4). The analyses comprised 40 plot–year observations in the pooled dataset and 20 observations in each season. Mean yield was 1352.8 ± 560.8 kg ha^{-1} in 2024 (range: 611–2444 kg ha^{-1}) and 3794.4 ± 496.9 kg ha^{-1} in 2025 (range: 2722–4500 kg ha^{-1}); the observed yield ranges did not overlap. The correlations between the 20 temporal metrics and yield are given in Table 6.

Table 6. Pearson correlations between temporal vegetation index metrics and yield in the 190–245 DAS window. Each cell reports Pearson's r, the unadjusted two-sided *p*-value, and the Benjamini–Hochberg-adjusted q-value (r; *p*; q). n = 40 for pooled correlations and N = 20 for each season. The BH correction was applied separately within each dataset subset to all 210 unique off-diagonal tests of the corresponding correlation matrix, comprising 190 inter-predictor pairs among the 20 metrics and 20 predictor–yield pairs.

Temporal Index Metric	Pooled (N = 40), r; *p*; q	2024 (N = 20), r; *p*; q	2025 (N = 20), r; *p*; q
NDVI_t_mean	0.874; 1.72×10^{-13}; 3.32×10^{-13}	0.144; 0.545; 0.566	0.062; 0.797; 0.885
NDVI_t_median	0.860; 1.16×10^{-12}; 2.09×10^{-12}	0.501; 0.0245; 0.0306	0.116; 0.627; 0.711
NDVI_t_std	−0.574; 1.07×10^{-4}; 1.51×10^{-4}	0.929; 3.43×10^{-9}; 1.06×10^{-8}	−0.381; 0.0975; 0.136
NDVI_t_auc	0.868; 4.29×10^{-13}; 8.04×10^{-13}	0.087; 0.715; 0.729	0.043; 0.856; 0.923
GNDVI_t_mean	0.851; 3.37×10^{-12}; 5.95×10^{-12}	0.061; 0.798; 0.806	−0.011; 0.962; 0.986
GNDVI_t_median	0.853; 2.88×10^{-12}; 5.13×10^{-12}	0.239; 0.310; 0.334	0.085; 0.722; 0.811
GNDVI_t_std	−0.524; 5.18×10^{-4}; 6.36×10^{-4}	0.853; 1.79×10^{-6}; 3.79×10^{-6}	−0.538; 0.0145; 0.0215
GNDVI_t_auc	0.845; 6.78×10^{-12}; 1.18×10^{-11}	0.004; 0.988; 0.988	−0.024; 0.921; 0.962
NDRE_t_mean	0.861; 1.06×10^{-12}; 1.94×10^{-12}	0.338; 0.145; 0.165	0.030; 0.900; 0.945
NDRE_t_median	0.868; 4.37×10^{-13}; 8.12×10^{-13}	0.487; 0.0294; 0.0361	0.165; 0.486; 0.583
NDRE_t_std	−0.097; 0.552; 0.595	0.725; 2.97×10^{-4}; 5.03×10^{-4}	−0.392; 0.0874; 0.123
NDRE_t_auc	0.848; 4.75×10^{-12}; 8.31×10^{-12}	0.258; 0.272; 0.296	−0.006; 0.980; 0.986
OSAVI_t_mean	0.829; 4.01×10^{-11}; 6.89×10^{-11}	−0.027; 0.911; 0.915	−0.401; 0.0798; 0.113
OSAVI_t_median	0.858; 1.44×10^{-12}; 2.58×10^{-12}	0.115; 0.628; 0.646	0.039; 0.869; 0.927
OSAVI_t_std	0.401; 0.0104; 0.0116	0.497; 0.0257; 0.0319	−0.714; 4.04×10^{-4}; 8.17×10^{-4}
OSAVI_t_auc	0.824; 6.78×10^{-11}; 1.16×10^{-10}	−0.066; 0.783; 0.794	−0.401; 0.0795; 0.113
WDRVI_t_mean	0.881; 6.28×10^{-14}; 1.23×10^{-13}	0.243; 0.301; 0.326	0.060; 0.803; 0.887
WDRVI_t_median	0.869; 3.65×10^{-13}; 6.97×10^{-13}	0.480; 0.0323; 0.0395	0.135; 0.569; 0.657
WDRVI_t_std	−0.454; 0.0033; 0.0038	0.855; 1.56×10^{-6}; 3.37×10^{-6}	−0.515; 0.0202; 0.0293
WDRVI_t_auc	0.877; 1.17×10^{-13}; 2.28×10^{-13}	0.181; 0.444; 0.468	0.042; 0.860; 0.923

In the pooled analysis, the temporal means, medians, and AUC-derived metrics of all five vegetation indices were positively correlated with yield ($r = 0.824$–0.881; BH-adjusted $q \leq 1.16 \times 10^{-10}$). The standard deviation metrics showed a different pattern: yield was negatively correlated with the temporal standard deviations of NDVI ($r = -0.574$, $p = 1.07 \times 10^{-4}$, $q = 1.51 \times 10^{-4}$), GNDVI ($r = -0.524$, $p = 5.18 \times 10^{-4}$, $q = 6.36 \times 10^{-4}$), and WDRVI ($r = -0.454$, $p = 0.0033$, $q = 0.0038$), positively correlated with that of OSAVI ($r = 0.401$, $p = 0.0104$, $q = 0.0116$), and not significantly correlated with that of NDRE ($r = -0.097$, $p = 0.552$, $q = 0.595$).

The season-specific associations were less uniform. In 2024, the temporal standard deviations of all five indices were positively correlated with yield ($r = 0.497$–0.929; $q = 1.06 \times 10^{-8}$–0.0319), as were the temporal medians of NDVI ($r = 0.501$, $p = 0.0245$, $q = 0.0306$), NDRE ($r = 0.487$, $p = 0.0294$, $q = 0.0361$), and WDRVI ($r = 0.480$, $p = 0.0323$, $q = 0.0395$); the other 12 metrics did not reach $q < 0.05$. In 2025, only the temporal standard deviations of GNDVI ($r = -0.538$, $p = 0.0145$, $q = 0.0215$), OSAVI ($r = -0.714$, $p = 4.04 \times 10^{-4}$, $q = 8.17 \times 10^{-4}$), and WDRVI ($r = -0.515$, $p = 0.0202$, $q = 0.0293$) were significantly (negatively) correlated with yield at $q < 0.05$; the remaining 17 metrics had $q > 0.05$. The pattern of yield associations, therefore, differed between the pooled and the season-specific analyses (Figure 3). Because the pooled analysis combines two seasons that differ strongly in yield and index level, its correlations should not be interpreted as within-season relationships in both years.

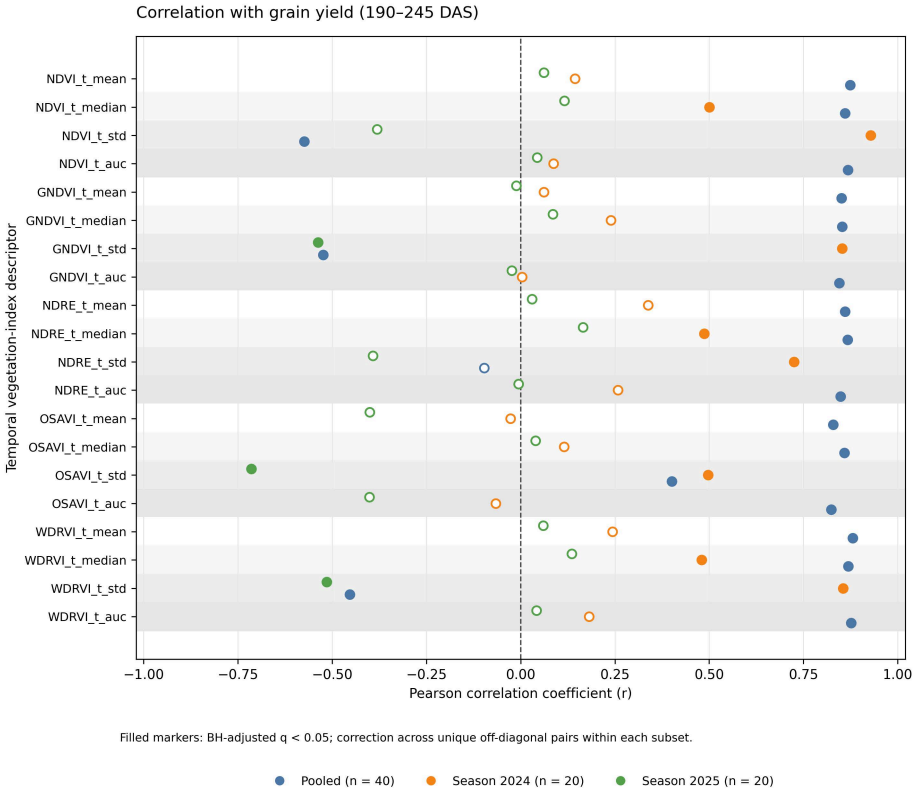

Figure 3. Pearson correlations between the 20 temporal descriptors and grain yield for the pooled data ($n = 40$) and for each season ($n = 20$). Filled markers indicate Benjamini–Hochberg-adjusted $q < 0.05$. Pooled correlations are dominated by the difference between the seasons.

4.2. Single-Index Models for Plot-Level Yield Prediction

Five predictor sets were evaluated (NDVI, GNDVI, NDRE, OSAVI, and WDRVI), each comprising the temporal mean, median, standard deviation, and AUC-derived metric of one index. Nine regression algorithms were compared using nested cross-validation with

five seeds and five plot-grouped outer folds per seed (25 outer splits), and hyperparameters were tuned in three-fold grouped inner cross-validation. Both observations from a given plot were kept in the same outer fold. No missing values were present in the yield or predictor columns.

Across the five index-specific predictor sets, the model with the lowest mean pooled RMSE was K-nearest neighbour (KNN) for WDRVI (RMSE = 432.7 ± 9.9 kg ha^{-1}; R^2 = 0.893 ± 0.005), followed by KNN for NDVI (447.8 ± 11.9 kg ha^{-1}; R^2 = 0.886 ± 0.006), support vector regression (SVR) for GNDVI (544.6 ± 21.2 kg ha^{-1}; R^2 = 0.831 ± 0.013), random forest for OSAVI (612.2 ± 45.7 kg ha^{-1}; R^2 = 0.786 ± 0.032), and KNN for NDRE (629.2 ± 19.2 kg ha^{-1}; R^2 = 0.774 ± 0.014). These are descriptive rankings by mean pooled RMSE among the evaluated models. The pooled plot–bootstrap intervals and the season-specific performance are shown in Table 7. The pooled out-of-fold RMSE of all nine algorithms for each vegetation index is shown in Supplementary Material Figure S1.

Table 7. Performance of the lowest-mean-RMSE model within each vegetation index predictor set. Values are means across the five cross-validation seeds, and ± denotes the across-seed standard deviation. Bracketed intervals are 95% percentile intervals from 2000 plot-level bootstrap replicates. Season-specific values are means across seeds. Model rankings are descriptive and do not constitute formal tests of differences between models.

Index Set	Model	Pooled R^2, Mean ± SD [95% CI]	Pooled RMSE (kg ha^{-1}), Mean ± SD [95% CI]	Pooled NRMSE (%)	2024 R^2; RMSE (kg ha^{-1})	2025 R^2; RMSE (kg ha^{-1})
WDRVI	KNN	0.893 ± 0.005 [0.836, 0.937]	432.7 ± 9.9 [332.6, 531.5]	16.8	0.702; 295.0	−0.217; 533.3
NDVI	KNN	0.886 ± 0.006 [0.827, 0.931]	447.8 ± 11.9 [345.0, 549.4]	17.4	0.709; 289.9	−0.340; 560.4
GNDVI	SVR	0.831 ± 0.013 [0.690, 0.928]	544.6 ± 21.2 [359.3, 742.0]	21.2	0.540; 368.7	−0.945; 675.2
OSAVI	Random forest	0.786 ± 0.032 [0.618, 0.897]	612.2 ± 45.7 [420.8, 825.6]	23.8	−0.036; 553.7	−0.890; 664.9
NDRE	KNN	0.774 ± 0.014 [0.637, 0.875]	629.2 ± 19.2 [473.8, 787.7]	24.5	0.000; 545.2	−1.105; 701.1

The pooled results did not consistently represent performance within each season. For the lowest-RMSE model of each index set, the 2024 R^2 values ranged from −0.036 to 0.709, whereas all corresponding 2025 R^2 values were negative (−1.105 to −0.217). In 2024, the best models of the NDVI, WDRVI, and GNDVI sets had RMSE values of 289.9–368.7 kg ha^{-1}, compared with 571.2 kg ha^{-1} for the year-specific mean reference, whereas those of the OSAVI and NDRE sets (553.7 and 545.2 kg ha^{-1}) were close to the reference. In 2025, none of the five models was better than the reference (RMSE = 533.3–701.1 versus 519.8 kg ha^{-1}). For example, the WDRVI–KNN model had R^2 = 0.702 and RMSE = 295.0 kg ha^{-1} in 2024, but R^2 = −0.217 and RMSE = 533.3 kg ha^{-1} in 2025. The year-specific mean reference had a pooled R^2 of 0.830 and RMSE of 546.2 kg ha^{-1}, and the global-mean reference a pooled R^2 of −0.013 and RMSE of 1334.3 kg ha^{-1} (Table 8). The paired plot–bootstrap estimate of the pooled RMSE difference between WDRVI–KNN and the year-specific mean reference was −113.5 kg ha^{-1} (model minus reference; 95% CI: −224.9 to 7.2 kg ha^{-1}), and the interval included zero (Figure 4). The pooled results, therefore, do not establish a clear improvement over the year-specific mean reference, and the large, non-overlapping yield ranges of the two seasons mean that pooled performance is not evidence of consistently strong within-season prediction.

Table 8. Pooled and within-season performance of the lowest-mean-RMSE model for each leakage-controlled selector (ENet, Elastic Net stability selection; GA, consensus genetic algorithm), alongside the reference models. Values are means over the five cross-validation seeds, and ± denotes the across-seed standard deviation. Bracketed intervals are 95% percentile intervals from 2000 plot-level bootstrap replicates. Season-specific metrics are means across seeds. The table reports the lowest-mean-pooled-RMSE algorithm for each selector and is not a formal comparison between selectors.

Selector/Reference	Model	Pooled R^2, Mean ± SD [95% CI]	Pooled RMSE (kg ha^{-1}), Mean ± SD [95% CI]	Pooled NRMSE (%)	2024 R^2; RMSE (kg ha^{-1})	2025 R^2; RMSE (kg ha^{-1})
ENet	KNN	0.819 ± 0.022 [0.715, 0.897]	563.1 ± 34.6 [429.7, 698.5]	21.9	0.084; 522.4	−0.545; 597.9
GA	KNN	0.881 ± 0.008 [0.819, 0.928]	457.5 ± 15.0 [364.7, 552.4]	17.8	0.509; 380.6	−0.161; 520.2
Year-specific mean reference	Ref_year_mean	0.830 ± 0.006 [0.753, 0.898]	546.2 ± 9.8 [425.9, 650.3]	21.2	−0.092; 571.2	−0.153; 519.8
Global mean reference	Ref_global_mean	−0.013 ± 0.009 [−0.045, −0.007]	1334.3 ± 5.9 [1231.8, 1433.0]	51.8	−5.061; 1345.6	−6.461; 1323.0

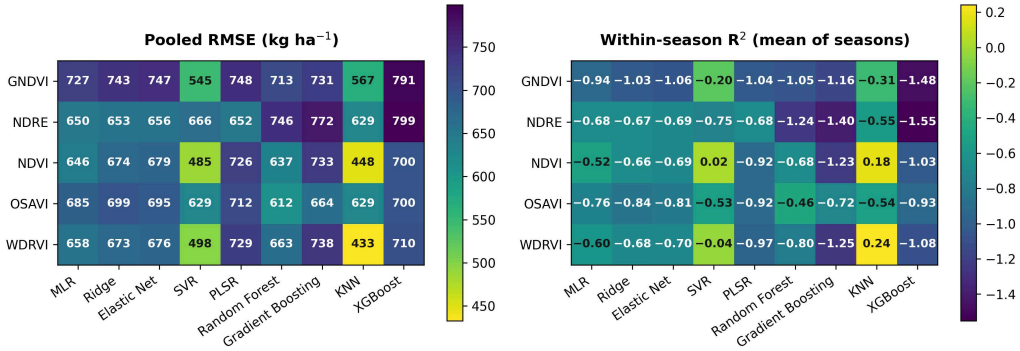

Figure 4. Performance of single-index models across five vegetation index predictor sets and nine regression algorithms. The left panel shows pooled out-of-fold RMSE (kg ha^{-1}), and the right panel shows the arithmetic mean of the season-specific R^2 values for 2024 and 2025. Readers should note that averaging season-specific R^2 values can mask strong performance asymmetries between individual years; disaggregated, season-specific R^2 and RMSE values are provided in Tables 7 and 8. Values are means across five repeats of plot-grouped cross-validation.

Leave-one-year-out (LOYO) validation showed poor transfer between seasons.

When a season was held out, the training-season mean predicted the yield with an RMSE of 2502 kg ha^{-1} (2024 held out) and 2489 kg ha^{-1} (2025 held out); 95% and 96% of the corresponding mean squared error, respectively, is due to the difference in yield level between the seasons. Across the nine algorithms and five index sets, the LOYO RMSE ranged from 2068 to 2795 kg ha^{-1} when 2024 was held out and from 1992 to 4776 kg ha^{-1} when 2025 was held out, that is, from about 17% below to 12% above and from 20% below to 92% above the RMSE of the training-season mean. The corresponding R^2 values were between −25.2 and −13.3 (2024 held out) and between −96.2 and −15.9 (2025 held out). They are strongly negative because the models did not reproduce the level of the held-out season. The favourable pooled cross-validation metrics, therefore, did not translate into reliable prediction when an entire season was excluded from training; inference about transferability is limited to the two seasons available here.

Permutation-based importance was exploratory and varied across index sets. Among the lowest-pooled-RMSE models, the temporal standard deviation had the largest mean permutation importance for NDVI and WDRVI, whereas the temporal median ranked highest for GNDVI, NDRE, and OSAVI. The comparatively large fold-to-fold variation of these importance estimates warrants cautious interpretation.

4.3. Leakage-Controlled Predictor Selection and Multi-Index Models

The 20 candidate predictors represented the temporal mean, median, standard deviation, and AUC-derived metric of each of five vegetation indices. Elastic Net bootstrap stability selection (ENet) and a consensus genetic algorithm (GA) each selected eight predictors within the outer training data. Seven regression algorithms were evaluated for each selected set, using five seeds and five plot-grouped outer folds per seed (25 outer splits); hyperparameter tuning was performed within the outer training data, and the two observations of each plot remained together in each split.

Across the 25 outer folds, the GA stopped by the patience criterion in all runs (median generations_run, 25; range, 10–39), and no run reached the 50-generation limit. The convergence curves of the individual outer folds are shown in Supplementary Material Figure S2. The best penalised inner-CV objective (mean RMSE + 0.5 × the standard deviation of RMSE) ranged from 325.8 to 667.6 kg ha^{-1}, with 968–3424 candidate evaluations per fold; the fold-specific predictor sets varied, as no predictor was selected in all 25 folds (Table 9).

Table 9. Most frequently selected predictors across the 25 outer folds (ENet, Elastic Net stability selection; GA, consensus genetic algorithm). Selection frequency is the proportion of outer folds in which a predictor was included in the eight-predictor set; it is not a *p*-value or a bootstrap selection probability.

Selector	Predictor	Selected Folds	Frequency
ENet	NDVI_t_mean	25/25	100%
ENet	NDVI_t_median	25/25	100%
ENet	NDRE_t_median	25/25	100%
ENet	WDRVI_t_mean	25/25	100%
ENet	WDRVI_t_median	25/25	100%
ENet	WDRVI_t_auc	25/25	100%
ENet	NDVI_t_auc	20/25	80%
ENet	OSAVI_t_median	14/25	56%
GA	OSAVI_t_mean	19/25	76%
GA	NDVI_t_mean	18/25	72%
GA	WDRVI_t_std	18/25	72%
GA	OSAVI_t_auc	17/25	68%
GA	NDVI_t_std	14/25	56%
GA	GNDVI_t_auc	14/25	56%
GA	WDRVI_t_auc	14/25	56%
GA	WDRVI_t_mean	13/25	52%

The two selection procedures produced different patterns of predictor stability (Table 9). Elastic Net selected 6 predictors in all 25 outer folds: NDVI_t_mean, NDVI_t_median, NDRE_t_median, WDRVI_t_mean, WDRVI_t_median, and WDRVI_t_auc. NDVI_t_auc was selected in 20 folds (80%) and OSAVI_t_median in 14 folds (56%). With the genetic algorithm, the most frequently selected predictors were OSAVI_t_mean (19/25 folds, 76%), NDVI_t_mean and WDRVI_t_std (18/25 each, 72%), and OSAVI_t_auc (17/25, 68%); no single predictor occurred in every outer fold. The selection frequencies of all 20 candidate predictors are given in Supplementary Material Table S1 and Figure S3.

KNN had the lowest mean pooled RMSE for both leakage-controlled selection procedures. With Elastic Net selection, KNN achieved a pooled RMSE of 563.1 ± 34.6 kg ha^{-1}, an R^2 of 0.819 ± 0.022, and an NRMSE of 21.9%; with genetic algorithm selection, the corresponding values were 457.5 ± 15.0 kg ha^{-1}, 0.881 ± 0.008, and 17.8%. The genetic algorithm result was numerically better in pooled cross-validation, but no paired comparison between the two selection procedures was calculated, so the difference is descriptive. Within each selector, the paired plot–bootstrap interval for the RMSE difference between

KNN and the next-lowest-RMSE model (SVR) included zero (ENet: −97.1 kg ha^{-1}, 95% CI −194.3 to 9.7; GA: −75.2 kg ha^{-1}, 95% CI −181.2 to 36.1), so the results do not establish that KNN outperformed the runner-up with either selector. Pooled and within-season metrics of all seven algorithms are given in Supplementary Material Table S2a, and the paired plot-bootstrap RMSE differences between all pairs of algorithms in Table S2b.

Season-specific performance was less consistent than pooled performance. For the KNN model with Elastic Net selection, R^2 was 0.084 in 2024 and −0.545 in 2025; with genetic algorithm selection, it was 0.509 and −0.161, respectively. The year-specific mean reference had a pooled R^2 of 0.830 (RMSE = 546.2 kg ha^{-1}), whereas the global-mean reference had a pooled R^2 of −0.013 (RMSE = 1334.3 kg ha^{-1}); the year-specific mean reference itself had negative within-season R^2 in both years (−0.092 in 2024 and −0.153 in 2025). Within 2024, the KNN model with genetic algorithm selection had a lower RMSE than the reference (380.6 versus 571.2 kg ha^{-1}), whereas within 2025 its RMSE was equal to that of the reference (520.2 versus 519.8 kg ha^{-1}) and that of the Elastic-Net-selected model was higher (597.9 kg ha^{-1}). Given the large, non-overlapping seasonal yield ranges, the high pooled R^2 values should not be interpreted as evidence of equally strong prediction within each season—pooled performance is influenced by the separation between seasons. Observed and predicted yields of the WDRVI-KNN and GA-KNN models form two distinct seasonal clusters (Supplementary Material Figure S4), and the RMSE of individual outer test folds, each containing eight records, varied considerably among folds (Supplementary Material Figure S5).

Leave-one-year-out validation showed poor transfer between seasons (Figure 5a–c, Table 10). For the KNN model with Elastic Net selection, the LOYO RMSE was 2435 kg ha^{-1} when 2024 was held out and 2458 kg ha^{-1} when 2025 was held out, compared with 2502 and 2489 kg ha^{-1} for the training-season mean, i.e., reductions of only 2.7% and 1.3%; with genetic algorithm selection, the RMSE was 2491 and 2333 kg ha^{-1} (reductions of 0.4% and 6.3%). The corresponding R^2 values were −18.8 and −24.8 (Elastic Net) and −19.8 and −22.2 (genetic algorithm). Across all 7 algorithms and both selectors, 9 out of 28 model–season combinations had a lower RMSE than the training-season mean. Because only two seasons were available, this assessment does not support generalisation to other years.

Table 10. LOYO performance of the lowest-pooled-RMSE model of each predictor set and selector compared with the training-season mean (RYM).

Predictor Set/Selector	Model	2024 Held Out: RMSE (kg ha^{-1})	Difference to RYM (%)	R^2	2025 Held Out: RMSE (kg ha^{-1})	Difference to RYM (%)	R^2
Training-season mean (RYM)	–	2502	–	−20.0	2489	–	−25.4
WDRVI	KNN	2288	−8.6	−16.5	2410	−3.2	−23.8
NDVI	KNN	2288	−8.6	−16.5	2524	+1.4	−26.2
GNDVI	SVR	2476	−1.1	−19.5	2024	−18.7	−16.5
OSAVI	RF	2788	+11.4	−25.0	1995	−19.8	−16.0
NDRE	KNN	2408	−3.8	−18.4	2443	−1.9	−24.4
Elastic Net selection	KNN	2435	−2.7	−18.8	2458	−1.3	−24.8
Genetic algorithm selection	KNN	2491	−0.4	−19.8	2333	−6.3	−22.2

Random forest feature-importance summaries based on permutation, Gini, SHAP, and RFE methods were also produced (Supplementary Material Figure S6). These method-specific results are exploratory and are reported separately from the primary performance comparison.

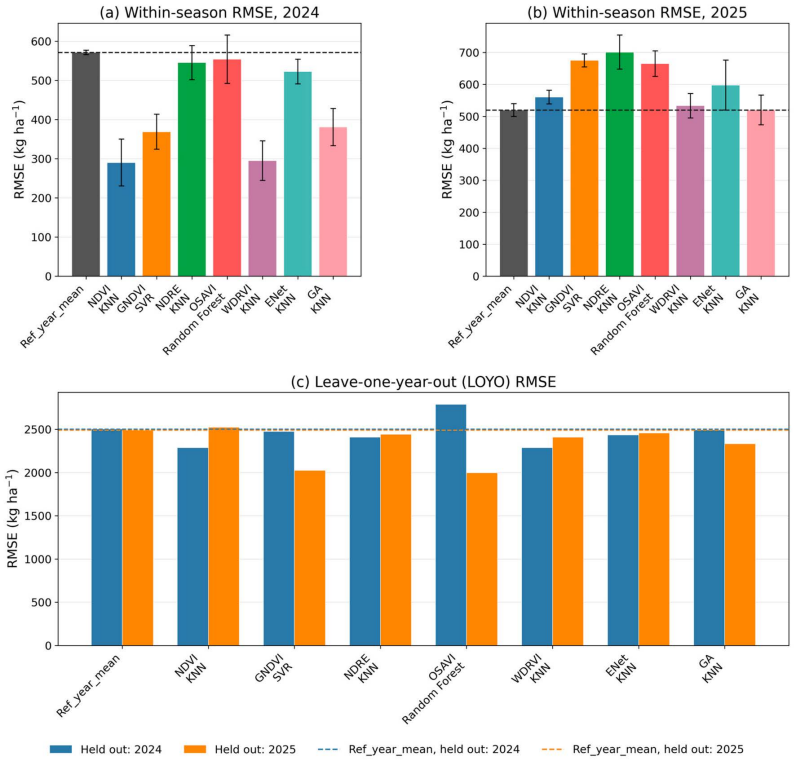

Figure 5. Within-season and leave-one-year-out (LOYO) performance of the training-season mean-yield reference (Ref_year_mean = RYM) and cross-validation-selected models. (**a**,**b**) RMSE in 2024 and 2025 for RYM, the model with the lowest pooled RMSE for each single-index predictor set (NDVI, GNDVI, NDRE, OSAVI, and WDRVI), and the lowest-pooled-RMSE model for each selection procedure (ENet, Elastic Net stability selection; GA, consensus genetic algorithm). Error bars in (**a**,**b**) indicate the standard deviation across five repeats of plot-grouped cross-validation. (**c**) LOYO RMSE for the same model configurations when 2024 or 2025 was held out; bars distinguish the held-out season. LOYO values are single estimates for each held-out season and have no repeat-based standard deviation. Dashed lines indicate the corresponding RYM baseline (RMSE = 2502 and 2489 kg ha^{-1} for 2024 and 2025 held out, respectively). The numerical values are given in Table 10.

4.4. Plot-Level Associations with Manure Treatment

The analysis included 40 plot–year observations from 20 plots: 20 observations in each of 2024 and 2025, with 10 control and 10 manured plots per season. All 40 yield records and the 20 vegetation index metrics matched the treatment table by season and plot identifier. Treatment assignment was unchanged for each plot across the two seasons and was systematic rather than randomised: within each group of four consecutive plot identifiers, the first two plots were controls and the other two were manured.

The direction of the observed yield difference varied between the seasons (Table 11). In 2024, mean yield was 1516.6 ± 613.4 kg ha^{-1} in control plots and 1188.9 ± 477.4 kg ha^{-1} in manured plots, a difference of −327.7 kg ha^{-1} (95% CI −846.4 to 191.0; Welch $p = 0.200$; Hedges' g = −0.57). In 2025, the respective means were 3644.5 ± 603.2 and 3944.3 ± 328.0 kg ha^{-1}, a difference of +299.8 kg ha^{-1} (95% CI −166.2 to 765.8; Welch $p = 0.189$; Hedges' g = 0.59). Both confidence intervals included zero. The opposite directions of the differences and the large difference in yield between the seasons make an unstratified pooled comparison unsuitable as a summary of the treatment association.

Differences between the treatment groups were also observed in the vegetation-index metrics (Supplementary Material Table S3a,b; effect sizes in Figure S7). In 2024, none of the 20 metrics remained significant after Holm correction, although 17 had unadjusted Welch *p*-values below 0.05; the same 17 also met the less conservative Benjamini–Hochberg

criterion ($q < 0.05$). In all metrics with $q < 0.05$, the manured group had lower values than the control group. In 2025, NDRE_t_std was the only metric that remained significant after either correction: its mean was 0.0420 ± 0.0106 in control plots and 0.0242 ± 0.0042 in manured plots (difference −0.0178; 95% CI −0.0257 to −0.0099; Holm-adjusted $p = 0.0074$; Hedges' g = −2.11). This is a treatment-group association in one season and does not by itself establish a corresponding yield response or a mechanism linking the index to yield.

Table 11. Seasonal plot-yield summaries by treatment group. Differences are manured minus control. Welch tests are unadjusted group comparisons and are not causal treatment tests.

Season	Control, *n*; Mean ± SD (kg ha^{-1})	Manured, *n*; Mean ± SD (kg ha^{-1})	Difference (95% CI), kg ha^{-1}	Welch *p*	Hedges' g
2024	10; 1516.6 ± 613.4	10; 1188.9 ± 477.4	−327.7 (−846.4, 191.0)	0.200	−0.57
2025	10; 3644.5 ± 603.2	10; 3944.3 ± 328.0	+299.8 (−166.2, 765.8)	0.189	+0.59

The contrasts within groups of four consecutive plots describe differences between the two treatment groups at neighbouring positions, but treatment was not randomised within the groups and was tied to the same plot positions in both seasons (Figure 6). Treatment and position are, therefore, confounded, and these contrasts cannot separate a manure effect from positional differences. The paired *t* test and the Wilcoxon test also gave different results in 2024 ($p = 0.017$ and $p = 0.125$, respectively; five groups); in 2025, the corresponding *p*-values were 0.245 and 0.313 (Supplementary Material Table S4 and Figure S8). These tests are, therefore, reported as descriptive sensitivity analyses. All treatment contrasts are interpreted as associations, not as causal effects; the repetition of the experiment over two seasons does not remove the confounding of treatment and position.

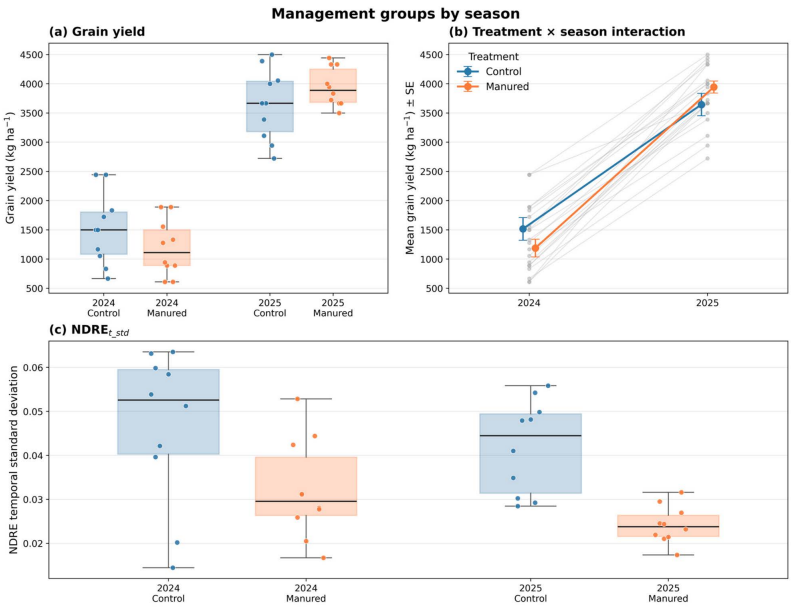

Figure 6. Management groups. (**a**) Grain yield of control and manured plots in 2024 and 2025 (box plots with individual plot points), (**b**) treatment x season interaction showing mean yield ± standard error, with individual plot trajectories connected across seasons, and (**c**) temporal standard deviation of NDRE (NDRE_t_std). NDRE_t_std was the only core metric exhibiting a significant season-specific Welch comparison after Holm correction (in 2025; descriptive comparison).

5. Discussion

5.1. Pooled and Season-Specific Signals

Every model and predictor set had negative R^2 in leave-one-year-out validation. A more informative comparison is with the training-season mean, which predicted the held-out season with an RMSE of about 2.5 t ha^{-1}. The selected multi-index models reduced this error by at most 6.3%, and the best single-index models by about 17–20%, whereas the least successful models exceeded it by up to 92%. The difference of 2.44 t ha^{-1} between the seasonal means accounts for 95–96% of the squared error of the training-season mean, so the models did not reproduce the shift in yield level between seasons, and the descriptors could not replace information on the season itself. With two seasons, this validation is an extrapolation in both directions; the result, therefore, shows that two seasons are insufficient to calibrate a transferable model, and it is not a statement about the algorithms themselves.

The mathematical relationship formulated in Section 3.4 directly explains the extreme LOYO performance. The observed mean yield difference between the two seasons ($\Delta = 2.44$ t ha^{-1}) was four to five times larger than the within-season standard deviations ($\sigma = 0.50$–0.56 t ha^{-1}). Substituting these empirical values into the theoretical expression yields $-\left(\frac{2.44}{0.56}\right)^2 \approx -19$, which closely matches the baseline R^2 of the training-season mean (-20.0 and -25.4 for 2024 and 2025 held out, respectively; Table 10). This confirms that 95–96% of the squared error in LOYO predictions is attributable strictly to the inter-seasonal-level shift.

The main result of this study is the contrast between pooled and season-specific analyses. Several temporal vegetation index metrics were strongly associated with yield across the 40 plot–year observations (r up to 0.88), but the yields of 2024 and 2025 occupied non-overlapping ranges, and the within-season associations were less consistent (Table 6). In 2024, the temporal standard deviations and some temporal medians were positively related to yield, whereas in 2025 only three standard deviation metrics reached $q < 0.05$, and their sign was negative. Pooled correlations and pooled prediction scores, therefore, partly reflect the separation between the seasons rather than a stable index–yield relationship. The year-specific mean reference illustrates this: it used no spectral information, yet it reached a pooled R^2 of 0.830, while its within-season R^2 was negative (Table 8).

Within 2024, the best single-index models and the genetic-algorithm-selected KNN model clearly improved on this reference (RMSE of about 290–380 versus 571 kg ha^{-1}; Tables 7 and 8), which indicates that canopy variation among plots carried yield-relevant information in that season. Within 2025, no model improved on the reference. A possible explanation is that yield was limited by different processes in the two seasons and that the spectral signal captured these processes to different degrees; the present data cannot test this. Growth-stage dependence of the index–yield relationship is a common finding in cereals [3,8,16], and weather-related NDVI dynamics have been described for winter triticale [25], which is consistent with the sensitivity of triticale spectral response to water supply [24]. The two seasons also differed in the timing of precipitation (Tables 1 and 2), and the same days-after-sowing window covered different calendar periods and probably different developmental stages in the two seasons. With two seasons, however, weather effects cannot be separated from other season effects.

Studies on other crops likewise report that the best stage and sensor combination varies. In winter wheat, prediction depended on growth stage and improved when multispectral data were combined with RGB imagery and canopy height [46], cover-crop biomass estimates depended on season and species, and optical–radar fusion improved them [87], and multispectral–thermal and multi-sensor fusion improved estimates of water and nitrogen status [88,89]. These outcomes are not yield benchmarks, but together with the earlier

finding that crop height complements spectral information [5], they suggest that structural or complementary sensor data may add information beyond the spectral indices used here.

5.2. Model Performance and Validation

Nested, plot-grouped cross-validation gave pooled R^2 values of up to 0.89 for some single- and multi-index models. KNN had the lowest mean pooled RMSE for three of the five single-index sets and for both selectors, but the paired intervals did not establish superiority over the runner-up, and the two selectors were not formally compared. Because the seasons form two well-separated groups, a local-average method, such as KNN, may partly reproduce the seasonal level, and the pooled ranking of algorithms should not be over-interpreted. Values of this size are comparable in magnitude to R^2 of up to 0.90 reported for wheat breeding lines [45], but crop, yield range, predictor set, and validation design differ, and a vegetation index regression in winter wheat achieved an RMSE of 972 kg ha^{-1} without improvement from machine learning models [44]. In our data, the regularised linear and tree-based models did not outperform KNN, and deep learning models were not tested because the 40 observations are too few for them.

With two seasons, this validation is an extrapolation in both directions, and the strongly negative values mainly reflect the difference of about 2.4 t ha^{-1} between the seasonal means; the result, therefore, shows that two seasons are insufficient to calibrate a transferable model, and it is not a statement about the algorithms themselves. Limited transferability across seasons, sites, and cultivars is a recurrent finding: temporal validation gave lower accuracy than spatial validation in district-level wheat prediction [14], five-fold accuracies fell in across-year prediction in wheat breeding [59], and domain shifts reduced the transferability of rapeseed models [62] and of soybean models between continents [65]. Cross-year prediction was improved by combining AUC of vegetation index time series with domain adaptation in cotton [90], by adding climate, evapotranspiration, and soil data in county-level wheat prediction (held-out-year R^2 of 0.71 and 0.63) [91], and by local fine-tuning in field-scale wheat studies [63,64,92]. Compact sets of vegetation indices transferred between potato and sweet potato [93], transfer learning in agriculture has been reviewed [94], and a lightweight alternative to deep architectures was proposed for domain transfer under limited data [43]. These studies differ in crop, scale, sensors, and validation, so scores cannot be compared directly, but they agree that temporal and independent validation should be reported and that local calibration with several seasons is needed. A linear model with a season term and a within-season leave-plot-out validation, which would complement the present design, were not evaluated.

5.3. Indices, Predictors, and Their Interpretation

The NDVI and WDRVI sets gave the lowest pooled errors among the five single-index sets, and the NDRE set the highest (Table 7). The red-edge index, therefore, did not outperform the broadband indices in this dataset, in contrast to reports of its value at advanced canopy stages in wheat [7,8,20]. Possible reasons are the use of plot means and a fixed window that includes late stages, but these were not tested, and the differences among indices were not formally compared.

The Elastic Net selection repeatedly chose the same NDVI and WDRVI mean and median descriptors and one NDRE median, whereas the genetic algorithm gave a less stable set, and no predictor occurred in every fold (Table 9). Selection frequency describes stability within the training folds and not causal importance. Importance rankings from random forest are further affected by the strong correlation among the descriptors [51,52], and a UAV-yield review noted that the best models and growth periods vary with the selected features [9]. The temporal standard deviation had the strongest association with yield in

2024 (r up to 0.93) but the opposite sign for several indices in 2025; it describes the variation among flights, depends on the position of the window, and should not be treated as a general predictor. The normalised AUC is extremely closely related to the temporal mean, that is, for all indices r exceeds 0.99, showing that the four descriptors of an index overlap substantially in information.

Weather and soil variables were not modelled. They are common yield prediction features [41] and have been combined with NDVI time series [12,13], but with two seasons any season-level covariate is collinear with the season itself and cannot be estimated. Their use requires more seasons.

5.4. Comparison of the Management Areas

The direction of the yield difference between the management groups reversed between seasons (−328 and +300 kg ha^{-1}), and both confidence intervals included zero (Table 11). In 2024, all vegetation index metrics with $q < 0.05$ had lower values in the manured group, and in 2025 only the temporal standard deviation of NDRE remained significant after correction. UAV indices can register differences in nutrient supply [11,30,31], and a previous study at this site examined the effect of manure on yield and quality [29]. In the present design, however, the assignment followed a fixed pattern of plots, was not randomised, and was not replicated, so treatment is confounded with plot position. The differences are, therefore, descriptive and cannot be attributed to the manure.

5.5. Limitations and Outlook

The evidence is limited to 1 site, 1 cultivar, 20 fixed plots, and 2 seasons. The window was defined in days after sowing and not by observed phenology, random assignment of plots to folds does not remove spatial autocorrelation, the bootstrap intervals do not include the variability of retraining, and the radiometric calibration, the grain moisture, and the treatment layout need to be documented. Further research should (1) add site-years, so that leave-one-year-out validation and season effects can be estimated, ideally with four to five seasons or more; (2) record phenology and use stage-specific windows and indices around heading, flowering, and grain filling; (3) add structural predictors, such as canopy height [5], texture, or within-plot heterogeneity, from the pixel statistics that were archived but not used; (4) include weather and water-stress indicators once enough seasons are available; (5) test local calibration and transfer learning [63,64,92]; (6) use a replicated, randomised treatment design, so that effects of manure can be separated from location.

An additional limitation is that the nested plot-grouped cross-validation did not reserve spatially contiguous test blocks. The control and poultry manure treatments occupied separate adjacent 2 × 5 areas; therefore, holding out either entire area would also hold out one treatment regime and confound management with spatial position. Although a balanced column-wise spatial holdout could be defined from the 4 × 5 plot layout, it was not included in the present analysis. Consequently, the better within-season predictive performance observed in 2024 has not been tested for transfer to a contiguous spatial block. Future work should evaluate this using prespecified spatial blocks that retain both management regimes in the training and test data.

6. Conclusions

This study asked whether seasonal descriptors of multi-temporal UAV multispectral imagery can explain plot-level grain yield of winter triticale grown in an organic system, and how far this relationship holds when the plot and season structure of the data is respected in validation. Twenty descriptors (temporal mean, median, standard deviation, and normalised area under the curve of five vegetation indices) were derived from seven

flights per season for twenty fixed plots in two seasons. Single-index and multi-index models were evaluated by nested, plot-grouped cross-validation and by leave-one-year-out validation against season-mean reference predictions.

The descriptors were strongly associated with yield in the pooled data, and single-index models based on WDRVI and NDVI descriptors reached pooled R^2 of about 0.89 and RMSE of about 430–450 kg ha^{-1}. Most of this performance reflects the difference between the two seasons, because a season-mean reference alone gave a pooled R^2 of 0.83. The informative result lies within the seasons: in 2024, the best models reduced the error relative to this reference by roughly one-third to one-half, which shows that canopy variation among plots carried yield-relevant information, whereas in 2025 no model improved on it. In leave-one-year-out validation, the lowest model RMSE (about 2.0 t ha^{-1}) was only 20% below that of the training-season mean (2.5 t ha^{-1}), and R^2 was negative for all models, which shows that two seasons are insufficient to calibrate a transferable model.

The descriptors can, therefore, support the description and ranking of plot conditions within a season, but the data do not support yield forecasting for an unseen season. The comparison of the management areas showed no consistent yield difference and, because treatment is confounded with plot position, it cannot be attributed to the manure.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/rs18223993/s1, Table S1: Complete predictor selection frequencies; Table S2a: Pooled and within-season performance of the leakage-controlled multi-index models; Table S2b: Paired plot-bootstrap differences in pooled RMSE between models; Table S3a: Seasonal treatment-group summaries for the 20 vegetation-index metrics (2024); Table S3b: Seasonal treatment-group summaries for the 20 vegetation-index metrics (2025); Table S4: Contrasts within groups of four consecutive plots; Figure S1: Pooled out-of-fold RMSE of the single-index models; Figure S2: Convergence of the genetic algorithm; Figure S3: Selection frequency of the 20 candidate predictors; Figure S4: Observed versus predicted grain yield for the WDRVI–KNN and GA–KNN models; Figure S5: Distribution of RMSE over the outer test folds; Figure S6: Random forest predictor importance; Figure S7: Effect sizes of the 20 temporal descriptors between the management groups; Figure S8: Contrasts within groups of four consecutive plots.

Author Contributions: Conceptualisation, G.D., B.G., J.C., E.K.K., B.S., Z.K., S.S., Z.M., I.A. and Z.T.; methodology, G.D., B.G., J.C., E.K.K., B.S. and Z.T.; software, G.D., B.G. and Z.T.; validation, G.D., B.G. and Z.T.; formal analysis, G.D.; investigation, G.D., B.G. and Z.T.; resources, G.D., B.G. and Z.T.; data curation, G.D., B.G. and Z.T.; writing—original draft preparation, G.D., B.G., J.C., E.K.K., B.S., Z.K. and Z.T.; writing—review and editing, G.D., B.G., J.C., E.K.K., B.S., Z.K., S.S., Z.M., I.A. and Z.T.; visualisation, G.D. and Z.T.; supervision, G.D., B.G., J.C., E.K.K., B.S., Z.K., S.S., Z.M., I.A. and Z.T.; project administration, B.G., Z.K., S.S., Z.M. and I.A.; funding acquisition, Z.K., S.S., Z.M. and I.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Research, Development and Innovation Office of Hungary (NKFIH), grant number 2022-1.2.2-TÉT-IPARI-UZ-2022-00023.

Data Availability Statement: The Python processing code and scripts developed for multi-temporal UAV index extraction and yield modelling are publicly available at: https://github.com/gdrephub/triticale-yield-uav-modeling_DG (accessed on 6 November 2026), and archived in Zenodo at: https://doi.org/10.5281/zenodo.24017104 (accessed on 6 November 2026). HungaroMet Nonprofit Ltd. 150 years of climate data series Nyíregyháza: https://www.met.hu/eghajlat/magyarorszag_eghajlata/150_eves_eghajlati_adatsorok/Nyiregyhaza/adatok/napi_adatok/index.php (accessed on 8 July 2026). Further inquiries can be directed to the corresponding author.

Acknowledgments: Supported by the University of Debrecen Program for Scientific Publication. We acknowledge the Digital Government Development and Project Management Ltd. for awarding us access to the Komondor HPC facility based in Hungary. During the preparation and revision of this manuscript, the authors used Claude (Anthropic, Claude Sonnet 5.5) and ChatGPT (OpenAI, GPT-6.1 Sol) for English-language editing and improving textual clarity. The authors reviewed and edited the

AI-assisted text and take full responsibility for the content of this publication. We express our thanks to Krisztina Dezső for her help in organising and verifying a substantial part of the literature sources.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area under the curve
BBCH	Biologische Bundesanstalt, Bundessortenamt und Chemische Industrie (growth-stage scale)
BH	Benjamini–Hochberg
CV	Cross-validation
DAS	Days after sowing
GA	Genetic algorithm
GB	Gradient boosting
GNDVI	Green normalised difference vegetation index
GNSS	Global Navigation Satellite System
KNN	k-nearest neighbour
LAI	Leaf area index
LOYO	Leave-one-year-out
MAE	Mean absolute error
NDRE	Normalised difference red-edge index
NDVI	Normalised difference vegetation index
NRMSE	Normalised root mean square error
OSAVI	Optimised soil-adjusted vegetation index
PLSR	Partial least squares regression
RF	Random forest
RFE	Recursive feature elimination
RGB	Red–green–blue
RMSE	Root mean square error
RTK	Real-time kinematic
SHAP	Shapley Additive exPlanations
SVR	Support vector regression
UAV	Unmanned aerial vehicle
UTM	Universal Transverse Mercator
WDRVI	Wide dynamic range vegetation index

References

1. Weiss, M.; Jacob, F.; Duveiller, G. Remote sensing for agricultural applications: A meta-review. *Remote Sens. Environ.* **2020**, *236*, 111402. [CrossRef]
2. Sishodia, R.P.; Ray, R.L.; Singh, S.K. Applications of Remote Sensing in Precision Agriculture: A Review. *Remote Sens.* **2020**, *12*, 3136. [CrossRef]
3. Panek, E.; Gozdowski, D.; Stępień, M.; Samborski, S.; Ruciński, D.; Buszke, B. Within-Field Relationships between Satellite-Derived Vegetation Indices, Grain Yield and Spike Number of Winter Wheat and Triticale. *Agronomy* **2020**, *10*, 1842. [CrossRef]
4. Chiu, M.S.; Wang, J. Local Field-Scale Winter Wheat Yield Prediction Using VENμS Satellite Imagery and Machine Learning Techniques. *Remote Sens.* **2024**, *16*, 3132. [CrossRef]
5. Yue, J.; Yang, G.; Li, C.; Li, Z.; Wang, Y.; Feng, H.; Xu, B. Estimation of Winter Wheat Above-Ground Biomass Using Unmanned Aerial Vehicle-Based Snapshot Hyperspectral Sensor and Crop Height Improved Models. *Remote Sens.* **2017**, *9*, 708. [CrossRef]
6. Fu, Z.; Jiang, J.; Gao, Y.; Krienke, B.; Wang, M.; Zhong, K.; Cao, Q.; Tian, Y.; Zhu, Y.; Cao, W.; et al. Wheat Growth Monitoring and Yield Estimation based on Multi-Rotor Unmanned Aerial Vehicle. *Remote Sens.* **2020**, *12*, 508. [CrossRef]
7. Zhang, D.; Qi, H.; Guo, X.; Sun, H.; Min, J.; Li, S.; Hou, L.; Lv, L. Integration of UAV Multispectral Remote Sensing and Random Forest for Full-Growth Stage Monitoring of Wheat Dynamics. *Agriculture* **2025**, *15*, 353. [CrossRef]
8. Camenzind, M.P.; Yu, K. Multi temporal multispectral UAV remote sensing allows for yield assessment across European wheat varieties already before flowering. *Front. Plant Sci.* **2024**, *14*, 1214931. [CrossRef] [PubMed]
9. Yuan, J.; Zhang, Y.; Zheng, Z.; Yao, W.; Wang, W.; Guo, L. Grain crop yield prediction using machine learning based on UAV remote sensing: A systematic literature review. *Drones* **2024**, *8*, 559. [CrossRef]

10. Kostić, M.M.; Aćin, V.; Mirosavljević, M.; Stamenković, Z.; Kešelj, K.; Ljubičić, N.; Scarfone, A.; Stanković, N.; Kovačević, D.B. Comparative Assessment of Remote and Proximal NDVI Sensing for Predicting Wheat Agronomic Traits. *Drones* **2025**, *9*, 641. [CrossRef]
11. Benincasa, P.; Antognelli, S.; Brunetti, L.; Fabbri, C.A.; Natale, A.; Sartoretti, V.; Modeo, G.; Guiducci, M.; Tei, F.; Vizzari, M. Reliability of NDVI derived by high resolution satellite and UAV compared to in-field methods for the evaluation of early crop n status and grain yield in wheat. *Exp. Agric.* **2017**, *54*, 604–622. [CrossRef]
12. Heremans, S.; Dong, Q.; Zhang, B.; Bydekerke, L.; Van Orshoven, J. Potential of ensemble tree methods for early-season prediction of winter wheat yield from short time series of remotely sensed normalized difference vegetation index and in situ meteorological data. *J. Appl. Remote Sens.* **2015**, *9*, 097095. [CrossRef]
13. Croci, M.; Ragazzi, M.; Grassi, A.; Impollonia, G.; Amaducci, S. Assessing the temporal transferability of machine learning models for predicting processing pea yield and quality using Sentinel-2 and ERA5-land data. *Smart Agric. Technol.* **2025**, *12*, 101207. [CrossRef]
14. Alpysbay, M.; Nurakynov, S.; Gapparov, A.; Kaldybayev, A. Crop-Masked Vegetation Indices and TerraClimate for District-Level Wheat Yield Prediction in Kazakhstan: SAVI Advantage, Climate Dominance, and Temporal Transferability Limits. *Agronomy* **2026**, *16*, 1264. [CrossRef]
15. Rouse, J.W., Jr.; Haas, R.H.; Schell, J.A.; Deering, D.W. Monitoring Vegetation Systems in the Great Plains with ERTS. In *Third NASA Earth Resources Technology Satellite Symposium*; U.S. Government Printing Office (GPO): Washington, DC, USA, 1974; Volume 1, pp. 309–317.
16. Hassan, M.A.; Yang, M.; Rasheed, A.; Yang, G.; Reynolds, M.; Xia, X.; Xiao, Y.; He, Z. A Rapid Monitoring of NDVI across the Wheat Growth Cycle for Grain Yield Prediction Using a Multi-Spectral UAV Platform. *Plant Sci.* **2019**, *282*, 95–103. [CrossRef] [PubMed]
17. Gitelson, A.A.; Kaufman, Y.J.; Merzlyak, M.N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. *Remote Sens. Environ.* **1996**, *58*, 289–298. [CrossRef]
18. Rondeaux, G.; Steven, M.; Baret, F. Optimization of soil-adjusted vegetation indices. *Remote Sens. Environ.* **1996**, *55*, 95–107. [CrossRef]
19. Gitelson, A.A. Wide Dynamic Range Vegetation Index for remote quantification of biophysical characteristics of vegetation. *J. Plant Physiol.* **2004**, *161*, 165–173. [CrossRef] [PubMed]
20. Ganeva, D.; Roumenina, E.; Dragov, R.; Taneva, K.; Nedyalkova, S.; Bozhanova, V.; Dimitrov, P. Distinguishing High- and Low-Yielding Durum Wheat Genotypes Using UAV Spectral and Textural Data. *Remote Sens.* **2026**, *18*, 2664. [CrossRef]
21. Benaragama, D.; Hussain, M.; Senetza, B.; Shirtliffe, S.; Willenborg, C. UAV-Based Multispectral Phenotyping and Machine-Learning Modeling Reveals Early Canopy Traits as Strong Predictors of Yield and Weed Competitiveness in Oat (*Avena sativa* L.). *Remote Sens.* **2026**, *18*, 1211. [CrossRef]
22. Aasen, H.; Honkavaara, E.; Lucieer, A.; Zarco-Tejada, P.J. Quantitative remote sensing at ultra-high resolution with UAV spectroscopy: A review of sensor technology, measurement procedures, and data correction workflows. *Remote Sens.* **2018**, *10*, 1091. [CrossRef]
23. Özer, E.; Keleş, R. Studies on Index Determination of Some Triticale Genotypes (×*Triticosecale* Witt.) Normalized Difference Vegetation Under Konya Rainfed Conditions. *J. Glob. Innov. Agric. Soc. Sci.* **2019**, *7*, 167–170. [CrossRef]
24. Munjonji, L.; Ayisi, K.; Vandewalle, B.; Dhau, I.; Boeckx, P.; Haesaert, G. Yield performance, carbon assimilation and spectral response of triticale to water stress. *Exp. Agric.* **2016**, *53*, 100–117. [CrossRef]
25. Bevz, S.Y.; Kozina, A.M.; Voblikova, T.V. Normalized Difference Vegetation Index (NDVI) in Assessment of Grain Crop State. *BIO Web Conf.* **2024**, *108*, 09003. [CrossRef]
26. Atanasov, A.I.; Atanasov, A.Z. Environmental Sensitivity of NDVI in Triticale (×*Triticosecale* Wittmack) Under Standardized Field Acquisition Protocols. *Agriculture* **2026**, *16*, 770. [CrossRef]
27. Noack, P.O. Estimating Triticale Dry Matter Yield in Parcel Plot Trials from Aerial and Ground Based Spectral Measurements. *IFAC-PapersOnLine* **2016**, *49*, 404–408. [CrossRef]
28. Atanasov, A.I.; Stoyanov, H.P.; Atanasov, A.Z.; Evstatiev, B.I. Analysis of Vegetation Dynamics and Phenotypic Differentiation in Five Triticale (×*Triticosecale* Wittm.) Varieties Using UAV-Based Multispectral Indices. *Agronomy* **2026**, *16*, 303. [CrossRef]
29. Kosztyuné Krajnyák, E.; Szabó, B.; Kovács, Z.; Makszim Györgyné Nagy, T.; Györgyiné Kovács, A.; Csabai, J.; Abdurahmanov, I.; Abdurakhmonov, Z. The effect of fermented poultry manure on the yield and quality parameters of triticale (×*Triticosecale* Wittmack) cultivated organic way on acidic sandy soil. *E3S Web Conf.* **2024**, *590*, 01009. [CrossRef]
30. Guo, J.; Zeng, X.; Ma, Q.; Yuan, Y.; Zhang, N.; Lin, Z.; Yin, P.; Yang, H.; Liu, X.; Zhang, F. UAV-Based Yield Prediction Based on LAI Estimation in Winter Wheat (*Triticum aestivum* L.) Under Different Nitrogen Fertilizer Types and Rates. *Plants* **2025**, *14*, 1986. [CrossRef] [PubMed]

31. Yang, M.; Hassan, M.A.; Xu, K.; Zheng, C.; Rasheed, A.; Zhang, Y.; Jin, X.; Xia, X.; Xiao, Y.; He, Z. Assessment of Water and Nitrogen Use Efficiencies Through UAV-Based Multispectral Phenotyping in Winter Wheat. *Front. Plant Sci.* **2020**, *11*, 927. [CrossRef] [PubMed]
32. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
33. Jeong, J.H.; Resop, J.P.; Mueller, N.D.; Fleisher, D.H.; Yun, K.; Butler, E.E.; Timlin, D.J.; Shim, K.-M.; Gerber, J.S.; Reddy, V.R.; et al. Random Forests for Global and Regional Crop Yield Predictions. *PLoS ONE* **2016**, *11*, e0156571. [CrossRef] [PubMed]
34. Crane-Droesch, A. Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environ. Res. Lett.* **2018**, *13*, 114003. [CrossRef]
35. Burdett, H.; Wellen, C. Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. *Precis. Agric.* **2022**, *23*, 1553–1574. [CrossRef]
36. Shokirov, S.; Abdurahmanov, I.; Mamatkulov, Z.; Abdurahmanov, Z.; Zarifboev, D.; Oymatov, R.; Kovács, Z.; Csabai, J.; Shiping, Y.; Khakberdiev, O. Leveraging Vegetation Indices and Random Forest for Soil Nutrient Monitoring in Winter Wheat. *J. Indian Soc. Remote Sens.* **2025**, *53*, 3711–3721. [CrossRef]
37. Zou, H.; Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2005**, *67*, 301–320. [CrossRef]
38. Shammi, S.A.; Huang, Y.; Feng, G.; Tewolde, H.; Zhang, X.; Jenkins, J.; Shankle, M. Application of UAV Multispectral Imaging to Monitor Soybean Growth with Yield Prediction through Machine Learning. *Agronomy* **2024**, *14*, 672. [CrossRef]
39. Jha, S.K.; Brinkhoff, J.; Robson, A.J.; Dunn, B.W. Multi-Source Time-Series Integration for Progressive In-Season Prediction of Rice Yield, Aboveground Biomass, and Harvest Index. *Remote Sens.* **2026**, *18*, 1785. [CrossRef]
40. Jabed, M.A.; Murad, M.A.A. Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability. *Heliyon* **2024**, *10*, e40836. [CrossRef] [PubMed]
41. van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [CrossRef]
42. Muruganantham, P.; Wibowo, S.; Grandhi, S.; Samrat, N.H.; Islam, N. A Systematic Literature Review on Crop Yield Prediction with Deep Learning and Remote Sensing. *Remote Sens.* **2022**, *14*, 1990. [CrossRef]
43. Pankajakshan, P.; Padmasanan, A.; Sundar, S. Deep architectures fail to generalize: A lightweight alternative for agricultural domain transfer in hyperspectral images. *Sensors* **2026**, *26*, 174. [CrossRef] [PubMed]
44. Zhou, X.; Kono, Y.; Win, A.; Matsui, T.; Tanaka, T.S.T. Predicting Within-Field Variability in Grain Yield and Protein Content of Winter Wheat Using UAV-Based Multispectral Imagery and Machine Learning Approaches. *Plant Prod. Sci.* **2021**, *24*, 137–151. [CrossRef]
45. Ali, N.; Mohammed, A.; Bais, A.; Berraies, S.; Ruan, Y.; Cuthbert, R.D.; Sangha, J.S. Field Scale Precision: Predicting Grain Yield of Diverse Wheat Breeding Lines Using High-Throughput UAV Multispectral Imaging. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2024**, *17*, 11419–11433. [CrossRef]
46. Chu, C.; Hu, Y.; Wu, W.; Li, Y.; Li, T.; Zhao, J.; Li, J.; Li, X.; Wang, Z. Enhanced wheat yield estimation with UAV multi-source remote sensing and optimized ensemble learning. *Smart Agric. Technol.* **2025**, *12*, 101486. [CrossRef]
47. Fei, S.; Hassan, M.A.; Xiao, Y.; Su, X.; Chen, Z.; Cheng, Q.; Duan, F.; Chen, R.; Ma, Y. UAV-based multi-sensor data fusion and machine learning algorithm for yield prediction in wheat. *Precis. Agric.* **2023**, *24*, 187–212. [CrossRef] [PubMed]
48. Meinshausen, N.; Bühlmann, P. Stability selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **2010**, *72*, 417–473. [CrossRef]
49. Leardi, R.; Boggia, R.; Terrile, M. Genetic algorithms as a strategy for feature selection. *J. Chemom.* **1992**, *6*, 267–281. [CrossRef]
50. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; Curran Associates Inc.: Red Hook, NY, USA, 2017; pp. 4768–4777.
51. Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* **2002**, *46*, 389–422. [CrossRef]
52. Strobl, C.; Boulesteix, A.-L.; Kneib, T.; Augustin, T.; Zeileis, A. Conditional variable importance for random forests. *BMC Bioinform.* **2008**, *9*, 307. [CrossRef] [PubMed]
53. Molnar, C.; König, G.; Bischl, B.; Casalicchio, G. Model-agnostic feature importance and effects with dependent features: A conditional subgroup approach. *Data Min. Knowl. Discov.* **2024**, *38*, 2903–2941. [CrossRef]
54. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
55. Habibi, L.N.; Matsui, T.; Tanaka, T.S.T. Critical evaluation of the effects of a cross-validation strategy and machine learning optimization on the prediction accuracy and transferability of a soybean yield prediction model using UAV-based remote sensing. *J. Agric. Food Res.* **2024**, *16*, 101096. [CrossRef]

56. Ambroise, C.; McLachlan, G.J. Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 6562–6566. [CrossRef] [PubMed]
57. Varma, S.; Simon, R. Bias in error estimation when using cross-validation for model selection. *BMC Bioinform.* **2006**, *7*, 91. [CrossRef] [PubMed]
58. Cawley, G.C.; Talbot, N.L.C. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.
59. Juliana, P.; Montesinos-López, O.A.; Crossa, J.; Mondal, S.; González Pérez, L.; Poland, J.; Huerta-Espino, J.; Crespo-Herrera, L.; Govindan, V.; Dreisigacker, S.; et al. Integrating Genomic-Enabled Prediction and High-Throughput Phenotyping in Breeding for Climate-Resilient Bread Wheat. *Theor. Appl. Genet.* **2019**, *132*, 177–194. [CrossRef] [PubMed]
60. Zheng, H.; Ji, W.; Wang, W.; Lu, J.; Li, D.; Guo, C.; Yao, X.; Tian, Y.; Cao, W.; Zhu, Y.; et al. Transferability of Models for Predicting Rice Grain Yield from Unmanned Aerial Vehicle (UAV) Multispectral Imagery across Years, Cultivars and Sensors. *Drones* **2022**, *6*, 423. [CrossRef]
61. Franch, B.; Moletto-Lobos, I.; Tarín-Mestre, J.; Mascolo, L.; Vermote, E.; Kalecinski, N.; Becker-Reshef, I.; San-Bautista, A.; Rubio, C.; San Francisco, S.; et al. The Yield Strikes Back: Enhancing the Transferability of Field Scale Wheat and Barley Yield Models by Leveraging Sentinel-1/2. *Int. J. Appl. Earth Obs. Geoinf.* **2026**, *146*, 105140. [CrossRef]
62. Zhang, Y.; Niu, Y.; Cui, Z.; Chai, X.; Xu, L. Cross-Year Rapeseed Yield Prediction for Harvesting Management Using UAV-Based Imagery. *Remote Sens.* **2025**, *17*, 2010. [CrossRef]
63. Kaushal, S.; Billah, M.M.; Singh, M.; Koupal, D.; Sitaula, P.R.; Halder, J.; Sidhu, J.; Maimaitijiang, M.; Sehgal, S.K. UAV-based deep transfer learning to improve grain yield prediction in winter wheat across temporal and spatial variability. *Plant Phenome J.* **2026**, *9*, e70085. [CrossRef]
64. Wang, J.; Liu, Y.; Xu, Y.; Wang, W.; Cai, D.; Yan, H. Cross-Field Transfer of UAV-Based Winter Wheat Yield Prediction: Model Transferability and Sample Efficiency. *Remote Sens.* **2026**, *18*, 3484. [CrossRef]
65. Skobalski, J.; Sagan, V.; Alifu, H.; Al Akkad, O.; Lopes, F.A.; Grignola, F. Bridging the gap between crop breeding and GeoAI: Soybean yield prediction from multispectral UAV images with transfer learning. *ISPRS J. Photogramm. Remote Sens.* **2024**, *210*, 260–281. [CrossRef]
66. Zhang, P.; Lu, B.; Shang, J.; Tan, C.; Xu, Q.; Shi, L.; Jin, S.; Wang, X.; Jiang, Y.; Yang, Y.; et al. TKSF-KAN: Transformer-enhanced oat yield modeling and transferability across major oat-producing regions in China using UAV multisource data. *ISPRS J. Photogramm. Remote Sens.* **2025**, *224*, 166–186. [CrossRef]
67. Csorba, P.; Ádám, S.; Bartos-Elekes, Z.; Bata, T.; Bede-Fazekas, Á.; Czúcz, B.; Csima, P.; Csüllög, G.; Fodor, N.; Frisnyák, S.; et al. Landscapes. In *National Atlas of Hungary—Natural Environment*, 1st ed.; Kocsis, K., Gercsák, G., Horváth, G., Keresztesi, Z., Nemerkényi, Z., Eds.; MTA CSFK Geographical Institute: Budapest, Hungary, 2018; pp. 112–129.
68. Gáti, B. Comparison of vegetation indices obtained by drone and satellite. *E3S Web Conf.* **2024**, *590*, 01008. [CrossRef]
69. HungaroMet Nonprofit Ltd. 150 Years of Climate Data Series, Nyíregyháza (Daily Data). 2026. Available online: https://www.met.hu/eghajlat/magyarorszag_eghajlata/150_eves_eghajlati_adatsorok/Nyiregyhaza/adatok/napi_adatok/index.php (accessed on 8 July 2026).
70. Kosztyuné Krajnyák, E.; Henzsel, I.; Shokirov, S.; Abdurahmanov, I.; Akbarov, M.; Mamatkulov, Z.; Kuziev, U.; Khasanov, K.; Teshaev, N.; Khakberdiev, O.; et al. Analysis of Crop Elements of Triticale Treated with Poultry Manure. In *Fenntartható Tájgazdálkodási Tudományos Műhely Konferenciája 2026: Innovatív megoldások a XXI. század mezőgazdaságában*; Nyíregyházi Egyetem Műszaki És Agrártudományi Intézet: Nyíregyháza, Hungary, 2026; pp. 44–52. [CrossRef]
71. Maes, W.H. Practical Guidelines for Performing UAV Mapping Flights with Snapshot Sensors. *Remote Sens.* **2025**, *17*, 606. [CrossRef]
72. Gitelson, A.A.; Merzlyak, M.N. Quantitative estimation of chlorophyll-a using reflectance spectra: Experiments with autumn chestnut and maple leaves. *J. Photochem. Photobiol. B* **1994**, *22*, 247–252. [CrossRef]
73. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1995**, *57*, 289–300. [CrossRef]
74. Welch, B.L. The generalization of 'Student's' problem when several different population variances are involved. *Biometrika* **1947**, *34*, 28–35. [CrossRef] [PubMed]
75. Virtanen, P.; Gommers, R.; Oliphant, T.E.; Haberland, M.; Reddy, T.; Cournapeau, D.; Burovski, E.; Peterson, P.; Weckesser, W.; Bright, J.; et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nat. Methods* **2020**, *17*, 261–272. [CrossRef] [PubMed]
76. Hedges, L.V. Distribution theory for Glass's estimator of effect size and related estimators. *J. Educ. Stat.* **1981**, *6*, 107–128. [CrossRef]
77. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
78. Wilcoxon, F. Individual comparisons by ranking methods. *Biom. Bull.* **1945**, *1*, 80–83. [CrossRef]

79. Hoerl, A.E.; Kennard, R.W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **1970**, *12*, 55–67. [CrossRef]
80. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
81. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130. [CrossRef]
82. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
83. Cover, T.; Hart, P. Nearest neighbor pattern classification. *IEEE Trans. Inf. Theory* **1967**, *13*, 21–27. [CrossRef]
84. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016*; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
85. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
86. Efron, B. Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **1979**, *7*, 1–26. [CrossRef]
87. Jennewein, J.S.; Lamb, B.T.; Hively, W.D.; Thieme, A.; Thapa, R.; Goldsmith, A.; Mirsky, S.B. Integration of Satellite-Based Optical and Synthetic Aperture Radar Imagery to Estimate Winter Cover Crop Performance in Cereal Grasses. *Remote Sens.* **2022**, *14*, 2077. [CrossRef]
88. Wang, J.; Lou, Y.; Wang, W.; Liu, S.; Zhang, H.; Hui, X.; Wang, Y.; Yan, H.; Maes, W.H. A Robust Model for Diagnosing Water Stress of Winter Wheat by Combining UAV Multispectral and Thermal Remote Sensing. *Agric. Water Manag.* **2024**, *291*, 108616. [CrossRef]
89. Wang, J.; Wang, W.; Liu, S.; Hui, X.; Zhang, H.; Yan, H.; Maes, W.H. UAV-Based Multiple Sensors for Enhanced Data Fusion and Nitrogen Monitoring in Winter Wheat across Growth Seasons. *Remote Sens.* **2025**, *17*, 498. [CrossRef]
90. Sun, Y.; Lin, Z.; Guo, W. Integrating Growth Curves and Data Transformation for Robust Cross-Year Cotton Yield Prediction Using UAV Imagery and Machine Learning. *Smart Agric. Technol.* **2026**, *14*, 102238. [CrossRef]
91. Khosravani Shariati, S.A.; Abbasi, A. Machine Learning-Based Winter Wheat Yield Prediction Using Multisource Data. *Agric. Water Manag.* **2025**, *322*, 109951. [CrossRef]
92. Bueechi, E.; Reuß, F.; Pikl, M.; Homolova, L.; Lukas, V.; Trnka, M.; Dorigo, W. Making Optimal Use of Limited Field-Scale Data for Crop Yield Forecasting Using Transfer Learning and Sentinel-1 and 2 Data. *Smart Agric. Technol.* **2025**, *12*, 101567. [CrossRef]
93. Yadav, S.A.; Huang, Y.; Zhu, K.Q.; Haque, R.; Young, W.; Harvey, L.; Hall, M.; Zhang, X.; Wijewardane, N.K.; Qin, R.; et al. Deep Transfer Learning for UAV-Based Cross-Crop Yield Prediction in Root Crops. *Remote Sens.* **2025**, *17*, 4054. [CrossRef]
94. Hossen, M.I.; Awrangjeb, M.; Pan, S.; Al Mamun, A. Transfer Learning in Agriculture: A Review. *Artif. Intell. Rev.* **2025**, *58*, 97. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.