Research Square

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

Deep time-series generative/predictive exploratory study of site-specific crop yield using satellite imagery

Rosie Finnegan

2251306@swansea.ac.uk

Swansea University

Sara Sharifzadeh

Swansea University

Alberto Hornero

Instituto de Agricultura Sostenible

Hamidreza Eskandari

Swansea University

Xianghua Xie

Swansea University

Research Article

Keywords:

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11569341/v1

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

Additional Declarations: No competing interests reported.

Deep time-series generative/predictive exploratory study of site-specific crop yield using satellite imagery

Rosie Finnegan[1*], Sara Sharifzadeh[1], Alberto Hornero[2], Hamidreza Eskandari[3], Xianghua Xie[1]

[1*]School of Mathematics and Computer Science, Swansea University, Fabian Way, Crymlyn Burrows, Swansea, SA1 8EN, Wales, United Kingdom.
[2]Institute for Sustainable Agriculture (IAS), Spanish National Research Council (CSIC), Alameda del Obispo, S/N, Córdoba, 14004, Spain.
[3]School of Management, Swansea University, Fabian Way, Crymlyn Burrows, Swansea, SA1 8EN, Wales, United Kingdom.

*Corresponding author(s). E-mail(s): r.h.finnegan2251306@swansea.ac.uk;
Contributing authors: sara.sharifzadeh@swansea.ac.uk; alberto.hornero@csic.es; hamidreza.eskandari@swansea.ac.uk; xianghua.xie@swansea.ac.uk;

Abstract

Purpose: Pixel-level crop yield estimation using satellite imagery is a less explored problem with limited comparison of deep time-series models and their generalisability. Reliable prediction is constrained by sparse ground truth, spatial autocorrelation, imbalanced yield distributions, and temporal inconsistencies from cloud removal. We investigate how temporal alignment and feature pruning influence within-field interpolation versus spatial block generalisation. Furthermore, we introduce a new time-series generative augmentation pipeline to address yield imbalance and investigate its impact on prediction accuracy and robustness in a comparative predictive modelling framework. **Methods:** Using 20,170 winter wheat pixel-year observations (2017–2021; 19 field-years) at Witzwil, Switzerland, we benchmark Random Forest (RF), a Convolutional Neural Network–Attention–Gated Recurrent Unit hybrid (CNN-A-GRU), and a selective state-space model (Mamba). We compare seven-pass Consistent Timestamps

via Single Closest Pass (CTSCP) against 90-day seasonal averaging, prune features from $\boldsymbol{K = 20}$ to **1**, and introduce a regime-stratified temporal conditional generative adversarial network (T-CGAN) for extreme yield-bin augmentation. **Results:** RF achieved the highest within-field accuracy ($\boldsymbol{R^2 = 0.901}$), but its performance declined substantially on average across unseen spatial blocks within the same farm ($\boldsymbol{R^2 = 0.329}$). In contrast, CNN-A-GRU demonstrated stronger average within-farm spatial generalisation ($\boldsymbol{R^2 = 0.444}$), improving to **0.485** with T-CGAN augmentation, while Mamba improved from **0.434** to **0.464**. Consistent temporal sampling improved prediction accuracy across all models ($\boldsymbol{\Delta R^2 \approx +0.022}$), while feature pruning maintained competitive performance using only six variables. **Conclusion:** High within-field accuracy does not guarantee spatial transferability; sequence models with discrete phenological sampling generalise better on average across unseen blocks within the same farm. Generative augmentation further improves average prediction accuracy under within-farm spatial distribution shift.

Keywords: Pixel-Level Crop Yield Estimation, Sentinel-2, Generative Adversarial Networks, Spatiotemporal Feature Selection, Spatial Cross-Validation, SHAP

1 Introduction

High-resolution (10 m) pixel-level crop yield forecasting from Earth observation (EO) data is a cornerstone of site-specific crop management, enabling within-field delineation of management zones, variable-rate nitrogen application, targeted scouting for abiotic or biotic stress, and parametric crop insurance assessment (Copernicus Climate Change Service 2024; Filippi et al. 2019; Perich et al. 2023). However, translating satellite reflectance into reliable agronomic decisions requires quantifying predictive uncertainty under both extreme weather anomalies and spatial distribution shift. In practice, pixel-level yield ground truth from combine harvesters is rarely publicly available due to commercial privacy restrictions (Perich et al. 2023; Paudel et al. 2023a), while optical Sentinel-2 observations are frequently disrupted by cloud cover at critical phenological stages (Copernicus 2025). These constraints create a challenging small-data setting for high-resolution forecasting. In this study, we use 20,170 Winter Wheat (WW) samples from five growing seasons at Witzwil, Switzerland (Perich et al. 2023).

Several methodological challenges arise from this setting. First, generative time-series models for data augmentation and balancing yield data can suffer from mode collapse and temporal instability when applied to heterogeneous agricultural data (Sampath et al. 2021; Yoon et al. 2019; Mogren 2016). Second, the relative performance of sequence models (Mamba, CNN-A-GRU) and strong tabular approaches such as Random Forest (RF) under genuine small-data and data shift conditions remains insufficiently characterised (Oikonomidis et al. 2022; Wang et al. 2023; Gu and Dao 2024). Third, fixed calendar windows and coarse seasonal aggregation can obscure phenological differences between years (Gao et al. 2024; Finnegan et al. 2025).

Finally, high-dimensional multispectral and environmental inputs may contain redundant variables that can hinder efficient modelling (Meghraoui et al. 2024; Paudel et al. 2023a).

Despite recent advancements, existing methods face three open agro-informatics challenges:

(i) **Phenological Signal Dilution:** Widely used fixed calendar windows or broad seasonal composites average out brief, critical developmental stages—such as booting (ear swelling within the flag-leaf sheath) and anthesis (flowering)—diluting the spectral signals required to estimate final grain yield.

(ii) **Balancing Realism and Variety in Synthetic Time Series:** Targeted synthetic augmentation can help address data scarcity in extreme yield regimes, such as hail-induced canopy collapse (lodging of above-ground crop foliage) or peak canopy saturation (where dense leaf cover causes optical reflectance to plateau). However, unconstrained time-series Generative Adversarial Networks (GANs) risk introducing erratic step-to-step jitter into sensitive spectral trajectories—particularly the Red Edge (`B06`) and near-infrared (NIR, `B08`) bands associated with chlorophyll dynamics and Leaf Area Index (LAI) (Frampton et al. 2013)—which can destabilise downstream sequence models.

(iii) **Over-Optimistic Spatial Validation:** Splitting neighbouring pixels from the same field between training and test sets introduces spatial autocorrelation leakage, inflating within-field test accuracy and masking substantial performance drops when models are evaluated on geographically separate farm blocks (Roberts et al. 2017; Filippi et al. 2025).

This research seeks to answer the primary question: How can reliable pixel-level crop-yield models be developed when labelled observations are limited, temporally irregular, and spatially heterogeneous, while maintaining predictive performance across unseen farm blocks? We approach this through three targeted sub-questions:

Which temporal representation best preserves informative biophysical signals—particularly Red Edge and NIR reflectance—from irregular satellite sequences?

Can targeted generative augmentation reduce prediction error in scarce extreme yield bins (e.g., 0–2 t/ha crop failure and 12–14 t/ha peak yield) without degrading global accuracy?

Do within-field predictive improvements transfer across geographically separated blocks within a farm?

The main contributions of this paper are:

1. **Methodological contribution (Regime-stratified T-CGAN):** Building on conditional time-series GANs (Ramponi et al. 2018), we design an autoencoder-based, equilibrium-regularised Temporal Conditional GAN (T-CGAN) incorporating Conditional Layer Normalisation, latent manifold mixup, and regime-specific temporal and statistical constraints, evaluating when synthetic extreme-bin trajectories outperform Gaussian noise injection.
2. **Architectural benchmark (Tabular vs. Sequence and State-Space Models):** We propose a multi-head Convolutional Neural Network–Attention–Gated Recurrent Unit hybrid (CNN-A-GRU) and provide, to our knowledge, the first

pixel-level crop yield evaluation of selective state-space models (Mamba), showing that while Random Forest (RF) excels at within-field interpolation, sequence models generalise better on average across unseen spatial blocks within the same farm.

3. **Spatiotemporal and biophysical representation analysis:** We introduce Consistent Timestamps via Single Closest Pass (CTSCP) to preserve discrete phenological transitions under structured cloud missingness, combine SHapley Additive exPlanations (SHAP) directional attribution with systematic feature pruning ($K = 20$ down to 1), and quantify the relative contributions of site-level meteorology versus pixel-level spectral reflectance.
4. **Empirical evaluation of the spatial interpolation trap:** Through 5-fold contiguous within-year spatial block cross-validation (CV), we demonstrate that high coordinate-grouped within-field accuracy ($R^2 \approx 0.90$) masks substantial performance degradation under spatial distribution shift in tree ensembles (Train $R^2 = 0.984 \rightarrow$ Test $R^2 = 0.329$), whereas sequence architectures maintain higher mean within-farm spatial generalisation across unseen blocks (Test $R^2 = 0.434$–0.444 unaugmented; 0.464–0.485 with fold-isolated T-CGAN regularisation).

The remainder of the paper is organised as follows: Section 2 reviews related work; Section 3 describes the dataset; Section 4 details the methodology; Section 5 reports the results; Section 6 discusses the findings; Section 7 concludes the paper.

2 Related work

2.1 Predictive Architectures and Feature Selection

Recent studies highlight the necessity of balancing predictive accuracy with computational efficiency in crop yield forecasting (Filippi et al. 2019; Oikonomidis et al. 2022; Singh et al. 2025; Bisht 2023). Moreover, apparent model performance in data-driven yield studies depends not only on architectural choice but also on validation design: non-spatial splits frequently produce over-optimistic accuracy estimates when nearby observations are spatially autocorrelated, requiring a strict distinction between within-field interpolation and out-of-block spatial prediction (Filippi et al. 2025; Roberts et al. 2017). RF (Breiman 2001) and kernel-based methods remain highly robust tabular baselines for high-dimensional, small-data regimes where attention-based Transformers often struggle to resolve non-linear biophysical manifold boundaries without overfitting (Maxwell et al. 2018; Singh et al. 2025; Finnegan et al. 2026). Whereas our preliminary work on this site (Finnegan et al. 2026) examined unsupervised manifold structure (PCA/t-SNE) and static baselines (RF, SVM, standard Transformer) under a single coordinate-grouped split, the present study introduces hybrid recurrent-attention (CNN-A-GRU) and state-space (Mamba) sequence architectures, regime-stratified T-CGAN tail augmentation, 10-repetition feature pruning, and 5-fold contiguous within-year spatial block validation. For temporal modelling, recurrent architectures combined with convolutional feature extraction (e.g., CNN-LSTM, CNN-GRU) effectively capture complementary spatiotemporal dynamics (Wang et al. 2023; Vani and Guruprakash 2024), while Transformers model long-range dynamics at the cost of

quadratic attention complexity (Vaswani et al. 2017). Recently, selective state-space models like Mamba have provided a linear-time alternative for sequence modelling (Gu and Dao 2024). While Mamba has demonstrated state-of-the-art performance in livestock forecasting (Kim 2026), its application to pixel-level crop yield prediction remains unexplored. To manage the high dimensionality of Sentinel-2 and environmental inputs, model-agnostic game-theoretic methods like SHAP (Lundberg and Lee 2017) are increasingly used to quantify feature contributions and isolate compact, efficient variable subsets (Paudel et al. 2023a).

2.2 Temporal Alignment and Phenological Representation

Irregular satellite observations caused by cloud occlusion are commonly addressed using fixed calendar windows or seasonal temporal aggregation (Gao et al. 2024; Misra et al. 2020). However, rigid compositing methods can dilute or obscure brief, high-frequency phenological transitions between growing seasons (Gao et al. 2024; Misra et al. 2020). From a data-theoretic perspective, missing observations rarely occur completely at random, but instead exhibit structured missingness patterns that can bias downstream analysis if treated merely as random noise (Johansson Fernstad 2019; Johansson Fernstad and Johansson Westberg 2021). In optical agricultural remote sensing, cloud-induced data gaps around key developmental transitions complicate inter-annual phenological comparison (Misra et al. 2020). Phenological alignment provides a structural alternative by selecting temporally comparable developmental passes across years (Finnegan et al. 2025, 2026), prompting our evaluation of Consistent Timestamps via Single Closest Pass (CTSCP) combined with local spatial imputation against standard seasonal averaging.

2.3 Generative Augmentation under Data Scarcity

High-resolution pixel-level yield prediction is fundamentally constrained by limited ground-truth observations (Oikonomidis et al. 2022), a challenge traditionally approached via weak supervision, spatial disaggregation, or broad empirical augmentation (Paudel et al. 2023b; Iwana and Uchida 2021). To address extreme data scarcity directly, GAN-based architectures (e.g., C-RNN-GAN, TimeGAN) have been applied to synthesise agricultural time series and augment irregularly sampled sequences (Pothapragada and R 2025; Mogren 2016; Yoon et al. 2019; Ramponi et al. 2018; Rodríguez-Suárez et al. 2022). However, synthetic agricultural sequences must strictly preserve both feature covariance and biological temporal structure. Generating diverse samples for minority yield extremes without inducing mode collapse or temporal instability motivates the constrained, equilibrium-regularised T-CGAN developed in this study. Building on the conditional time-series GAN formulation (T-CGAN) introduced by Ramponi et al. (2018), we develop an autoencoder-based, equilibrium-regularised T-CGAN architecture tailored to stratified agricultural yield regimes.

Table 1 Synthesis of related literature in crop yield estimation, missingness structure, spatial validation, and time-series augmentation, detailing each study's methodology, scope relevant to this study, and the corresponding approach evaluated in this work.

Study	Method & Focus	Scope and Limitation Relevant to This Study	Approach in This Work
Perich et al. (2023)	Pixel-level (10 m) Sentinel-2 wheat yield mapping via RF, CNN, and RNN.	Focuses on whole-field and random splits; does not evaluate extreme yield-bin augmentation or state-space models under spatial block shift.	Benchmark CNN-A-GRU and Mamba with T-CGAN tail augmentation and 5-fold spatial block CV.
Filippi et al. (2019)	Multi-layered grain yield forecasting in *Precision Agriculture* using RF ensembles.	Models seasonal summaries across farms rather than intra-seasonal multi-pass pixel trajectories.	Preserve discrete intra-seasonal transitions via 7 CTSCP windows and sequence architectures.
Filippi et al. (2025)	Critical review of crop-yield modelling, validation design, and spatial autocorrelation.	Establishes principles for distinguishing interpolation from spatial prediction; does not benchmark deep sequence or generative models.	Empirically contrast coordinate-grouped interpolation against 5-fold contiguous spatial block CV across RF, CNN-A-GRU, and Mamba.
Paudel et al. (2023a)	Regional yield forecasting via LSTM, 1D-CNN, and GBDT with SHAP (Lundberg and Lee 2017).	Operates at coarse regional scale with fixed calendar windows rather than 10 m pixel resolution.	Operate at 10 m pixel scale using CTSCP windows and evaluate feature pruning from $K = 20$ down to 1.
Paudel et al. (2023b)	Weakly supervised DNN for sub-regional yield from coarse labels.	Relies on spatial disaggregation from regional statistics rather than combine-harvester pixel ground truth.	Train directly on combine-harvester pixel ground truth and regularise tails via T-CGAN.
Wang et al. (2023)	Hybrid CNN-GRU for county-level wheat estimation using composites.	County-scale composites without multi-head temporal self-attention or pixel-level spatial block validation.	Introduce multi-head self-attention pooling (CNN-A-GRU) over 10 m CTSCP windows.
Johansson Fernstad (2019); Johansson Fernstad and Johansson Westberg (2021); Misra et al. (2020)	Formalisation of structured missingness patterns and review of Sentinel-2 phenological gaps.	Establish general missing-data theory and satellite phenology challenges without benchmarking deep yield sequence architectures.	Address structured Sentinel-2 cloud missingness via CTSCP alignment and within-year spatial K-NN ($K = 1$).
Finnegan et al. (2026)	Witzwil wheat manifold analysis (PCA/t-SNE) benchmarking RF, SVM, and Transformer.	Uses coordinate-grouped within-field splitting; does not evaluate contiguous spatial block shift or minority yield-bin synthesis.	Extend with 5-fold contiguous spatial block CV (Roberts et al. 2017), Mamba, and T-CGAN tail synthesis.
Gu and Dao (2024); Kim (2026)	Selective state-space model (Mamba) with input-dependent step (Δ_t) for sequence/livestock modelling.	Evaluated on text, genomics, and 1D livestock telemetry; unexplored for multi-band satellite crop yield regression.	Provide a pixel-level crop yield benchmark of Mamba across contrasting seasonal extremes and spatial blocks.
Mogren (2016); Yoon et al. (2019)	C-RNN-GAN and TimeGAN for continuous time-series synthesis.	Unconditioned on continuous regression targets or temporal sampling metadata in extreme distribution tails.	Combine Autoencoder bottleneck, Conditional LayerNorm, LSGAN (Mao et al. 2017), $\mathcal{L}_{TV}$, and Soft-DTW conditioned on $(\tilde{y}, \mathbf{X}^{\text{meta}})$.
Pothapragada and R (2025)	Stacked CNN with GAN augmentation and XAI for regional maize yield.	Augments static seasonal summaries rather than multi-temporal spectral trajectories.	Synthesise multi-pass trajectories preserving inter-band covariance ($\mathcal{L}_{Cov}$) and temporal continuity.
This Study	**RF, CNN-A-GRU, Mamba (10 m pixel)**	**Single-site case study (19 wheat field-years across 17 rotated parcels)**	**CTSCP + Feature Pruning ($K = 20 \rightarrow 1$) + Regime-Stratified T-CGAN + 5-Fold Within-Year Spatial Block CV**

3 Dataset

This study builds upon the open-access, pixel-level Sentinel-2 multispectral (Copernicus 2025) and combine-harvester yield benchmark dataset curated by Gregor Perich and colleagues (Perich et al. 2023) at the Witzwil agricultural site in Switzerland, which has a long history of cultivation on predominantly peat-based soils (Kellerhals 1925). While the broader Witzwil cereal archive comprises 54,098 pixel-year observations across 54 cereal field-years (2017–2021), this study focuses on the 20,170 pixel-year observations of monocrop Winter Wheat (WW) spanning 19 field-year parcels (`FIDs`) across 17 distinct physical fields. Cloud-obscured observations were removed, resulting in irregular temporal sampling concentrated in spring and summer when crop development is most active (USDA 2026). Each sample represents a 10 m $\times$ 10 m Sentinel-2 pixel (0.01 ha). For primary within-field model development, pixels were partitioned using a coordinate-grouped 70/15/15 (training/validation/test) split via `GroupShuffleSplit`, ensuring that repeated observations of the same coordinate across seasons remain within a single partition, while a 5-fold contiguous within-year spatial block cross-validation was used to evaluate geographic extrapolation (Section 4.7).

3.1 Study Area and Climatological Context

The Witzwil site experiences substantial interannual climatic variability, providing contrasting environmental extremes for model evaluation. The year 2020 tied as the warmest year on record in Switzerland since 1864, driven by historically warm winter and spring seasons (Bader et al. 2021). This thermal excess triggered rapid early-season growing degree day (GDD) accumulation, accelerating early vegetative development. Although this phase was immediately followed by a severe and persistent spring drought during critical stem elongation and booting stages, final crop yields remained exceptionally high (8.04 t/ha). This decoupling of moisture deficits from final yield outcomes suggests a non-linear micro-climatic interaction, potentially wherein early thermal accumulation advanced vegetative rooting and phenological timing prior to upper-soil moisture depletion (Perich et al. 2023; Bader et al. 2021). Such counter-intuitive climate–yield dynamics underscore the necessity of non-linear predictive architectures capable of interpreting complex spectral trajectories over simple linear meteorological indicators.

Conversely, 2021 brought a cold, delayed spring followed by one of the wettest summers north of the Alps in Swiss meteorological history, causing widespread waterlogging that culminated in a historic series of convective supercells between June 18 and July 31, 2021 (Bader et al. 2022). These storms produced severe hailstorms with return periods exceeding 70 to 100 years, mechanically damaging maturing wheat canopies and driving a historic CHF 110 million payout by the Swiss Hail Insurance Company (*Schweizer Hagel*) (Kopp et al. 2023; Portmann et al. 2024). In our experimental framework, these contrasting seasonal extremes directly populate the peak-productivity (12–14 t/ha, predominantly 2020) and catastrophic-failure (0–2 t/ha, predominantly 2021 hail and waterlogging damage) distribution tails targeted by our T-CGAN augmentation and mid-June CTSCP phenological windows.

3.2 Dataset Size Information

The 20,170 cloud-free WW pixel-year observations span 19 distinct field-year parcels across the five harvest seasons: 1,638 pixels (2 fields) in 2017, 6,240 (7 fields) in 2018, 4,340 (4 fields) in 2019, 4,969 (4 fields) in 2020, and 2,983 (2 fields) in 2021. Across the five-year rotation, these observations map to 16,465 unique 10 m $\times$ 10 m spatial coordinates (`coord_id` in EPSG:32632); no parcel grew winter wheat in two consecutive seasons, and only two parcels (P36 in 2017/2019 and P21 in 2019/2021, totalling 3,705 coordinates) returned to winter wheat after a one-year break crop. Mean yield was 8.04 t/ha in 2020 compared with 4.52 t/ha in 2021, with annual yields approximately following a unimodal distribution (Figure 1). Importantly, these 20,170 pixel-year observations do not represent 20,170 independent environments: because each pixel covers 0.01 ha across a cumulative cultivated area of 201.7 ha ($\approx$ 40.3 ha/year across 19 field-years), pixels within the same field-year share identical meteorological forcing, sowing schedules (reflected in days after sowing, `das`), and spatially continuous peat-soil properties. Consequently, the effective environmental sample size is governed by the number of distinct field-years rather than the raw pixel count, creating strong spatial autocorrelation and a severe small-sample bottleneck in the severe low-yield (0–2 t/ha) and peak saturation (12–14 t/ha) tails (where fewer than 60 to 260 real training pixels exist per fold). Of the 20,170 pixels, 8,581 possess complete seven-pass unoccluded spectral sequences prior to K-NN ($K = 1$) spatial imputation.

4 Methods

Based on the identified gaps, the methodologies are designed for this study. The methodological framework is organised into two distinct stages (Figure 2): Stage 1 (Sections 4.1–4.5) details spatio-temporal data engineering, feature selection, development of a new time-series generative augmentation model, and predictive model design, while Stage 2 (Sections 4.6–4.9) defines the comparative benchmarking, spatial domain-shift evaluation, and statistical testing protocols.

4.1 Input Features and Engineering

Input features were divided into biophysical core features (Sentinel-2 spectral variables and climatic data) ($\mathbf{s}_t$) and temporal metadata ($\mathbf{m}_t$):

$$\mathbf{x}_t = [\mathbf{s}_t, \mathbf{m}_t] \tag{1}$$

The metadata comprised days after sowing (`das`), temporal pass interval (`time_delta`), and cyclical day-of-year (DOY) representations (`sin_doy`, `cos_doy`). The `time_delta` variable explicitly represents irregular sampling intervals.

4.2 Temporal Feature Selection

Two temporal representation strategies were evaluated to quantify the operational trade-off between discrete phenological pass selection and coarse calendar compositing. Consistent Timestamps via Single Closest Pass (CTSCP) selected the observation closest to the centre of each predefined 10-day window ($\pm$5 days), producing seven

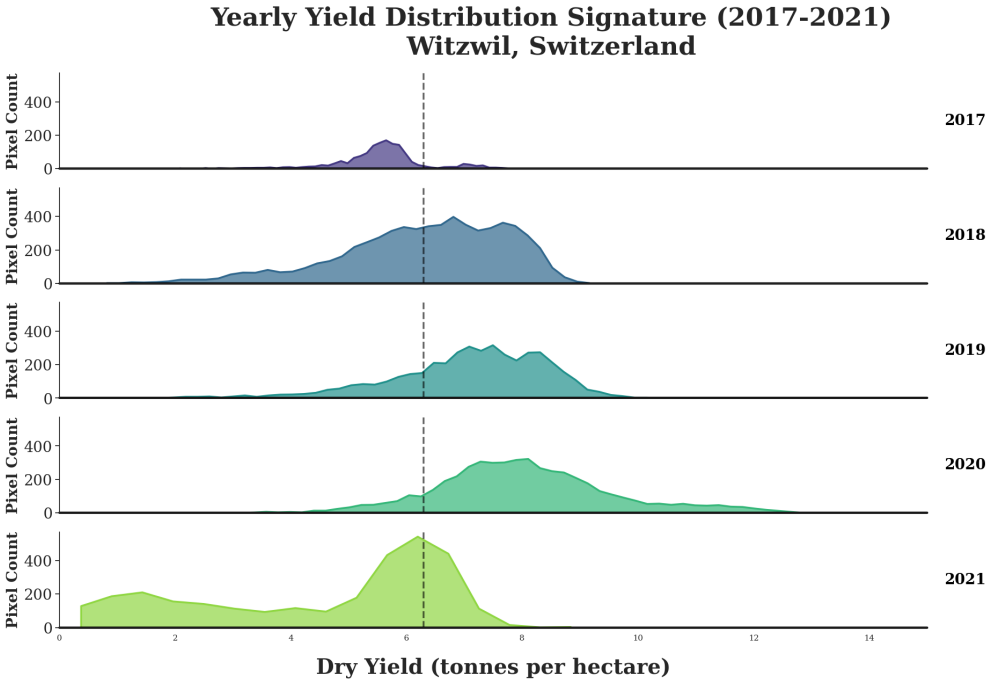


Fig. 1 Ridgeline distribution of pixel-level Winter Wheat yields at the Witzwil study site (2017–2021). The vertical stacking illustrates the distinct yearly phenological signatures, highlighting the significant yield depression in 2021 as compared to the high-productivity year of 2020. The dotted vertical line represents the mean yield across all years in the dataset (6.29 t/ha).

aligned timestamps across years. Conversely, 90-day Seasonal Averaging divided the calendar year into three coarse 90-day periods starting on DOY 1 (1 January) and averaged all valid observations within each period. The annual crop cycle was partitioned into three 90-day phenological quadrants mapping to distinct developmental phases: Q1 [DOY 1-90] capturing winter dormancy and tillering; Q2 [DOY 91-180] capturing stem elongation and flowering; and Q3 [DOY 181-270] capturing grain filling and harvest.

CTSCP serves as the baseline and as the primary data representation for all predictive and generative experiments.

To guarantee zero spatial imputation leakage during within-field benchmarking (Sections 5.1–5.5), where adjacent 10 m pixels may reside in different partitions, models were evaluated strictly on the unoccluded 8,581-pixel CTSCP subset requiring no K-NN fallback. Conversely, for the 5-fold contiguous spatial block cross-validation (Section 5.6), the full 20,170-pixel dataset was utilised to preserve contiguous field geometry. Missing single-pass spectral covariates (never target yield values) were imputed prior to block partitioning from the single nearest unoccluded 10 m pixel (K-NN, $K = 1$ in Euclidean coordinate space) within the same harvest year. Crucially, this pre-partition imputation cannot leak training information into the test set

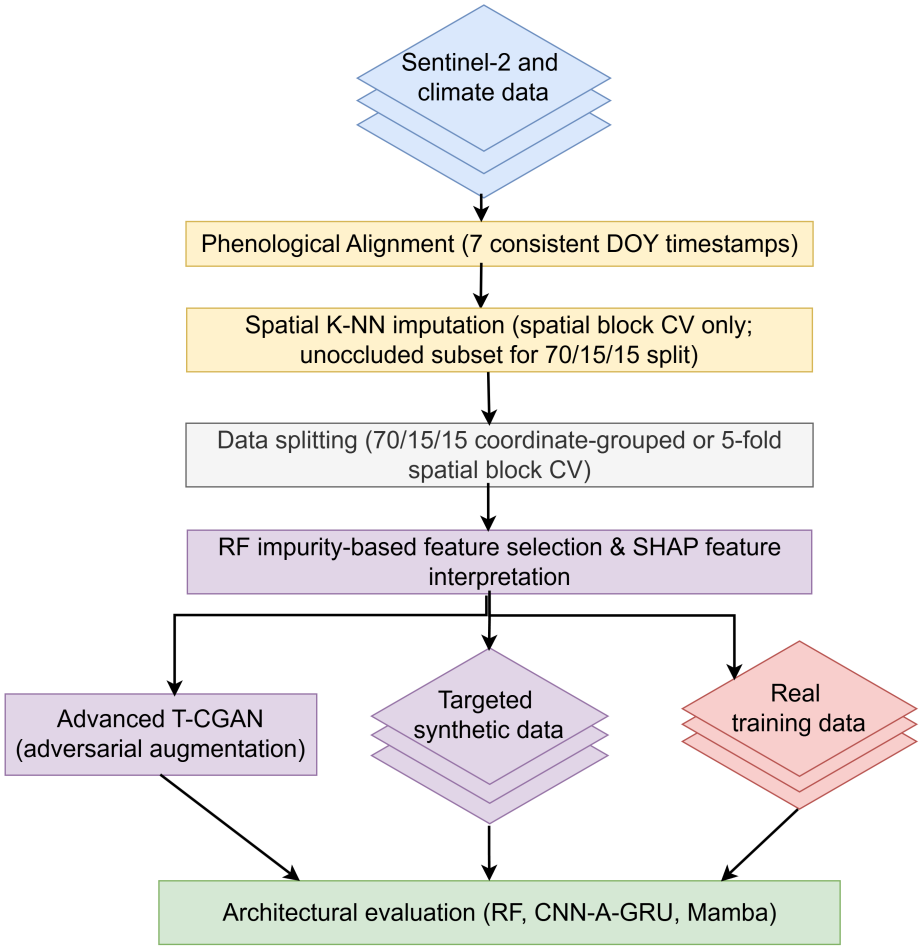


Fig. 2 Conceptual methodological workflow. Within-field benchmarking uses the 8,581 unoccluded CTSCP sequences (70/15/15 coordinate-grouped split, no imputation), whereas 5-fold contiguous within-year spatial block cross-validation uses all 20,170 pixels with within-year spatial K-NN ($K = 1$) spectral covariate imputation. Following data partitioning, all feature scaling, RF impurity-based feature selection ($K = 20 \rightarrow 1$), SHAP directional interpretation, and regime-stratified T-CGAN synthesis are fitted strictly on the training partition.

or affect test performance: because the 60% training block and 20% test block are separated by a contiguous 20% validation buffer spanning hundreds of metres (and frequently occupying separate field parcels; Figures 5–6), nearest-neighbour ($K = 1$) lookup never links training and test pixels. At most, a single row of 10 m pixels along an internal parcel boundary between the validation buffer and test block can share a missing single-pass reflectance value with an adjacent validation pixel (which is never used for parameter fitting), and any resulting perimeter smoothing is negligible and conservative. Across all experiments, feature scalers, tree-based/SHAP rankings, and

T-CGAN generators were fitted exclusively on the training partition of each split or fold.

4.3 Spectral and Climate Feature Selection

Feature analysis was conducted strictly on the training partition using two complementary methods: SHAP for directional feature interpretation and RF impurity importance for nested feature selection. First, SHAP (Lundberg and Lee 2017) was fitted on the training partition to interpret directional variable attributions and timestamp-specific phenological drivers on CTSCP-aligned sequences (Figure 8), complementing prior unsupervised manifold analyses on this site (Finnegan et al. 2026). Second, for the systematic feature-pruning sensitivity experiments (Table 3), the 20 input variables were ranked on the training partition using RF variance-reduction (impurity) importance, and nested subsets from the $K = 20$ baseline down to $K = 1$ (each retained across all $T = 7$ timestamps, yielding $7 \times K = 7K$ flattened feature-timestamp inputs for RF) were evaluated across all three architectures.

4.4 Data Augmentation

4.4.1 Generative Adversarial Network and the Proposed T-CGAN

Due to the imbalanced yield distribution over train samples causing higher errors in the extreme regions, augmentation of time-series data used for prediction is considered. To prevent data leakage and preserve regime-specific biophysical characteristics, independent conditional generators were trained strictly on the training partition within four target minority yield bins (0–2, 2–4, 10–12, and 12–14 t/ha; 2 t/ha bin width). Within each target bin b containing n_b real training sequences, inputs are partitioned into $d_{\text{core}} = 16$ standardised biophysical variables ($\mathbf{X}^{\text{core}} \in \mathbb{R}^{T \times d_{\text{core}}}$) and $d_{\text{meta}} = 4$ temporal metadata variables ($\mathbf{X}^{\text{meta}} \in \mathbb{R}^{T \times d_{\text{meta}}}$), alongside continuous Min–Max normalised target yield values $\tilde{y} \in \mathbb{R}$ (rescaled from physical t/ha to the continuous interval $0 \leq \tilde{y} \leq 1$ to stabilise generator conditioning). During synthesis, novel core trajectories $\hat{\mathbf{X}}^{\text{core}}$ are generated via latent convex interpolation (manifold mixup) between random real training pairs (i, j) within bin b:

$$\mathbf{z}_{\text{syn}} = \lambda \boldsymbol{\mu}_i + (1 - \lambda)\boldsymbol{\mu}_j + \sigma_{\text{jitter}} \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}, \quad \hat{\mathbf{X}}^{\text{core}} = G_{\text{decode}}\big(\mathbf{z}_{\text{syn}},\, \tilde{y}_{\text{syn}},\, \mathbf{X}_i^{\text{meta}}\big) \tag{2}$$

where $(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) = G_{\text{encode}}(\mathbf{X}_i^{\text{core}})$, $\lambda \sim \mathcal{U}(0.6, 0.95)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\tilde{y}_{\text{syn}} = \lambda \tilde{y}_i + (1 - \lambda)\tilde{y}_j \in \mathbb{R}$ is the corresponding continuous convex interpolation between the two real continuous yields $\tilde{y}_i$ and $\tilde{y}_j$ (which is subsequently inverse-scaled back to physical tonnes per hectare, y_{syn} in t/ha, before training the downstream regression models). Synthetic samples are appended exclusively to the training partition, leaving validation and test sets strictly unaugmented. In our volume-sweep experiments (25%, 50%, and 100% volume), synthetic sequences equal to $0.25n_b$, $0.50n_b$, or $1.00n_b$ (where n_b is the number of real training samples in minority bin b) were generated and appended to the training set.

Architectural Evolution and the Autoencoder Bottleneck. Developing a generator capable of balancing trajectory diversity with biophysical realism in scarce

extreme yield regimes required five progressive design stages, which also form the basis of the generative ablation study in Section 5.5. First, **GAN 1 (Multimodal Transformer)** was initially considered to jointly capture local 2D spatial context and 1D meteorological series via cross-attention, but its high parameter complexity relative to the small sample sizes in extreme yield bins caused severe structural alignment instability during training. Second, **GAN 2 (Direct 1D-GAN; Ablation Model 1)** simplified the task to 1D pixel-level sequences generated directly from random Gaussian noise via an unencoded 1D-CNN generator; however, mapping unstructured noise directly to 16-variable biophysical trajectories without latent anchoring resulted in mode collapse and unrealistic inter-band correlations. Third, **GAN 3 (Stratified Autoencoder; Ablation Model 2)** was introduced to overcome mode collapse by stratifying training into discrete 2 t/ha yield regimes and embedding a variational Conv1D–GRU–MultiheadAttention Encoder–Decoder bottleneck (Figure 3), allowing the generator to interpolate between encoded real biophysical representations rather than sampling from unconstrained noise. Fourth, **GAN 4 (Feature-Matched Autoencoder)** incorporated discriminator Feature Matching ($\mathcal{L}_{FM}$) to stabilise adversarial updates and preserve higher-order temporal feature statistics across passes. Finally, the **Proposed T-CGAN (Ablation Model 5)** was developed to enforce both continuous target conditioning and smooth phenological transitions, integrating Conditional Layer Normalisation (projecting embedded continuous yield values into affine scale $\boldsymbol{\gamma}$ and shift $\boldsymbol{\beta}$ parameters within the decoder normalisation layer) alongside composite statistical ($\mathcal{L}_{Curve}, \mathcal{L}_{Cov}, \mathcal{L}_{Bound}$) and temporal ($\mathcal{L}_{TV}, \mathcal{L}_{DTW}$) regularisation (with intermediate Ablation Models 3 and 4 isolating $\mathcal{L}_{TV}$ and $\mathcal{L}_{DTW}$, respectively). Across these stages, to prevent discriminator overpowering, training targets an approximate adversarial equilibrium (Goodfellow et al. 2014; Mirza and Osindero 2014) using one-sided discriminator real-target smoothing (setting the adversarial target for real sequences to 0.9 rather than 1.0) and latent perturbation noise (ϵ_{scale}) (Hyland et al. 2017; Sun et al. 2020).

Internal Sequential Structure of the Encoder, Decoder, and Discriminator. To explicitly preserve intra-seasonal temporal dynamics across the $T = 7$ phenological passes rather than treating observations as static tabular vectors, the Encoder, Decoder, and Discriminator are constructed around sequential recurrent-attention layers (Figure 3). Inside the **Encoder** (G_{encode}), the input trajectory $\mathbf{X}^{\text{core}} \in \mathbb{R}^{B\times T\times d_{\text{core}}}$ is first processed by a 1D convolutional layer (`Conv1D`) to extract local cross-band features at each pass, followed by a Gated Recurrent Unit (`GRU`) and temporal `MultiheadAttention` pooling that compress sequential state transitions across the season into the latent Gaussian parameters $(\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2) \in \mathbb{R}^{B\times d_z}$. Inside the **Decoder** (G_{decode}), the interpolated latent vector $\mathbf{z}_{\text{syn}}$ is repeated across all $T = 7$ timestamps and concatenated at each time step t with that pass's specific temporal sampling metadata $\mathbf{X}_t^{\text{meta}}$ (`das`, `time_delta`, `sin_doy`, `cos_doy`) to anchor the irregular inter-pass intervals. This sequence is then modulated via Conditional Layer Normalisation (using affine scale $\boldsymbol{\gamma}$ and shift $\boldsymbol{\beta}$ projected from the continuous yield value $\tilde{y}_{\text{syn}}$) and decoded sequentially through a multi-layer `GRU`, temporal `MultiheadAttention`, and a linear output projection to reconstruct ordered 7-pass trajectories $\hat{\mathbf{X}}^{\text{core}} \in \mathbb{R}^{B\times T\times d_{\text{core}}}$. Symmetrically, the **Discriminator** (D) processes the

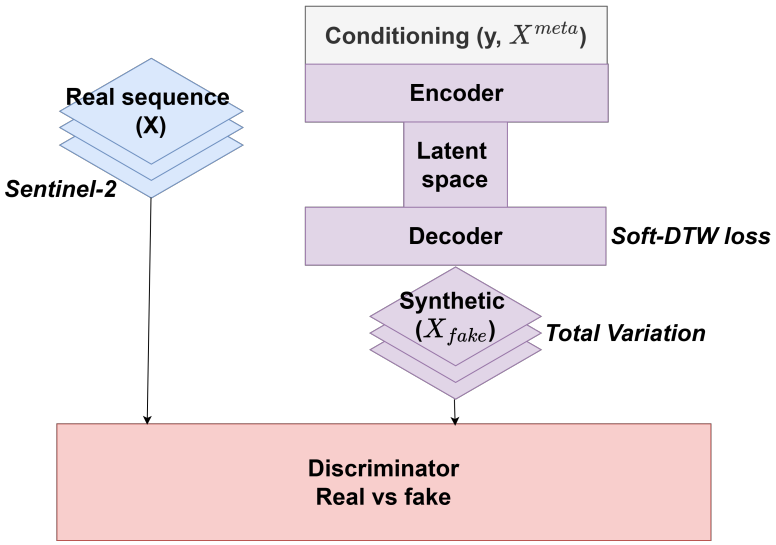


Fig. 3 Architecture of the proposed equilibrium-regularised T-CGAN. Both the Encoder and Decoder incorporate sequential `Conv1D`, `GRU`, and temporal `MultiheadAttention` layers conditioned on continuous target yield ($\tilde{y}$) via Conditional Layer Normalisation and pass-specific temporal metadata ($\mathbf{X}^{\text{meta}}$) across all $T = 7$ timestamps.

concatenated sequence $[\mathbf{X}^{\text{core}}, \mathbf{X}^{\text{meta}}, \tilde{y}]$ through sequential `Conv1D`, `GRU`, and temporal attention-pooling layers ($\mathbf{f}_D(\cdot) \in \mathbb{R}^{B\times 64}$), evaluating trajectory-level temporal realism and step-to-step phenological transitions across all seven passes simultaneously.

Let $\mathbf{x}_{\text{real}}, \mathbf{x}_{\text{fake}} \in \mathbb{R}^{B\times T\times d_{\text{core}}}$ denote real and synthetic core feature batches of batch size B over $T = 7$ timestamps. To evaluate the contribution of each architectural constraint in the ablation study (Section 5.5, Tables 5 and 6), the generator designs progressively incorporate the following four groups of objective terms:

1. Adversarial and Feature Matching Losses

Minimax GAN Loss ($\mathcal{L}_{BCE}$): Serves as the sole training objective in the unencoded Direct 1D-GAN (**GAN 2 / Ablation Model 1**) and as the base adversarial loss in the intermediate Autoencoder ablation tiers (**GAN 3 / Ablation Models 2–4**):

$$\mathcal{L}_{D\text{-BCE}} = -\mathbb{E}[\log D(\mathbf{x}_{\text{real}})] - \mathbb{E}[\log(1 - D(\mathbf{x}_{\text{fake}}))], \quad \mathcal{L}_{G\text{-BCE}} = -\mathbb{E}[\log D(\mathbf{x}_{\text{fake}})] \tag{3}$$

Least Squares GAN ($\mathcal{L}_{LSGAN}$) and Feature Matching ($\mathcal{L}_{FM}$): Feature Matching was introduced in **GAN 4** to stabilise intermediate discriminator representations $\mathbf{f}_D(\cdot) \in \mathbb{R}^{B\times 64}$, and is combined with Least Squares GAN loss (Mao et al. 2017) (replacing $\mathcal{L}_{BCE}$ and smoothing the discriminator's real-sequence target from 1.0 to 0.9 in $D(\mathbf{x}_{\text{real}})$ to prevent vanishing gradients) in the final **Proposed T-CGAN**

(Ablation Model 5):

$$\mathcal{L}_{D\text{-LSGAN}} = \frac{1}{2}\mathbb{E}\big[(D(\mathbf{x}_{\text{real}}) - 0.9)^2\big] + \frac{1}{2}\mathbb{E}\big[D(\mathbf{x}_{\text{fake}})^2\big]\,,$$
$$\mathcal{L}_{G\text{-LSGAN}} = \frac{1}{2}\mathbb{E}\big[(D(\mathbf{x}_{\text{fake}}) - 1)^2\big] \tag{4}$$

$$\mathcal{L}_{FM} = \text{MSE}(\mathbf{f}_D(\mathbf{x}_{\text{fake}}),\, \mathbf{f}_D(\mathbf{x}_{\text{real}})) \tag{5}$$

2. Autoencoder Structural Constraints

Reconstruction and Kullback–Leibler Loss ($\mathcal{L}_{Rec}, \mathcal{L}_{KL}$): Omitted in the unencoded Direct GAN (**GAN 2 / Model 1**), and introduced in the Stratified Autoencoder (**GAN 3 / Model 2**) and retained across all subsequent autoencoder-based architectures (**Models 2–5**) to anchor synthetic trajectories to real encoded latent posteriors $\boldsymbol{\mu}, \log \boldsymbol{\sigma}^2 \in \mathbb{R}^{B \times d_z}$ ($d_z = 128$):

$$\mathcal{L}_{Rec} = \text{MSE}(\mathbf{x}_{\text{fake}}, \mathbf{x}_{\text{real}}), \quad \mathcal{L}_{KL} = -\frac{1}{2Bd_z}\sum_{i=1}^{B}\sum_{k=1}^{d_z}\left(1 + \log(\sigma_{i,k}^2) - \mu_{i,k}^2 - \sigma_{i,k}^2\right) \tag{6}$$

3. Statistical and Boundary Constraints

Integrated specifically into the **Proposed T-CGAN (Ablation Model 5)** to ensure that generated sequences respect empirical seasonal distributions, inter-band spectral collinearity, and physical sensor limits:

Curve Matching Loss ($\mathcal{L}_{Curve}$): Matches batch-wise temporal mean $\boldsymbol{\mu}_{\mathbf{x}} \in \mathbb{R}^{T \times d_{\text{core}}}$ and standard deviation $\boldsymbol{\sigma}_{\mathbf{x}} \in \mathbb{R}^{T \times d_{\text{core}}}$ across the batch dimension:

$$\mathcal{L}_{Curve} = \text{MSE}(\boldsymbol{\mu}_{\mathbf{x}_{\text{fake}}}, \boldsymbol{\mu}_{\mathbf{x}_{\text{real}}}) + \text{MSE}(\boldsymbol{\sigma}_{\mathbf{x}_{\text{fake}}}, \boldsymbol{\sigma}_{\mathbf{x}_{\text{real}}}) \tag{7}$$

Covariance Loss ($\mathcal{L}_{Cov}$): Preserves inter-band spectral relationships across batch-time flattened matrices ($N = B \cdot T$), where $\mathbf{C}(\mathbf{x}) = \frac{1}{N-1}\tilde{\mathbf{x}}^\top\tilde{\mathbf{x}} \in \mathbb{R}^{d_{\text{core}} \times d_{\text{core}}}$ denotes the centred sample covariance matrix:

$$\mathcal{L}_{Cov} = \frac{1}{d_{\text{core}}^2}\sum_{u=1}^{d_{\text{core}}}\sum_{v=1}^{d_{\text{core}}}\left|C_{u,v}(\mathbf{x}_{\text{real}}) - C_{u,v}(\mathbf{x}_{\text{fake}})\right| \tag{8}$$

Boundary Loss ($\mathcal{L}_{Bound}$): A rectified linear penalty preventing biophysical excursions beyond observed batch extrema ($\mathbf{x}_{\text{real}}^{\min}, \mathbf{x}_{\text{real}}^{\max} \in \mathbb{R}^{T \times d_{\text{core}}}$):

$$\mathcal{L}_{Bound} = \frac{1}{B}\sum_{i=1}^{B}\sum_{t=1}^{T}\sum_{d=1}^{d_{\text{core}}}\Big(\max\big(0,\, x_{\text{fake},i,t,d} - x_{\text{real},t,d}^{\max}\big) + \max\big(0,\, x_{\text{real},t,d}^{\min} - x_{\text{fake},i,t,d}\big)\Big) \tag{9}$$

4. Temporal Sequence Regularisation

To isolate the impact of temporal smoothing versus elastic phase alignment on downstream sequence models, these two temporal constraints are first evaluated individually on top of the Stratified Autoencoder in **Ablation Model 3** (Autoencoder + $\mathcal{L}_{TV}$) and **Ablation Model 4** (Autoencoder + $\mathcal{L}_{DTW}$), and then combined jointly in the **Proposed T-CGAN (Ablation Model 5)**:

Total Variation Loss ($\mathcal{L}_{TV}$): Evaluated in **Model 3** and **Model 5** to penalise erratic step-to-step jumps across adjacent passes:

$$\mathcal{L}_{TV} = \frac{1}{B(T-1)d_{\text{core}}} \sum_{i=1}^{B} \sum_{t=2}^{T} \left\| \mathbf{x}_{\text{fake},i,t} - \mathbf{x}_{\text{fake},i,t-1} \right\|_1 \tag{10}$$

Soft-DTW Loss ($\mathcal{L}_{DTW}$): Evaluated in **Model 4** and **Model 5**, computing the batch-averaged differentiable dynamic time warping alignment cost ($\gamma = 0.1$) over alignment matrices $\mathcal{A}_{T,T}$ and pairwise squared Euclidean cost matrix $\mathbf{\Delta}(\mathbf{x}_{\text{fake},i}, \mathbf{x}_{\text{real},i}) \in \mathbb{R}^{T \times T}$ (Cuturi and Blondel 2017):

$$\mathcal{L}_{DTW} = \frac{1}{B} \sum_{i=1}^{B} \text{soft-min}_{\gamma} \left\{ \langle \mathbf{A}, \mathbf{\Delta}(\mathbf{x}_{\text{fake},i}, \mathbf{x}_{\text{real},i}) \rangle : \mathbf{A} \in \mathcal{A}_{T,T} \right\} \tag{11}$$

Summary of Ablation Model Objectives. In summary, the five generative tiers evaluated in Section 5.5 optimise: **Model 1 (Direct GAN / GAN 2):** $\mathcal{L}_{G\text{-BCE}}$ only; **Model 2 (Autoencoder / GAN 3):** $\mathcal{L}_{G\text{-BCE}} + w_{\text{rec}}\mathcal{L}_{Rec} + w_{\text{kl}}\mathcal{L}_{KL}$; **Model 3 (Autoencoder + TV):** Model 2 objective $+w_{\text{tv}}\mathcal{L}_{TV}$; **Model 4 (Autoencoder + DTW):** Model 2 objective $+w_{\text{dtw}}\mathcal{L}_{DTW}$; and **Model 5 (Proposed T-CGAN):** the full composite generator objective $\mathcal{L}_G = \mathcal{L}_{G\text{-LSGAN}} + w_{\text{rec}}\mathcal{L}_{Rec} + w_{\text{kl}}\mathcal{L}_{KL} + w_{\text{fm}}\mathcal{L}_{FM} + w_{\text{cov}}\mathcal{L}_{Cov} + w_{\text{dtw}}\mathcal{L}_{DTW} + w_{\text{tv}}\mathcal{L}_{TV} + w_{\text{curve}}\mathcal{L}_{Curve} + w_{\text{bound}}\mathcal{L}_{Bound}$, with regime-stratified weights defined in Section 4.9.

4.4.2 Gaussian Noise

To establish a benchmark against generative methods, isotropic Gaussian noise injection was evaluated across targeting strategies (Global, Minority bins, Bin 0), sweeping varying volumes (10%, 25%, 50%) and intensities (2.5%, 5.0%, 10.0%).

4.5 Proposed Spatiotemporal Hybrid Architecture (CNN-A-GRU)

To jointly model spectral interactions and phenological temporal transitions, we propose a hybrid deep spatiotemporal architecture (CNN-A-GRU Pipeline 5), illustrated in Figure 4.

The architecture comprises three core processing stages:

1. **Spectral Feature Extraction:** A 1D-CNN base layer operates across the $d = 20$ aligned input variables at each timestamp, projecting raw reflectance and climatic indicators into high-level spatial representations.

2. **Recurrent Temporal Encoding:** A multi-layer Gated Recurrent Unit (GRU) processes the sequence of extracted spectral vectors, capturing sequential phenological memory and hidden temporal dynamics $\mathbf{H} \in \mathbb{R}^{B\times 7\times 128}$.
3. **Multi-Head Attention Pooling:** Rather than relying on simple last-step or mean pooling, temporal hidden states are routed through a multi-head self-attention mechanism (Query, Key, Value). This dynamically computes attention weights across the seven phenological timestamps, prioritising critical grain-filling stages before projection through two dense layers for final yield regression.

4.6 Predictive Baselines and Comparative Framework

During preliminary model screening on the validation partition (extending prior baseline evaluations of Support Vector Machines and standard Transformers on this split; Finnegan et al. 2026), we evaluated Random Forest (RF) (Breiman 2001), CNN-LSTM (Shi et al. 2015; Hochreiter and Schmidhuber 1997), Attention-GRU, seven structural variants of our proposed CNN-A-GRU (Pipelines P1–P7, detailed in Appendix A), and selective state-space models (Mamba) (Gu and Dao 2024). To maintain a focused comparative benchmark across the three primary methodological paradigms, the strongest architecture from each paradigm—RF (tabular ensemble), CNN-A-GRU Pipeline 5 (hybrid recurrent-attention), and Mamba (selective state-space)—was carried forward for the full 10-repetition temporal, feature-pruning, generative, and 5-fold spatial cross-validation experiments.

4.7 Spatial Generalisation

The coordinate-grouped 70/15/15 split was used for primary model development and within-field interpolation benchmarking. To mitigate spatial autocorrelation leakage and evaluate within-farm geographic extrapolation (Roberts et al. 2017; Filippi et al. 2025), models were also evaluated using 5-fold contiguous within-year spatial block cross-validation (Figures 5 and 6). At Witzwil, winter wheat is rotated across 17 distinct parcels over the 19 field-years, with as few as two wheat parcels cultivated in 2017 and 2021 and intervening non-cereal break crops unrecorded. Consequently, partitioning the entire farm boundary into fixed multi-year geographic boxes first would leave several box-year combinations devoid of wheat pixels.

Instead, spatial partitioning was performed on an annual basis. Within each harvest year, active winter wheat pixels were identified first, ordered geographically from west to east along their projected UTM Easting and Northing coordinates (x, y) in EPSG:32632, and divided into five equal-sized rolling contiguous blocks. Each fold allocated 60% of that year's pixels to training, an adjacent 20% buffer block to validation, and 20% to a geographically isolated test block (illustrated for the representative 2018 harvest season in Figures 5 and 6). Furthermore, the two repeated parcels (P36 in 2017/2019 and P21 in 2019/2021) were locked into the same block index across their respective seasons, ensuring that a training coordinate in one year never appears in another year's validation or test set. Across each of the five cross-validation folds, model performance was then evaluated over the test blocks of all five harvest seasons ($N = 20,170$ total pixels), with each test pixel weighted equally.

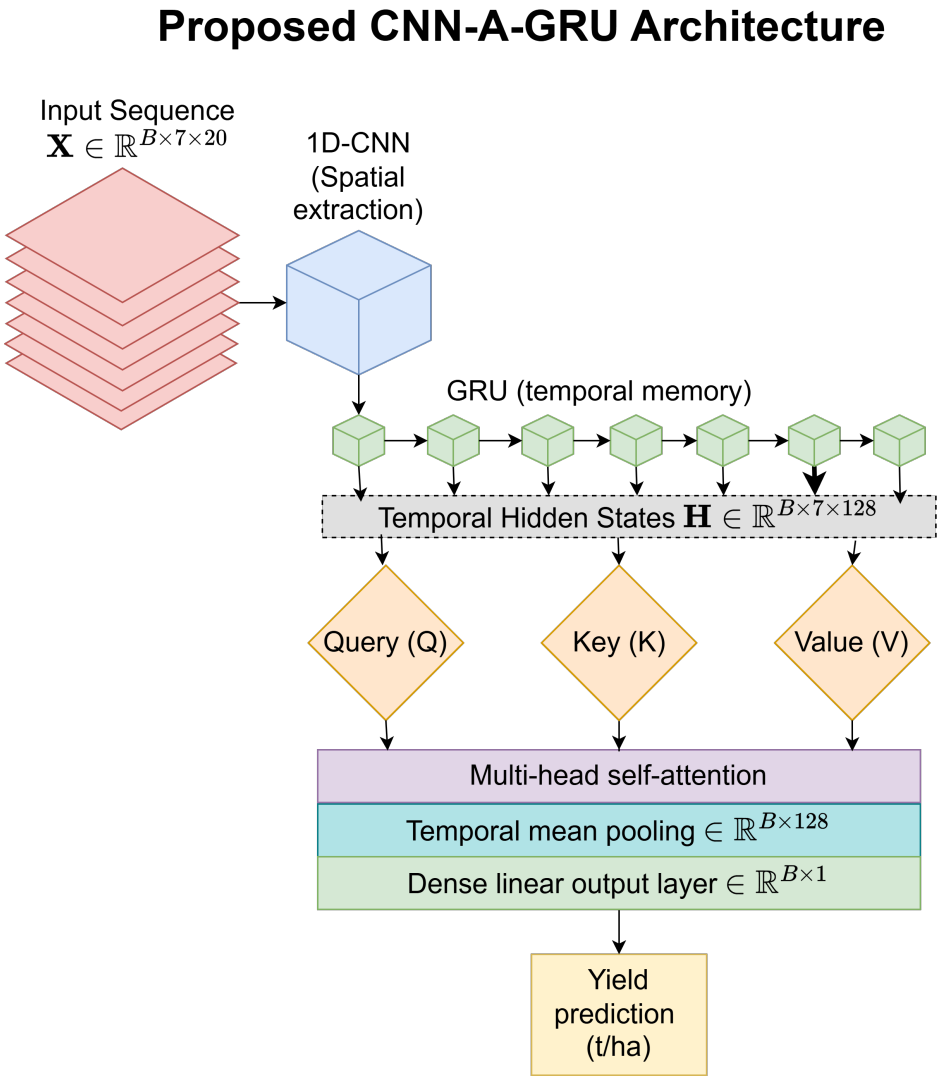


Fig. 4 Architecture of the proposed CNN-A-GRU model.

4.8 Control of Training Stochasticity and Statistical Testing

A fixed seed (42) was used for data partitioning. Within-field experiments were repeated across $n = 10$ random seeds (42–51), whereas the computationally intensive 5-fold spatial block cross-validation used 5 random seeds (42–46) per fold, yielding 25 fold-seed evaluations per model. Results are reported as mean ± standard deviation (σ) alongside 95% confidence intervals ($\mathrm{CI}_{95\%}$). Pairwise differences were evaluated using two-sided Welch's t-tests (allowing unequal variances across model families), supplemented by paired t-tests, non-parametric Wilcoxon signed-rank tests,

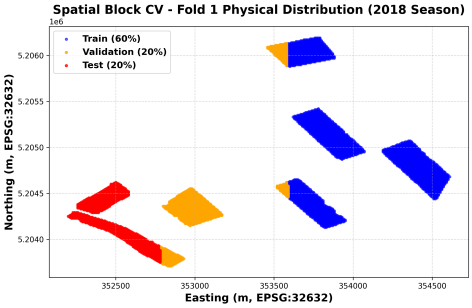


Fig. 5 Annual contiguous spatial block allocation for Fold 1, illustrated for the 2018 harvest season (7 active winter wheat fields; 6,240 pixels). Within each year, the 60% training block (blue) is geographically separated from that year's 20% test block (red) by a contiguous 20% validation buffer (orange).

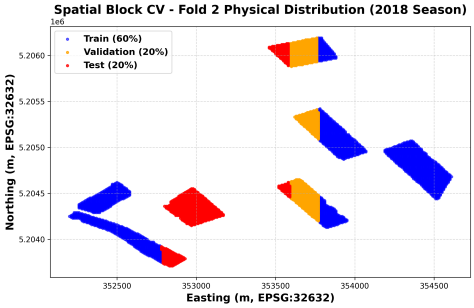


Fig. 6 Annual contiguous spatial block allocation for Fold 2 (illustrated for the 2018 harvest season), demonstrating how the rolling 20% test and validation windows shift eastward across that year's active wheat parcels.

and Cohen's d standardised effect sizes. Train–test yield distributional shift across spatial blocks was quantified via the two-sample Kolmogorov–Smirnov (KS) statistic ($KS = \sup_y |F_{\text{train}}(y) - F_{\text{test}}(y)| \in [0, 1]$, where 0 indicates identical distributions and larger values indicate stronger distributional divergence). We explicitly distinguish between two levels of statistical inference: within the fixed 70/15/15 split, tests across the $n = 10$ seeds quantify algorithmic stability against weight initialisation and mini-batch stochasticity rather than variance across independent farms. In the 5-fold spatial block cross-validation, primary geographic inference is evaluated strictly across the $n = 5$ independent spatial fold means; paired tests across the $n = 25$ clustered fold-seed evaluations are reported as a supplementary measure of seed-level regularisation consistency across runs, and do not constitute 25 independent geographic replications.

4.9 Implementation Details

Models were implemented in PyTorch 2.0 (Ansel et al. 2024) and trained on the Sunbird Supercomputing Cluster using NVIDIA V100/A100 GPUs. Local prototyping used an NVIDIA RTX 4080. Predictive models used AdamW (Loshchilov and Hutter 2017), a learning rate of 0.001, weight decay of 0.01, batch size 32, MSE loss, and early stopping with patience 70. The T-CGAN was trained for 600 epochs with a generator learning rate of 2×10^{-4}, a discriminator learning rate of 5×10^{-5}, and latent dimension 128. To account for differing biophysical dynamics at the distribution tails, loss weights were stratified between low-yield failure bins (0–4 t/ha: $w_{rec} = 20, w_{kl} = 0.1, w_{fm} = 100, w_{cov} = 50, w_{dtw} = 10, w_{tv} = 0.1, w_{curve} = 1, w_{bound} = 5$) and high-yield canopy saturation bins (10–14 t/ha: $w_{rec} = 50, w_{kl} = 0.1, w_{fm} = 50, w_{cov} = 50, w_{dtw} = 25, w_{tv} = 0.5, w_{curve} = 2, w_{bound} = 10$). RF used 300 estimators and square-root feature selection. For the computationally intensive 5-fold spatial block cross-validation across all 20,170 pixels, fold-isolated T-CGANs were trained for 300 epochs with adaptive learning-rate equilibrium scheduling and annealed instance noise

($\sigma = 0.1$), and sequence models were trained with standardised targets for up to 100 epochs (batch size 64, early-stopping patience 25).

5 Results

This section is structured as follows: Section 5.1 presents the model benchmarking results; Section 5.2 evaluates temporal selection results; Section 5.3 analyses SHAP attributions and systematic feature pruning; Section 5.4 details the baseline impact of statistical noise; Section 5.5 evaluates the generative ablation study; and Section 5.6 outlines spatial cross-validation. The dataset's baseline predictive difficulty is established by the raw yield's global standard deviation ($SD_{yield} = 1.9215$ t/ha).

5.1 Models: Comparative Evaluation

Seven CNN-A-GRU structural variants (Appendix A) were initially screened on the validation partition. Pipeline **5** (sequential 1D-CNN, GRU, and multi-head self-attention pooling) consistently achieved the lowest validation loss and highest test generalisation ($R^2 = 0.8849 \pm 0.0059$, RMSE $= 0.6792 \pm 0.0174$ t/ha), and was therefore selected as the primary CNN-A-GRU architecture (hereafter referred to simply as CNN-A-GRU) for all subsequent experiments.

The baseline predictive capability of the flattened tabular approach is visualised in Figure 7, where predictions from a representative cross-validation repetition ($R^2 = 0.9008$, RMSE $= 0.6309$ t/ha) tightly cluster along the ideal diagonal across the primary yield distribution (roughly 4 to 10 t/ha), with slight heteroscedasticity emerging only at the extreme upper and lower yield limits. Across the full 10-repetition baseline evaluation, Random Forest (RF) achieved a global mean R^2 of 0.9006 ± 0.0005 ($\text{CI}_{95\%} : [0.9002, 0.9010]$, RMSE $= 0.6314 \pm 0.0016$ t/ha, MAE $= 0.4430$ t/ha), statistically significantly outperforming both CNN-A-GRU ($R^2 = 0.8849 \pm 0.0059$, $\text{CI}_{95\%} : [0.8807, 0.8891]$, MAE $= 0.4762$ t/ha; Welch's $t = 8.38, p = 1.39 \times 10^{-5}$, Cohen's $d = 3.75$) and Mamba ($R^2 = 0.8791 \pm 0.0040$, $\text{CI}_{95\%} : [0.8763, 0.8819]$, RMSE $= 0.6964 \pm 0.0115$ t/ha, MAE $= 0.4884$ t/ha; Welch's $t = 16.87, p = 2.82 \times 10^{-8}$, Cohen's $d = 7.54$). Furthermore, CNN-A-GRU significantly outperformed Mamba under the baseline within-site protocol (Welch's $t = 2.57, p = 0.0206$, Cohen's $d = 1.15$).

5.2 Impact of Temporal Selection Strategies

CTSCP identified seven consistent timestamps centred on DOYs 70, 110, 130, 150, 170, 195 and 205. As shown in Table 2, discrete phenological alignment via CTSCP achieved a baseline R^2 of 0.9006 ± 0.0005 for RF, 0.8849 ± 0.0059 for CNN-A-GRU, and 0.8791 ± 0.0040 for Mamba. In contrast, 90-day seasonal averaging significantly degraded predictive performance across all three architectures, reducing RF to $R^2 = 0.8781 \pm 0.0005$ ($\Delta R^2 = -0.0225$; Welch's $t = 100.62, p = 3.24 \times 10^{-26}$), CNN-A-GRU to $R^2 = 0.8627 \pm 0.0020$ ($\Delta R^2 = -0.0222$; $t = 11.27, p = 2.14 \times 10^{-7}$), and Mamba to $R^2 = 0.8574 \pm 0.0015$ ($\Delta R^2 = -0.0217$; $t = 16.06, p = 3.18 \times 10^{-9}$). These consistent drops confirm that both tabular and sequence-based models lose critical

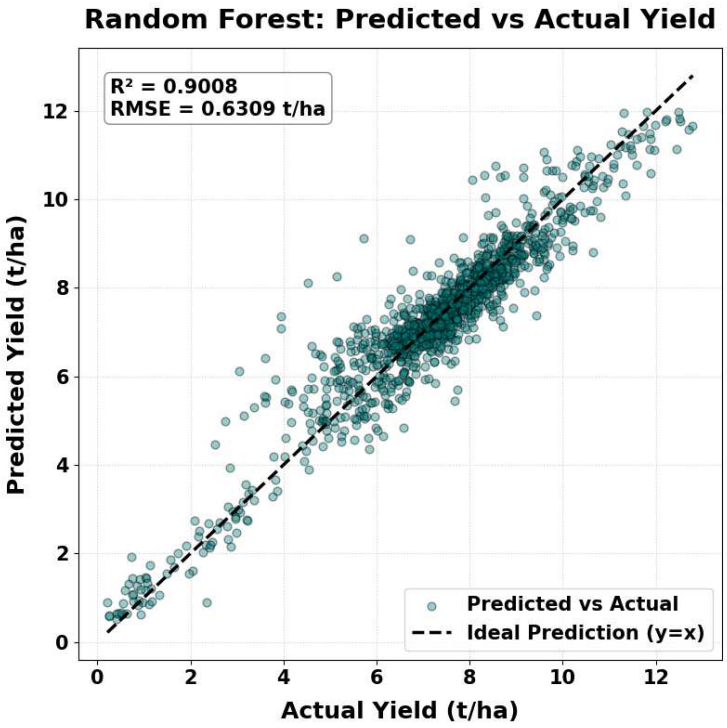


Fig. 7 Actual versus predicted crop yield (t/ha) using the optimal Random Forest architecture on the baseline dataset. The metrics displayed within the figure (R^2 = 0.9008, RMSE = 0.6309 t/ha) represent a single representative cross-validation fold (Rep 1). This closely mirrors the 10-repetition global average (R^2 = 0.9006, RMSE = 0.6314 t/ha). Predictions tightly cluster along the ideal 1:1 diagonal (dashed line), indicating robust baseline performance across the primary yield distribution prior to any generative augmentation.

predictive information when seven discrete intra-seasonal passes are compressed into three coarse 90-day calendar composites.

5.3 Spectral and Climate Feature Selection

SHAP rankings revealed a marked structural divergence between tabular and sequential architectures (Figure 8). RF placed the greatest weight on cumulative and seasonal thermal/hydrological summaries (`cum_rain`, `gdd`, `min_temp`, and `max_temp`) while ranking daily rainfall (`day_rain`) last when aggregated across the season. Conversely, CNN-A-GRU exploited time-specific sequential context, isolating daily precipitation at timestamp 5 (`day_rain_t5`, corresponding to mid-June anthesis/grain-filling around DOY 170) as its most influential directional feature, followed by mid-to-late season NIR (`B08_t4`, `B08_t3`, `B8A_t4`), SWIR (`B11_t1`), and Red Edge (`B05`, `B07`, `B06`) reflectance passes (Frampton et al. 2013).

Systematic 10-repetition feature pruning from $K = 20$ down to $K = 1$ (Table 3) showed that removing low-importance temporal metadata (`das`, `cos_doy`, `sin_doy`,

Table 2 Comparison of CTSCP (Baseline) against 90-day Seasonal Averaging using R^2 (Mean $\pm\sigma$) across $n = 10$ repetitions, with two-sided Welch's t-test statistics (t, p-value) quantifying seed-level separability on the fixed split.

Strategy / Metric	**RF (R^2 $\pm\sigma$)**	**CNN-A-GRU (R^2 $\pm\sigma$)**	**Mamba (R^2 $\pm\sigma$)**
CTSCP (Baseline)	$\mathbf{0.9006 \pm 0.0005}$	$\mathbf{0.8849 \pm 0.0059}$	$\mathbf{0.8791 \pm 0.0040}$
Seasonal Averaging	0.8781 ± 0.0005	0.8627 ± 0.0020	0.8574 ± 0.0015
Welch's t-statistic	$t = 100.62$	$t = 11.27$	$t = 16.06$
p-value (Two-sided)	$p = 3.24 \times 10^{-26}$	$p = 2.14 \times 10^{-7}$	$p = 3.18 \times 10^{-9}$

`time_delta`) and collinear visible/Red-Edge bands slightly improved test generalisation across all three architectures: RF peaked at $K = 18$ ($R^2 = 0.9011 \pm 0.0005$), CNN-A-GRU peaked at $K = 14$ ($R^2 = 0.8908 \pm 0.0044, p = 0.0223$ vs. $K = 20$), and Mamba peaked at $K = 12$ and $K = 10$ ($R^2 = 0.8851 \pm 0.0025, p = 0.0010$ vs. $K = 20$). All three models maintained near-baseline accuracy across $K = 7$–10 (at $K = 9$, retaining `cum_rain`, `max_temp`, `day_rain`, `min_temp`, `avg_temp`, `B08`, `B11`, `B12`, and `B05`: RF $R^2 = 0.8998 \pm 0.0006$, CNN-A-GRU 0.8844 ± 0.0046, and Mamba 0.8830 ± 0.0030) and remained viable down to $K = 6$ ($R^2 = 0.8635$–0.8767), where NIR reflectance (`B08`) was the sole remaining satellite band alongside the five meteorological variables. Removing `B08` at $K = 5$ caused an immediate, simultaneous drop across all three architectures ($R^2 = 0.7014$ for RF, 0.7025 for CNN-A-GRU, and 0.6999 for Mamba; $p < 10^{-16}$): without at least one pixel-resolved spectral band, models could no longer distinguish within-field spatial variability and predicted only field-year meteorological means.

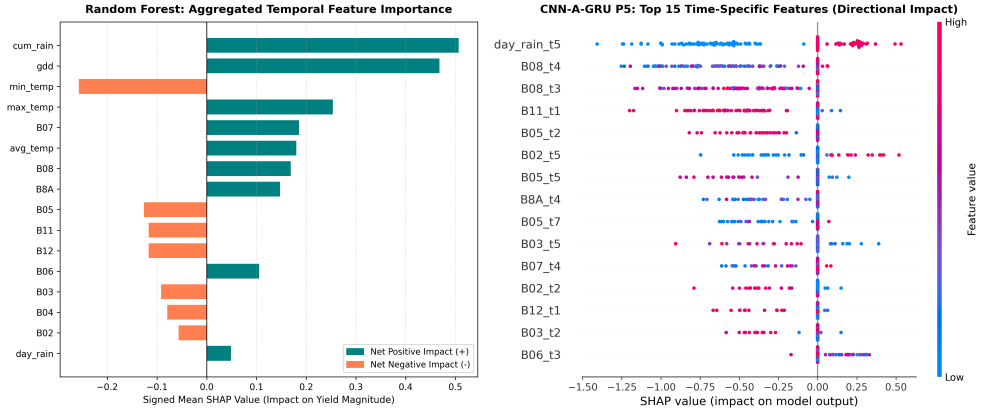


Fig. 8 Comparison of SHAP feature importance. **Left (RF):** The flattened tabular architecture relies predominantly on seasonal thermal and hydrological summaries (`cum_rain`, `gdd`, `min_temp`, `max_temp`), ranking daily rainfall (`day_rain`) last. **Right (CNN-A-GRU):** The sequence model resolves time-specific phenological drivers, prioritising mid-to-late season precipitation (`day_rain_t5` at DOY 170) alongside key NIR (`B08_t4`, `B08_t3`, `B8A_t4`), SWIR (`B11_t1`), and Red Edge (`B05`, `B07`, `B06`) reflectance passes.

Table 3 Feature pruning sensitivity analysis under the CTSCP strategy across $n = 10$ repetitions (Mean $\pm\sigma$). Subsets retain the top K variables ranked by training-set impurity importance.

	Random Forest		CNN-A-GRU		Mamba	
Feature Subset (K)	**R^2 ($\pm\sigma$)**	**RMSE**	**R^2 ($\pm\sigma$)**	**RMSE**	**R^2 ($\pm\sigma$)**	**RMSE**
20 (All Features Baseline)	0.9006 ± 0.0005	0.6314	0.8849 ± 0.0059	0.6792	0.8791 ± 0.0040	0.6964
18 (RF Peak: −das, −cos_doy)	$\mathbf{0.9011 \pm 0.0005}$	**0.6298**	0.8889 ± 0.0035	0.6674	0.8795 ± 0.0054	0.6951
15 (−sin_doy, −B03, −B8A)	0.9008 ± 0.0004	0.6310	0.8844 ± 0.0040	0.6810	0.8785 ± 0.0036	0.6981
14 (CNN Peak: −B02)	0.9009 ± 0.0004	0.6306	$\mathbf{0.8908 \pm 0.0044}$	**0.6618**	0.8811 ± 0.0034	0.6906
12 (Mamba Peak: −B06)	0.9010 ± 0.0003	0.6301	0.8880 ± 0.0047	0.6701	$\mathbf{0.8851 \pm 0.0033}$	**0.6789**
10 (−B07, −B04)	0.8999 ± 0.0003	0.6337	0.8858 ± 0.0045	0.6769	$\mathbf{0.8851 \pm 0.0025}$	**0.6789**
9 (−gdd; 5 Met. + 4 Spec.)	0.8998 ± 0.0006	0.6339	0.8844 ± 0.0046	0.6809	0.8830 ± 0.0030	0.6850
8 (−B05 Red Edge)	0.8980 ± 0.0006	0.6397	0.8834 ± 0.0049	0.6837	0.8841 ± 0.0041	0.6818
7 (−B12 SWIR)	0.8942 ± 0.0006	0.6514	0.8815 ± 0.0038	0.6893	0.8800 ± 0.0054	0.6938
6 (−B11; 5 Met. + B08)	0.8767 ± 0.0005	0.7033	0.8657 ± 0.0056	0.7339	0.8635 ± 0.0047	0.7399
5 (−B08; 5 Met. Only)	0.7014 ± 0.0001	1.0945	0.7025 ± 0.0017	1.0925	0.6999 ± 0.0040	1.0972
2 (cum_rain, max_temp)	0.7014 ± 0.0001	1.0945	0.7021 ± 0.0008	1.0932	0.6995 ± 0.0036	1.0979
1 (cum_rain Only)	0.7014 ± 0.0001	1.0945	0.6966 ± 0.0066	1.1032	0.6813 ± 0.0124	1.1306

5.4 Data Augmentation

5.4.1 Gaussian Noise Baseline

Across the 27-configuration single-seed screening grid search (seed = 42), isotropic Gaussian noise provided only a negligible uptick for RF, moving R^2 from the 10-repetition baseline of 0.9006 ± 0.0005 to 0.9009 and RMSE from 0.6314 ± 0.0016 to 0.6306 t/ha with 2.5% noise intensity applied at 50% volume to minority-bin samples (Table 4). Conversely, for the sequence models, even the least disruptive Gaussian noise configurations failed to improve upon the 10-repetition baseline beyond natural seed-to-seed variance ($R^2 = 0.8866$, RMSE = 0.6745 t/ha vs. baseline $R^2 = 0.8849 \pm 0.0059$, RMSE $= 0.6792 \pm 0.0174$ t/ha for CNN-A-GRU, within 1σ; and degrading from $R^2 = 0.8791 \pm 0.0040$, RMSE $= 0.6964 \pm 0.0115$ t/ha to $R^2 = 0.8777$, RMSE = 0.7003 t/ha for Mamba), indicating that unstructured random jitter does not resolve sequential tail scarcity.

5.5 Generative Ablation Study

To quantify T-CGAN constraint impacts, an ablation study evaluated RF, CNN-A-GRU and Mamba models across progressively constrained generative architectures: a Direct Generator (Model 1, ablating the Encoder), a Stratified Autoencoder (Model 2), variants isolating Total Variation (Model 3) or Soft Dynamic Time Warping (Model 4), and the Proposed T-CGAN combining all constraints with regime-stratified weights (Model 5).

While global metrics (Table 5) confirm that synthetic augmentation leaves overall within-site R^2 statistically unchanged ($p \geq 0.258$), they mask localised impacts at the distribution tails. Tracking bin-wise RMSE as synthetic additions scale from 0% to 25%, 50%, and 100% of the real training sample count within each target minority bin (Figure 9 and Table 6) reveals distinct, architecture-dependent responses in the extreme yield regimes.

Table 4 Summary of optimal isotropic Gaussian noise configurations from the 27-combination single-seed screening grid search (seed = 42) compared against the 10-repetition unaugmented baseline (Mean $\pm\,\sigma$ in t/ha for RMSE). Grid parameters denote: (i) **Target subset** (*Global*: noise applied across all training samples; *Minority*: noise applied strictly to the sparse extreme yield bins, 0–4 and 10–14 t/ha); (ii) **Volume (Vol)**: percentage of noisy synthetic copies appended relative to the real training sample count in the target subset $(10\%, 25\%, 50\%)$; and (iii) **Intensity (Int)**: standard deviation of the zero-mean Gaussian perturbation $\mathcal{N}(0, \sigma^2_{\text{noise}})$ scaled as a percentage of each feature's training-set standard deviation $(2.5\%, 5.0\%, 10.0\%)$. Because the 27-configuration grid sweep was conducted as a single-seed screening pass (seed = 42) and even the best augmented point estimates remained within 1σ of (or degraded below) the 10-repetition baseline, full 10-seed evaluation was reserved for the generative T-CGAN ablation study (Tables 5–6).

		R^2		RMSE (t/ha)	
Model	**Optimal Grid Config.**	**Base ($\pm\sigma$)**	**Aug.**	**Base ($\pm\sigma$)**	**Aug.**
RF	Minority, 50% Vol, 2.5% Int	0.9006 ± 0.0005	0.9009	0.6314 ± 0.0016	0.6306
CNN-A-GRU	Global, 25% Vol, 10.0% Int	0.8849 ± 0.0059	0.8866	0.6792 ± 0.0174	0.6745
Mamba	Global, 50% Vol, 5.0% Int	0.8791 ± 0.0040	0.8777	0.6964 ± 0.0115	0.7003

Table 5 Global predictive performance across the generative ablation study and augmentation volume sweep ($n = 10$ repetitions). Values denote R^2 (Mean $\pm\sigma$). Constrained T-CGAN augmentation preserves global manifold stability across all three architectures ($p \geq 0.258$). Bold values denote the highest mean R^2 in each column.

Tier / Vol	**Configuration**	**RF**	**CNN-A-GRU**	**Mamba**
Baseline (0%)	Real Data Only	$\mathbf{0.9006 \pm 0.0005}$	0.8849 ± 0.0059	0.8791 ± 0.0040
Model 1	Direct GAN (GAN 2)	0.9003 ± 0.0007	0.8863 ± 0.0044	0.8791 ± 0.0046
Model 2	Autoencoder (GAN 3)	0.9003 ± 0.0005	$\mathbf{0.8869 \pm 0.0059}$	0.8798 ± 0.0056
Model 3	Autoencoder + TV	0.9002 ± 0.0005	0.8856 ± 0.0047	$\mathbf{0.8811 \pm 0.0037}$
Model 4	Autoencoder + DTW	0.8998 ± 0.0005	0.8867 ± 0.0045	0.8764 ± 0.0053
Model 5 (25%)	Proposed T-CGAN	$\mathbf{0.9006 \pm 0.0006}$	0.8849 ± 0.0041	0.8798 ± 0.0032
Model 5 (50%)	Proposed T-CGAN	0.9005 ± 0.0006	0.8836 ± 0.0039	0.8779 ± 0.0053
Model 5 (100%)	Proposed T-CGAN	0.9003 ± 0.0007	0.8837 ± 0.0046	0.8791 ± 0.0053

Evaluating the 10-repetition statistics at the distribution margins reveals three key patterns (Figure 10 and Table 6). First, Random Forest extracted statistically significant boundary refinement from the T-CGAN trajectories in both extreme regimes: at 100% augmentation volume, RF RMSE in the catastrophic failure bin (0–2 t/ha) dropped by 6.75% from 0.409 ± 0.007 to 0.382 ± 0.008 t/ha (exact mean 0.3815 t/ha; Welch's $t = -8.40, p < 0.0001$; paired $p = 0.0001$, Cohen's $d = -3.76$), while in the peak productivity bin (12–14 t/ha), RMSE dropped from 0.866 ± 0.015 to 0.839 ± 0.016 t/ha at 50% volume ($t = -3.77, p = 0.0014$) and 0.843 ± 0.006 t/ha at 100% volume ($t = -4.40, p = 0.0009$; paired $p = 0.0037, d = -1.97$).

Second, for CNN-A-GRU, the primary utility of T-CGAN augmentation occurred in the canopy-saturated peak productivity regime (12–14 t/ha), where dense leaf area index (LAI) masks late-season biomass accumulation (Perich et al. 2023). CNN-A-GRU's unaugmented baseline exhibited elevated error in this regime (RMSE = 1.008 ± 0.086 t/ha), which decreased by 12.69% to 0.880 ± 0.168 t/ha at 50% T-CGAN

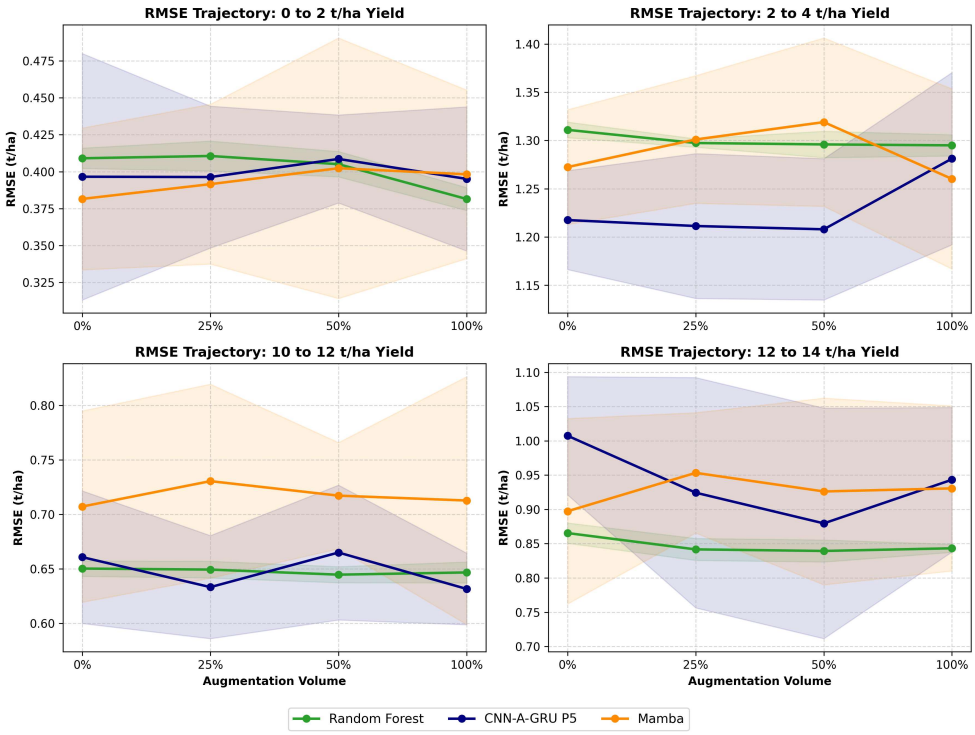


Fig. 9 RMSE trajectory (Mean $\pm 1\sigma$ shaded bands across $n = 10$ seeds) across target minority yield bins as Proposed T-CGAN synthetic volume scales from 0% to 100%.

volume (borderline significant under paired testing across seeds at $p = 0.0495$, and $p = 0.0513$ under Welch's test; $d = -0.96$), though this gain was non-monotonic at 100% volume (0.943 ± 0.105 t/ha, $p = 0.1539$). In the 0–2 t/ha bin, while CNN-A-GRU's mean RMSE reduction was minimal ($0.397 \rightarrow 0.395$ t/ha at 100% volume, and 0.385 t/ha under Model 4), T-CGAN augmentation substantially stabilised training across random initialisations, reducing seed-to-seed RMSE standard deviation by 41.5% (from $\sigma = 0.083$ to 0.049 t/ha at 100% volume, and 0.030 t/ha at 50% volume).

Third, Mamba exhibited the lowest unaugmented baseline error among the sequence models in both extreme bins (0.382 ± 0.048 t/ha at 0–2 t/ha and 0.897 ± 0.135 t/ha at 12–14 t/ha), and within-site synthetic augmentation did not further reduce its extreme-bin RMSE ($p \geq 0.289$), whereas its primary generative gains emerged under contiguous spatial block extrapolation (Section 5.6).

5.6 Spatial Generalisation: Block Cross-Validation

To evaluate out-of-distribution geographic extrapolation across all 20,170 pixels while ensuring every harvest season is represented in each partition despite annual crop rotation, 5-fold contiguous spatial block cross-validation was conducted within each harvest year (Figures 5 and 6). As summarised in Table 7, evaluating models

Table 6 Bin-wise Root Mean Square Error (RMSE, Mean $\pm\sigma$ in t/ha across $n = 10$ repetitions; **lower values indicate better accuracy**) for extreme yield regimes across Proposed T-CGAN augmentation volumes (0%, 25%, 50%, 100%) and intermediate ablation tiers. Because narrow 2 t/ha yield bins have restricted target variance that makes within-bin R^2 ill-conditioned, performance is reported strictly as physical error in t/ha (ranging from 0.382 to 1.008 t/ha). Bold values denote the lowest RMSE in each column, and asterisks denote statistically significant error reductions relative to the 0% baseline ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

Part A: Catastrophic Failure Regime (Lowest Yields: 0–2 t/ha) — RMSE (t/ha)

Tier / Vol	Configuration	RF RMSE	CNN-A-GRU RMSE	Mamba RMSE
Baseline (0%)	Real Data Only	0.409 ± 0.007	0.397 ± 0.083	$\mathbf{0.382 \pm 0.048}$
Model 2	Autoencoder (GAN 3)	0.405 ± 0.007	0.397 ± 0.061	0.412 ± 0.055
Model 3	Autoencoder + TV	0.403 ± 0.008	0.413 ± 0.058	0.396 ± 0.052
Model 4	Autoencoder + DTW	0.402 ± 0.007	$\mathbf{0.385 \pm 0.054}$	0.386 ± 0.049
Model 5 (25%)	Proposed T-CGAN	0.411 ± 0.010	0.396 ± 0.048	0.392 ± 0.054
Model 5 (50%)	Proposed T-CGAN	0.405 ± 0.009	0.409 ± 0.030	0.402 ± 0.088
Model 5 (100%)	Proposed T-CGAN	$\mathbf{0.382 \pm 0.008}^{***}$	0.395 ± 0.049	0.398 ± 0.057

Part B: Canopy Saturation Regime (Peak Productivity: 12–14 t/ha) — RMSE (t/ha)

Tier / Vol	Configuration	RF RMSE	CNN-A-GRU RMSE	Mamba RMSE
Baseline (0%)	Real Data Only	0.866 ± 0.015	1.008 ± 0.086	$\mathbf{0.897 \pm 0.135}$
Model 2	Autoencoder (GAN 3)	0.862 ± 0.014	0.880 ± 0.142	0.973 ± 0.118
Model 3	Autoencoder + TV	0.864 ± 0.015	0.907 ± 0.135	0.985 ± 0.122
Model 4	Autoencoder + DTW	0.861 ± 0.014	0.947 ± 0.128	0.943 ± 0.115
Model 5 (25%)	Proposed T-CGAN	$0.842 \pm 0.016^{**}$	0.924 ± 0.168	0.953 ± 0.088
Model 5 (50%)	Proposed T-CGAN	$\mathbf{0.839 \pm 0.016}^{**}$	$\mathbf{0.880 \pm 0.168}^{*}$	0.926 ± 0.136
Model 5 (100%)	Proposed T-CGAN	$0.843 \pm 0.006^{***}$	0.943 ± 0.105	0.931 ± 0.120

on geographically isolated blocks induced a substantial generalisation gap relative to coordinate-grouped within-field partitioning, though all architectures converged strongly on the training partitions (Train $R^2 = 0.841$–0.990).

Contiguous spatial evaluation reversed the architectural hierarchy—not only relative to the 8,581-pixel within-field benchmark, but directly between the training and test partitions of the same 20,170-pixel spatial block dataset (Table 7, Part B). Within the 20,170-pixel spatial folds, Random Forest achieved the highest training fit (Train $R^2 = 0.984 \pm 0.003$, Train RMSE = 0.237 t/ha) alongside dominating coordinate-grouped within-field interpolation ($R^2 = 0.9006$), yet degraded to the lowest accuracy on geographically separated test blocks (Test $R^2 = 0.329 \pm 0.363$, failing in Fold 5 at $R^2 = -0.313$ where distributional divergence reached $KS = 0.2438$), indicating high sensitivity to spatially structured distribution shift. Conversely, despite lower training-block fit (Train $R^2 = 0.841$–0.855), the regularised deep sequence models generalised substantially better on average across unseen blocks within the farm: on real data alone, CNN-A-GRU achieved a mean Test R^2 of 0.444 ± 0.203 (reaching $R^2 = 0.611$ in Fold 2 and 0.623 in Fold 4) and Mamba achieved 0.434 ± 0.223 ($R^2 = 0.643$ in Fold 2).

Furthermore, fold-isolated T-CGAN augmentation increased mean spatial block R^2 for CNN-A-GRU from 0.444 to 0.485 ± 0.185 and reduced Test RMSE from 1.303

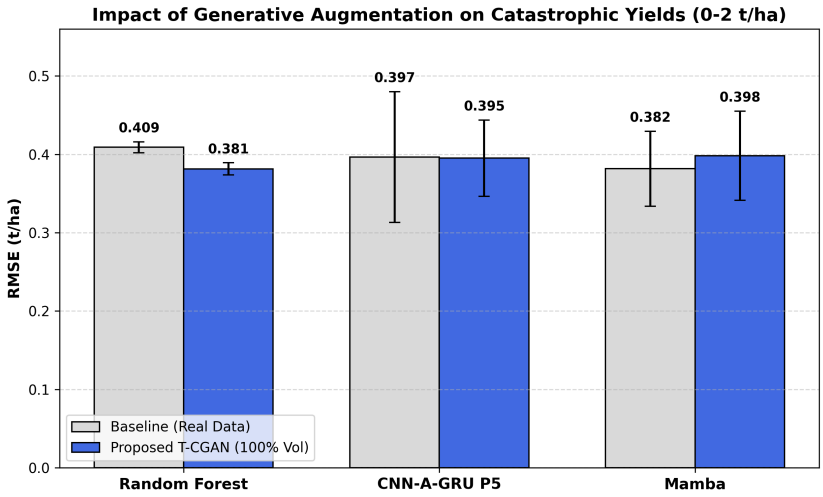


Fig. 10 Direct comparison of baseline (0%) versus fully augmented (100% volume) RMSE (Mean $\pm$ 1σ) within the catastrophic low-yield bin (Bin 0: 0–2 t/ha). Proposed T-CGAN augmentation significantly reduces RF error (0.4091 $\rightarrow$ 0.3815 t/ha, displayed as 0.381 under half-to-even plot rounding and 0.382 t/ha in Table 6; $p < 0.0001$) and compresses CNN-A-GRU seed-to-seed variance by 41.5% (σ : 0.083 $\rightarrow$ 0.049 t/ha).

to 1.259 t/ha, with improvements in four of the five geographic blocks—most notably Fold 2 (0.611 $\rightarrow$ 0.691) and the high-divergence Fold 5 (0.181 $\rightarrow$ 0.281). While this difference was not statistically significant across the $n = 5$ independent fold means alone (paired $t = 2.032, p = 0.112$), the stronger result across the $n = 25$ clustered fold-seed evaluations (paired $t = 2.842, p = 0.0090$; Wilcoxon $p = 0.0173$) indicates consistent seed-level regularisation across repeated runs within those four blocks rather than 25 independent geographic replications. For Mamba, T-CGAN augmentation similarly lifted mean Test R^2 from 0.434 to 0.464 $\pm$ 0.175, largely by rescuing Fold 5 (0.069 $\rightarrow$ 0.236; fold-mean $p = 0.342$, $n = 25$ paired $t = 1.388, p = 0.178$). Compared against Random Forest ($R^2 = 0.333 \pm 0.360$), both T-CGAN-augmented CNN-A-GRU ($n = 25$ paired $t = 3.163, p = 0.0042$) and Mamba ($t = 2.826, p = 0.0094$) achieved higher mean spatial generalisation, primarily by avoiding the Fold 5 failure exhibited by RF ($R^2 = -0.321$).

5.7 Error Stratification Across Yield Distributions

Under the coordinate-grouped 70/15/15 within-field split ($N = 8,581$ pixels; $n = 10$ repetitions) at 100% Proposed T-CGAN augmentation volume (Figure 11), test RMSE across the full yield distribution was highest in the mid-low transition bins of 2–4 t/ha (1.26–1.30 t/ha) and 4–6 t/ha (0.98–1.07 t/ha) across all three architectures, compared with 0.48–0.60 t/ha in the dense central distribution (6–10 t/ha). In contrast, the regularised 0–2 t/ha failure regime maintained substantially lower absolute RMSE (0.382–0.398 t/ha across architectures). This elevated error in the 2–6 t/ha transition regime arises because moderate-to-low yields at Witzwil frequently result from

Table 7 5-Fold Within-Year Spatial Block Cross-Validation across all 20,170 pixels ($n = 25$ evaluations per row: 5 folds × 5 seeds). Spatial partitioning into training (60%), validation (20%), and test (20%) blocks is performed independently for each harvest year's active wheat fields, with repeated coordinates locked to the same block across seasons so that no coordinate in any year's training set ever appears in another year's test set. Reported test metrics represent averages across all five harvest seasons (2017–2021) in which each test pixel is weighted equally. **Part A** reports fold-level Test R^2 (averaged over 5 seeds) alongside Train–Test Kolmogorov–Smirnov divergence (KS). **Part B** reports 5-fold aggregate Training and Test metrics (Mean $\pm\sigma$ across the 5 spatial folds). Bold values denote the best test performance in each column.

Part A: Fold-Wise Spatial Test R^2 (and Fold Train–Test KS Divergence)

Model	Strategy	Fold 1 ($KS = 0.13$)	Fold 2 ($KS = 0.11$)	Fold 3 ($KS = 0.27$)	Fold 4 ($KS = 0.13$)	Fold 5 ($KS = 0.24$)
Random Forest	Real Only	0.235	**0.716**	0.505	0.502	−0.313
Random Forest	T-CGAN Aug.	0.260	0.654	0.514	0.560	−0.321
CNN-A-GRU	Real Only	0.262	0.611	0.541	**0.623**	0.181
CNN-A-GRU	T-CGAN Aug.	0.284	0.691	0.548	0.622	**0.281**
Mamba	Real Only	**0.344**	0.643	**0.553**	0.562	0.069
Mamba	T-CGAN Aug.	0.332	0.648	0.543	0.561	0.236

Part B: 5-Fold Aggregate Performance Summary

		Training Partition		Spatial Test Partition		
Model	**Strategy**	**Mean R^2**	**RMSE**	**Test R^2 ($\pm\sigma$)**	**RMSE**	**MAE**
Random Forest	Real Only	0.984	0.237	0.329 ± 0.363	1.383	1.028
Random Forest	T-CGAN Aug.	0.990	0.225	0.333 ± 0.360	1.377	1.027
CNN-A-GRU	Real Only	0.841	0.756	0.444 ± 0.203	1.303	0.958
CNN-A-GRU	**T-CGAN Aug.**	0.902	0.704	$\mathbf{0.485 \pm 0.185}$	**1.259**	**0.923**
Mamba	Real Only	0.855	0.723	0.434 ± 0.223	1.305	0.958
Mamba	T-CGAN Aug.	0.896	0.724	0.464 ± 0.175	1.287	0.930

late-season lodging or localised hail damage (particularly in 2021) occurring after canopies have already developed normal mid-season vegetative vigour (high NIR/Red-Edge reflectance through DOY 150–170), creating spectral ambiguity between healthy and partially damaged stands. By contrast, a global within-field RMSE of 0.63 t/ha ($\approx 10\%$ relative error against the 6.29 t/ha site mean) is well below the ~ 1.92 t/ha natural field variability, providing sufficient precision for delineating within-field management zones when local calibration data are available.

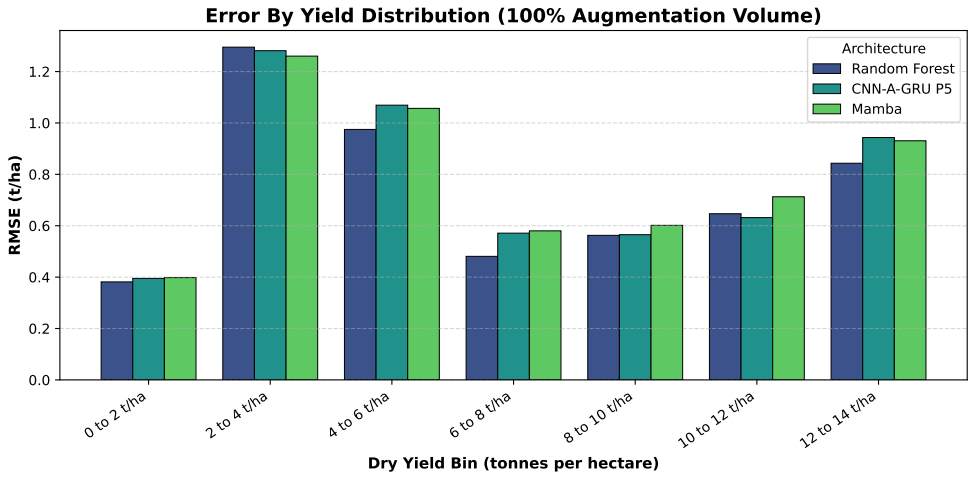


Fig. 11 Error stratification across all seven yield bins (RMSE in t/ha, averaged across $n = 10$ random seeds) under the coordinate-grouped 70/15/15 within-field split ($N = 8,581$ pixels) at 100% Proposed T-CGAN augmentation volume. The grouped bar chart highlights elevated error in the mid-low transition bins (2–4 and 4–6 t/ha) across all architectures, contrasted against lower absolute error in the dense central distribution (6–10 t/ha) and the regularised catastrophic failure margin (0–2 t/ha).

6 Discussion

6.1 Generative Augmentation and the Fidelity–Diversity Tradeoff

Rather than acting as a universal performance enhancer, T-CGAN augmentation produces selective, data- and architecture-dependent improvements governed by the amount of available training data and how each model family processes temporal sequences, based on the differences in the level of train data in each bin and the observed behaviour of models in Figure 9. Global within-site R^2 remains unchanged across all architectures ($p \geq 0.258$) because central yield bins (4–10 t/ha) are already densely populated. At the sparse distribution tails, Random Forest benefits most directly in both extreme regimes (0–2 and 12–14 t/ha) because it evaluates flattened feature-timestamp combinations independently and is therefore insensitive to minor sequential discontinuities. Conversely, for CNN-A-GRU, within-site synthetic utility is concentrated in reducing peak-yield (12–14 t/ha) error and compressing 0–2 t/ha initialisation variance, whereas Mamba already achieves the lowest unaugmented sequence error in both tails and shows no further within-site tail reduction from synthetic data, benefiting instead when extrapolated across contiguous spatial blocks (Table 7). Notably, the ablation tiers (Table 6) show that while individual constraints such as Soft-DTW (Model 4 at 0–2 t/ha) or the Autoencoder bottleneck alone

(Model 2 at 12–14 t/ha) can aid CNN-A-GRU in isolated bins, the full composite T-CGAN objective (Model 5) is required to simultaneously deliver tail error reductions for RF across both extremes while preventing global degradation in sequence models.

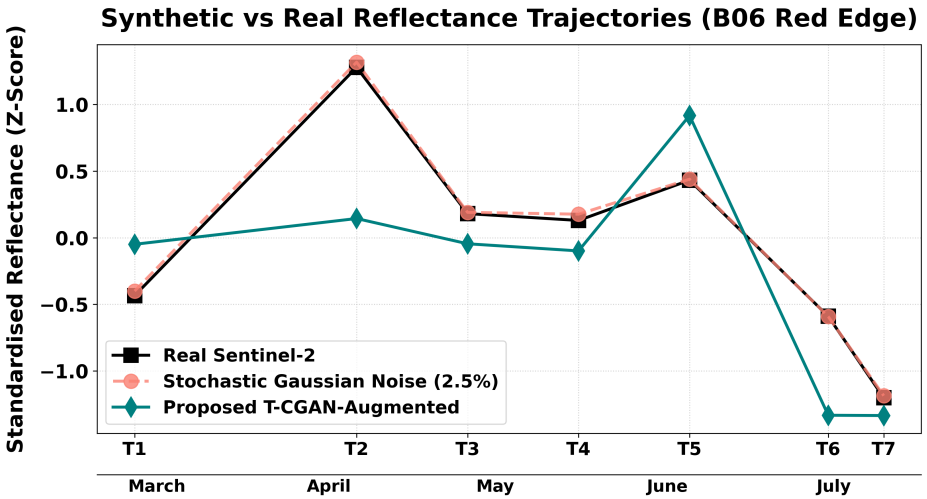


Fig. 12 Comparison of generated versus real temporal trajectories for the B06 (Red Edge) band within the catastrophic failure yield regime (0–2 t/ha). Reflectance values are plotted as standardised Z-scores, where 0 represents the training-set reference mean for a given timestamp, and positive or negative values indicate reflectance above or below that reference mean, respectively. Whereas 2.5% Gaussian jitter remains tethered to the single reference pixel, the Proposed T-CGAN generates novel trajectories conditioned on the target yield regime (0–2 t/ha) and sampling metadata rather than reconstructing a single test pixel, following a distinct path within the low-yield empirical distribution while maintaining piecewise smoothness between passes.

Figure 12 illustrates why unconstrained noise fails for sequence architectures within the low-yield regime (0–2 t/ha). Isotropic Gaussian jitter (2.5%) merely perturbs an existing Sentinel-2 observation without expanding the underlying empirical distribution. In contrast, T-CGAN latent mixup generates distinct multi-pass trajectories conditioned on the target yield regime and sampling metadata while preserving piecewise temporal smoothness between passes. Because Red Edge (`B06`) and NIR (`B08`) reflectance are jointly influenced by multiple interacting biophysical factors—including chlorophyll content, canopy architecture, soil background, and phenological stage (Frampton et al. 2013)—differences between real and synthetic Z-score trajectories represent statistical variations within the empirical reflectance manifold rather than validated physiological changes. From a sequence-modelling perspective, avoiding artificial step-to-step jitter in these reflectance trajectories prevents erratic state updates in recurrent and state-space layers.

6.2 Spatiotemporal and Biophysical Feature Optimisation

Seven-pass CTSCP phenological sampling consistently outperformed coarse three-period seasonal averaging across all evaluated architectures. Averaging across 90-day

windows smooths critical developmental inflection points and penalises sequence models—including selective state-space architectures like Mamba—that rely on precise trajectory shape. Retaining phenologically aligned discrete observations is therefore essential when modelling irregular optical time series. While CTSCP outperformed seasonal averaging, two structural factors contribute to this gap: first, CTSCP retains seven discrete timestamps compared with three broad composites; second, the third calendar quadrant (Q3: DOY 181–270) extends beyond typical late-July harvest dates (DOY $\approx$ 205), incorporating post-harvest stubble reflectance that dilutes late-season grain-filling signals. Future work will isolate pure alignment effects by comparing phenologically aligned and unaligned sequences of identical length truncated strictly at harvest.

Moreover, the 20 input variables are not statistically independent: neighbouring Sentinel-2 bands (e.g., Red Edge `B05`–`B07`, NIR `B08`/`B8A`, and SWIR `B11`/`B12`) exhibit strong spectral collinearity, while thermal and precipitation metrics covary across the season (Frampton et al. 2013; Perich et al. 2023). This inter-feature redundancy explains why pruning from 20 down to 10–18 features slightly improves test generalisation across all three architectures (Table 3), and why compact subsets of $K = 7$–10 variables preserve full baseline accuracy ($R^2 = 0.880$–0.900), consistent with the nine-component plateau observed in prior unsupervised manifold analyses on this site (Finnegan et al. 2026). Furthermore, the simultaneous drop of RF, CNN-A-GRU, and Mamba from $R^2 = 0.864$–0.877 at $K = 6$ to $R^2 \approx 0.700$–0.703 (RMSE $\approx$ 1.09 t/ha) at $K = 5$ reflects the transition from pixel-resolved modelling to field-year lookup: because meteorological variables are recorded at the site weather-station level and are therefore identical across all pixels within a field-year on a given pass date, removing the last spectral band (`B08`) at $K = 5$ reduces all three models to predicting field-year averages (for comparison, a simple group-mean reference calculation predicting the average yield of each of the 19 field-years achieves $R^2 = 0.6550$, RMSE = 1.1285 t/ha, whereas predicting the 5 harvest-year means achieves $R^2 = 0.3571$). Consequently, the $R^2 \approx 0.70$ plateau at $K \leq 5$ captures combined field-year identity (synoptic weather, parcel soil baseline, sowing schedule, and management across the 19 field-years), whereas at least one pixel-resolved reflectance band (`B08` NIR at $K = 6$, supplemented by SWIR `B11`–`B12` and Red Edge `B05` at $K = 9$) is required to resolve the remaining within-field spatial variability. Similarly, while SHAP rankings (Figure 8) show that CNN-A-GRU assigns high weight to `day_rain_t5` (around DOY 170) alongside mid-to-late season NIR and Red Edge passes, `day_rain_t5` is derived from the site weather station and is therefore constant across all pixels sharing a pass date, acting as an indicator of year- and field-specific mid-June storm events (such as the June 2021 convective storms) rather than pixel-specific hydrology.

6.3 Architectural Resilience to Climate Extremes

Predictive accuracy at the distribution margins carries disproportionate real-world significance. The 0–2 t/ha bin represents catastrophic canopy collapse or severe abiotic stress—such as the 2021 Witzwil hail events—that typically triggers structural interventions and crop insurance payouts (Kopp et al. 2023). Conversely, the 12–14 t/ha

range represents near-perfect genetic potential under optimal micro-climatic conditions. Successfully modelling these extremes provides far more value for agricultural risk assessment than marginal improvements in average-yield years. In a site-specific management workflow, accurately resolving the 0–2 t/ha failure tail identifies persistent waterlogged or hail-damaged sub-field zones where late-season agrochemical inputs would be economically and environmentally wasted, whereas resolving the 12–14 t/ha saturation tail delineates high-potential zones where canopy reflectance normally saturates. Although this study evaluates prediction accuracy and model stability rather than directly measuring fertiliser savings or environmental outcomes in the field, reducing error and run-to-run variance in extreme yield zones is a practical prerequisite before yield maps can reliably guide variable-rate input applications.

RF is the strongest operational model for within-field interpolation ($R^2 = 0.9006$) and provides a simple non-parametric tabular baseline, though it exhibits lower average performance when transferred across contiguous spatial blocks within the farm, potentially because it relies on static flattened feature-timestamp splits. Conversely, CNN-A-GRU and Mamba process the season as an ordered sequence: multi-head attention pooling in CNN-A-GRU weights informative late-season phenological windows, while Mamba's input-dependent step size (Δ_t) provides a theoretical mechanism for modulating recurrent state transitions ($\mathbf{A}_t = e^{\mathbf{A}\Delta_t}$) during abrupt canopy disruptions, although direct inspection of internal Δ_t activations across stress years remains a subject for future interpretability work. Although selective state-space models such as Mamba were originally designed for long-context efficiency (Gu and Dao 2024), evaluating Mamba on short, irregular phenological sequences ($T = 7$) tests whether input-dependent state gating (Δ_t) offers inductive bias advantages over recurrent gates (GRU) under spatial covariate shift; future work should benchmark these sequence models against tuned gradient-boosted decision trees (e.g., XGBoost, LightGBM) and longer full-season pass sequences.

6.4 The Interpolation Trap and Spatial Transferability

Comparing coordinate-grouped within-field partitioning ($R^2 \approx 0.90$ on the 8,581 unoccluded pixels) against contiguous spatial block cross-validation ($R^2 = 0.329$–0.485 across all 20,170 pixels)—as well as comparing training versus test partitions within the 20,170-pixel spatial folds themselves—exposes the interpolation trap inherent in spatially autocorrelated agricultural modelling (Roberts et al. 2017; Filippi et al. 2025). Under coordinate-grouped within-field splitting, Random Forest achieves high interpolation accuracy (Test $R^2 = 0.901$) because adjacent 10 m training and test pixels within the same field-year share identical meteorological forcing, sowing dates, and local soil characteristics; within the spatial block protocol, RF similarly achieves near-perfect fit on the training blocks (Train $R^2 = 0.984$) but degrades sharply on geographically separated test blocks (Test $R^2 = 0.329$, dropping to -0.313 in Fold 5). In contrast, although RF still achieves the highest accuracy in Fold 2 ($R^2 = 0.716$), CNN-A-GRU (Train $R^2 = 0.841$–0.902, Test $R^2 = 0.444$–0.485) and Mamba (Train $R^2 = 0.855$–0.896, Test $R^2 = 0.434$–0.464) perform substantially better on average when tested on unseen blocks within the same farm across the exact same 20,170-pixel folds. One possible explanation for this higher average performance is that sequence

architectures capture ordered intra-seasonal spectral transitions (alongside stronger regularisation and lower training-block overfitting than RF) rather than relying on static feature-timestamp splits, although this mechanism was not directly isolated in our experiments. Moreover, whereas fold-isolated T-CGAN augmentation yielded negligible global R^2 change under within-field splitting, it increased mean spatial block R^2 for CNN-A-GRU across four of the five geographic blocks (+0.041 mean R^2); while not statistically significant across the $n = 5$ independent fold means ($p = 0.112$), the consistent gain across all $n = 25$ fold-seed evaluations ($p = 0.0090$) suggests that synthetic minority-bin regularisation helps stabilise recurrent training under spatial distribution shift. Nevertheless, two scope limitations must be emphasised: first, while the wider Witzwil cereal archive (54,098 pixel-years across 54 cereal field-years) confirms that winter wheat was rotated across 17 parcels without consecutive wheat monoculture and occasionally alternated with other small-grain cereals, non-cereal break crops in intervening seasons are unrecorded, meaning that the spatial covariate shift across blocks ($KS = 0.11$–0.27) likely reflects both local soil/topographic variation and unmodelled preceding-crop effects; second, because all observations originate from a single peat-soil farm at Witzwil, this evaluation quantifies within-farm block transferability rather than cross-farm generalisation, confirming that operational deployment across distinct agro-ecosystems requires local calibration or domain adaptation.

7 Conclusions and Future Work

We benchmarked RF, CNN-A-GRU and Mamba for pixel-level winter wheat yield estimation at a single Swiss site, alongside temporal alignment, feature reduction and generative augmentation. Under coordinate-grouped within-field evaluation, RF achieved the highest predictive accuracy ($R^2 \approx 0.90$) and offers a simple, effective tabular baseline. Across all architectures, preserving discrete phenological passes via CTSCP consistently outperformed coarse seasonal averaging. Systematic feature pruning improved or preserved baseline accuracy down to compact subsets ($K = 7$–18) and remained viable with a single near-infrared band (`B08` at $K = 6$), before dropping sharply at $K \leq 5$ when all pixel-level spectral bands were removed.

Within the coordinate-grouped split, T-CGAN augmentation preserved global R^2 across all architectures. Its within-site utility was confined to the sparse distribution tails, significantly reducing RF error in both the 0–2 and 12–14 t/ha extremes, reducing CNN-A-GRU peak-yield error (12–14 t/ha) while compressing its 0–2 t/ha seed-to-seed variance, and leaving Mamba's within-site tail errors unchanged. Ablation results confirm that combining the autoencoder bottleneck with statistical and temporal constraints (TV and Soft-DTW) best balances tail refinement with global manifold stability, although elevated errors persist in the 2–6 t/ha transition bins.

Finally, 5-fold contiguous within-year spatial block cross-validation demonstrated that high within-field interpolation accuracy overestimates geographic transferability. Under contiguous within-farm spatial block shift, RF degraded sharply on average across unseen blocks (Test $R^2 = 0.329$), whereas CNN-A-GRU and Mamba retained stronger average within-farm spatial generalisation (reaching mean Test $R^2 = 0.485$ and 0.464, respectively, with fold-isolated T-CGAN regularisation). Future work will

test leave-one-year-out and domain-adaptation strategies, incorporate multi-annual climate and rotation history, align observations on thermal time rather than calendar days, evaluate synthetic sequences directly (e.g., train-on-synthetic, test-on-real), and inject agronomically-informed hydrological proxies (e.g., Hargreaves-Samani ET_0) to constrain late-season drought dynamics.

Acknowledgements. All authors wish to thank Gregor Perich and the team who collected the Swiss data and made it available.

Declarations

- Funding: Rosie Finnegan was funded by Swansea University Research Excellence Scholarship RS228-SURES8.
- Competing interests: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
- Ethics approval and consent to participate: Not applicable
- Consent for publication: Not applicable
- Data availability: The Sentinel-2 MSI data utilised in this study are available through the Copernicus Data Space Ecosystem. The Swiss yield and climate datasets were kindly provided by the original authors (Perich et al. 2023) via the ETH Zürich Research Collection repository [https://doi.org/10.3929/ethz-b-000596665].
- Materials availability: Not applicable
- Code availability: The PyTorch and Scikit-Learn implementation code developed for this study will be made publicly available upon publication.
- Author contribution: Rosie Finnegan: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Writing – original draft, Visualisation. Sara Sharifzadeh: Securing of funding, Conceptualisation, Supervision, Methodology, Writing – review and editing. Alberto Hornero, Hamidreza Eskandari, Xianghua Xie: Supervision, Methodology, Writing – review and editing.

Appendix A CNN-A-GRU Pipeline Configurations

Across all variants, 1D-CNN layers extract spectral–spatial feature representations while GRU blocks model temporal sequential dynamics:

P1 (Sequential): Sequential CNN feature extraction feeding into a GRU, followed by single-head Attention pooling.

P2 (Parallel Late Fusion): Parallel branches where the CNN extracts spectral features and the GRU processes the temporal sequence; outputs are concatenated before final prediction.

P3 (GRU-First): The temporal sequence is processed first by the GRU, followed by spatial feature extraction by the CNN, then pooling by an Attention mechanism.

P4 (Cascaded CNN): Cascaded multi-stage convolutions feeding into a GRU and Attention layer.

P5 (Multi-Head Attention – Primary): Sequential CNN feature extraction and GRU temporal encoding, followed by multi-head Attention pooling across timestamps.

P6 (Parallel Early Fusion): Parallel CNN and GRU representations fused prior to multi-head Attention pooling.

P7 (Stochastic Gated Attention): Multi-head Attention augmented with stochastic gating (dropout on attention weights) to regularise temporal timestamp selection.

References

Ansel J, Yang E, He H, et al (2024) PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In: Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24). ACM, New York, NY, USA, pp 929–947, https://doi.org/10.1145/3620665.3640366

Bader S, Collaud Coen M, Duguay-Tetzlaff A, et al (2021) Klimareport 2020. Tech. rep., Federal Office of Meteorology and Climatology MeteoSwiss, Zürich, URL https://www.meteoswiss.admin.ch

Bader S, Bircher S, Collaud Coen M, et al (2022) Klimareport 2021. Tech. rep., Federal Office of Meteorology and Climatology MeteoSwiss, Zürich, URL https://www.meteoswiss.admin.ch

Bisht B (2023) Yield prediction using spatial and temporal deep learning algorithms and data fusion. Master's thesis, University of Ottawa, URL https://ruor.uottawa.ca

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

Copernicus (2025) Sentinel-2. https://documentation.dataspace.copernicus.eu/Data/SentinelMissions/Sentinel2.html, accessed: 15 April 2025

Copernicus Climate Change Service (2024) Global Climate Highlights 2023. https://climate.copernicus.eu/global-climate-highlights-2024, accessed: 19 January 2026

Cuturi M, Blondel M (2017) Soft-DTW: a Differentiable Loss Function for Time-Series. In: Proceedings of the 34th International Conference on Machine Learning (ICML)

Filippi P, Jones EJ, Wimalathunge NS, et al (2019) An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. Precision Agriculture 20(5):1015–1029. https://doi.org/10.1007/s11119-018-09628-4

Filippi P, Han SY, Bishop TFA (2025) On crop yield modelling, predicting, and forecasting and addressing the common issues in published studies. Precision

Agriculture 26(1):8. https://doi.org/10.1007/s11119-024-10212-2

Finnegan R, Metcalfe J, Sharifzadeh S, et al (2025) Temporal Optimisation of Satellite Image-Based Crop Mapping: A Comparison of Deep Time Series and Semi-Supervised Time Warping Strategies. IET Computer Vision 19(1)

Finnegan R, Sharifzadeh S, Eskandari H, et al (2026) Identifying the biophysical manifold: Why boundary-based models outperform transformers for pixel-level crop yield prediction in site-specific small-data regimes. In: Proceedings of SPIE Sensors + Imaging, SPIE

Frampton WJ, Dash J, Watmough G, et al (2013) Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation. ISPRS Journal of Photogrammetry and Remote Sensing 82:83–92

Gao J, Hu W, Chen Y (2024) Revisiting PCA for Time Series Reduction in Temporal Dimension. arXiv preprint arXiv:241219423 2412.19423

Goodfellow IJ, Pouget-Abadie J, Mirza M, et al (2014) Generative Adversarial Nets. arXiv preprint: arxivorg/abs/14062661

Gu A, Dao T (2024) Mamba: Linear Time-Sequence Modelling with Selective State Spaces. arXiv preprint arXiv:231200752 2312.00752

Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hyland SL, Esteban C, Rätsch G (2017) Real-Valued (Medical) Time-Series Generation with Recurrent Conditional GANs. arXiv preprint arxivorg/pdf/170602633

Iwana B, Uchida S (2021) An empirical survey of data augmentation for time series classification with neural networks. PLoS ONE 16

Johansson Fernstad S (2019) To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization. Information Visualization 18(2):230–250. https://doi.org/10.1177/1473871618785387

Johansson Fernstad S, Johansson Westberg J (2021) To explore what isn't there—glyph-based visualization for analysis of missing values. IEEE Transactions on Visualization and Computer Graphics 28(10):3513–3529. https://doi.org/10.1109/TVCG.2021.3065124

Kellerhals O (1925) Strafanstalt Witzwil: Ein Beispiel von Innenkolonisation durch die Arbeit von Gefangenen und Arbeitslosen. Ins - Buchdrückerei F. Dätwiler, Ins, Switzerland

Kim J (2026) PLF-Mamba: Analyzing Individual Milk Yield Dynamics Under Data Scarcity Using Selective State Space Models. Agriculture 16(3)

Kopp J, Schröer K, Schwierz C, et al (2023) The summer 2021 Switzerland hailstorms: weather situation, major impacts and unique observational data. RMetS Weather 78(7)

Loshchilov I, Hutter F (2017) Decoupled Weight Decay Regularization. arXiv preprint arXiv:171105101

Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems (NeurIPS), pp 4765–4774

Mao X, Li Q, Xie H, et al (2017) Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp 2794–2802, https://doi.org/10.1109/ICCV.2017.304

Maxwell AE, Warner TA, Fang F (2018) Implementation of machine-learning classification in remote sensing: an applied review. International Journal of Remote Sensing 39(9):2784–2817. https://doi.org/10.1080/01431161.2018.1433343

Meghraoui K, Sebari I, Pilz J, et al (2024) Applied Deep Learning-Based Crop Yield Prediction: A Systematic Analysis of Current Developments and Potential Challenges. Technologies 12(4). https://doi.org/10.3390/technologies12040043

Mirza M, Osindero S (2014) Conditional Generative Adversarial Nets. arXiv:1411.1784

Misra G, Cawkwell F, Wingler A (2020) Status of phenological research using Sentinel-2 data: A review. Remote Sensing 12(17):2760. https://doi.org/10.3390/rs12172760

Mogren O (2016) C-RNN-GAN: Continuous recurrent neural networks with adversarial training. In: 30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain

Oikonomidis AS, Catal C, Kassahun A (2022) Deep learning for crop yield prediction: a systematic literature review. New Zealand Journal of Crop and Horticultural Science https://doi.org/10.1080/01140671.2022.2032213

Paudel D, de Wit A, Boogaard H, et al (2023a) Interpretability of deep learning models for crop yield forecasting. Comput Electron Agric 206. https://doi.org/10.1016/j.compag.2023.107663

Paudel D, de Wit A, Boogard H, et al (2023b) A weakly supervised framework for high-resolution crop yield forecasts. Environmental Research Letters 18(9). https://doi.org/10.1088/1748-9326/acf50e

Perich G, Turkoglu MO, Graf LV, et al (2023) Pixel-based yield mapping and prediction from Sentinel-2 using spectral indices and neural networks. Field Crops Res 292

Portmann R, Schmid T, Villiger L, et al (2024) Modelling crop hail damage footprints with single-polarization radar: the roles of spatial resolution, hail intensity, and cropland density. Natural Hazards and Earth System Sciences 24(7)

Pothapragada IS, R S (2025) GANs for data augmentation with stacked CNN models and XAI for interpretable maize yield prediction. Smart Agricultural Technology 11

Ramponi G, Protopapas P, Brambilla M, et al (2018) T-CGAN: Conditional Generative Adversarial Network for Data Augmentation in Noisy Time Series with Irregular Sampling. arXiv preprint arXiv:181108295

Roberts DR, Bahn V, Ciuti S, et al (2017) Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40(8):913–929. https://doi.org/10.1111/ecog.02881

Rodríguez-Suárez B, Quesada-Barriuso P, Argüello F (2022) Design of cgan models for multispectral reconstruction in remote sensing. Remote Sensing 14(4):816. https://doi.org/10.3390/rs14040816

Sampath V, Maurtua I, Aguilar Martín J, et al (2021) A survey on generative adversarial networks for imbalance problems in computer vision tasks. Journal of Big Data https://doi.org/10.1186/s40537-021-00414-0

Shi X, Chen Z, Wang H, et al (2015) Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In: Advances in Neural Information Processing Systems (NIPS 2015)

Singh A, Awasthi AK, Badola U, et al (2025) A Novel Hybrid Approach to Crop Yield Prediction: Combining Deep Learning Efficiency with Statistical Precision. International Journal of Computing and Digital Systems https://doi.org/10.12785/ijcds/1571032941

Sun H, Deng Z, Chen H, et al (2020) Decision-Aware Conditional GANs for Time Series Data. arXiv preprint: arxivorg/abs/200912682

USDA (2026) Switzerland Production. https://ipad.fas.usda.gov/countrysummary/Default.aspx?id=SZ, accessed: 16 Jan 2026

Vani B, Guruprakash CD (2024) Ia-mstm: Improved adaptive moment estimation enabled mixed attention-based long short-term memory for crop yield prediction. In: IEEE International Conference, https://doi.org/10.1109/nkcon62728.2024.10775210

Vaswani A, Shazeer N, Parmar N, et al (2017) Attention Is All You Need. In: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA

Wang J, Wang P, Tian H, et al (2023) A deep learning framework combining CNN and GRU for improving wheat yield estimates using time series remotely sensed multi-variables. Computers and Electronics in Agriculture 206:107705. https://doi.org/10.1016/j.compag.2023.107705

Yoon J, Jarrett D, van der Schaar M (2019) Time-series Generative Adversarial Networks. In: 33rd Conference on Neural Information Processing Systems (NIPS 2019), Vancouver, Canada