applied sciences

MDPI

Article

Evaluating Reliability-Gated Multimodal Fusion for Winter Wheat Increment Forecasting

Murat Gökşin Bakır [1,*] and Yücel Batu Salman [2]

[1] Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Bahçeşehir University, Çırağan Caddesi Osmanpaşa Mektebi Sokak No: 4–6, Beşiktaş, 34353 Istanbul, Türkiye
[2] Department of Software Engineering, Faculty of Engineering and Natural Sciences, Bahçeşehir University, Çırağan Caddesi Osmanpaşa Mektebi Sokak No: 4–6, Beşiktaş, 34353 Istanbul, Türkiye; batu.salman@bau.edu.tr
* Correspondence: muratgoksin.bakir@bau.edu.tr

Featured Application

Headline UAV experts use 31 patch statistics, not a CNN on raw four-band images (the CNN was not re-fit after spatial screening). Among the evaluated gate variants, availability-only gating (availability plus agronomic context) offers a simpler alternative to the full reliability gate. The method forecasts an increment given a measured current state y_t; biophysics expert withdrawal is not a sensor-free next-state forecast. Concatenated walk-forward R^2 is negative for every method, so field-out cross-validation must not be read as prospective next-campaign skill.

Abstract

Agricultural sensors are informative at different times and drop out for different reasons: a UAV flight is canceled, a weather window is empty, a ground visit is skipped. A fusion rule that cannot see those conditions must average over them. This paper tests a reliability-conditioned mixture of experts for winter wheat leaf area index (LAI) and biomass increments: modality weights may depend on availability, input typicality and expert uncertainty, optionally together with agronomic context. The scientific contribution is not a new sensor or a new increment target. It is a comparative test of that reliability hypothesis against availability-only gating, nonnegative stacking and nested expert-dropout, under nested field-grouped cross-validation, systematic expert withdrawal, and concatenated walk-forward. Typicality and uncertainty do not earn a consistent field-level role. Availability plus agronomic context is the simpler alternative among the gates. On fully observed ΔLAI, a nonnegative stacker and UAV patch statistics beat the full gate; the gap is consistent with a high-capacity gate over-fitting a small out-of-fold expert table rather than with unused reliability information. Reconstruction $\hat{y}_{t+1} = y_t + \widehat{\Delta y}$ uses the measured current state, so the task is incremental change given y_t. Concatenated walk-forward R^2 is negative for every method because campaign increment means are non-stationary; field-out scores are therefore not a temporal forecasting skill. Consecutive elementary sampling unit (ESU) visits are kept when their 20 m windows overlap ($n = 105$ LAI, $n = 141$ biomass). Paired tests are exploratory and unadjusted.

Keywords: multimodal data fusion; dynamic gating; unmanned aerial vehicle (UAV) multispectral imagery; winter wheat; growth-increment forecasting; missing-modality robustness; nested cross-validation

Check for updates

Academic Editor: Xiaoming Zhang

Received: 3 October 2026
Revised: 20 October 2026
Accepted: 23 October 2026
Published: 27 October 2026

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

1. Introduction

Precision agriculture is not short of sensors. A drone returns Green, Red, Red-Edge and Near-Infrared reflectance at centimeter resolution; a ceptometer returns leaf area index (LAI) and absorbed radiation; a laboratory returns nitrogen concentration; a station streams temperature, radiation and soil moisture; a farm manager records variety, sowing density and predecessor crop. The applied difficulty is that those streams are rarely present together when a decision is due. A UAV sortie is canceled by wind or cloud; a November weather window is empty; a destructive biomass cut is too costly to repeat on every plot; nitrogen or growth-regulator timing cannot wait for a complete research campaign. On commercial winter wheat, the operational question is therefore not only "which sensor is best on a fully observed research plot?" but "if this sensor is missing, can the remaining streams still support an increment forecast, and should the manager trust that forecast?" Fusion architectures that assign one global weight per modality cannot answer that question: they average over campaigns in which the useful sensor changes.

The difficulty is that these streams are informative at different times and in different amounts, and that fusion rules rarely acknowledge it.

Consider the evidence that motivates this paper. On the spatially screened TS2AgroBg increments, unmanned aerial vehicle (UAV) patch statistics explain $R^2 = 0.719$ of ΔLAI and a nonnegative stacker reaches 0.744, while the full gate is behind both (0.571). For biomass, no statistically significant difference was detected between the gate and a ridge stacker (RMSE 102.4 vs. 105.2; cluster-bootstrap ΔRMSE CI -18 to $+15$). Availability-only gating is numerically ahead of the full reliability gate on fully observed biomass (RMSE 101.1 vs. 102.4). Small pooled LAI improvements of the full gate over availability-only appeared in exploratory bootstrap analyses, but field-level Wilcoxon tests did not establish improvement across fields (unadjusted). A single global weight per modality—the ridge stacker of classical decision-level fusion [1,2]— is required to average over campaign heterogeneity. It cannot express "trust the canopy image now, trust the phenological score in three weeks, and ignore the spectrum once the canopy has closed".

We therefore replace the fixed combination rule with a gate. Writing $\hat{y}_m$ for the prediction of the expert that sees modality m only, the forecast is

$$\widehat{\Delta y}_{t\to t+1} = \sum_{m=1}^{M} \alpha_m(z_t, \Delta t, a_m, q_m, u_m)\hat{y}_m + \beta_0, \qquad \sum_m \alpha_m = 1, \quad \alpha_m \geq 0, \tag{1}$$

where $\widehat{\Delta y}_{t\to t+1}$ is the predicted increment of LAI or dry biomass between consecutive campaigns, $\hat{y}_m$ is the prediction of the expert that sees modality m only, α_m is that expert's fusion weight, and β_0 is a learned intercept. The weight α_m is allowed to depend on the phenological (Zadoks) state z_t at issue time, the forecast horizon Δt, and three reliability scalars for modality m: availability a_m, input typicality q_m and predictive uncertainty u_m. Vectors are written in bold: $\boldsymbol{\alpha}$, $\mathbf{a}$, $\mathbf{q}$ and $\mathbf{u}$ collect the M modality-wise scalars; scalars (z_t, Δt, $\hat{y}_m$, β_0) are not bold. The weights are produced by a small network, $\boldsymbol{\alpha} = \mathrm{softmax}[g_\theta(z_t, \Delta t, \mathbf{a}, \mathbf{q}, \mathbf{u})]$, so that the fusion mechanism itself is a function of crop development and sensor reliability rather than a constant of the dataset.

The hypothesis that typicality and uncertainty should help is agronomic, not only statistical. A winter wheat sensor is informative only in a window: UAV structure matters after tillering and less once the canopy is closed and yellowing; a November meteorological gap is not a random missing-at-random hole but a seasonal outage; a plot with atypical weediness or damage is exactly the plot on which a global stacker, fitted to ordinary canopies, is most likely to be wrong. Mixture-of-experts theory then says that the gate should down-weight an expert that is unavailable, that is looking at an input far

from its training cloud, or that is internally disagreed [3–5]. Availability a_m encodes the first of those conditions. Typicality q_m is a cheap, fold-local proxy for the second. Uncertainty u_m (tree-spread or a heteroscedastic σ_m) is a proxy for the third. That is the methodological warrant for putting q_m and u_m into the gate. It is a testable hypothesis, not a claim that those proxies will improve point forecasts on this sample. The experiments ask whether they do. Retaining q_m and u_m in the framework is therefore a way to test—and, on this sample, largely reject—the agronomic warrant, not a claim that those scalars earn a pooled accuracy gain. Availability-only gating remains the simpler default among the gates (Section 6.3); the extra signals stay in the architecture so that a later, field-calibrated score can replace the present heuristics without changing the fusion interface. The construction is a mixture of experts [3,6] that can also see agronomic context, in the spirit of dynamic and uncertainty-aware multimodal fusion [4,7]. Phenology-aware multimodal models already exist. Liu et al. [8] identify soybean stages with a dynamic gate; Huang et al. [9] forecast strawberry yield from phenology-informed vision–weather series; Wang et al. [10] estimate maize LAI with stacking and phenology-specific feature sensitivity. Increment-style growth forecasts are not new either: Senti et al. [11] track winter wheat LAI, including day-to-day ΔLAI. Once y_t is measured, $\hat{y}_{t+1} = y_t + \widehat{\Delta y}$ is just the increment written as a next state. Hierarchical and centralized or decentralized federation have already been tried for agricultural yield [12,13].

The research question is explicit: under what conditions, if any, do typicality and uncertainty improve fusion beyond availability alone? Availability-only gating still receives agronomic context $\mathbf{c}$ (Section 4.1): it is availability plus context, not availability in isolation. The gate may also see stage and horizon. We do not assume that the extra signals will help. Component ablations (Sections 4.1 and 6.2) isolate availability, typicality, uncertainty and optional context; they do not identify a reliability block that consistently improves point forecasts. Paired field tests of the full gate versus availability-only gating are in Section 6.3. The stress tests are withheld experts and increment forecasts that still need a measured y_t. Negative results are retained: they are the evidence, not a prelude to a performance claim. The practical scope is therefore narrower than a general remote-forecasting claim: given a current in situ measurement, predict the change to the next campaign.

Two design choices run through the rest of the paper.

Forecast the increment, and do not read R^2 across scales.

Dry above-ground biomass rises by almost two orders of magnitude between November and June, so $\mathrm{Var}(y_{t+1}) \gg \mathrm{Var}(\Delta y)$ on pooled campaigns. When the current state y_t is observed and clipping in Equation (5) is inactive,

$$\hat{y}_{t+1} - y_{t+1} = \widehat{\Delta y} - \Delta y, \tag{2}$$

the left-hand side is the absolute next-state error and the right-hand side is the increment error, so the two residuals are identical. Their R^2 values differ because the denominators $\sum(y_{t+1} - \bar{y}_{t+1})^2$ and $\sum(\Delta y - \overline{\Delta y})^2$ differ. A mean-increment baseline that predicts the training set $\overline{\Delta y}$ and then adds the measured y_t therefore does not ignore the crop: it uses the current state and a constant growth step. It reaches $R^2 = 0.675$ on absolute next-campaign biomass and $R^2 = -0.009$ on the increment, which is expected when the target variance changes, not a split of skill into "seasonal progression". We also report persistence ($\widehat{\Delta y} = 0$) and a mean growth-rate baseline. We did not fit a calendar-only or stage-only predictor of Δy, so we do not claim to have measured seasonal predictability beyond those controls.

Use the imagery.

Averaging a hyperspectral signature down to four band means is a stand-in for a drone, not a use of one. For every elementary sampling unit (ESU), we cut the matching 20 m × 20 m window $X_{\mathrm{UAV}} \in \mathbb{R}^{H\times W\times 4}$ from the four-band orthomosaics and encode it. The fusion problem is then an image together with spectra, weather series and tabular records, not four extra columns.

We contribute:

1. A spatially screened increment $\Delta y_{t\to t+1}$, kept when consecutive 20 m windows overlap, and reconstructed as $\hat{y}_{t+1} = y_t + \widehat{\Delta y}$ from the measured y_t. Screening is not a check that the plot is the same: slivers with $f < 0.10$ remain under any-overlap. The majority-overlap sensitivity ($f \geq 0.5$) is the check that more than half the window is shared (Section 3.3). Headline UAV experts use 31 patch statistics; the CNN on raw four-band images was not re-fit after screening.
2. A comparative evaluation of a reliability-conditioned mixture-of-experts (Equation (1)) against availability-only gating, nonnegative stacking and nested expert-dropout under every withheld modality, asking when typicality and uncertainty help. Component ablations, not a single full-versus-stacker contrast, are used to attribute any change to availability, typicality, uncertainty or context.
3. Nested field-grouped cross-fitting, field-bootstrap intervals, and exploratory unadjusted paired tests. No primary hypothesis was prespecified. Field-out scores are not interpreted as temporal forecasting skill; that claim is reserved for concatenated walk-forward (Section 6.4).

2. Related Work and Positioning

2.1. From Fixed to Dynamic Fusion

The classical taxonomy separates fusion by abstraction level—data, feature and decision—and by uncertainty calculus [2,14–16]. Decision-level fusion trains a specialist per modality and combines the outputs by voting, stacking [1] or Bayesian model averaging. Our nonnegative and ridge stackers come from that stacking literature. How well stacking survives a missing expert depends on the implementation—zeroing a weight is not the same as dropping a sample— which is why we test sensor loss rather than assume robustness. The usual limitation remains: the combination rule is fitted once for the whole dataset.

Adaptive mixtures of local experts [3] and their modern sparsely gated descendants [6] instead learn an input-dependent gate. Dynamic multimodal fusion [7] adapts computation per sample; trusted multi-view classification [4] weights views by evidential uncertainty. Ma et al. [17] learn from severely incomplete modalities with a missing-feature reconstruction network and Bayesian meta-learning. That is a reconstruction method. Reliability-gated fusion (SARG) does not fill in a withheld modality; it passes an availability vector $\mathbf{a}$ ($a_m \in [0,1]$ is the finite-feature fraction), together with typicality and uncertainty, to the gate. Heteroscedastic heads [5] supply the per-expert uncertainty the gate can use. Stage, horizon and calendar features can also enter the gate. We try that option on TS2AgroBg; it does not improve accuracy.

2.2. Agricultural Multimodal Learning

Retrieval of LAI, fAPAR, cover fraction and biomass from Sentinel-2 and UAV multispectral imagery is mature [18], including on the Bulgarian winter wheat campaigns that produced TS2AgroBg [19–21]. Deep multimodal fusion of UAV imagery with structural and thermal features has been shown to improve soybean yield prediction [22]. Multitemporal UAV multispectral and digital vegetation indices have been used to predict rice

grain yield [23]; that study does not use structural or thermal fusion. Convolutional models also operate directly on drone mosaics [24,25]. Phenology is now routinely an input or a target of multimodal models: Liu et al. [8] fuse machine learning and deep learning pathways with a dynamic gate to identify soybean stages; Huang et al. [9] couple visual phenology with weather embeddings for strawberry yield; Wang et al. [10] report phenology-specific SHAP sensitivity in a stacked maize LAI estimator; and Senti et al. [11] forecast winter wheat LAI trajectories, including day-to-day increments, from satellite and meteorological time series. Phenology-aware fusion and increment-style growth forecasts are already in the literature. What we add is the reliability signals, the expert-withdrawal tests, and increment forecasts that use fewer ground-measured covariates—not a claim that we introduced phenology or increments.

2.3. Positioning

On the same Bulgarian landscape, the usual retrieval question is, "what is LAI now?". We ask how much canopy will develop before the next campaign. Liu et al. [8] predict soybean phenological stage; we treat phenology as an optional gate input, not as the thing to predict. Huang et al. [9] forecast strawberry yield; we forecast a next-campaign LAI or biomass increment given the measured y_t. Wang et al. [10] estimate maize LAI now, with stacking and phenology-specific sensitivity; we forecast an increment and then drop sensors. Turning off agronomic context in the SARG gate does not reduce pooled increment skill on this sample. Against increment-style growth forecasts [11], the increment is a way to hold current state fixed, not what we claim to have invented; a growth-rate ablation then asks the harder per-day question. Cross-silo sharing in agri-food data [26] and hierarchical or centralized/decentralized agricultural federation [12,13] are already discussed; Yang et al. [27] survey federated learning in general. Appendix B reports a neural optimization run that standardizes the concatenated training-client table in one place. That is not a field-local pipeline, we did not re-fit it after screening, and we do not treat it as a contribution.

The scientific contribution relative to those families is therefore comparative, not architectural. Adaptive mixtures of experts [3,6] already learn an input-dependent $\boldsymbol{\alpha}$; dynamic multimodal fusion [7] already adapts computation per sample; trusted multi-view classification [4] already weights views by evidential uncertainty; heteroscedastic heads [5] already supply a σ_m; and SMIL [17] already handles incomplete modalities, but by reconstructing the missing view. SARG does not introduce a new expert class, a new uncertainty calculus, or a missing-view imputer. It asks a narrower empirical question regarding winter wheat increments: if the gate is allowed to see explicit availability, typicality and tree-spread uncertainty—without filling in a withheld sensor—do those extra scalars improve fusion beyond availability plus agronomic context, stacking, and nested expert-dropout? The answer on this sample is largely no (Sections 6.2 and 6.3). That negative comparative result, together with the spatially screened increment protocol and the expert-withdrawal tests, is the contribution. It is not a claim that reliability-gated fusion is a new mixture-of-experts architecture, nor that it outperforms uncertainty-aware fusion in general.

3. Materials and Methods

3.1. Dataset: TS2AgroBg

We use the public TS2AgroBg dataset of Dimitrov et al.: georeferenced UAV multispectral imagery, ground spectra, crop biophysics and biochemistry, soil, phenology, management records and station weather from commercial winter wheat in Bulgaria [19,28]. It

is released on Zenodo under CC BY 4.0. Figure 1 shows the source families and the keys that join them.

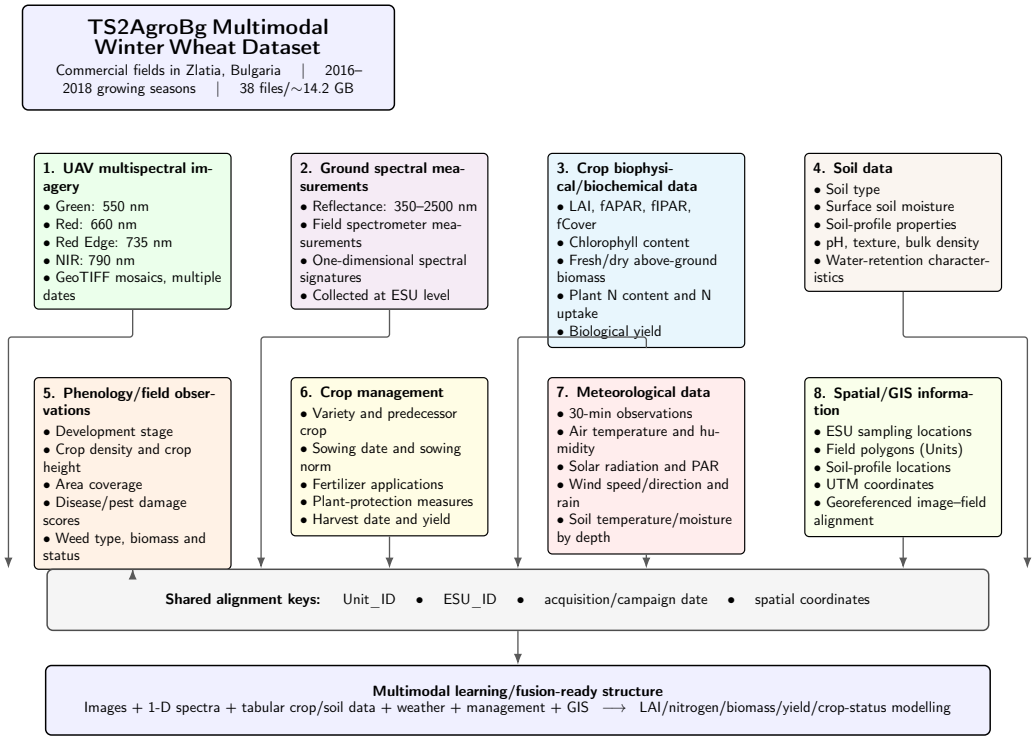


Figure 1. Overview of the TS2AgroBg dataset and its multimodal structure. The dataset combines unmanned aerial vehicle (UAV) multispectral imagery, ground spectral measurements, crop biophysical and biochemical variables, soil properties, phenological observations, crop-management records, meteorological measurements, and geospatial information. The modalities can be linked through common field identifiers, sampling-unit identifiers, acquisition dates, and geographic coordinates, enabling multimodal fusion for crop-status modeling.

TS2AgroBg covers two growing seasons (2016/17 and 2017/18) on commercial winter wheat of the Zlatia test site, collected under ESA PECS contract 4000117474/16/NL/NDe. Elementary Sampling Units are 20 m × 20 m plots surveyed in six agricultural units. Campaigns are November 2016, March, April, May and June 2017, plus November 2017 and April 2018, giving 206 ESU visits over 60 spatially recurring plots. Phenological stages follow the decimal code of Zadoks et al. [29], and soil types follow WRB [30].

Six agricultural units, twelve season–unit clients.

The dataset labels six commercial units per season. Those labels are not the same physical fields reused in 2017/18. Winter wheat is rotated, and the unit numbers are reassigned: S16-17_U*k* and S17-18_U*k* occupy different cadastral parcels (Figure 2, Table 1). The nearest same-label pair is S16-17_U2 and S17-18_U2 at 2.92 km; U1 is 4.94 km apart across seasons, and units 4–6 of 2017/18 lie more than 15 km north of their 2016/17 namesakes. No cross-season client pair is closer than 2.9 km. The twelve season–unit identifiers are therefore twelve distinct places, and holding one out of training never leaks the same physical field from the other season. Within a season, the units of a farm cluster can be as close as 90–230 m; grouped cross-validation still holds the entire unit out. The grouping key is this season–unit identifier, matching the data-governance boundary of a farm operator in a given year.

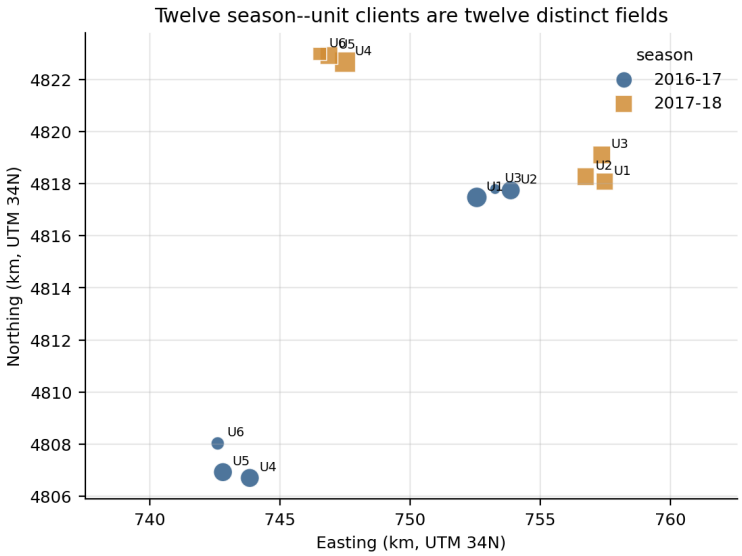


Figure 2. Centroid of each season–unit client in UTM 34N. Marker size is the number of ESU plots. Units 1–6 in 2016/17 and units 1–6 in 2017/18 are different commercial fields; the shared integer is a seasonal label, not a persistent parcel identifier.

Table 1. Season–unit clients. Coordinates are mean ESU easting and northing (UTM 34N). Distance is to the nearest other client.

Client	Season	Unit	Plots	Visits	Easting (m)	Northing (m)	Dist. (m)
S16-17_U1	2016/17	1	7	35	752,572	4,817,469	121
S16-17_U2	2016/17	2	6	30	753,874	4,817,740	145
S16-17_U3	2016/17	3	2	10	753,275	4,817,791	121
S16-17_U4	2016/17	4	6	29	743,851	4,806,706	228
S16-17_U5	2016/17	5	6	30	742,819	4,806,929	228
S16-17_U6	2016/17	6	3	12	742,618	4,808,030	340
S17-18_U1	2017/18	1	5	10	757,480	4,818,074	106
S17-18_U2	2017/18	2	5	10	756,748	4,818,263	106
S17-18_U3	2017/18	3	5	10	757,367	4,819,098	533
S17-18_U4	2017/18	4	7	14	747,511	4,822,664	185
S17-18_U5	2017/18	5	5	10	746,893	4,822,915	90
S17-18_U6	2017/18	6	3	6	746,536	4,822,968	90

3.2. UAV Patch Extraction

The public release ships the imagery as seventeen archives (one per season, unit group and flight date), each holding four single-band reflectance GeoTIFFs—Green, Red, Red Edge and Near-Infrared—at 0.27 m ground sampling distance in EPSG:32634, totaling approximately 14 GB. Rather than unpacking them, the rasters are read in place through the GDAL zip layer and only the ESU windows are materialized. For each visit, the 20 m square centered on the recorded ESU centroid is read as a 73×73 native window per band, the sensor no-data value and non-physical reflectances are masked, and the result is NaN-aware average-pooled to a $32 \times 32 \times 4$ tensor

$$X_{\mathrm{UAV}} \in \mathbb{R}^{32\times32\times4}, \qquad \{G, R, RE, NIR\}, \tag{3}$$

so X_{UAV} is a four-band reflectance patch of height and width 32 (Green, Red, Red-Edge, Near-Infrared). The tensor is retained only when at least 25% of the pixels are valid in all four bands. This yields 161 of 206 ESU visits with imagery. The 45 missing cases are genuine acquisition gaps rather than processing failures: there was no June 2017 flight, and units 1–3 were not flown in November 2017. Missing UAV is not a nuisance: it is one of the availability conditions the gate is meant to handle.

Figure 3 shows extracted patches. The phenological progression is unmistakable—mottled, partly bare soil in November (patch-mean NDVI $\approx$ 0.55) closing to a dense uniform canopy in April (NDVI $\approx$ 0.94)—and sprayer tramlines are resolved as dark lines, which confirms both the georeferencing and the presence of real sub-plot texture.

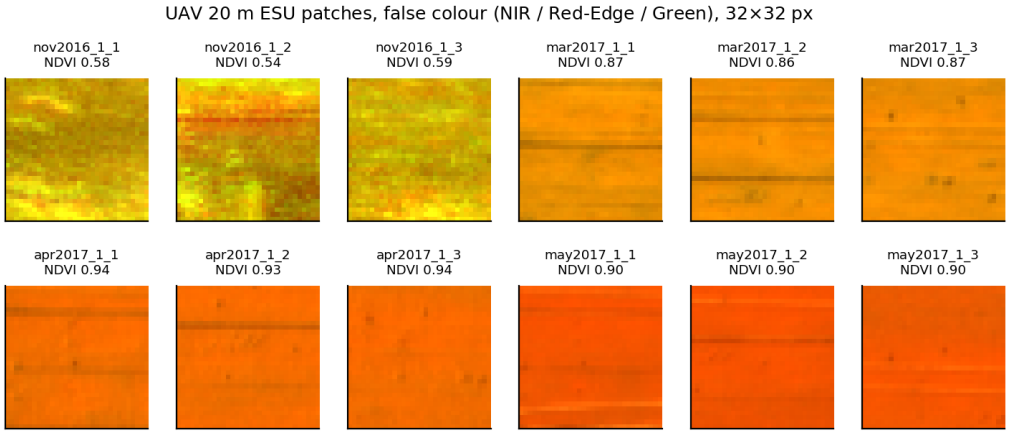


Figure 3. Extracted 20 m UAV patches for ESUs of unit 1 across four campaigns, shown as false color (NIR/Red-Edge/Green) at the 32 $\times$ 32 resolution used by the image encoder. Canopy closure between November and April is visible, as are sprayer tramlines within the plot.

Each patch contributes two representations. The first is a set of 31 patch-level statistics: per-band mean, standard deviation and 10th/90th percentiles; the means and standard deviations of NDVI, NDRE, GNDVI, CIRE and SAVI computed pixel-wise; the within-patch NDVI coefficient of variation, inter-percentile range and mean gradient magnitude as measures of canopy patchiness; and the valid-pixel fraction, which doubles as the UAV term in the typicality score q_{UAV}. The second is the raw tensor, consumed by the convolutional encoder of Section 4.1. Reporting both is deliberate: they turn out to behave very differently at this sample size (Section 6.2). Headline ExtraTrees experts use the 31 statistics, not the CNN. That choice is deliberate and limits generalizability in a stated way. A compact CNN on the 32 $\times$ 32 $\times$ 4 tensor was trained on the older unfiltered sample (Appendix A) and did not beat the statistics (0.661 $\rightarrow$ 0.632 on that historical ΔLAI comparison). With roughly ninety training patches, the convolutional encoder is underdetermined; the 31 handcrafted summaries already encode mean color, dispersion, index moments and texture that an ExtraTrees expert can use without a second nested fit. We therefore kept the statistics as the main UAV representation after spatial screening and did not re-fit the CNN. The implication is that the headline UAV skill ($R^2 = 0.719$ on screened ΔLAI) is the skill of those summaries, not of a learned image encoder. Transfer to other sensors, resolutions or crops would need the CNN—or another image encoder—to be refit on a sample that supports it. We do not treat the 31-statistic expert as a general UAV fusion recipe.

3.3. Unification of the Remaining Modalities

The remaining streams have to be joined on plot and date before they are features. Each ESU identifier is parsed into a stable plot key (season, unit, ESU), and modalities are joined on that key and campaign date. The spectral modality integrates ASD FieldSpec reflectance over Sequoia/Sentinel-2 analogue windows (Green 540–570 nm, Red 640–680 nm, Red-Edge 710–740 nm, NIR 770–810 nm), 50 nm optical bins from 400–1000 nm, and the indices NDVI, NDRE, GNDVI, EVI2, SAVI, CIRE, SR, MTCI and NDWI. Phenology contributes density, height, visually assessed cover, damage and weediness scores, a mid-range Zadoks stage and days after sowing. Biophysics contributes

LAI (measured in situ; estimation errors and sampling are reviewed by Weiss et al. [31]), fIPAR, fAPAR, fCover, chlorophyll, fresh and dry biomass and nitrogen concentration and uptake, always of the observation campaign and never of the target campaign. Soil contributes campaign ESU gravimetric moisture and nearest-profile topsoil clay, organic carbon, pH and bulk density from campaign pits [19,28], not a public soil map. Weather is summarized from two on-site meteorological stations (half-hourly records into 7- and 30-day windows plus growing degree-days since sowing), not a reanalysis or GIS weather layer. Station coverage is incomplete and the gaps are systematic, not scattered at random (Table 2). Every November 2016 visit lacks growing degree-days and both the 7- and 30-day meteorological windows (missingness 1.0). Half of the April 2017 visits lack the 7- and 30-day windows (missingness 0.5) while GDD is present. The remaining campaigns are complete. Empty weather features are median-imputed inside each training fold only, so a November 2016 row receives the training-fold median of later, mostly spring, campaigns. That fill can pull the weather expert toward the climatology of observed windows and can create a campaign-level imputation signature that trees can split on. Availability a_m is computed from the raw (pre-imputation) matrix, so November 2016 weather still enters the gate with $a_m \approx 0$; imputation fills the expert features, it does not hide the gap from the gate. Persistence-by-client is not identified under grouped CV: a held-out unit has no same-client training row. We therefore compared fold-local medians with two alternatives that remain usable at test time, fold-local mean imputation and a five-neighbor KNN imputer, applied only to the weather expert (Table 3). On ΔLAI, the three fills agree to about 0.01 RMSE (SARG 1.069/1.074/1.066). On ΔAGBd, KNN and mean move the full gate from 102.4 to 101.7 and 101.9; availability-only stays near 101–104. Weather-expert withdrawal under the median fill still moves RMSE only weakly (1.069 → 1.063; 102.4 → 101.4). The November 2016 climatology-pull is not rescued by changing the imputer.

Table 2. Fraction of increment rows with a missing weather feature, by campaign, before fold-local median imputation. November 2016 is a complete outage of the meteorological windows and of growing degree-days. April 2017 is a half-campaign gap in the 7- and 30-day windows. Other campaigns are complete. Source: `weather_missing_by_campaign.csv`.

Campaign	GDD	T 7-Day	P 7-Day	T 30-Day	P 30-Day
November 2016	1.00	1.00	1.00	1.00	1.00
April 2017	0.00	0.50	0.50	0.50	0.50
March 2017, May 2017, June 2017, November 2017, April 2018	0.00	0.00	0.00	0.00	0.00

Table 3. Spatial and weather-imputation sensitivities on the screened sample. Farm-cluster: leave-one-cluster-out over the four same-season groups with centroid distance < 2 km. Weather: fold-local median (headline), mean, or five-neighbor KNN on the weather expert only; other experts stay median. Withdrawn weather uses the missing-expert encoding of Section 4.1.

Protocol	Method	ΔLAI		ΔAGBd	
		RMSE	R^2	RMSE	R^2
Leave-one-farm-cluster-out					
	Full SARG	1.276	0.390	161.6	0.661
	Availability-only	1.281	0.385	141.0	0.742
	NNLS stacker	1.368	0.299	156.6	0.682
	Ridge stacker	1.358	0.309	173.2	0.611
Weather imputer (full gate/availability-only/weather withheld)					
Median		1.069/1.080/1.063	—	102.4/101.1/101.4	—
KNN		1.074/1.071/1.062	—	101.7/102.0/101.9	—
Mean		1.066/1.073/1.061	—	101.9/104.1/98.3	—

Neighboring agricultural units lie 90–230 m apart within a season (Table 1; nearest-ESU distance, not centroid distance). Grouped cross-validation holds a unit out but still

trains on its farm-cluster neighbor, which shares the same station weather and similar soils. Field-cluster bootstrap and the field Wilcoxon test account for within-field row dependence; they do not remove between-field spatial autocorrelation. Client-mean residual correlations among same-campaign farm-cluster pairs (centroid distance < 2 km) are not significant at unadjusted 0.05 for SARG, the nonnegative stacker or ridge (ΔLAI near-pair Pearson $p = 0.17, 0.55, 0.96$; ΔAGBd $p = 0.067, 0.59, 0.73$). That diagnostic does not license a claim of spatial independence: the effective sample of independent landscapes is closer to a handful of farm clusters than to twelve clients, and we did not compute a distance-weighted Moran's I on ESU-level residuals. Leave-one-farm-cluster-out (four clusters defined by centroid distance < 2 km) is the stricter spatial check (Table 3). Unit-out CV overstates skill: ΔLAI RMSE rises from 1.07 to 1.28 and ΔAGBd from 102 to 141–162. Availability-only remains level with the full gate on LAI (1.281 vs. 1.276, $p = 0.88$) and is ahead on biomass (141.0 vs. 161.6, $p = 0.25$). Geographic generalization claims are therefore limited to held-out units inside the same landscape. Management contributes sowing density, field area, variety and predecessor codes and campaign index. The forecast gap Δt is not an input to the management expert—it would let that expert infer how much growth time remains—and is supplied only to the gate as context. Median imputation is fitted inside each training fold only, with empty features preserved so that an absent modality never silently drops a sample.

Spatial screening of Δy.

The increment $\Delta y = y_{p,t+1} - y_{p,t}$ mixes locations if consecutive visits do not sample the same ground. The published column dictionary states that an ESU is a 20 m $\times$ 20 m plot whose center is `Point_X_UTM34/Point_Y_UTM34` (WGS84/UTM 34N). UAV patches use the same square. Two axis-aligned 20 m windows share some ground if and only if $|\Delta x| < 20$ m and $|\Delta y| < 20$ m, equivalently overlap fraction $f = (20 - |\Delta x|)(20 - |\Delta y|)/400 > 0$. That any-overlap screen is plot geometry, not an RMSE-tuned cutoff. It is adequate only as a filter for identifier jumps: a pair with $f = 0.09$ (less than 10% shared area) is two mostly distinct plots, and we do not treat it as a repeated measurement of the same canopy. We keep those slivers in the headline sample because dropping them would be a second, error-tuned rule, whereas the keep-rule we prespecified is "any shared ground". The check that more than half the window is common is the majority-overlap sensitivity ($f \geq 0.5$), which we now treat as the companion table to the headline $n = 105/141$ results, not as an appendix afterthought. A sliver of shared ground is still two mostly distinct plots. That geometry affects the increment itself, not only the sample size. When f is small, $y_{p,t+1} - y_{p,t}$ mixes two mostly distinct 20 m windows, so part of Δy is a change of place rather than a change of canopy. Any-overlap therefore leaves residual label noise in the headline $n = 105/141$ tables; it is a filter for identifier jumps, not a guarantee that consecutive visits measure the same plot. Majority overlap ($f \geq 0.5$) is the companion that requires more than half the window to be shared. Concatenated walk-forward is now also repeated under that stricter rule ($n = 75$ LAI, $n = 106$ biomass); every R^2 remains negative, so the temporal caveat is not an any-overlap artefact. Forecast reliability should still be read with the residual label-noise caveat on the headline any-overlap tables: a low-overlap pair can inflate increment RMSE even when the fusion rule is otherwise well specified. Table 4 reports displacement and overlap-fraction quantiles among pairs that pass the any-overlap screen; Table 5 is the majority-overlap companion. Median displacement is 3.8 m (LAI) and 2.9 m (dry above-ground biomass, AGBd); the 90th percentiles are 9.4 m and 9.3 m. Median overlap fraction is 0.75 (LAI) and 0.82 (biomass). Fourteen LAI and seventeen biomass pairs that overlap at all have $f < 0.5$.

Table 4. Consecutive-visit spatial audit. Any-overlap: $f > 0$. Majority: $f \geq 0.5$ (more than half the 20 m window shared). Quantiles are among any-overlap pairs. Displacement in metres; f is unitless overlap fraction.

	Pairs	Any-Overlap	Majority	Disp. q_{50}	Disp. q_{90}	f q_{50}	f min	$f < 0.5$
ΔLAI	108	105	91	3.84	9.39	0.75	0.09	14
ΔAGBd	146	141	124	2.86	9.34	0.82	0.05	17

Table 5. Companion sensitivity to the headline any-overlap sample: majority overlap ($f \geq 0.5$; more than half the 20 m window shared). This is the table to read when asking whether slivers with $f < 0.10$ drive the ranking. Nested field-grouped cross-fitting on `v20260920_majority_overlap`. Fully observed models and biophysics-withdrawal comparators only. Concatenated walk-forward under the same $f \geq 0.5$ rule is reported in Section 6.4 ($n = 75$ LAI, $n = 106$ biomass), not in this table. This rule is geometry, not RMSE-tuned.

Method	ΔLAI ($n = 91$)		ΔAGBd ($n = 124$)	
	RMSE	R^2	RMSE	R^2
NNLS stacker	0.854	0.743	115.1	0.831
Dropout nested	0.839	0.752	123.9	0.805
Ridge stacker	0.928	0.697	112.5	0.839
SARG availability only	0.934	0.693	98.7	0.876
SARG no reliability	0.949	0.683	98.3	0.877
SARG full	1.001	0.647	101.6	0.869
Biophysics withheld, SARG	1.049	0.612	107.3	0.853
Biophysics withheld, avail. only	0.958	0.677	111.1	0.843
Biophysics withheld, dropout nested	0.963	0.674	129.3	0.787
Biophysics withheld, ridge masked	1.010	0.641	180.5	0.586
Biophysics withheld, NNLS masked	1.134	0.547	246.1	0.229

A stricter, independently justified rule requires majority shared area, $f \geq 0.5$: more than half of the 20 m window is common ground. That is still geometry, not a threshold chosen because it improves RMSE. It leaves $n = 91$ LAI and $n = 124$ biomass transitions (Table 5; run `v20260920_majority_overlap`). That sensitivity repeats fully observed fusion and missing-modality comparisons (Table 5). Concatenated walk-forward under the same $f \geq 0.5$ rule is in Section 6.4 ($n = 75$ LAI, $n = 106$ biomass); every R^2 remains negative. The main fully observed and gate-versus-availability findings hold: the full gate is not the most accurate fully observed model (nested dropout 0.839 on ΔLAI; no-reliability 98.3 on ΔAGBd); availability-only remains numerically ahead of or comparable to the full gate (LAI 0.934 vs. 1.001, $p = 0.38$; biomass 98.7 vs. 101.6, $p = 0.57$); under biomass biophysics, withdrawal SARG still has the lowest pooled RMSE (107.3 vs. availability-only 111.1, nested dropout 129.3, masked ridge 180.5, nonnegative least squares (NNLS) 246.1; vs. availability $p = 0.85$). Two rankings change: nested dropout overtakes NNLS on fully observed ΔLAI, and availability-only is numerically ahead of SARG under LAI biophysics withdrawal (0.958 vs. 1.049). The campaign documentation does not describe relocating an ESU between visits. Five November 2016 → March 2017 jumps of 41–600 m fail even any-overlap (3/108 LAI, 5/146 biomass) and are identifier discontinuities relative to sibling-ESU medians (5–10 m). Tree-expert fusion, missing-modality and walk-forward tables use the any-overlap screened sample ($n = 105/141$); majority-overlap walk-forward is the companion in Section 6.4. Neural federation and the reduced-input ladder were not re-fit after screening (Appendixes A and B). Pair manifests are in the Zenodo reproducibility deposit [32].

3.4. Growth-Increment Targets

Let $x_{p,t}$ be the unified vector of plot p at campaign t. For a state variable $y \in \{\text{LAI}, \text{AGBd}\}$, the primary target is the increment to the next campaign,

$$\Delta y_{t \to t+1} = y_{p,t+1} - y_{p,t}, \tag{4}$$

where $y_{p,t}$ is the measured LAI or dry biomass of plot p at campaign t and $y_{p,t+1}$ is the measured value at the next campaign. The absolute forecast is recovered by

$$\hat{y}_{p,t+1} = \max\left(0,\ y_{p,t} + \widehat{\Delta y}_{t \to t+1}\right). \tag{5}$$

The $\max(0, \cdot)$ clip prevents a non-physical negative state; $y_{p,t}$ on the right-hand side is the ground measurement, not an estimate. Equation (5) is not a remote-state estimator: $y_{p,t}$ is the measured current LAI or biomass. The model never predicts y_{t+1} from covariates alone. A UAV, weather or soil observation at t can enter $\widehat{\Delta y}$, but the absolute next state is recovered only after adding the ground measurement y_t. Historical reduced-input rungs in Appendix A change which covariates enter $\widehat{\Delta y}$, not whether y_t is observed. The task is therefore incremental change given a current measurement, not remote forecasting of future crop status in the absence of a ground visit at t. That restriction is a genuine limit on operational scope: a no-ground-sampling deployment still requires y_t for reconstruction, even if the increment experts drop the current-biophysics modality. Transitions are Nov $\to$ Mar, Mar $\to$ Apr, Apr $\to$ May and May $\to$ Jun in 2016/17 and Nov $\to$ Apr in 2017/18. After excluding pairs whose 20 m windows do not overlap, $n = 105$ LAI and $n = 141$ biomass increments remain on the any-overlap screen ($n = 91/124$ under majority overlap).

Because campaign gaps are unequal—November $\to$ March is months, April $\to$ May is weeks—the increment still contains a horizon-length effect. Median Δt is 22–39 d on spring transitions and 131–146 d on the two November departures (Section 6.5). We therefore also report a growth-rate ablation

$$r_y = \frac{y_{t+1} - y_t}{\Delta t}, \qquad \widehat{\Delta y} = \hat{r}_y\, \Delta t, \tag{6}$$

where r_y is the per-day growth rate, $\hat{r}_y$ its prediction, and Δt the known gap in days. Multiplying by Δt maps the rate back onto the increment scale so that RMSE remains comparable to Equation (4). Predicting r_y asks how rapidly the crop is developing rather than how much it will change over an interval of known length. It does not replace the increment task, and it does not substitute for scoring the increment models themselves inside short versus long intervals. SARG already receives Δt in the gate context $\mathbf{c}$, so a horizon-stratified breakdown of the same field-out predictions (Section 6.5) is the within-framework check: it asks whether the full reliability gate, which can see both Δt and the reliability signals, is ahead of availability-only or the stackers on short and long transitions separately.

The default accuracy protocol is field-out GroupKFold over the twelve season–unit clients of Table 1: no physical field contributes training and test rows in the same fold. Walk-forward and cold-start splits (Section 6.4) are a different protocol and are labeled as such; they are not field-grouped cross-validation. Unless noted, results use random seed 42. Point estimates of RMSE and R^2 from field-out folds are accompanied by 95% percentile intervals from a field-level (cluster) bootstrap [33]. Methods are compared paired on the same held-out fields, $\Delta\text{RMSE}_k = \text{RMSE}_{k,\text{comp}} - \text{RMSE}_{k,\text{SARG}}$, with a Wilcoxon signed-rank test on the field-level differences. Those tests are exploratory: no primary hypothesis was prespecified, and p-values are unadjusted for the number of modalities, targets, gate configurations and withdrawal scenarios. The availability-only comparison in Section 6.3 alone reports eight Wilcoxon tests; a Bonferroni threshold of $0.05/8 = 0.00625$ leaves none of them significant. Counting also the fully observed method rankings, seven withheld

experts, two targets, typicality and uncertainty correlations, and the horizon strata produces several dozen comparisons, so isolated unadjusted $p < 0.05$ values—including the biomass biophysics-withdrawal contrast versus NNLS ($p = 0.002$)—must not be read as confirmatory. Ordinary row bootstrap is not used: rows inside a field share soil, cultivar and management and are not independent [34].

4. Reliability-Gated Multimodal Fusion

Figure 4 is the evaluated tree-expert protocol. Each ExtraTrees encoder returns a prediction and an uncertainty. The headline UAV expert uses the 31 patch statistics of Section 3.2, not a convolutional encoder on the raw $32 \times 32 \times 4$ four-band image. A compact CNN on that tensor is a neural alternative that was fitted only on the older unfiltered sample and was not re-fit after spatial screening; it is not used in the spatially screened increment tables of Sections 6.1–6.4. The gate turns reliability signals—and, if we switch it on, agronomic context—into convex weights; the weighted sum plus a learned intercept β_0 is an increment, added to the measured current state. A federated optimization run with centralized preprocessing is summarized in Appendix B and reported in the Zenodo deposit [32]; it is not part of the main protocol.

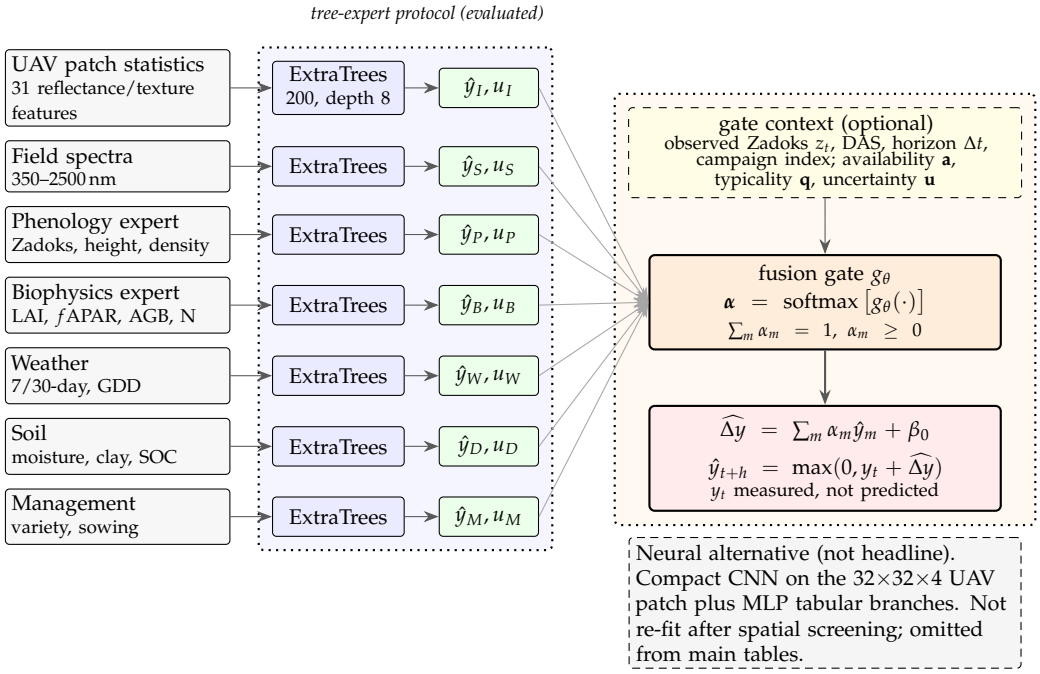


Figure 4. Tree-expert reliability-gated fusion used in the main tables. Each ExtraTrees expert (200 trees, depth 8) consumes one modality—UAV as 31 patch statistics, not a CNN—and returns a prediction $\hat{y}_m$ and an uncertainty u_m. The gate g_θ maps reliability signals and, optionally, agronomic context (observed Zadoks z_t, days after sowing, horizon Δt, campaign index; Δt is not an input to the management expert) to a convex weight vector $\boldsymbol{\alpha}$. The forecast includes a learned intercept β_0 and is added to the measured current state. A compact CNN on the $32 \times 32 \times 4$ UAV patch is a neural alternative that was not re-fit after spatial screening and is not a headline result.

4.1. Modality Encoders and Experts

Seven modalities are encoded as in Equation (7),

$$\begin{aligned} h_I &= E_I(X_{\text{UAV}}), & h_S &= E_S(X_{\text{FieldSpec}}), & h_P &= E_P(X_{\text{phenology}}), & h_B &= E_B(X_{\text{biophys}}), \\ h_W &= E_W(X_{\text{weather}}), & h_D &= E_D(X_{\text{soil}}), & h_M &= E_M(X_{\text{mgmt}}). \end{aligned} \tag{7}$$

$E_I, E_S, E_P, E_B, E_W, E_D, E_M$ are the seven modality encoders and $h_I, h_S, h_P, h_B, h_W, h_D, h_M$ are their latent representations (UAV imagery, field spectra, phenology, biophysics,

weather, soil, management). For ExtraTrees experts, those encoders are implicit: the "latent" is the 31-dimensional UAV statistic vector or the corresponding tabular block. For the neural alternative, each E is an explicit network and each h is a 12-dimensional vector. We use two encodings because they suit different sample sizes.

Randomized forest experts.

Each modality is given an extremely randomized trees regressor [35] with 200 trees of depth 8 after fold-local median imputation and standardization. The ensemble supplies both the prediction and its uncertainty: with T trees producing $\hat{y}_m^{(\tau)}$,

$$\hat{y}_m = \frac{1}{T}\sum_{\tau}\hat{y}_m^{(\tau)}, \qquad u_m = \frac{1}{s_m}\sqrt{\frac{1}{T}\sum_{\tau}\left(\hat{y}_m^{(\tau)} - \hat{y}_m\right)^2}, \tag{8}$$

where s_m is the median training spread of modality m so that u_m is comparable across modalities. For this family, the UAV modality enters as the 31 patch statistics of Section 3.2. Table 6 lists the feature space of each ExtraTrees expert after the fold-local imputer and standardizer: dimensionality is unchanged because empty columns are kept and no principal-component or other reduction is applied. Categorical farm records enter as integer codes from `pandas.factorize` on lower-cased strings, not as one-hot vectors; ExtraTrees then treat those codes as ordered numeric splits. Sentinel-2 analogue windows, profile sand and silt, wind speed and the unused 30-day weather summaries are computed in the unification step but are not passed to any expert. Harvest yield and harvest date are excluded from every increment expert. Forecast horizon Δt is excluded from the management expert and is supplied only to the gate. The gate's own Zadoks principal-stage one-hot encoding is context, not an expert input.

End-to-end neural experts.

The same seven modalities are alternatively encoded by a single network. The image branch E_I is a compact CNN—three stride-2 convolutions with 12, 20 and 24 channels, batch normalization, global average pooling and a linear projection to a 12-dimensional latent—applied to the raw $32 \times 32 \times 4$ patch; the six tabular branches are two-layer perceptrons into the same latent space. Each latent receives a heteroscedastic head [5] returning a mean and a log-variance,

$$(\mu_m, \log\sigma_m^2) = \mathrm{head}_m(h_m), \qquad u_m = \sigma_m, \tag{9}$$

where μ_m is the expert mean, σ_m is the learned standard deviation, and h_m is the modality latent. The uncertainty consumed by the gate is therefore learned rather than derived from ensemble spread. Training minimizes the fused squared error plus the mean per-expert Gaussian negative log-likelihood, which keeps the individual σ_m meaningful, using AdamW [36]. The network is trained in one place: an inner 20% validation split and early stopping (patience 60). It is not the federated trainer, and we did not re-fit it on the screened sample.

4.2. Reliability Signals

For every sample and modality, the gate receives three scalars besides the prediction. Availability $a_m \in [0, 1]$ is the fraction of finite raw features in modality m, which is 0 for an ESU with no UAV coverage and 1 for a complete record. Input typicality q_m combines an in-distribution score with an explicit completeness term,

$$q_m = \underbrace{\frac{1}{d_m}\sum_{j=1}^{d_m}\exp\left(-|\tilde{x}_{mj}|\right)}_{\text{typicality}} \cdot \eta_m, \tag{10}$$

The first factor averages a Laplace-like kernel over the d_m standardized features of modality m; $\tilde{x}_{mj}$ is the training-standardized value of feature j and η_m is the valid-pixel fraction for the UAV modality, the spectrum-present indicator for the spectral modality, and 1 otherwise. After fold-local standardization, the training mean of each coordinate is the origin, so $|\tilde{x}_{mj}|$ is a coordinate-wise z-score and $\exp(-|\tilde{x}_{mj}|)$ is a Laplace-like kernel that maps that score to $(0, 1]$. Averaging the kernels treats features equally and is bounded, cheap and closed-form inside every nested fold: it needs no extra model, no covariance inverse and no held-out density estimator. Completeness η_m prevents a missing UAV or spectrum from looking typical merely because imputed zeros sit near the origin.

Table 6. Feature space of each ExtraTrees expert on the screened run. Counts are columns after fold-local median imputation and standardization (no reduction). Integer factorize codes are not one-hot. Horizon Δt is gate context only.

Expert	d_m	Features Used	Encoding/Exclusions
UAV	31	Per-band mean, SD, q_{10}/q_{90} (G, R, RE, NIR); mean and SD of NDVI, NDRE, GNDVI, CIRE, SAVI; NDVI CV, inter-percentile range, mean gradient magnitude; `uav_valid_frac`, `uav_available`	All numeric. Completeness columns also enter q_{UAV}. The raw $32 \times 32 \times 4$ tensor is not an ExtraTrees input.
Spectral	26	Sequoia-analogue Green, Red, Red-Edge, NIR; NDVI, NDRE, GNDVI, EVI2, SAVI, CIRE, SR, MTCI, NDWI; `spec_available`; twelve 50 nm bins on 400–1000 nm	All numeric. Sentinel-2 analogue windows ($b3$–$b8$) are computed but not passed to the expert.
Biophysics	9	LAI, fIPAR, fAPAR, fCover, chlorophyll, dry and fresh wheat biomass, N concentration, N uptake	Current-campaign values only. Target-campaign and harvest quantities are excluded.
Phenology	7	Density, height, visually assessed cover, damage, weediness, mid-range Zadoks, days after sowing	Numeric scores as published. Observation date is used only to compute days after sowing.
Soil	6	Gravimetric moisture 0–5 and 0–30 cm; nearest-pit topsoil clay, SOC, pH, bulk density	Numeric. Profile sand and silt are computed but excluded.
Weather	10	7-day mean, min and max air temperature, precipitation, PAR, relative humidity and 5 cm soil moisture; 30-day mean temperature and precipitation; GDD since sowing	Numeric station summaries. Wind speed and the unused 30-day summaries are computed in the daily table but excluded.
Management	6	Sowing density, field area, variety code, predecessor code, season code, campaign index	Variety and predecessor: integer `factorize` of lower-cased strings (not one-hot). Season: binary 2017/18 indicator. Campaign index: integer order. Δt, yield and harvest date are excluded.

That construction is a heuristic, not a claim that mean-centered absolute deviation is an optimal out-of-distribution score. Alternatives include a Mahalanobis distance to the training cloud [37], isolation forests [38], k-nearest-neighbor density and reconstruction error from an autoencoder. We did not compare those scores. Mahalanobis distance is unstable at the spectral dimensionality ($d = 26$) with a few dozen training rows per inner fold; isolation forests and autoencoders would add a second fitted model inside every nested split. Equation (10) was chosen as a fold-local, scale-free, missingness-aware proxy that the gate can consume without extra hyperparameters. Whether the proxy tracks expert error remains an empirical question, reported for q_m in Section 6.2. We do not claim an advantage over those alternatives.

Uncertainty u_m is given by Equation (8) or by the heteroscedastic head. It is tree-to-tree disagreement (or a learned σ_m), not a calibrated predictive interval. Section 6.2 reports the missing grouped-inner Spearman diagnostic of u_m versus expert $|\text{error}|$ on the screened sample. Row-level associations appear for several modalities; field-level tests mostly do not reject a zero median correlation. The signal is therefore not established as a field-level indicator of forecast unreliability, and it remains an additional gate feature rather than a validated reliability score.

4.3. The Agronomic Context Vector

The optional context vector **c** supplied to the gate at forecast time comprises: Zadoks z_t scaled by 100 (or the thermal-time proxy $10 + 55\min(\text{GDD}/1600, 1)$ when z_t is absent); days after sowing; horizon Δt; campaign index; $\sin(2\pi\,\text{DOY}/365)$ and $\cos(2\pi\,\text{DOY}/365)$; the interaction $(z_t/100)(\Delta t/100)$; and a one-hot encoding of Zadoks principal stage over $\{<25, 25\text{–}31, 32\text{–}44, 45\text{–}59, \geq 60\}$. That vector is independent of the phenology expert. Withholding the phenology expert at inference replaces its prediction with the training-fold mean increment $\overline{\Delta y}_{\text{tr}}$ and sets $(a_m, q_m, u_m) = (0, 0, 3)$, where 3 is a fixed missing-expert sentinel inside the $[0, 5]$ clip of Tables 7–9. On the scaled target used to train the gate, $(\Delta y - \bar{y}_{\text{tr}})/s_{\text{tr}}$, that mean increment is zero; on the original increment scale, it is not. The procedure does not blank observed Zadoks, days after sowing or campaign index from **c**, so it is phenology-expert withdrawal, not complete loss of phenological information. The ablation labeled "no stage context" in Appendix A sets `use_context=False`: it removes this entire vector, not only z_t. Calendar, horizon and campaign features are then removed together with phenology. When ground phenology scoring is unavailable in the historical reduced-input configuration of Appendix A, z_t is replaced by the thermal-time proxy so the remaining context stays defined. Gate ablation switches are listed in Table 8.

Table 7. Tree-expert SARG configuration used in the spatially screened increment tables of Sections 6.1–6.4 (`v20260920_spatial_verified`). Inner out-of-fold preprocessing is fold-local: each inner complement fits its own median imputer and standardizer; raw tree-spread is then divided by the median out-of-fold spread of that modality on the outer training fold and clipped to $[0, 5]$. A withheld expert is not that clip: its u_m is the fixed sentinel 3 in Table 9. Deployed experts are refit on the full outer training fold with a new imputer/scaler; test-time u_m uses the same fold-level median.

Item	Setting
Outer split	GroupKFold, 6 folds over 12 season–unit clients
Inner expert OOF	GroupKFold, 5 folds on outer-training clients
Gate width	one hidden layer, 16 tanh units; input $[\mathbf{c}; \mathbf{a}; \mathbf{q}; \mathbf{u}]$
Experts	ExtraTrees, 200 trees, depth 8, min. leaf 2
UAV encoding	31 patch statistics (CNN alternative not re-fit)
Optimizer	AdamW; learning rate 5×10^{-3}
Weight decay	3×10^{-3} on the context trunk only
Training duration	at most 1500 epochs; early stopping patience 200
Model selection	hold out $\approx 1/4$ of training clients; restore best validation MSE
Target scaling	$(\Delta y - \bar{y}_{\text{tr}})/s_{\text{tr}}$ inside the gate; invert on output
Temperature	$\tau = \exp(\log \tau)$ clamped to $[0.2, 5]$
Intercept	learned β_0; not part of the softmax
Warm-start $\boldsymbol{\pi}$	intercept-free NNLS, L1-normalized to the simplex, then log
Entropy regularizer	off (0)
Seed	42

The gate is a one-hidden-layer network with a learned global preference $\boldsymbol{\pi}$ and a learned temperature τ,

$$\boldsymbol{\alpha} = \text{softmax}\left[\frac{1}{\tau}\left(W_2 \tanh\left(W_1[\,\mathbf{c}; \mathbf{a}; \mathbf{q}; \mathbf{u}\,] + b_1\right) + b_2 + \boldsymbol{\pi}\right)\right], \tag{11}$$

where $\mathbf{c}$ is the agronomic context vector, $\mathbf{a}, \mathbf{q}, \mathbf{u}$ are the stacked availability, typicality and uncertainty vectors, W_1, b_1, W_2, b_2 are the gate weights and biases, $\boldsymbol{\pi}$ is a learned global preference over experts, τ is a learned softmax temperature, and $\boldsymbol{\alpha}$ is the resulting weight vector on the simplex. Bold symbols are vectors; W_1 and W_2 are matrices. The forecast is $\widehat{\Delta y} = \sum_m \alpha_m \hat{y}_m + \beta_0$, mapped to the absolute scale by Equation (5). Two implementation details are essential.

Table 8. Gate-input ablations, not expert-withdrawal tests. Agronomic context **c** includes observed Zadoks, days after sowing, campaign index and horizon. The "no stage context" row zeros **c** entirely (historical unfiltered sample only). It is the complete-absence-of-context condition and is distinct from phenology-expert withdrawal in Table 9. Availability + typicality and availability + uncertainty are evaluated on the screened sample in Section 6.2. A disabled block is replaced by zeros, not removed from the network.

Configuration	Context c	Availability a	Typicality q	Uncertainty u
Full gate (SARG)	yes	yes	yes	yes
Availability + typicality	yes (kept)	yes	yes	zeroed
Availability + uncertainty	yes (kept)	yes	zeroed	yes
Availability-only	yes (kept)	yes	zeroed	zeroed
No-reliability gate	yes	zeroed	zeroed	zeroed
No stage context (historical)	zeroed	yes	yes	yes

Table 9. Inference-time expert (sensor) withdrawal, not context ablation. Phenology-expert withdrawal removes the phenology expert only; observed Zadoks remains in gate context **c**. It is not the complete absence of agronomic context (Table 8). Biophysics-expert withdrawal removes LAI/fAPAR/AGB/N from the increment predictor; measured y_t is still required by Equation (5). This is expert withdrawal, not a fully sensor-free next-state forecast.

Quantity	Representation When Withheld	Zero Weight?
Expert prediction $\hat{y}_m$	training-fold mean increment $\overline{\Delta y}_{\text{tr}}$	no; softmax may still assign mass
Availability a_m	0	—
Typicality q_m	0	—
Uncertainty u_m	fixed sentinel 3 (inside the $[0, 5]$ clip; not the range maximum)	—
Phenology expert	as above; height, density, observed Zadoks dropped from that expert	observed Zadoks, DAS and campaign index remain in **c**
Biophysics expert	as above; current LAI/fAPAR/AGB/N dropped from that expert	measured y_t still added in Equation (5); not a sensor-free next-state forecast

1. The gate is trained on out-of-fold expert predictions. If it saw in-sample predictions, every expert would look almost perfect and there would be nothing to learn. Meta-features are generated by a five-fold split within the training fold.
2. That inner split is field-grouped, matching the outer protocol. A protocol diagnostic that compared inner-split policies on an earlier SARG implementation (before Δt was removed from the management expert) found random-inner RMSE = 0.860 (R^2 = 0.716) versus grouped-inner 0.989 (R^2 = 0.625) on ΔLAI (Figure 5; [34,39]). That 0.989 is not the gate in the main tables. After restricting Δt to gate context, the historical unfiltered grouped-inner SARG was RMSE = 1.045 (R^2 = 0.581). Primary fusion, missing-modality and walk-forward tables use the spatially screened run `v20260920_spatial_verified` (SARG ΔLAI RMSE 1.069, R^2 = 0.571). Random inner folds are reported only as an optimistic sensitivity. The gate warm-start is not the NNLS comparator of Section 5. The comparator is unconstrained-sum non-negative linear regression with an intercept (scikit-learn, nonnegative coefficients, intercept on). The warm-start instead solves intercept-free $\min_{w \geq 0} \|\hat{Y}w - \Delta y\|_2$ on the inner out-of-fold expert matrix $\hat{Y}$, L1-normalizes w onto the simplex, and sets

$\pi_m = \log\big(\mathrm{clip}(w_m, 10^{-3})\big)$. Softmax of $\boldsymbol{\pi}$ therefore starts at a fixed convex combination; β_0 is a separate learned intercept, initialized at zero. Weight decay is applied to the context trunk only, so $\boldsymbol{\pi}$ and τ are not pulled toward uniform weights.

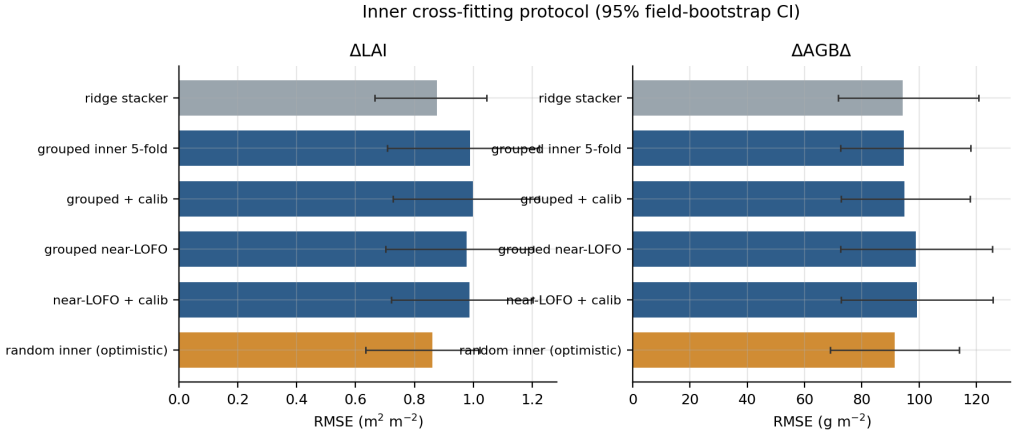


Figure 5. Inner-split protocol diagnostic on an earlier SARG implementation (management expert still saw Δt). Random inner folds (orange) make that model look stronger than a ridge stacker on ΔLAI; grouped inner folds and the ridge comparator (grey/blue in the legend) do not. This figure is a protocol diagnostic on an earlier unfiltered implementation. Primary tables use `v20260920_spatial_verified`.

4.4. Centralized Tree-Expert Protocol

The main tables use ExtraTrees experts and a small gate (Algorithm 1). Appendix B has a neural federated run that imputes and standardizes the concatenated training-client table in one place. That is not field-local training, and we did not re-fit it after screening.

Algorithm 1 Centralized tree-expert SARG used in the spatially screened increment tables of Sections 6.1–6.4.

Require: Training fold $\mathcal{D}_{\mathrm{tr}}$ grouped by season–unit clients

1: Field-grouped five-fold split of $\mathcal{D}_{\mathrm{tr}}$; fit one ExtraTrees expert per modality on each inner complement; record $(\hat{y}_m, a_m, q_m, u_m)$ out of fold.
2: Warm-start $\boldsymbol{\pi}$ from non-negative least squares of Δy on those $\hat{y}_m$; fit g_θ on the out-of-fold signals; refit experts on all of $\mathcal{D}_{\mathrm{tr}}$.
3: Predict $\widehat{\Delta y} = \sum_m \alpha_m \hat{y}_m + \beta_0$, then $\hat{y}_{t+1} = \max(0, y_t + \widehat{\Delta y})$ using measured y_t.

5. Experimental Setup

All models share the unified table, the seed, and GroupKFold over the twelve season–unit clients unless a different protocol is named. Compared methods on the screened sample are: data-level (early) fusion with gradient-boosted trees [40]; a ridge stacker over the seven ExtraTrees experts; a nonnegative linear stacker (NNLS: unconstrained-sum nonnegative regression with an intercept); SARG; an availability-only gate; a no-reliability gate; and a ridge trained with expert-column dropout. Availability-only gating still receives $\mathbf{c}$; it is availability plus agronomic context (Table 8). NNLS masking zeros the dropped coefficient and keeps the intercept; remaining weights are not L1-renormalized. Ridge masking zeros and L1-renormalizes. Dropout probability is selected in $\{0.15, 0.30, 0.45\}$ by inner GroupKFold on the training clients of each outer fold; for every fold selected, $p = 0.15$. Feature-level networks and the CNN branch were not re-fit after spatial screening. Seven single-modality ExtraTrees baselines use the same splits. A growth-rate ablation trains the same models on r_y and maps back with Equation (6). That ablation, and absolute-trained counterparts, belong to the historical

unfiltered sample; Figure 6 shows learned fusion models that are increment-trained on the screened run, together with the baselines of Section 6.1. Horizon-stratified RMSE of those same increment-trained field-out forecasts is reported in Section 6.5. Tree-spread u_m versus expert |error| uses the same ExtraTrees experts and grouped inner folds (Section 6.2); typicality q_m is scored on that same protocol. The four gate-signal configurations in Section 6.2 retrain only the gate on stored out-of-fold expert signals, keeping agronomic context on. Leave-one-farm-cluster-out and the weather-imputer alternatives use the same experts and clients (Table 3). Majority-overlap walk-forward repeats Section 6.4 on pairs with $f \geq 0.5$.

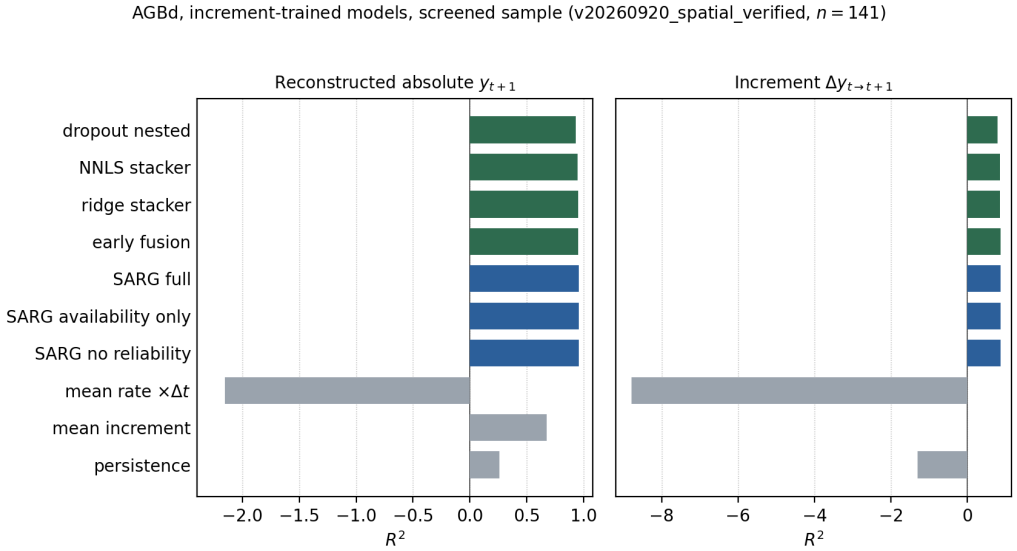


Figure 6. Dry above-ground biomass on the spatially screened sample (`v20260920_spatial_verified`, $n = 141$). All learned fusion models shown are increment-trained; baseline definitions follow Section 6.1. Bar colors distinguish methods as shown in the legend. Absolute-trained and growth-rate variants from the historical unfiltered sample are archived with the Supplementary Materials [32] (Appendix A). R^2 differs because $\mathrm{Var}(y_{t+1}) \neq \mathrm{Var}(\Delta y)$. The mean-increment baseline adds the training-set $\overline{\Delta y}$ to measured y_t; it is not a calendar-only model and does not ignore the crop.

Gate ablations and expert-withdrawal encodings are defined in Tables 8 and 9. Those two tables are different experiments and must not be read as one robustness claim. Resilience to loss of an expert (Table 9; Section 6.3) withholds one modality specialist at inference: the expert prediction is replaced by the training-fold mean increment and $(a_m, q_m, u_m) = (0, 0, 3)$. Agronomic context $\mathbf{c}$ stays on, so phenology-expert withdrawal is not the complete absence of Zadoks, days after sowing or campaign index. Resilience to the absence of agronomic context (Table 8, row "no stage context") zeros the entire vector $\mathbf{c}$—stage, horizon, calendar and campaign features together—while the seven experts remain. That ablation was fitted only on the historical unfiltered sample (Appendix A) and is not a missing-sensor test. Unavailable experts do not receive exactly zero weight. Softmax maps a finite logit vector to the interior of the simplex: if ℓ_m is the gate logit for the withheld expert, then $\alpha_m = e^{\ell_m} / \sum_k e^{\ell_k} > 0$ unless $\ell_m \to -\infty$. The encoding $(a_m, q_m, u_m) = (0, 0, 3)$ with $\hat{y}_m = \overline{\Delta y}_{\mathrm{tr}}$ is a missing-expert flag, not a hard mask. The gate can learn to send most of the mass elsewhere, but whatever mass remains multiplies the training-fold mean increment, not a live sensor prediction. Stacker masking is different: NNLS zeros the dropped coefficient; ridge zeros and L1-renormalizes. That is why a withheld expert can still move the fused forecast.

Numbers in the abstract, in Sections 6.1–6.4 and in Section 8 come only from the spatially screened run `v20260920_spatial_verified` ($n = 105$ LAI, $n = 141$ biomass, any-overlap). The inner-split diagnostic (Figure 5) and the unfiltered historical analyses in Appendixes A and B (gate weights, reduced-input ladder, CNN, federation) sit outside those claims. Additional tables, prediction files and sensitivity grids are provided in the Supplementary Materials [32].

Metrics are RMSE, MAE and R^2, with field-bootstrap 95% intervals and paired per-field Wilcoxon tests against the ridge stacker. Mean absolute percentage error is deliberately not reported for increment targets, because increments cross zero and the statistic diverges. Implementation is Python 3.13.14 with scikit-learn 1.8.0, XGBoost 3.4.1, PyTorch 2.12 (CPU) and rasterio 1.5.1; patch extraction over the seventeen flight archives takes about eleven minutes and is cached.

Sample size, capacity and overfitting.

The screened increment tables use $n = 105$ LAI and $n = 141$ biomass rows, grouped into twelve season–unit clients. That is a small meta-learning sample. The full gate is a 16-unit network with a learned prior and temperature, fitted on out-of-fold expert predictions from those clients; NNLS has eight free coefficients; a UAV ExtraTrees expert sees 31 statistics and no meta-layer. The practical safeguard is nested field-grouped cross-fitting (Figure 5): random inner folds made an earlier SARG look strongest on ΔLAI (RMSE $= 0.860$), whereas grouped inner folds reversed that ranking (0.989 on the same earlier implementation). Primary tables use the stricter protocol. We do not claim that $n = 105/141$ is large enough for a high-capacity gate; the ranking itself (NNLS and UAV ahead of SARG on fully observed ΔLAI) is the overfitting diagnostic. The recommendation follows: use the nonnegative stacker or the UAV unimodal expert on a fully observed LAI campaign, and availability-only gating among the gates when a sensor may be absent.

Hyperparameter protocol.

Headline ExtraTrees experts use 200 trees, depth 8 and minimum leaf size 2. The gate is one hidden layer of 16 tanh units. Nested dropout p is selected in $\{0.15, 0.30, 0.45\}$ inside each outer training fold; every fold selected $p = 0.15$. The inner-split policy (Figure 5) remains a protocol decision: random inner folds change the LAI ranking and are not used in the headline tables. On top of that protocol, we now re-run the fully observed full-gate versus availability-only comparison on the same spatially screened GroupKFold over a $3 \times 3 \times 3$ grid: gate width $\{8, 16, 32\}$, ExtraTrees depth $\{4, 8, 12\}$ and seeds $\{42, 7, 123\}$ (Table 10). That grid is a gate-only sensitivity, not a reprint of the headline joint fit. Experts are fit once per (depth, seed, outer fold); the gate is then retrained with `refit_gate` at each width, including the availability-only ablation. Nested dropout p is not part of this grid.

Because the gate is retrained, the grid cell at width 16, depth 8 and seed 42 is not the headline experiment. On ΔLAI, the RMSEs match the headline availability comparison in Section 6.3 (1.069 versus 1.080) but the field p differs (0.52 versus 0.47). On ΔAGBd, the gate-only refit reports SARG 99.1 versus availability-only 101.3, whereas the headline joint fit is 102.44 versus 101.14 ($p = 0.23$) and availability-only is ahead. The first row of Table 10 is that joint-fit headline; it is not marked as a grid cell. Across the 27 gate-only cells, SARG has the lower pooled RMSE in only eight LAI and eight biomass settings. Mean ΔRMSE is $+0.036$ on ΔLAI and $+1.91$ on ΔAGBd, so availability-only is better on average. Two unadjusted field Wilcoxon tests on ΔLAI reach 0.05 (width 8, seed 42, depths 8 and 12; ΔRMSE ≈ -0.01); none survive $0.05/27 \approx 0.0019$, and no biomass field test reaches unadjusted 0.05. Changing the seed moves availability-only ΔLAI RMSE from 1.08 to 0.99, a larger shift than width or depth. The central conclusion is unchanged: typicality and uncertainty do not earn a consistent benefit beyond availability-only gating.

Table 11 reports wall-clock fit time on the same nested grouped folds, now with identical ExtraTrees experts (200 trees, depth 8, minimum leaf 2) for both families. The seven deployed experts are fit once per outer fold; SARG then pays nested out-of-fold expert fits plus the 16-unit gate, while the stackers pay only ridge and NNLS on those same expert means. Times are means over the six outer folds on Windows 11, a Qualcomm ARMv8 eight-core CPU, 15.6 GB RAM and Python 3.13. Peak working set during the earlier audit was 835 MB (ΔLAI) and 919 MB (ΔAGBd). Almost all of the extra SARG cost is the nested ExtraTrees loop (7.8/8.3 s), not the gate (0.91/0.74 s). Stacker meta-fitting on the shared experts is <0.01 s. That extra nested-fit cost is not justified by a consistent accuracy gain. Inference on the same test fold, after warmup, is 0.12 s for SARG (includes per-tree spread) and <0.001 s for the linear combo of the same expert means. An independent 200/8 stacker that grows its own trees (omitted from Table 11; wall-clock times below) fits in 2.2/2.0 s and infers in 0.51/0.70 s.

Table 10. Fully observed increment RMSE on the spatially screened GroupKFold. The first row is the headline joint fit from Sections 6.2 and 6.3 (width 16, depth 8, seed 42). All other rows are a gate-only $3 \times 3 \times 3$ sensitivity: experts are fit once per (depth, seed, fold) and the gate is retrained with `refit_gate`. Those rows are not the headline experiment; the width-16/depth-8/seed-42 gate-only cell reverses the biomass ranking (99.1 vs. 101.3) relative to the joint fit (102.44 vs. 101.14). $\Delta = \mathrm{RMSE}_{\mathrm{SARG}} - \mathrm{RMSE}_{\mathrm{avail}}$; a negative Δ favors the full gate. Wilcoxon p is the unadjusted field test on the twelve season–unit clients. The full 27-cell grid is in the Zenodo deposit: SARG has lower pooled RMSE in 8/27 LAI and 8/27 biomass cells; mean Δ is +0.036 and +1.91.

Setting	ΔLAI			ΔAGBd		
	SARG	Avail.	Δ (p)	SARG	Avail.	Δ (p)
Joint fit (Section 6.3)	1.069	1.080	−0.010 (0.47)	102.4	101.1	+1.30 (0.23)
Width 8	1.069	1.079	−0.010 (0.005)	105.0	100.9	+4.08 (0.47)
Width 16	1.069	1.080	−0.012 (0.52)	99.1	101.3	−2.17 (1.00)
Width 32	1.072	1.082	−0.010 (0.79)	100.1	101.6	−1.56 (0.68)
Depth 4	1.016	1.036	−0.021 (0.68)	106.3	103.3	+2.94 (0.20)
Depth 8	1.069	1.080	−0.012 (0.52)	99.1	101.3	−2.17 (1.00)
Depth 12	1.057	1.069	−0.012 (0.13)	99.1	101.3	−2.22 (0.91)
Seed 7	1.064	0.990	+0.074 (0.11)	99.9	101.7	−1.73 (0.62)
Seed 42	1.069	1.080	−0.012 (0.52)	99.1	101.3	−2.17 (1.00)
Seed 123	1.068	1.017	+0.051 (0.38)	99.1	98.4	+0.69 (0.30)

Width, depth and seed rows are gate-only refits. Width rows hold depth 8 and seed 42; depth rows hold width 16 and seed 42; seed rows hold width 16 and depth 8.

Table 11. Wall-clock fit time per outer GroupKFold fold with matched ExtraTrees experts (200 trees, depth 8, min. leaf 2) on the spatially screened sample (Windows 11, Qualcomm ARMv8 8-core, 15.6 GB RAM, Python 3.13). Means over six folds. Deployed experts are shared. SARG total = nested out-of-fold ExtraTrees + deployed experts + gate. Stacker total = the same deployed experts + ridge and NNLS on those expert means. Inference on the same test fold (median of five repeats after warmup): SARG 0.12 s (includes tree-spread); linear combo of the same means < 0.001 s. Peak working set 835/919 MB.

Target	Experts	SARG OOF	Gate	SARG Total	Stacker Meta	Stacker Total
ΔLAI	1.44	7.75	0.91	10.1	<0.01	1.45
ΔAGBd	1.43	8.33	0.74	10.5	<0.01	1.43

Use of generative artificial intelligence.

Following MDPI's authorship and ethics policies on generative artificial intelligence (GenAI) [41,42], we disclose the following. During the preparation of this manuscript, the authors used a large language model writing assistant to propose LATEX edits. No GenAI tool generated data, figures, experimental results, statistical analyses, or study design. Su-

perficial grammar and formatting edits are not declared, as MDPI does not require that class of use to be declared. The authors reviewed every proposed change, take full responsibility for the originality, validity and integrity of the content, and do not list any GenAI tool as an author.

6. Results

6.1. Increments Versus Absolute Future States

Table 12 reports increment-trained models scored on both scales: the implied absolute state is Equation (5), not a separately trained absolute-state model.

Absolute next-campaign biomass can look almost solved ($R^2 = 0.956$ for increment-trained SARG in Table 12). That number is not increment skill. On pooled campaigns, $\mathrm{Var}(y_{t+1})$ is much larger than $\mathrm{Var}(\Delta y)$, so the same residuals produce a higher absolute R^2. A baseline that predicts the training-set mean increment and adds the measured y_t already reaches $R^2 = 0.675$ on absolute biomass and $R^2 = -0.009$ on ΔAGBd. Persistence and a mean growth-rate $\times\ \Delta t$ baseline are in Table 12 for both targets (biomass mean-rate RMSE 869, $R^2 = -8.79$ on ΔAGBd). We did not train a calendar-only or stage-only Δy model. On the increment scale, no statistically significant difference was detected between the gate and the ridge stacker (RMSE 102.4 vs. 105.2; cluster-bootstrap ΔRMSE CI −18 to +15). The median per-field ΔR^2 is 0.016, which is about the size of the field-to-field spread on this sample—we do not call 0.01 noise as a general rule.

Leaf area index really does move: persistence $R^2 = -0.586$ on the absolute scale. That is not proof there is no calendar shortcut; we never fitted a calendar-only model. The climatological increment has $R^2 = -0.010$ on ΔLAI. The nonnegative stacker is the most accurate fusion model on fully observed ΔLAI (RMSE 0.827, $R^2 = 0.744$), ahead of nested dropout (0.856), early fusion (0.887), the ridge stacker (0.913) and SARG (1.069). Figure 7 shows the corresponding full-gate increment and reconstruction scatter from the same out-of-fold predictions.

Table 12. Increment-trained models on the spatially screened sample, scored on both scales. RMSE is identical across the two score columns when y_t is observed and clipping is inactive; R^2 differs because $\mathrm{Var}(y_{t+1}) \neq \mathrm{Var}(\Delta y)$. Run `v20260920_spatial_verified`.

Method	Absolute y_{t+1}		Increment Δy	
	RMSE	**R^2**	**RMSE**	**R^2**
Leaf area index, $n = 105$				
Persistence	2.240	−0.586	2.240	−0.881
Mean increment	1.641	0.149	1.641	−0.010
Mean rate $\times\ \Delta t$	2.227	−0.568	2.227	−0.859
NNLS stacker	0.827	0.784	0.827	0.744
Dropout nested (p inner-selected)	0.856	0.769	0.856	0.726
Early fusion	0.887	0.752	0.887	0.705
Ridge stacker	0.913	0.737	0.913	0.688
SARG	1.069	0.639	1.069	0.571
Dry above-ground biomass ($\mathrm{g\,m^{-2}}$), $n = 141$				
Persistence	421	0.260	421	−1.298
Mean increment	279	0.675	279	−0.009
Mean rate $\times\ \Delta t$	869	−2.155	869	−8.790
SARG no reliability	101	0.957	101	0.868
SARG availability only	101	0.957	101	0.867
SARG full	102	0.956	102	0.864
Early fusion	104	0.955	104	0.860
Ridge stacker	105	0.954	105	0.857
NNLS stacker	110	0.950	110	0.843
Dropout nested	126	0.934	126	0.794

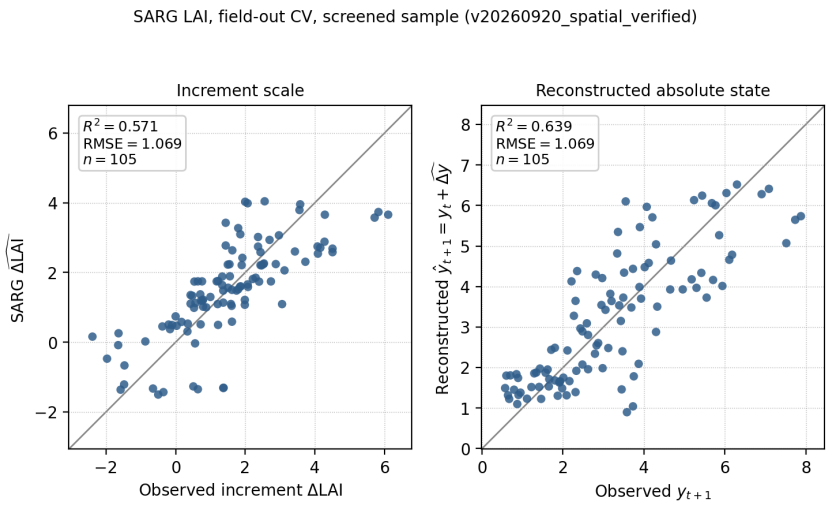


Figure 7. Leaf-area increment forecasts of the full gate under field-out cross-validation on the spatially screened sample (`v20260920_spatial_verified`; $n = 105$), regenerated from the same out-of-fold predictions as Table 12. (**left**) increment scale ($R^2 = 0.571$). (**right**) reconstruction to the absolute next campaign state by Equation (5) ($R^2 = 0.639$).

6.2. Fusion Strategies and Gate Ablations

Fusion strategies are compared on the increment, still under nested field-grouped cross-fitting on the screened sample. The full gate is not the most accurate fully observed model. On ΔLAI, UAV patch statistics remain the strongest single modality ($R^2 = 0.719$); the nonnegative stacker reaches 0.744, nested dropout 0.726, the ridge stacker 0.688, and SARG 0.571. On ΔAGBd, no statistically significant difference was detected between SARG and the ridge stacker (RMSE 102.4 vs. 105.2; cluster-bootstrap ΔRMSE CI −18 to +15; median per-field $\Delta R^2 = 0.016$). Availability-only (availability plus agronomic context) and no-reliability gates are slightly better than the full gate on fully observed biomass (RMSE 101.1 and 100.9 vs. 102.4; Wilcoxon $p = 0.23$ vs. availability-only). Forecasting the absolute next state is a fair task if one says so; a higher absolute R^2 is still not increment skill.

No consistent benefit beyond availability was established.

Dropping typicality and uncertainty, or dropping all reliability inputs, does not hurt fully observed accuracy (LAI 0.571 → 0.569; biomass 0.864 → 0.868). The comparison that tests the reliability claim is the full gate versus availability-only gating (availability plus agronomic context) (Section 6.3), not versus a fixed stacker. Field Wilcoxon tests and cluster-bootstrap intervals of pooled ΔRMSE answer different questions and need not agree. None of the field Wilcoxon tests reach unadjusted 0.05. The LAI bootstrap intervals exclude zero at the fourth decimal place, a difference of about 0.01 RMSE, while the field tests do not ($p = 0.47$ and 0.34). Those tests are exploratory and unadjusted; a nonsignificant field test is not equivalence. Row-wise Spearman tests of q_m against expert |error| ignore the fact that rows share a field. On biomass, soil and weather look convincing at row level ($p < 10^{-5}$) and disappear once each field is one observation ($p = 0.73$ and 0.16). UAV typicality does not track error.

Component attribution.

The four gate blocks can be switched independently (Table 8). Table 13 now separates typicality from uncertainty on the screened sample, with agronomic context kept in every row. Adding typicality alone does not move RMSE (LAI 1.080 → 1.076, $p = 0.34$; biomass 101.14 → 101.12, $p = 0.85$). Adding uncertainty alone, or both signals, can lower pooled RMSE after the gate is retrained (LAI 1.069; biomass 99.6 and 99.3), but ev-

ery field Wilcoxon test versus availability-only fails to reject ($p \geq 0.52$). Those full-gate biomass numbers are gate-only refits on stored expert signals; the headline joint fit remains 102.4 versus availability-only 101.1 ($p = 0.23$). Neither reliability signal contributes independently at field level. The historical `use_context=False` ablation likewise left pooled skill unchanged (Appendix A). The operative simpler alternative among the gates is availability plus context.

Table 13. Four-way gate-signal ablation on the spatially screened GroupKFold. Agronomic context **c** is on in every row. Experts are fit once per outer fold; only the gate is retrained. $\Delta = \text{RMSE} - \text{RMSE}_{\text{avail}}$; a negative Δ value favors the named configuration. Wilcoxon p is the unadjusted field test versus availability-only. Headline joint-fit SARG biomass RMSE is 102.4 (headline fusion comparison below); the 99.3 full-gate cell is a gate-only refit and does not replace that table.

Configuration	ΔLAI			ΔAGBd		
	RMSE	Δ	*p*	RMSE	Δ	*p*
Availability-only	1.080	—	—	101.1	—	—
Availability + typicality	1.076	−0.004	0.34	101.1	−0.01	0.85
Availability + uncertainty	1.069	−0.011	0.73	99.6	−1.56	1.00
Full SARG	1.069	−0.011	0.52	99.3	−1.81	0.97

Uncertainty versus expert error.

Table 14 is the independent check that was previously missing for u_m. On ΔLAI, weather, spectral, phenology and soil tree-spread correlate with |error| at the row level ($\rho = 0.38, 0.30, 0.27, 0.25$), but every field-level Wilcoxon test of the per-field Spearman coefficients fails to reject zero ($p \geq 0.13$). On ΔAGBd the row-level associations are stronger for UAV, biophysics, spectra, soil and phenology ($\rho = 0.42$ to 0.31); the field-level tests are unadjusted $p = 0.039$ (UAV) and $p = 0.010$ (phenology) and are not significant for the other modalities. Weather and management tree-spread do not track biomass error even at the row level. Taken with the gate ablations, u_m can correlate with expert error in pooled rows and still fail as a field-level reliability input that improves fusion. It remains an additional meta-feature, not a validated unreliability score.

Table 14. Held-out diagnostic of ExtraTrees tree-spread u_m against expert |error| on grouped-inner out-of-fold predictions (`v20260920_spatial_verified`). This is the check that u_m is not only a gate input. Row p-values treat visits as independent and are optimistic. Field p is an unadjusted Wilcoxon test of the per-field Spearman coefficients against zero. A positive ρ value means larger spread with larger error. Field-level evidence is weak, so u_m is interpreted cautiously as a reliability signal, not as a validated unreliability score.

Target	Modality	*n*	Row ρ	Row *p*	Median Field ρ	Field *p*
ΔLAI	UAV	105	0.13	0.20	−0.05	0.64
	Spectral	105	0.30	0.0016	0.26	0.22
	Biophysics	105	0.15	0.12	0.20	0.59
	Phenology	105	0.27	0.0047	0.20	0.42
	Soil	105	0.25	0.0089	−0.37	0.13
	Weather	105	0.38	6.5×10^{-5}	−0.14	0.84
	Management	105	0.20	0.038	0.34	0.63
ΔAGBd	UAV	141	0.42	1.5×10^{-7}	0.42	0.039
	Spectral	141	0.38	3.0×10^{-6}	0.43	0.063
	Biophysics	141	0.41	3.5×10^{-7}	0.31	0.12
	Phenology	141	0.31	1.9×10^{-4}	0.38	0.010
	Soil	141	0.32	1.0×10^{-4}	0.15	0.16
	Weather	141	0.07	0.38	−0.04	0.22
	Management	141	−0.08	0.32	0.41	0.63

Typicality versus expert error.

Table 15 is the matching held-out check for q_m, using the same ExtraTrees experts and grouped-inner protocol as Table 14. If the kernel worked as a reliability score, typicality would fall as |error| rises (negative ρ). On ΔLAI, only soil and weather have that sign at row level ($\rho = -0.37$ and -0.21); both field tests fail to reject zero ($p = 0.30$ and 0.64). On ΔAGBd, the same two modalities, plus phenology and management, are negative at row level; the only unadjusted field test below 0.05 is management ($p = 0.031$), and it does not survive a seven-modality correction. UAV typicality is indistinguishable from zero on both targets. Taken with the four-way gate ablation (Table 13), q_m is a heuristic input, not a validated typicality score.

Table 15. Held-out diagnostic of input typicality q_m against expert |error| on grouped-inner out-of-fold predictions (`v20260920_spatial_verified`), the same protocol as Table 14. A negative ρ is the expected direction (more typical inputs, smaller error). Row p-values treat visits as independent. Field p is an unadjusted Wilcoxon test of the per-field Spearman coefficients against zero.

Target	Modality	n	Row ρ	Row p	Median Field ρ	Field p
ΔLAI	UAV	105	0.01	0.89	0.09	0.61
	Spectral	105	−0.06	0.52	0.04	0.44
	Biophysics	105	0.10	0.29	0.20	0.48
	Phenology	105	−0.01	0.89	−0.10	0.56
	Soil	105	−0.37	1.2×10^{-4}	0.09	0.30
	Weather	105	−0.21	0.028	−0.31	0.64
	Management	105	0.17	0.090	0.03	0.84
ΔAGBd	UAV	141	0.03	0.74	0.06	0.55
	Spectral	141	−0.07	0.44	0.13	0.22
	Biophysics	141	−0.04	0.63	−0.14	0.28
	Phenology	141	−0.20	0.017	−0.07	0.78
	Soil	141	−0.38	4.0×10^{-6}	−0.02	0.73
	Weather	141	−0.38	2.6×10^{-6}	−0.12	0.16
	Management	141	−0.27	0.0014	−0.36	0.031

Stage-conditioned weights on the screened model.

Table 16 reports mean softmax weights of the final SARG model on held-out clients of the screened GroupKFold. On ΔLAI, the mass sits on UAV and phenology at tillering and stem elongation (UAV 0.33–0.38, phenology 0.32–0.34); November 2017 shifts toward management (0.41) as UAV falls from 0.40 to 0.22. On ΔAGBd, phenology dominates (0.48–0.65), except in November 2017, when management takes 0.64. Soil and weather stay near zero in every stratum. That pattern is readable, and it matches the older unfiltered table (Appendix A), but it is not a stable phenological interpretation that would justify the full reliability gate: the same November shift appears where availability-only is already as accurate, and turning context off on the unfiltered sample did not reduce pooled skill. We therefore treat the weights as a description of what this gate learned, not as evidence that typicality or uncertainty track stage.

Gate-weight, context-ablation and CNN analyses on the older unfiltered sample are in Appendix A. Table 17 reports the corresponding screened fusion comparison.

A neural federated run on the older unfiltered sample is in Appendix B; we did not re-fit it after screening, and it is a different protocol. Figure 8 is Table 17 drawn as bars. CNN and federated models are left off because they were not re-fit after screening.

Table 16. Mean softmax gate weights of the final screened SARG model on held-out GroupKFold clients (`v20260920_spatial_verified`). Strata with fewer than eight rows are omitted. Columns sum to one up to rounding. The table is descriptive; Table 13 is the accuracy test.

Stratum (n)	UAV	Spec.	Bioph.	Phen.	Soil	Weath.	Mgmt
ΔLAI							
Z < 25 tillering (75)	0.33	0.02	0.06	0.32	0.00	0.03	0.24
Z32–44 stem elong. (30)	0.38	0.02	0.05	0.34	0.00	0.03	0.16
November 2016 (17)	0.40	0.04	0.04	0.37	0.00	0.03	0.11
November 2017 (28)	0.22	0.00	0.08	0.27	0.00	0.02	0.41
ΔAGBd							
Z < 25 tillering (81)	0.05	0.00	0.07	0.48	0.01	0.00	0.39
Z32–44 stem elong. (30)	0.02	0.00	0.10	0.60	0.01	0.00	0.26
Z≥60 anthesis (30)	0.02	0.00	0.10	0.65	0.01	0.00	0.21
November 2016 (21)	0.02	0.00	0.11	0.64	0.02	0.00	0.21
November 2017 (30)	0.09	0.00	0.02	0.25	0.00	0.00	0.64

Table 17. Increment forecasts on the spatially screened sample (`v20260920_spatial_verified`), nested field-grouped cross-fitting. The UAV row is 31 patch statistics. CNN and feature-level networks were not re-fit after spatial screening (Appendix A) and are not headline results.

Method	Family	ΔLAI ($n = 105$)		ΔAGBd ($n = 141$)	
		RMSE	R^2	RMSE	R^2
NNLS stacker	decision	0.827	0.744	110	0.843
Dropout nested (p inner-selected)	decision	0.856	0.726	126	0.794
Early fusion	data-level	0.887	0.705	104	0.860
Ridge stacker	decision	0.913	0.688	105	0.857
Dropout $p = 0.30$ (fixed)	decision	0.965	0.651	145	0.729
SARG availability only	gated	1.080	0.563	101	0.867
SARG no reliability	gated	1.073	0.569	101	0.868
SARG full	gated	1.069	0.571	102	0.864
Unimodal: UAV patch statistics	single	0.865	0.719	149	0.712
Unimodal: phenology	single	0.947	0.664	119	0.817
Unimodal: current biophysics	single	1.055	0.583	108	0.849
Unimodal: management	single	1.023	0.608	106	0.853
Unimodal: weather	single	1.229	0.434	147	0.720
Unimodal: plot-mean spectra	single	1.529	0.123	263	0.106
Unimodal: soil	single	1.667	−0.042	271	0.048

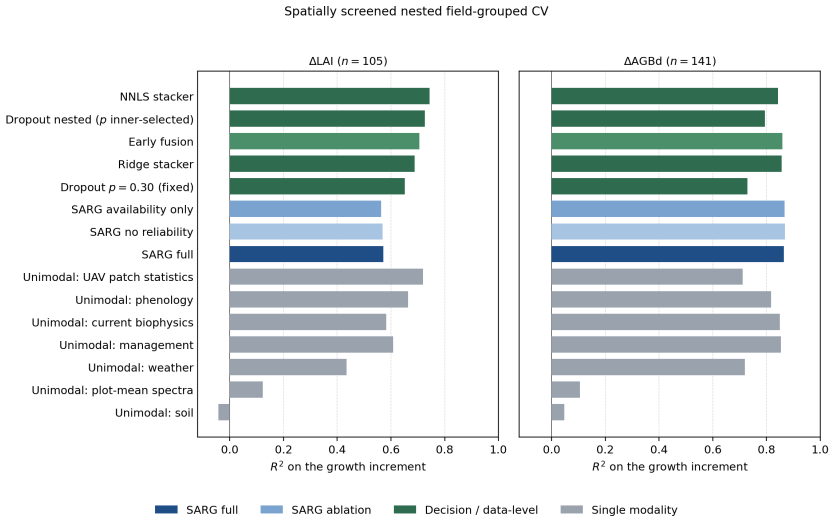


Figure 8. Increment-forecast R^2 on the spatially screened sample (`v20260920_spatial_verified`), matching Table 17. CNN and federated models were not re-fit after screening (Appendixes A and B) and are omitted from this figure. Bars are nested field-grouped point estimates, not field-bootstrap intervals.

6.3. Robustness to Missing Modalities

Table 18 withholds one expert at inference, still using the same ExtraTrees experts from `v20260920_spatial_verified`. Biophysics-expert withdrawal drops current LAI/fAPAR/AGB/N from the increment predictor only; Equation (5) still adds the measured y_t. That test is therefore not a fully sensor-free next-state forecast. Phenology-expert withdrawal does not remove observed Zadoks from gate context **c** (Table 9); it is not complete phenology-sensor loss. The reliability claim is again full gate versus availability-only gating (availability plus agronomic context), not versus a stacker. The nonnegative stacker zeros the dropped coefficient and keeps the intercept (no L1 renormalization). The masked ridge zeros and L1-renormalizes. Dropout p is chosen in $\{0.15, 0.30, 0.45\}$ by inner GroupKFold on the training clients of each outer fold; every fold picked $p = 0.15$.

Table 18. Absolute RMSE when one expert is withheld at inference. Phenology-expert withdrawal does not remove observed Zadoks from gate context. Same experts for every column. NNLS masking: zero dropped weights, keep intercept, no renormalization. Nested dropout: p selected inside training folds. Run `v20260920_spatial_verified`.

Withheld Expert	SARG	Avail. Only	Ridge Mask	NNLS Mask	Drop. Nested
ΔLAI					
None	1.069	1.080	0.913	0.827	0.856
UAV	1.178	1.216	1.399	1.726	1.183
Biophysics expert	1.072	1.081	0.972	1.033	0.941
Phenology expert	1.430	1.429	0.923	0.827	0.897
Weather	1.063	1.077	1.054	0.827	0.820
Soil	1.070	1.079	0.930	0.854	0.827
Spectra	1.064	1.074	0.906	0.844	0.855
Management	0.961	0.974	0.948	0.827	0.876
$\Delta AGBd$ ($\mathrm{g\,m^{-2}}$)					
None	102.4	101.1	105.2	109.9	126.0
UAV	109.8	113.3	147.5	162.6	128.6
Biophysics expert	107.0	112.0	149.5	224.9	128.8
Phenology expert	205.8	207.2	212.3	163.4	130.1
Weather	101.4	100.4	317.9	109.9	121.1
Soil	102.7	101.6	107.9	113.6	114.1
Spectra	104.0	101.1	113.6	109.8	112.5
Management	141.0	130.4	122.2	109.9	124.5

On fully observed ΔLAI, NNLS (0.827) and nested dropout (0.856) beat SARG (1.069). Withholding the phenology expert does not move NNLS (those weights are 0), while SARG rises to 1.430. Withholding UAV, nested dropout and SARG are similar (1.183 vs. 1.178); NNLS collapses to 1.726.

On ΔAGBd, availability-only gating is numerically ahead of the full gate when every sensor is present (101.14 vs. 102.44; field Wilcoxon $p = 0.23$; pooled ΔRMSE +1.30, cluster-bootstrap CI −3.41 to +6.14). Withholding the current biophysics expert, SARG has the lowest pooled point-estimate RMSE (107.01) against availability-only 111.99 ($p = 0.42$; ΔRMSE −4.98, CI −18.37 to +2.64), nested dropout 128.8 ($p = 0.34$), masked ridge 149.5 ($p = 0.077$) and NNLS 224.9 ($p = 0.002$). No field-level Wilcoxon comparison versus availability-only was significant at the unadjusted 0.05 level. Phenology-expert withdrawal leaves the two gates numerically close (205.79 vs. 207.20, $p = 0.97$). Management-expert withdrawal favors availability-only numerically (130.39 vs. 140.98, $p = 0.20$). The gate does reroute when an expert is withheld. Figure 9 plots the biomass block of Table 18. The LAI bootstrap-versus-Wilcoxon split is the same as in Section 6.2. Versus nested dropout under phenology-expert withdrawal the comparison favors dropout (130 vs. SARG 206, $p = 0.016$ unadjusted). Masked-ridge weather-expert withdrawal remains unstable (105 → 318) because weather coefficients are signed; NNLS assigns weather weight 0 so that withdrawal is a no-op.

Single-expert withdrawal does not describe a campaign in which two sensors fail together—a canceled UAV flight on a day with an empty weather window, or a missed ground-spectra visit together with a skipped canopy description. After the same nested ExtraTrees fit, we therefore also score four simultaneous dual-withdrawal pairs at inference: UAV+weather, UAV+spectral, biophysics+phenology, and weather+soil (Table 19). The encoding is the same as in Table 9: each withheld expert is replaced by the training-fold mean increment and $(a_m, q_m, u_m) = (0, 0, 3)$. Reconstruction still adds the measured y_t. Dual withdrawal is an inference-time stress test, not a training objective; the gate was not trained to expect those pairs. Table 19 is the resulting scorecard. Those four pairs are not an exhaustive 2^7 missingness design, and a dual-missing RMSE is not a claim that the gate was trained for simultaneous sensor loss. On ΔLAI, withdrawing UAV together with weather or spectra collapses the nonnegative stacker (1.726 and 1.821) while both gates remain near 1.17–1.24; availability-only is slightly ahead of the full gate. Withholding weather and soil together barely moves any LAI method that already ignored those experts. Biophysics-plus-phenology withdrawal still favors the stackers on LAI (1.009–1.033 versus SARG 1.377). On ΔAGBd, the numerical exception is UAV+weather: SARG stays near the fully observed row (103.0 versus 102.4), while availability-only rises to 120.8 and the masked ridge to 275.8. That is a point-estimate, not a field-tested win, and it does not repeat under biophysics-plus-phenology, where both gates jump above 230 g m^{-2}. Dual withdrawal therefore shows that a gate can reroute when two experts vanish; it does not establish a consistent benefit of typicality and uncertainty over availability-only gating. Table 20 collects the paired field tests of that contrast.

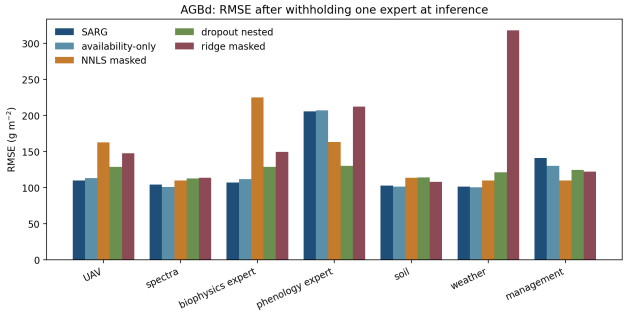


Figure 9. Biomass increment RMSE after withholding one expert at inference. Phenology-expert withdrawal leaves observed Zadoks in the gate context. Comparators are availability-only gating (availability plus agronomic context), masked ridge, nonnegative stacker and nested dropout ridge.

Table 19. Absolute RMSE when two experts are withheld simultaneously at inference on the spatially screened sample (`v20260920_spatial_verified`). Same ExtraTrees experts as Table 18. Dual withdrawal is an inference-time encoding (Table 9), not a training objective. NNLS masking zeros both dropped weights and keeps the intercept; ridge zeros and L1-renormalizes. The four pairs are a planned stress set, not an exhaustive missingness design.

Withheld Pair	SARG	Avail. Only	Ridge Mask	NNLS Mask
Δ*LAI*				
None	1.069	1.080	0.913	0.827
UAV + weather	1.230	1.200	1.316	1.726
UAV + spectral	1.240	1.174	1.467	1.821
Biophysics + phenology	1.377	1.417	1.009	1.033
Weather + soil	1.063	1.075	1.103	0.854
Δ*AGBd* (g m^{-2})				
None	102.4	101.1	105.2	109.9
UAV + weather	103.0	120.8	275.8	162.6
UAV + spectral	104.6	108.4	134.3	163.6
Biophysics + phenology	236.3	233.7	440.0	323.4
Weather + soil	101.5	99.2	330.2	113.6

Table 20. Full reliability gate versus availability-only gating (availability plus agronomic context) on the spatially screened sample. Two summaries are reported and need not agree. Pooled ΔRMSE is $RMSE_{SARG} - RMSE_{avail}$ on the concatenated out-of-fold rows, using unrounded RMSEs. The 95% interval is a field-cluster bootstrap of that same pooled difference (2000 resamples of the twelve season–unit clients). Wilcoxon p is an unadjusted signed-rank test of the twelve field-level RMSE differences. A negative pooled ΔRMSE favors SARG. No field-level Wilcoxon comparison was significant at the unadjusted 0.05 level. These tests are exploratory and unadjusted, and they use only twelve season–unit groups. A nonsignificant Wilcoxon p is not equivalence and is not definitive robustness. Small pooled LAI improvements appeared in exploratory bootstrap analyses, but field-level Wilcoxon tests did not establish improvement across fields.

Target	Condition	SARG	Avail. Only	Wilcoxon p	Pooled ΔRMSE (CI)
ΔAGBd	Fully observed	102.44	101.14	0.23	+1.30 (−3.41, +6.14)
	Biophysics expert withheld	107.01	111.99	0.42	−4.98 (−18.37, +2.64)
	Phenology expert withheld	205.79	207.20	0.97	−1.41 (−6.90, +0.81)
	Management withheld	140.98	130.39	0.20	+10.59 (−0.35, +25.21)
ΔLAI	Fully observed	1.0694	1.0798	0.47	−0.0104 (−0.0262, −0.00010)
	Biophysics expert withheld	1.0725	1.0813	0.34	−0.0089 (−0.0219, −0.000070)
	Phenology expert withheld	1.4302	1.4293	0.97	+0.00093 (−0.00183, +0.00270)
	Management withheld	0.9613	0.9740	0.34	−0.0128 (−0.0311, −0.00022)

6.4. Walk-Forward Evaluation

Walk-forward split.

Issue time for campaign C is the last `obs_date` among ESUs visited at C, so the campaign is issued as one batch. LAI is measured in situ on the observation date [19,28]. Dry above-ground biomass is an oven-dried laboratory mass; TS2AgroBg does not record processing completion dates, so the protocol treats the next-campaign `obs_date` as the label-availability date. That equation is exact for LAI and optimistic for ΔAGBd if drying lags the field visit. Complete input campaigns stay together: every ESU visited at C is training or test, never split by row. At each issue time, the training set is every increment row, on any participating client, whose label date is $\leq$ that batch issue date. Imputers, scalers, ExtraTrees experts, stacker coefficients, dropout p (inner GroupKFold on those training clients) and the SARG gate are fit on that training set only. No future labels, future covariates or issue-time test rows enter preprocessing or tuning. Using April $\rightarrow$ May labels to train a May $\rightarrow$ June forecast is permitted because those May measurements are already in hand under the observation-date convention; mixing April $\rightarrow$ May and May $\rightarrow$ June rows of the same campaign in one split is not. After spatial screening, 2016/17 yields two LAI tests (March $\rightarrow$ April, $n = 30$; April $\rightarrow$ May, $n = 30$) and three biomass tests (those plus May $\rightarrow$ June, $n = 30$). November $\rightarrow$ March ($n = 17$ LAI, $n = 21$ biomass kept) has no prior increment labels anywhere and is excluded from within-client temporal evaluation; those labels become training at the March issue. 2017/18 clients have only November $\rightarrow$ April ($n = 28$ LAI, $n = 30$ biomass) and are cold-start tests at the November 2017 issue, trained on all 2016/17 labels (available by June 2017 under the same convention). Table 21 lists the screened counts; row-level assignments and predictions are in the Zenodo reproducibility deposit [32].

Concatenated walk-forward scores on the screened sample are target-specific (Table 22). Pooled RMSE is computed on the concatenated test predictions, not by averaging per-issue R^2. On ΔLAI ($n = 88$), the ridge stacker (2.190) is numerically ahead of the mean-increment baseline (2.234) and of SARG (2.365). On ΔAGBd ($n = 120$), the mean increment (368) beats SARG (427), NNLS (423), ridge (428) and persistence (452). Every concatenated R^2 is negative. Negative R^2 means worse than predicting the test-set mean increment; it does not mean that every deployable baseline is better. Ridge therefore records a small numerical

LAI RMSE improvement over the mean-increment baseline, while practical or statistically supported temporal forecasting skill remains unestablished on this sample.

The uniformly negative concatenated R^2 in Table 22 is not a software error and is not specific to SARG. Coefficient of determination is $1 - \sum_i (y_i - \hat{y}_i)^2 / \sum_i (y_i - \bar{y}_{\text{test}})^2$. A negative value means the method is worse than predicting the test-set mean increment. Three structural facts produce that outcome here. First, the increment mean is non-stationary across campaigns: April → May biomass change is a different population from November → March, so a model trained on earlier transitions is not aiming at the test mean that R^2 uses as its reference. Second, the mean-increment baseline itself uses the training mean, not the test mean, and is therefore also negative (-0.62 on ΔLAI, -0.61 on ΔAGBd). Third, each issue fold is small ($n = 28$–30) and the concatenated pool mixes short spring gaps with a 2017/18 cold-start, so residuals are large relative to the pooled test variance. RMSE remains the fairer ranking statistic on that table: ridge is slightly ahead of the training-mean baseline on ΔLAI (2.190 vs. 2.234) and the training-mean baseline is ahead of every learned method on ΔAGBd (368). Negative R^2 therefore diagnoses distribution shift and a moving increment mean, not a sign error in the implementation.

A neural federated walk-forward run on the unfiltered sample is in Appendix B; we did not re-fit it after screening.

Field-out GroupKFold (Tables 12 and 17) and concatenated walk-forward answer different questions. The former holds out a season–unit client but still trains on other clients from the same campaign dates; it measures spatial transfer of an increment given y_t, not the ability to forecast a future campaign from earlier ones. Because every concatenated walk-forward R^2 is negative and SARG is behind the mean-increment baseline on ΔAGBd (427 vs. 368) and behind ridge on ΔLAI (2.365 vs. 2.190), the primary field-out scores must not be read as evidence of practical next-campaign forecasting skill. Repeating the protocol on majority-overlap pairs ($f \geq 0.5$; $n = 75$ LAI, $n = 106$ biomass) does not change that reading: every concatenated R^2 remains negative; ridge 2.29 still leads ΔLAI, availability-only (2.46) beats the full gate (2.86), and on biomass, the mean-increment baseline (377) again beats every learned method (≈ 436–443).

Table 21. Walk-forward transition counts on the spatially screened sample. Issue date is the last `obs_date` of the input campaign. November → March is training at the March issue, not a test fold.

Input → Target	Issue (Batch)	*n* LAI	*n* AGBd	2016/17 Role	2017/18 Role
Nov → Mar	Nov 2016	17	21	excluded (no prior labels)	—
Mar → Apr	Mar 2017	30	30	test; train on Nov → Mar	—
Apr → May	Apr 2017	30	30	test; train on prior two	—
May → Jun	May 2017	0	30	test (biomass only)	—
Nov → Apr	Nov 2017	28	30	—	cold-start test

Table 22. Concatenated walk-forward RMSE and R^2 on the spatially screened sample (`v20260920_spatial_verified`). Metrics are computed once on all test predictions, not averaged across issue dates. Every R^2 is negative because the increment mean is non-stationary and R^2 is scored against the test-set mean, whereas the mean-increment baseline uses the training mean; see the text.

Method	ΔLAI ($n = 88$)		ΔAGBd ($n = 120$)	
	RMSE	R^2	RMSE	R^2
Persistence	2.381	−0.84	452	−1.43
Mean increment	2.234	−0.62	368	−0.61
Ridge stacker	2.190	−0.56	428	−1.18
NNLS stacker	2.284	−0.69	423	−1.13
SARG	2.365	−0.82	427	−1.16

6.5. Horizon-Stratified Field-Out Scores

Table 23 splits the same nested field-out predictions as Table 17 by transition and by a short/long cut at $\Delta t = 80$ d (the gap between the spring campaigns and the two November departures). The gate already sees Δt in **c**, so this is a within-SARG check, not a new model. It does not replace the growth-rate ablation; it asks whether the increment ranking is an artifact of pooling unequal intervals.

It is not. On short ΔLAI ($n = 60$, median 28 d), the nonnegative stacker remains ahead (RMSE 0.966) of ridge (1.097) and of both gates (SARG 1.296, availability-only 1.311). On long ΔLAI ($n = 45$, median 145 d), ridge is ahead (0.583) and the two gates are tied (0.655 vs. 0.654). On short ΔAGBd ($n = 90$), availability-only and the no-reliability gate are numerically ahead of the full gate (107.5 and 107.4 vs. 110.5); on long ΔAGBd ($n = 51$), NNLS is ahead (78.2) of SARG (86.4) and availability-only (88.8). Per-transition rows show the same pattern: the mean increment is a poor short-horizon biomass predictor (April → May RMSE 437), which is exactly the horizon-length confounding the growth-rate ablation was meant to flag, yet replacing the target by r_y is unnecessary to see that SARG does not systematically beat availability-only inside any single interval. The full reliability modification therefore has no consistent within-horizon advantage either.

Table 23. Horizon-stratified increment RMSE on the spatially screened field-out predictions (`v20260920_spatial_verified`). Short/long uses $\Delta t = 80$ d. These are the same out-of-fold forecasts as Table 17, not a retrained multi-horizon model.

Target	Transition	*n*	Δ*t*	Mean Incr.	Ridge	NNLS	SARG	Avail.
ΔLAI	Apr → May	30	22	1.994	1.074	0.830	1.234	1.261
	Mar → Apr	30	35	2.146	1.120	1.085	1.354	1.360
	Nov → Mar	17	131	0.757	0.535	0.564	0.518	0.514
	Nov → Apr	28	146	0.746	0.610	0.613	0.726	0.726
	Short ($\Delta t < 80$)	60	28	2.072	1.097	0.966	1.296	1.311
	Long ($\Delta t \geq 80$)	45	145	0.750	0.583	0.595	0.655	0.654
ΔAGBd	Apr → May	30	22	437	146	121	97.5	97.0
	Mar → Apr	30	35	299	68.3	76.4	43.2	41.4
	May → Jun	30	39	201	122	161	159	153
	Nov → Mar	21	131	178	47.2	48.3	34.9	34.7
	Nov → Apr	30	146	152	97.4	93.6	109	112
	Short ($\Delta t < 80$)	90	35	327	117	124	111	107
	Long ($\Delta t \geq 80$)	51	145	163	80.6	78.2	86.4	88.8

7. Discussion

7.1. Why the Target Definition Is a Methodological Choice

The decision that matters most here is the target, not the gate. An R^2 of about 0.96 on next-campaign biomass and 0.68 for a mean-increment baseline that adds the measured y_t describe residuals of very different variances. They are not a split of skill into seasonal progression. When y_t is observed and clipping is off, $\hat{y}_{t+1} - y_{t+1} = \widehat{\Delta y} - \Delta y$: the errors are the same, and only $\mathrm{Var}(y_{t+1})$ versus $\mathrm{Var}(\Delta y)$ changes R^2. The mean-increment baseline still uses the crop. Persistence and a mean growth-rate × Δt baseline are in Table 12. We did not train a calendar-only, stage-only or horizon-only predictor of Δy, so we have not measured seasonal predictability beyond those controls. Anyone pooling campaigns should report both scales and those baselines. On this dataset, the increment also improved absolute-scale LAI, so the more honest parameterization was not the more pessimistic one.

That honesty cuts the other way as well. Because reconstruction always adds the measured y_t, a strong increment R^2 is still not evidence that a future canopy state can be read from UAV, weather or soil layers alone. The method addresses incremental change given a current ground visit. Remote next-state forecasting without y_t is a different task and is

not claimed. Horizon-stratified scores (Table 23) further show that pooling Nov → Mar with Apr → May mixes unequal Δt; the growth-rate ablation remains a complementary control, not a substitute for that within-SARG breakdown.

7.2. What the Gate Contributes, and What It Does Not

The gate did not win fully observed accuracy. On screened ΔLAI, the nonnegative stacker (RMSE 0.827) and nested dropout (0.856) beat SARG (1.069); UAV patch statistics alone reach $0.865/R^2 = 0.719$ and already beat the full gate. That ordering is the expected signature of a high-capacity meta-model on a small out-of-fold expert table, not of unused reliability information. The gate is a 16-unit network with a learned prior and temperature, fitted on roughly 100 inner-fold predictions; NNLS has 8 free coefficients (7 nonnegative weights plus an intercept); the UAV ExtraTrees expert sees 31 statistics and no meta-layer. Random inner folds made an earlier SARG look strongest on LAI (Figure 5); grouped inner folds reverse that ranking, which is the classical over-fit-in-model-selection pattern [39]. Typicality and uncertainty do not rescue the full gate once they are ablated (Section 6.2). Practical recommendation: on a fully observed campaign, use the nonnegative stacker or the UAV unimodal expert for ΔLAI, and treat availability-only gating as the default among the gates when a sensor may be absent. Do not deploy the full reliability gate for pooled accuracy on a sample of this size.

On screened ΔAGBd, no statistically significant difference was detected between SARG and the ridge stacker (RMSE 102.4 vs 105.2; cluster-bootstrap ΔRMSE CI −18 to +15). The median per-field ΔR^2 is 0.016, about the field-to-field spread—again, not a general claim that 0.01 is noise.

The reliability claim is tested by the full gate versus availability-only gating (availability plus agronomic context) (Table 20). On fully observed biomass, the pooled ΔRMSE is +1.30 in favor of availability-only (cluster-bootstrap CI −3.41 to +6.14; field Wilcoxon $p = 0.23$). Under withheld biophysics, SARG is numerically ahead (pooled ΔRMSE −4.98, CI −18.37 to +2.64; $p = 0.42$). Phenology-expert withdrawal is numerically close, with no statistically significant difference detected ($p = 0.97$); observed Zadoks remains in the gate context, so this is not complete phenology-sensor loss. Management-expert withdrawal leans toward availability-only ($p = 0.20$). Those summaries need not agree, and the LAI rows of Table 20 show they do not: the bootstrap sees a small pooled difference, the field Wilcoxon does not. Expert-withdrawal tests show that the gate reroutes; they do not establish a consistent benefit of typicality and uncertainty.

Comparisons with fixed stackers remain relevant as a second check, not as a substitute. Withholding the current-biophysics expert from biomass, SARG has the lowest pooled RMSE (107.0 vs. nested dropout 129, masked ridge 150, NNLS 225). The paired field test versus nested dropout is not statistically established ($p = 0.34$). Versus NNLS, the comparison is exploratory evidence favoring SARG ($p = 0.002$ unadjusted), subject to multiplicity. Versus the masked ridge, $p = 0.077$ unadjusted and does not clear 0.05. Phenology-expert withdrawal is the reverse: nested dropout 130 vs. SARG 206 ($p = 0.016$ unadjusted). On LAI, withholding the biophysics expert does not favor SARG (1.072 vs. nested dropout 0.941 and masked ridge 0.972). q_m is input typicality: row-wise Spearman tests ignore within-field dependence; soil and weather on biomass have row $p < 10^{-5}$ but field-level Wilcoxon $p = 0.73$ and 0.16. UAV typicality does not track error. The score is a Laplace-like average of absolute z-scores times a completeness factor (Equation (10)); we did not compare it with Mahalanobis distance, isolation forests or reconstruction error, and we do not claim it is a superior out-of-distribution detector. u_m is now checked against expert |error| (Table 14): several row-level Spearman coefficients are positive, but field-level tests mostly do not reject zero (unadjusted exceptions on biomass: UAV $p = 0.039$,

phenology $p = 0.010$). Combined with the ablations, neither signal is established as the component that improves fusion.

Mean softmax weights of the final screened SARG model (Table 16) are agronomically readable and still not an accuracy finding. On held-out ΔLAI, the mass sits on UAV and phenology at tillering and stem elongation (UAV 0.33–0.38, phenology 0.32–0.34); November 2017 shifts toward management (0.41) as UAV falls from 0.40 to 0.22. On ΔAGBd, phenology dominates (0.48–0.65), except in November 2017, when management takes 0.64. Soil and weather stay near zero in every stratum. The older unfiltered table (Appendix A) shows the same qualitative pattern. That pattern is consistent with a single-date UAV patch being more informative for leaf-area change than for biomass change, and with biomass tracking stage and management more closely than a four-band window. It is not a stable phenological interpretation that would justify the full reliability gate: availability-only is already as accurate on the same screened sample (Tables 13 and 20), turning context off on the unfiltered sample did not reduce pooled skill, and we did not see a stage-specific robustness or transfer gain. The architecture can depend on z_t; that is not a finding that stage inputs improve point forecasts here, and the title does not say they do. The context ablation and the end-to-end CNN were not re-fit after screening (Appendix A).

The evaluation choice is clearer than the accuracy ranking. The gate has to train on out-of-fold expert predictions, and those predictions have to come from grouped inner folds (Figure 5; [34,39]). Random inner folds made an earlier SARG look strongest on LAI; grouped folds reverse that ranking. Warm-starting π from the L1-normalized intercept-free NNLS of the out-of-fold experts, and leaving π and τ out of weight decay, still helps: otherwise, the softmax collapses toward uniform weights. That mapping is not the unconstrained-sum NNLS comparator.

7.3. Deployment Without a Frequent Ground Visit

Reconstruction Equation (5) always adds the measured current state y_t. The method is therefore an increment forecast given a ground visit, not a remote next-state product. That matches one operational setting and not another. After a scheduled LAI or biomass visit—growth-regulator timing, a nitrogen check, a destructive cut that will not be repeated on the same plot—the next-campaign change can be forecast from the remaining sensors, and a canceled UAV flight or an empty weather window can be flagged by availability. It does not match a setting in which no current LAI or biomass is measured. Frequent destructive biomass sampling is not required by the increment protocol itself: the cadence is the campaign gap (weeks to months), and y_t is whatever was measured at the last visit. Daily or weekly remote updates without that visit are a different task. The practical path we recommend, and did not evaluate, is to replace measured y_t with a retrieved current state (UAV or satellite LAI and biomass) and to report that pipeline separately from increment-given-measurement (Section 8). Until that replacement is scored, the honest deployment claim is narrower: given a current in situ measurement, fuse the available streams for the next increment, and prefer availability-only gating or the UAV unimodal expert over the full reliability gate on a sample of this size.

7.4. Imagery at Campaign Scale

On the screened sample, UAV patch statistics are still the best single LAI modality ($R^2 = 0.719$). A convolutional encoder trained end to end on the older unfiltered sample did not beat those statistics (Appendix A). That is a statement about this campaign size, not about deep image fusion in general, which works at larger scale [22,24].

7.5. Limitations

We have two seasons, six agricultural-unit labels and twelve distinct season–unit clients (206 ESU visits) on commercial winter wheat in one Bulgarian landscape. Grouped folds cannot separate variety from soil from management. The protocol does not establish transfer to other climates, irrigation regimes, soils, management systems or crop types. Weights, typicality kernels and UAV statistics fitted here would have to be re-estimated under the same nested field-grouped design on any new domain; we do not treat the present $\boldsymbol{\alpha}$ or RMSE ranking as portable. What may transfer is the evaluation recipe—spatially screened increments, availability-only versus full-gate ablation, expert withdrawal, and the refusal to read field-out R^2 as temporal skill—not the numerical results. Wilcoxon tests on twelve pairs are under-powered and unadjusted; we never named a primary hypothesis. Several dozen comparisons are reported across two targets, seven withdrawals, four gate configurations, seven modalities and the horizon strata; a Bonferroni threshold of $0.05/8 = 0.00625$ on the eight tests in Table 20 leaves none significant. Isolated unadjusted $p < 0.05$ values are exploratory. Because those field tests are under-powered and unadjusted, we do not treat a numerical RMSE edge as a meaningful improvement of reliability-gated fusion over availability-only gating or stacking. Dual-expert withdrawal (Table 19) is an inference-time stress set of four pairs, not a trained missingness model and not an exhaustive 2^7 design. Gate width, ExtraTrees depth and seed were re-run on the same grouped protocol (Table 10); that grid is a gate-only refit and is not the headline joint fit. The full-versus-availability conclusion does not reverse. Learning rate was not swept. Wall-clock cost is reported in Table 11 with matched ExtraTrees (200 trees, depth 8) on the same outer folds and the same deployed experts: the extra SARG cost is nested expert cross-fitting, not the 16-unit gate, and it is not justified by a consistent accuracy gain. Any-overlap of 20 m windows removes large jumps; it does not prove the same sampling area (Table 4). Majority overlap ($f \geq 0.5$) repeats the fully observed and missing-modality tables (Table 5) and, in this revision, concatenated walk-forward ($n = 75/106$); every majority walk-forward R^2 remains negative. Leave-one-farm-cluster-out and the weather-imputer alternatives are in Table 3. The four-way typicality/uncertainty ablation is Table 13. Under biomass biophysics withdrawal, SARG has the lowest pooled RMSE, but the field tests versus availability-only ($p = 0.42$) and nested dropout ($p = 0.34$) are not significant at unadjusted 0.05; versus masked ridge $p = 0.077$, versus NNLS $p = 0.002$ unadjusted. Most LAI increments sit in early stages. The UAV patches are single-date windows. We did not model harvest yield. Uncertainty u_m is now checked against expert |error| (Table 14); the association is row-level for several modalities and is not established at field level after multiplicity. Typicality q_m is checked on the same screened protocol (Table 15): soil and weather have the expected negative row-level sign, but field tests do not reject zero after multiplicity, and UAV typicality does not track error. The score remains a heuristic z-score kernel, not a comparison with Mahalanobis or isolation-forest alternatives. Field-out CV is spatial transfer of an increment given y_t, not temporal skill (Section 6.4). The historical typicality diagnostic (Appendix A) is the older unfiltered companion for q_m; field-level correlations there also mostly do not reject zero. The forecasts are not distribution-calibrated; conformal intervals on the increment are the obvious next step. The older within-field 30% row split is not a prospective test. Neighbouring units are 90–230 m apart (nearest ESU), so field-out CV still leaves spatial dependence among farm-cluster neighbours that share a weather station; client-mean residual correlations among those neighbors were not significant at unadjusted 0.05, and a distance-weighted Moran's I on ESU residuals was not computed. Leave-one-farm-cluster-out over the four 2 km clusters (Table 3) raises RMSE and does not produce a full-gate win. November 2016 weather is empty (Table 2); fold-local mean and KNN fills do not

reverse the availability-only ranking, and weather-expert withdrawal still moves RMSE only weakly. Continuity screening is in Section 3.3: 3/108 LAI and 5/146 biomass transitions were identifier jumps. The federated run in Appendix B standardizes concatenated training-client rows and was not re-fit after screening.

Tables 24 and 25 collect the headline claims on the spatially screened run, including those that go against the full gate.

Table 24. Headline findings on the spatially screened run `v20260920_spatial_verified`. Historical unfiltered analyses are in Appendixes A and B. Paired p-values are unadjusted.

Claim	What It Means	Evidence	Take-Home
Increments are spatially screened	Any-overlap of 20 m windows is not identical sampling area.	Displacement q_{50}/q_{90}: 3.84/9.39 m (LAI), 2.86/9.34 m (biomass). Overlap f $q_{50} = 0.75/0.82$; $f < 0.5$ for 14/17 retained pairs. Majority $f \geq 0.5$ field-out $n = 91/124$: nested dropout overtakes NNLS on ΔLAI; availability-only remains ahead of the full gate.	Headline tables use any-overlap $n = 105/141$. Majority walk-forward ($n = 75/106$) still has every concatenated $R^2 < 0$.
Increment and absolute R^2 use different denominators	Reconstruct $\hat{y}_{t+1} = y_t + \widehat{\Delta y}$ with measured y_t.	Same residuals; different $\mathrm{Var}(y_{t+1})$ vs. $\mathrm{Var}(\Delta y)$. Mean increment: absolute biomass $R^2 = 0.675$, ΔAGBd −0.009.	Persistence and mean-rate baselines are in Table 12.
Typicality and uncertainty are unproven	The test of reliability conditioning is full SARG versus availability-only (Tables 13 and 20). Typicality and uncertainty are now switched independently, with context kept.	Typicality alone: LAI 1.076 vs. 1.080 ($p = 0.34$); biomass 101.12 vs. 101.14 ($p = 0.85$). Uncertainty alone and the full gate can lower pooled RMSE after a gate-only refit; no field Wilcoxon $p < 0.05$. Headline joint-fit biomass remains 102.4 vs. 101.1 ($p = 0.23$).	Neither reliability signal contributes independently at field level.
The full gate is not the best fully observed model	NNLS and nested dropout beat SARG on ΔLAI; on biomass, SARG and the ridge stacker are numerically close, with no statistically significant difference detected.	ΔLAI: NNLS 0.827/0.744; nested dropout 0.856; SARG 1.069/0.571. ΔAGBd: no-reliability 100.9; availability-only 101.1; SARG 102.4; ridge 105.2.	Fully observed accuracy does not justify the full reliability gate.
Biophysics withdrawal is exploratory	Compare SARG with availability-only, NNLS and nested dropout under every withheld modality.	Biomass biophysics expert: SARG 107.0 (lowest point estimate), avail. 112.0 ($p = 0.42$), dropout 128.8 ($p = 0.34$), masked ridge 149.5 ($p = 0.077$), NNLS 224.9 ($p = 0.002$). Phenology expert: dropout 130 vs. SARG 206 ($p = 0.016$).	Lowest pooled RMSE under biomass biophysics withdrawal is not a paired win versus availability-only.
Walk-forward skill is not established	Concatenate test predictions; include persistence and mean increment. Field-out CV is not this protocol.	Screened sample: LAI $n = 88$, ridge 2.190 vs. mean increment 2.234 vs. SARG 2.365; biomass $n = 120$, mean increment 368 vs. SARG 427. All concatenated $R^2 < 0$.	Ridge has a small numerical LAI RMSE gain; temporal skill is not established. Field-out scores must not be read as next-campaign forecasting skill. Counts: Tables 21 and 22. Majority-overlap walk-forward ($n = 75/106$) remains $R^2 < 0$.

Table 25. Headline findings on the spatially screened run (`v20260920_spatial_verified`), continued: diagnostics and cost. Paired p-values are unadjusted.

Claim	What It Means	Evidence	Take-Home
Horizon pooling is not the driver	Same field-out forecasts, split at $\Delta t = 80$ d and by transition (Table 23).	Short ΔLAI: NNLS 0.966 vs. SARG 1.296. Long ΔLAI: ridge 0.583 vs. SARG 0.655. Short ΔAGBd: avail. 107 vs. SARG 111. Long ΔAGBd: NNLS 78.2 vs. SARG 86.4.	Full reliability gating has no consistent within-horizon advantage. Growth-rate ablation remains complementary.
u_m and q_m are not validated field-level scores	Grouped-inner Spearman of tree-spread and typicality versus expert |error| (Tables 14 and 15).	u_m: row ρ up to 0.42 (biomass UAV); field exceptions UAV 0.039, phenology 0.010. q_m: soil/weather row $\rho \approx -0.38$; field tests mostly $p > 0.05$; biomass management $p = 0.031$ unadjusted.	Both signals can correlate in pooled rows and still fail as fusion reliability inputs.
Dual withdrawal is a stress set	Four inference-time pairs (Table 19); not a trained missingness model.	ΔLAI UAV+weather: avail. 1.200 vs. SARG 1.230 vs. NNLS 1.726. ΔAGBd UAV+weather: SARG 103.0 vs. avail. 120.8 vs. ridge 275.8. Biophysics+phenology biomass: both gates > 230.	Gating can reroute when UAV is among the missing pair; typicality and uncertainty are not established as the reason.
The full gate costs more to fit	Matched ExtraTrees (200/depth 8) on the same outer folds and the same deployed experts (Table 11).	Host: Windows 11, Qualcomm ARMv8 8-core, 15.6 GB RAM. Shared experts 1.4 s. SARG nested OOF 7.8/8.3 s plus gate 0.91/0.74 s (total 10.1/10.5 s). Stacker meta on those experts < 0.01 s (total 1.4 s).	Extra cost is the nested ExtraTrees loop, not the gate, and is not justified by a consistent accuracy gain.
Width, depth and seed do not reverse the gate ranking	Gate-only $3 \times 3 \times 3$ grid; not the headline joint fit (Table 10).	Headline joint fit: biomass SARG 102.4 vs. avail. 101.1 ($p = 0.23$). Gate-only default cell: 99.1 vs. 101.3 ($p = 1.00$). SARG lower RMSE in 8/27 LAI and 8/27 biomass cells. Mean ΔRMSE +0.036/+1.91.	The grid cell at width 16, depth 8, seed 42 is not the headline configuration. The ranking conclusion does not reverse.

8. Conclusions

Two evaluation protocols must not be collapsed. Nested field-out GroupKFold measures spatial transfer of an increment given a measured y_t. Concatenated walk-forward is the prospective next-campaign test; every R^2 on that table is negative (Table 22), so the field-out ranking is not evidence of temporal forecasting skill. That distinction is the first conclusion of this paper.

This study asked a bounded question on TS2AgroBg: under what conditions, if any, do typicality and uncertainty improve winter-wheat increment fusion beyond availability-only gating (availability plus agronomic context)? Consecutive visits were kept only when their 20 m windows overlap ($n = 105$ LAI, $n = 141$ biomass); slivers with $f < 0.10$ remain, and Table 5 is the majority-overlap companion. Headline UAV experts use 31 patch statistics, not a CNN. The full gate was compared with availability-only and no-reliability gates, nonnegative and ridge stackers, and nested expert-dropout, under nested field-grouped cross-fitting, expert withdrawal and concatenated walk-forward. Phenology-aware fusion

and increment targets already exist [8–11]; the test here is the extra reliability signals and the protocol. Paired tests on twelve season–unit groups are exploratory and unadjusted; nonsignificance is not equivalence. The evidence is summarized in Tables 24 and 25.

The first two rows of that table are protocol, not ranking. Any-overlap of 20 m windows is not a verified plot: median displacements are 3.84 m (LAI) and 2.86 m (biomass), yet 14 LAI and 17 biomass pairs share less than half the window ($f < 0.5$). Majority overlap ($f \geq 0.5$) was a sensitivity on $n = 91/124$, not the headline sample. The majority-overlap sensitivity was also extended to concatenated walk-forward evaluation ($n = 75$ LAI and $n = 106$ biomass). All corresponding R^2 values remained negative, confirming that stricter spatial continuity screening does not resolve the observed lack of prospective forecasting skill. Increment and reconstructed-absolute R^2 must be reported together because they normalize the same residuals by different variances and reconstruction adds the measured y_t. A mean-increment baseline already reaches absolute biomass $R^2 = 0.675$ while remaining near zero on ΔAGBd (-0.009); that is a change of denominator, not a split of skill into seasonal progression. Those rows also fix how the later ranking should be read: headline RMSE uses any-overlap $n = 105/141$, and a high reconstructed R^2 is not evidence that the increment itself was forecast well.

Typicality and uncertainty did not earn a consistent role. On fully observed biomass, availability-only gating had a slightly lower observed RMSE (101.1 versus 102.4), with no statistically significant field-level difference ($p = 0.23$). Small pooled LAI differences in exploratory bootstraps were not corroborated by field Wilcoxon tests. No field-level comparison of the full gate versus availability-only reached unadjusted 0.05 under the withdrawal conditions we report. Expert-withdrawal tests show that the gate can reroute; they do not show that the extra reliability signals help.

Fully observed accuracy likewise does not justify the full reliability architecture. A nonnegative stacker is more accurate on ΔLAI (RMSE 0.827 versus 1.069). On ΔAGBd, SARG and the ridge stacker are numerically close, with no statistically significant difference detected. Among the evaluated gate variants, availability-only gating therefore remains a simpler alternative. It is not identified as best under every withdrawal: withholding the current-biophysics expert leaves SARG with the lowest pooled biomass RMSE (107.0 versus availability-only 112.0, $p = 0.42$), while phenology-expert withdrawal favors nested dropout over SARG. Those p-values are unadjusted tests on twelve groups; they are not equivalence and not definitive robustness. Biophysics-expert withdrawal still reconstructs with measured y_t; it is not a sensor-free next-state forecast. Phenology-expert withdrawal is not complete loss of phenological information, because observed Zadoks remains in the gate context.

Concatenated walk-forward (Table 22) is the only prospective next-campaign protocol in the paper. It does not establish usable temporal skill ($n = 88$ LAI, $n = 120$ biomass). Ridge records a small numerical ΔLAI RMSE gain over the mean-increment baseline (2.190 versus 2.234); the mean increment is ahead on ΔAGBd (368 versus SARG 427); every concatenated R^2 is negative relative to the test-set mean. Field-out R^2 values in Tables 12 and 17 answer a different question and must not be cited as next-campaign forecasting capability.

The contribution is therefore comparative and methodological rather than a claim of operational advancement: spatially screened increments, an honest nested protocol, a specified expert feature space (Table 6), a direct test of reliability conditioning against availability-only gating, an uncertainty-versus-error diagnostic, a dual-expert withdrawal stress set (Table 19), and a horizon-stratified breakdown inside the same SARG forecasts. Relative to uncertainty-aware fusion and mixture-of-experts models, what is new is that comparative test—not a new expert class or a missing-view imputer. The scope is incremental change given a measured y_t on one winter-wheat landscape. Field-out accuracy

on that task is not temporal forecasting skill, and the present weights are not a portable recipe for other crops or climates. Recommended defaults on this sample are the nonnegative stacker or UAV expert for fully observed ΔLAI, and availability-only gating among the gates; the full reliability gate does not earn its extra fit cost (Table 11).

Prospects for future research.

The limited forecasting performance—especially the negative concatenated walk-forward R^2 and the lack of a consistent full-gate gain—points to specific modifications rather than to a larger version of the same gate. First, replace the heuristic typicality kernel and the uncalibrated tree-spread u_m with reliability scores that track expert |error| at field level—for example, fold-local Mahalanobis or isolation-forest scores [37,38] and conformal prediction intervals on Δy—and re-run the availability-only versus full-gate comparison under a prespecified primary hypothesis with multiplicity control. Until those scores are field-calibrated, default to availability-only gating or to the nonnegative stacker/UAV expert on fully observed ΔLAI. Second, treat heterogeneous horizons as part of the model rather than only as a post hoc split: multi-horizon heads, horizon-normalized targets used as the primary loss, or an explicit Δt-conditional expert, evaluated inside the same nested protocol. Campaign-mean non-stationarity is the walk-forward failure mode; a model that does not see the next campaign's increment mean cannot be rescued by a richer gate. Third, close the remote-forecasting gap that Equation (5) leaves open by replacing the measured y_t with a retrieved current state (UAV or satellite LAI and biomass) and reporting that pipeline separately from increment-given-measurement. Fourth, enlarge the sample beyond two seasons, one landscape and twelve season–unit clients, add later Zadoks stages and harvest yield, and re-fit the convolutional UAV encoder only when the plot count supports it. Transfer to other crops or climates requires that new evaluation, not a claim that the present weights generalize. Fifth, substitute campaign soil pits and station files with public soil and weather layers if the intended deployment is no-ground-sampling, and keep neighboring-unit spatial dependence in the validation design. Sixth, train for simultaneous sensor loss rather than only scoring it: nested dropout of modality pairs (the four pairs in Table 19 are a starting set) should be a training objective if dual-missing robustness is the operational requirement. Seventh, pre-register the primary contrast (we recommend full gate versus availability-only on a named target and a named withdrawal condition) before further exploratory withdrawal tables are interpreted as confirmatory.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/app162110634/s1, historical unfiltered gate-weight tables; typicality and uncertainty diagnostics; the four-way gate-signal ablation; screened stage-weight tables; farm-cluster and weather-imputer sensitivities; majority-overlap walk-forward; timing audit; reduced-input ladder; CNN UAV-encoder comparison; neural federated walk-forward results; prediction files; split manifests; and the $3 \times 3 \times 3$ hyperparameter-sensitivity tables on the spatially screened run. Historical unfiltered runs were not re-fit after spatial screening and are not headline results.

Author Contributions: Conceptualization, M.G.B. and Y.B.S.; methodology, M.G.B.; software, M.G.B.; validation, M.G.B.; formal analysis, M.G.B.; investigation, M.G.B.; data curation, M.G.B.; writing—original draft preparation, M.G.B.; writing—review and editing, M.G.B. and Y.B.S.; visualization, M.G.B.; supervision, Y.B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study uses the public TS2AgroBg dataset and does not involve new human or animal subjects.

Informed Consent Statement: Not applicable.

Data Availability Statement: The TS2AgroBg dataset analyzed in this study is openly available on Zenodo at https://doi.org/10.5281/zenodo.17475742 under the Creative Commons Attribution 4.0 International license [19,28]. Predictions, split manifests, preprocessing and gating source code, spatial-screening tables, and supplementary analyses supporting the results of this study are openly available on Zenodo at https://doi.org/10.5281/zenodo.23120842 [32].

Acknowledgments: The TS2AgroBg dataset analyzed in this study was produced independently of the present research by the TS2AgroBg project ("Testing Sentinel-2 vegetation indices for the assessment of the state of winter crops in Bulgaria"), ESA PECS contract no. 4000117474/16/NL/NDe, and is distributed via Zenodo under CC BY 4.0 [19,28]. During the preparation of this manuscript, the authors used a large language model writing assistant to propose LATEX edits. The authors reviewed and edited all output and take full responsibility for the content of this publication. No GenAI tool meets MDPI authorship criteria and none is listed as an author [41,42].

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

Abbreviations

The following abbreviations are used in this manuscript:

AGBd	dry above-ground biomass
ESU	elementary sampling unit
LAI	leaf area index
NNLS	nonnegative least squares
SARG	reliability-gated mixture of experts
UAV	unmanned aerial vehicle

Appendix A. Historical Unfiltered Analyses

Gate-weight tables, typicality–error scatter, the reduced-input ladder and the end-to-end CNN comparison use the older unfiltered sample. We did not re-fit them after spatial screening, and they are not part of the main results; the tables and figures are in the Zenodo deposit named in the Data Availability statement [32].

Screened-run stage weights of the final SARG model are in Table 16; they show the same qualitative pattern and are not a phenological accuracy claim. On that unfiltered sample, the biomass gate still favored phenology ($\alpha \approx 0.46$–0.59) with management at $\alpha \approx 0.30$–0.35. November 2017 shifted mass from phenology to management on biomass and from UAV to management on LAI (Table A1). The `use_context=False` ablation did not reduce pooled skill (ΔLAI R^2 0.581 $\rightarrow$ 0.580; biomass 0.851 $\rightarrow$ 0.870). Each reduced-input rung retrains SARG; reconstruction Equation (5) still uses measured y_t. The same 161 patches produced the best single LAI-increment modality as 31 statistics and did not improve as a CNN trained on roughly 90 examples (0.661 $\rightarrow$ 0.632). That CNN is the neural alternative labeled in Figure 4, not the ExtraTrees protocol of the main tables.

Table A1. Mean softmax gate weights on the historical unfiltered sample (grouped-inner SARG). These weights were not re-fit after spatial screening and are not headline results. Rows are phenological strata or campaigns with at least twenty observations. Columns sum to one up to rounding. The table is descriptive of what that older gate learned; turning context off on the same sample did not reduce pooled skill.

Stratum (n)	UAV	Spec.	Bioph.	Phen.	Soil	Weath.	Mgmt
ΔLAI							
Z < 25 tillering (78)	0.36	0.02	0.07	0.28	0.01	0.03	0.24
Z32–44 stem elong. (30)	0.41	0.02	0.06	0.31	0.01	0.03	0.17
November 2016 (20)	0.46	0.04	0.05	0.31	0.01	0.03	0.11
November 2017 (28)	0.23	0.00	0.09	0.25	0.00	0.02	0.40

Table A1. *Cont.*

Stratum (*n*)	UAV	Spec.	Bioph.	Phen.	Soil	Weath.	Mgmt
$\Delta AGBd$							
$Z < 25$ tillering (86)	0.02	0.00	0.16	0.46	0.00	0.00	0.35
Z32–44 stem elong. (30)	0.02	0.00	0.08	0.59	0.01	0.00	0.30
$Z \geq 60$ anthesis (30)	0.02	0.00	0.08	0.59	0.01	0.00	0.30
November 2016 (26)	0.03	0.00	0.08	0.65	0.01	0.00	0.23
November 2017 (30)	0.02	0.00	0.20	0.14	0.00	0.00	0.63

Appendix B. Historical Federated Optimization Simulation

A neural federated walk-forward and personalization run on the unfiltered historical sample is in the same Zenodo deposit [32]. We did not re-fit it after screening. Preprocessing still needs every training-client row in one place: medians, feature scales and image normalization are fitted on the concatenated training-client table (`PersonalizedFederatedFusion.fit` calls `_Preprocessor.fit` on that table). Local updates are then averaged with FedAvg [43], optionally leaving the gate ϕ_k local. That is federated optimization with centralized preprocessing, not a field-local training pipeline. Walk-forward neural federation did not produce a usable increment (2016/17 n-weighted RMSE 3.11–3.82 on ΔLAI; 481–513 g m^{-2} on ΔAGBd). The 2017/18 cold-start biomass case (FedAvg RMSE 124, $R^2 = -0.99$) remains worse than predicting the test-set mean.

References

1. Wolpert, D.H. Stacked Generalization. *Neural Netw.* **1992**, *5*, 241–259. [CrossRef]
2. Hall, D.L.; Llinas, J. An Introduction to Multisensor Data Fusion. *Proc. IEEE* **1997**, *85*, 6–23. [CrossRef]
3. Jacobs, R.A.; Jordan, M.I.; Nowlan, S.J.; Hinton, G.E. Adaptive Mixtures of Local Experts. *Neural Comput.* **1991**, *3*, 79–87. [CrossRef] [PubMed]
4. Han, Z.; Zhang, C.; Fu, H.; Zhou, J.T. Trusted Multi-View Classification. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 3–7 May 2021.
5. Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017; pp. 5574–5584.
6. Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; Dean, J. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
7. Xue, Z.; Marculescu, R. Dynamic Multimodal Fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*; IEEE: New York, NY, USA, 2023; pp. 2575–2584. [CrossRef]
8. Liu, Q.; Jing, H.; Wen, X.; Wu, X.; Yan, L.; Yang, Q.; Jia, S.; Guo, S.; Fan, F.; Jin, X. Soybean Phenological Stage Identification Based on Multimodal Data and a Dynamic Gating Fusion Model. *Smart Agric. Technol.* **2026**, *13*, 101827. [CrossRef]
9. Huang, Z.; Lee, W.S.; Ampatzidis, Y.; Agehara, S.; Peres, N.A. PheMuT: A Phenology-Informed, Multi-Modal Time-Series Model for Strawberry Yield Forecasting. *Comput. Electron. Agric.* **2026**, *244*, 111526. [CrossRef]
10. Wang, B.; Zhu, J.; Sun, S.; Zhang, L.; Yan, Y.; Wang, H.; Yang, W.; Lan, Y. Multimodal Remote Sensing Combination for Maize LAI Estimation: Stacking Model Development and Phenology-Specific Feature Sensitivity Analysis. *Artif. Intell. Agric.* **2026**, *16*, 873–888. [CrossRef]
11. Senti, D.; Turkoglu, M.O.; Volpi, M.; Aasen, H. Learning to Forecast Crop Growth from Earth Observation Data. *arXiv* **2026**, arXiv:2608.14281.
12. Abouaomar, A.; El Hanjri, M.; Kobbane, A.; Laouiti, A.; Nafil, K. Hierarchical Federated Learning for Crop Yield Prediction in Smart Agricultural Production Systems. In *Proceedings of the International Conference on Wireless Networks and Mobile Communications (WINCOM)*; IEEE: New York, NY, USA, 2025. [CrossRef]
13. Mukherjee, A.; Buyya, R. Federated Learning Architectures: A Performance Evaluation with Crop Yield Prediction Application. *Softw. Pract. Exp.* **2025**, *55*, 1165–1184. [CrossRef]
14. Khaleghi, B.; Khamis, A.; Karray, F.O.; Razavi, S.N. Multisensor Data Fusion: A Review of the State-of-the-Art. *Inf. Fusion* **2013**, *14*, 28–44. Erratum in *Inf. Fusion* **2013**, *14*, 562. [CrossRef]
15. Castanedo, F. A Review of Data Fusion Techniques. *Sci. World J.* **2013**, *2013*, 704504. [CrossRef] [PubMed]
16. Baltrušaitis, T.; Ahuja, C.; Morency, L.P. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 423–443. [CrossRef] [PubMed]

17. Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; Peng, X. SMIL: Multimodal Learning with Severely Missing Modality. *Proc. Aaai Conf. Artif. Intell.* 2021, *35*, 2302–2310. [CrossRef]
18. Verrelst, J.; Camps-Valls, G.; Muñoz-Marí, J.; Rivera, J.P.; Veroustraete, F.; Clevers, J.G.P.W.; Moreno, J. Optical Remote Sensing and the Retrieval of Terrestrial Vegetation Bio-Geophysical Properties—A Review. *ISPRS J. Photogramm. Remote Sens.* **2015**, *108*, 273–290. [CrossRef]
19. Dimitrov, P.; Roumenina, E.; Jelev, G.; Filchev, L.; Gikov, A.; Kamenova, I.; Ilieva, I.; Ganeva, D.; Kercheva, M.; Banov, M.; et al. In Situ Crop and Soil Data and UAV Imagery from Winter Wheat Fields in a Bulgarian Site. *Data* **2026**, *11*, 35. [CrossRef]
20. Dimitrov, P.; Kamenova, I.; Roumenina, E.; Filchev, L.; Ilieva, I.; Jelev, G.; Gikov, A.; Banov, M.; Krasteva, V.; Kolchakov, V.; et al. Estimation of Biophysical and Biochemical Variables of Winter Wheat through Sentinel-2 Vegetation Indices. *Bulg. J. Agric. Sci.* **2019**, *25*, 819–832.
21. Roumenina, E.; Jelev, G.; Dimitrov, P.; Filchev, L.; Kamenova, I.; Gikov, A.; Banov, M.; Krasteva, V.; Kercheva, M.; Kolchakov, V. Qualitative Evaluation and Within-Field Mapping of Winter Wheat Crop Condition Using Multispectral Remote Sensing Data. *Bulg. J. Agric. Sci.* **2020**, *26*, 1129–1142.
22. Maimaitijiang, M.; Sagan, V.; Sidike, P.; Hartling, S.; Esposito, F.; Fritschi, F.B. Soybean Yield Prediction from UAV Using Multimodal Data Fusion and Deep Learning. *Remote Sens. Environ.* **2020**, *237*, 111599. [CrossRef]
23. Zhou, X.; Zheng, H.B.; Xu, X.Q.; He, J.Y.; Ge, X.K.; Yao, X.; Cheng, T.; Zhu, Y.; Cao, W.X.; Tian, Y.C. Predicting Grain Yield in Rice Using Multi-Temporal Vegetation Indices from UAV-Based Multispectral and Digital Imagery. *ISPRS J. Photogramm. Remote Sens.* **2017**, *130*, 246–255. [CrossRef]
24. Nevavuori, P.; Narra, N.; Lipping, T. Crop Yield Prediction with Deep Convolutional Neural Networks. *Comput. Electron. Agric.* **2019**, *163*, 104859. [CrossRef]
25. van Klompenburg, T.; Kassahun, A.; Catal, C. Crop Yield Prediction Using Machine Learning: A Systematic Literature Review. *Comput. Electron. Agric.* **2020**, *177*, 105709. [CrossRef]
26. Durrant, A.; Markovic, M.; Matthews, D.; May, D.; Enright, J.; Leontidis, G. The Role of Cross-Silo Federated Learning in Facilitating Data Sharing in the Agri-Food Sector. *Comput. Electron. Agric.* **2022**, *193*, 106648. [CrossRef]
27. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 12:1–12:19. [CrossRef]
28. Dimitrov, P.; Roumenina, E.; Jelev, G.; Filchev, L.; Gikov, A.; Kamenova, I.; Ilieva, I.; Ganeva, D.; Kercheva, M.; Banov, M.; et al. *In-Situ and UAV Dataset with Crop and Soil Parameters Obtained from Winter Wheat Fields*; Zenodo (dataset), version 1.0.0, CC BY 4.0; CERN: Genève, Switzerland, 2025. [CrossRef]
29. Zadoks, J.C.; Chang, T.T.; Konzak, C.F. A Decimal Code for the Growth Stages of Cereals. *Weed Res.* **1974**, *14*, 415–421. [CrossRef]
30. IUSS Working Group WRB. *World Reference Base for Soil Resources 2014: International Soil Classification System for Naming Soils and Creating Legends for Soil Maps. Update 2015*; World Soil Resources Reports 106; Food and Agriculture Organization of the United Nations: Rome, Italy, 2015.
31. Weiss, M.; Baret, F.; Smith, G.J.; Jonckheere, I.; Coppin, P. Review of Methods for In Situ Leaf Area Index (LAI) Determination: Part II. Estimation of LAI, Errors and Sampling. *Agric. For. Meteorol.* **2004**, *121*, 37–53. [CrossRef]
32. Bakır, M.G.; Salman, Y.B. *Reliability-Gated Multimodal Fusion*; Zenodo (dataset), version 1, CC BY 4.0; CERN: Genève, Switzerland, 2026. [CrossRef]
33. Field, C.A.; Welsh, A.H. Bootstrapping Clustered Data. *J. R. Stat. Soc. Ser. B* **2007**, *69*, 369–390. [CrossRef]
34. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography* **2017**, *40*, 913–929. [CrossRef]
35. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely Randomized Trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
36. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
37. Lee, K.; Lee, K.; Lee, H.; Shin, J. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2018; Volume 31.
38. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation Forest. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*; IEEE: New York, NY, USA, 2008; pp. 413–422. [CrossRef]
39. Cawley, G.C.; Talbot, N.L.C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079–2107.
40. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; ACM: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
41. MDPI. MDPI's Updated Guidelines on Artificial Intelligence and Authorship. 2023. Available online: https://www.mdpi.com/about/announcements/5687 (accessed on 17 October 2026).

42. MDPI. Research and Publication Ethics. Section on Artificial Intelligence. 2026. Available online: https://www.mdpi.com/ethics (accessed on 17 October 2026).
43. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017*; Proceedings of Machine Learning Research; PMLR: New York, NY, USA, 2017; Volume 54, pp. 1273–1282.

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