AgriEngineering

MDPI

Article

Cross-Session Reconstruction and Environmental Limitation Screening of Greenhouse Tomato Leaf Photosynthesis from Gas-Exchange Data Using a CatBoost–ExtraTrees–RBF-SVR Stacked Ensemble

Guoqing Zhang [1], Shuping Zhang [2,*], Lili Tao [1], Yunlong Zhang [1], Jingbo Zhao [1,*] and Haimei Liu [1]

[1] School of Intelligent Manufacturing and Control Engineering, Shanghai Polytechnic University, Shanghai 201209, China; sspu1281960805@gmail.com (G.Z.); lltao@sspu.edu.cn (L.T.); thelong170@gmail.com (Y.Z.); hmliu@sspu.edu.cn (H.L.)

[2] School of Intelligent Medicine and Health Engineering, Shanghai Polytechnic University, Shanghai 201209, China

* Correspondence: spzhang@sspu.edu.cn (S.Z.); jbzhao@sspu.edu.cn (J.Z.)

Abstract

Reliable prediction and interpretation of photosynthetic rate are important for precision environmental management in greenhouse tomato production, but model stability is often limited by variable redundancy, measurement-session effects, and environmental heterogeneity. This study developed an integrated gas-exchange-data-based framework for key-factor selection, photosynthetic-rate reconstruction, cross-session validation, environmental correction, and physiology-informed limitation screening. Core predictors were first identified from high-dimensional gas-exchange variables using K-means clustering and random-forest importance analysis. A CatBoost–ExtraTrees–RBF-SVR stacked ensemble was then constructed to reconstruct the leaf photosynthetic rate from selected gas-exchange variables, and its cross-session generalization was evaluated using nested cross-validation and leave-one-file-out (LOFO) extrapolation. Environmental correction was further applied to improve cross-session comparability, and rule-based limitation screening was used to classify potential environmental constraints associated with reduced photosynthesis. The stacked model achieved an RMSE of 0.806 and an R^2 of 0.9918 under nested cross-validation, and an RMSE of 0.814 and an R^2 of 0.9917 under LOFO extrapolation. After environmental correction, the cross-session variability of the environmentally corrected photosynthetic indicator was reduced by 96.0%, reflecting improved cross-session comparability. Under deployable sensor inputs, performance decreased markedly (A1: RMSE = 4.578, R^2 = 0.735; A2: RMSE = 4.610, R^2 = 0.731), compared with the gas-exchange baseline (RMSE = 0.833, R^2 = 0.991). Thus, routine greenhouse sensor models should be regarded as approximate rather than high-accuracy substitutes for the gas-exchange-based model.

Keywords: greenhouse tomato; photosynthetic rate; CatBoost; ExtraTrees; RBF-SVR; stacked ensemble; leave-one-file-out validation; environmental correction; limitation screening; precision agriculture

Academic Editor: Stefano Benni

Received: 13 June 2026
Revised: 13 August 2026
Accepted: 2 September 2026
Published: 10 September 2026

1. Introduction

Photosynthesis is the fundamental physiological process underlying carbon assimilation, dry matter accumulation, and yield formation in crops [1,2], and it therefore represents

a central target for improving productivity and resource-use efficiency in protected horticulture [3]. In greenhouse systems, management practices such as supplemental lighting, CO_2 enrichment, ventilation, humidity regulation, and irrigation optimization ultimately act through their effects on leaf photosynthetic performance [4,5]. Accordingly, reliable prediction of photosynthetic rate under dynamically changing and strongly coupled greenhouse environments has become a prerequisite for intelligent sensing, precision regulation, and decision support in modern greenhouse production [6].

Considerable progress has been made in the quantitative characterization of photosynthesis in greenhouse tomato, with research gradually extending from leaf-level analysis to canopy-level assessment and management-oriented applications. Joubert et al. developed a dynamic leaf-level photosynthesis model for greenhouse tomato, demonstrating the feasibility of predicting tomato photosynthesis under naturally fluctuating irradiance [7]. Niu et al. further linked tomato photosynthesis modeling with greenhouse light-regulation decisions, showing the practical value of coupling photosynthesis prediction with environmental control [8]. Cheng et al. analyzed multienvironment-driven variations in net photosynthetic rate across developmental stages and established predictive models using machine-learning approaches, indicating that tomato photosynthesis is both stage-dependent and highly sensitive to environmental variation [9]. More recent studies on machine-learning-based photosynthetic-rate prediction, canopy photosynthesis, planting-layout evaluation, and digital-twin-enabled lighting control further suggest that protected-crop modeling is moving toward management-oriented applications [10–13] In parallel, studies on VPD effects, elevated CO_2 and light responses, and microclimate monitoring have reinforced the management relevance of the coupling between photosynthesis and the greenhouse environment [14–16].

For greenhouse leaf-level data, the first challenge is the redundancy, collinearity, and batch dependence of high-dimensional candidate variables. Photosynthetic rate is often affected jointly by diffusion-related derived variables and environmental variables, so the candidate inputs can be high-dimensional and strongly correlated. If all variables are directly used for model training without screening, incidental relationships specific to certain sessions, batches, or environmental conditions may be learned, rather than dominant physiological signals with real extrapolation value. Recent studies on feature selection and robust variable selection have indicated that variable treatment should not be regarded as a simple dimensionality-reduction step, but as a key procedure for improving model generalization and interpretive stability [17,18]. For tomato photosynthetic-rate prediction, this means that more representative and cross-session-stable predictors should be identified from high-dimensional candidate factors before formal modeling.

With the rapid development of multisource sensing and high-throughput phenotyping, machine learning has become an increasingly important methodological basis for photosynthesis prediction. Compared with conventional empirical regression or single process-based models, machine learning is better suited to handling high-dimensional inputs, complex nonlinear responses, and interactions among physiological and environmental variables [19]. It has already been applied to estimate photosynthesis-related traits from hyperspectral imagery, leaf reflectance information, and multisource sensing data [20,21]. However, existing studies have also repeatedly shown that high accuracy within the training dataset does not guarantee robust performance under new seasons, new platforms, or shifted sample distributions [22]. From the perspective of transferability, data-driven models remain highly vulnerable to dataset differences and distribution shift [23,24]. Therefore, building a machine-learning framework that balances predictive accuracy with cross-context robustness remains a key challenge in photosynthesis modeling.

Model evaluation is often mismatched with the grouped structure of greenhouse measurement data. Samples from the same file or measurement session are usually collected under similar temporal background, environmental conditions, instrument status, and operational settings. Therefore, if ordinary random splitting is used, information overlap between the training and test sets can easily be introduced [13,25]. Recent studies on machine-learning reproducibility and environmental modeling practice have shown that model performance can be systematically overestimated by data leakage [26,27]. Accordingly, common machine-learning pitfalls should be avoided, and extrapolation-based validation consistent with the data hierarchy should be adopted to obtain reliable conclusions on generalization [25,28]. For greenhouse tomato leaf photosynthesis data, only evaluation under strict separation of files or sessions can more realistically reflect predictive performance on unseen measurement batches.

From a technical perspective, the proposed framework was not derived from a single new algorithm, but integrates several established methodological components into a unified cross-session workflow. Its technical basis includes cluster- and tree-based feature screening for high-dimensional variables, heterogeneous ensemble learning for complementary nonlinear modeling, grouped and file-level validation for leakage-aware generalization assessment, conformal prediction for uncertainty quantification, limited-sample recalibration for adaptation to new sessions, and physiology-informed environmental correction and limitation screening. The methodological contribution of this study therefore lies primarily in integrating these techniques for greenhouse tomato gas-exchange reconstruction and physiological interpretation under strict cross-session validation.

However, an important methodological gap remains. Existing studies have mainly focused on photosynthetic-rate prediction under specific datasets or environmental conditions, while limited attention has been given to the combined challenges of leakage-free feature selection from highly redundant gas-exchange variables, strict generalization to unseen measurement sessions, and transformation of numerical predictions into cross-session-comparable and physiologically interpretable limitation information. Therefore, an integrated framework that jointly addresses cross-session reconstruction, leakage-aware validation, environmental correction, and physiology-informed limitation screening is still needed for greenhouse tomato gas-exchange data.

Based on the above background, a gas-exchange-data-based framework was developed in this study for cross-session reconstruction of leaf photosynthetic rate and environmental limitation screening in greenhouse tomato. First, to address the common problems of redundancy and unstable correlations in high-dimensional candidate variables, a combination of K-means clustering and random-forest importance analysis was used for structured dimensionality reduction and key-factor selection oriented to model generalization, so that more representative and robust predictive information could be extracted across sessions. Second, a stacked machine-learning model integrating CatBoost, ExtraTrees, and RBF-SVR was constructed to capture the main nonlinear relationships, supplementary environmental effects, and local residual corrections, and its extrapolation performance on unseen measurement batches was evaluated under a strict leave-one-file-out (LOFO) framework. Third, conformal prediction intervals, quality-weighted training, and few-shot recalibration were introduced into the extrapolation evaluation to better match the practical workflow of limited manual calibration followed by bulk automatic prediction. Finally, environmental correction and physiology-informed limitation screening were incorporated so that numerical predictions could be further transformed into environmentally corrected photosynthetic indicators and interpretable environmental constraint information, thereby providing quantitative support for supplemental lighting, VPD regulation, CO_2 fertilization, ventilation–humidity coordination, and irrigation management.

2. Materials and Methods

2.1. Overall Methodological Framework

The main objective was to establish a transferable analytical framework applicable to new sessions, as shown in Figure 1.

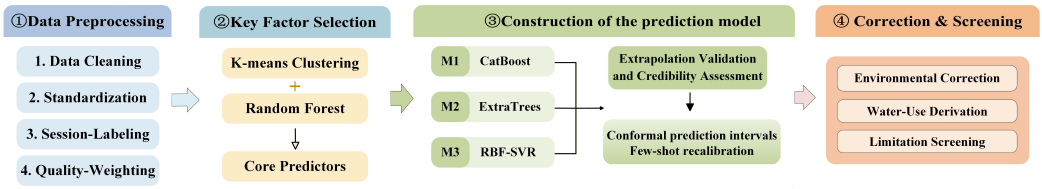

Figure 1. Methodology framework.

The methodological design of this study had three main features. First, representative key variables were selected from high-dimensional candidate factors before modeling through cluster-based identification and stability-based importance evaluation. Second, a leave-one-file-out extrapolative validation scheme was adopted for model evaluation, and conformal prediction intervals were further incorporated to quantify prediction reliability. Third, environmentally corrected photosynthetic indicators, water-use traits, and limitation-screening results were further derived from the prediction outputs, thereby extending the analysis from numerical prediction to physiological interpretation.

2.2. Data Source and Preprocessing

The data used in this study were obtained from the publicly available greenhouse tomato gas-exchange dataset reported by Manjarrez-Sanchez and Martinez-Carrillo [29]. The source data were collected by direct, non-destructive gas-exchange measurements of functional leaves from 100 greenhouse-grown tomato plants using an LI-6400XT Portable Photosynthesis System with OPEN software version 6.2. Measurements were conducted over 16 sampling days from November 2019 to January 2020, with measurements collected during approximately 6 h sampling periods. The source dataset was organized under two sampling schemes: one comprised measurements taken at approximately 1 min intervals from 100 plants in two modules, yielding approximately 200 observations per sampling day, whereas the second comprised approximately 50 observations per sampling day collected at intervals of about 5–15 min. The downloaded dataset comprised 26 Excel files containing photosynthetic, physiological, and environmental variables. All 26 files were included in the present study, and no complete measurement file or session was excluded. After unified import and format harmonization, the analytical dataset contained 2599 valid observations. The date-specific sampling schedule, associated measurement files, and sample sizes before and after preprocessing are summarized in Supplementary Table S3. Photosynthetic rate (Photo; μmol CO_2 m^{-2} s^{-1}) was defined as the response variable. The principal variables used in subsequent analyses included Cond (mol H_2O m^{-2} s^{-1}), Ci (μmol CO_2 mol^{-1}), Ci_Pa (Pa), PARi (μmol photons m^{-2} s^{-1}), VPD (kPa), Tleaf (°C), and CO_2S (μmol CO_2 mol^{-1}). Representative candidate-variable definitions and units are summarized in Table 1, while the complete set of candidate variables is provided in Supplementary Table S1.

Table 1. Representative candidate variables, definitions, and units.

No.	Variable	Definition	Unit
1	`Cond`	Conductance to H_2O	mol H_2O m^{-2} s^{-1}
2	`Ci`	Intercellular CO_2 concentration	μmol CO_2 mol^{-1}
3	`Trmmol`	Transpiration rate	mmol H_2O m^{-2} s^{-1}
4	`CO2S`	Sample-cell CO_2 concentration	μmol CO_2 mol^{-1}
		...	
46	`CndCO2`	Total conductance to CO_2	mol CO_2 m^{-2} s^{-1}
47	`Ci_Pa`	Intercellular CO_2 partial pressure	Pa
48	`Ci/Ca`	Ratio of intercellular to ambient CO_2 concentration	Dimensionless
49	`RHsfc`	Relative humidity at the leaf surface	%
50	`C2sfc`	CO_2 concentration at the leaf surface	μmol CO_2 mol^{-1}

Before modeling, all raw files underwent a unified data-cleaning and feature-engineering procedure. Abnormal trailing contents, blank formulas, and invalid cells were handled during file parsing; variable names with inconsistent capitalization or notation were harmonized; and the 26 files were labeled as Session 1–26. File-level metadata, including file name, data type, replicate ID, and date identifier, were extracted, and StableF-based sample-quality weights were subsequently generated. Missing values required by the machine-learning models were handled by median imputation within the corresponding modeling pipelines. In addition, sample-level quality weights were assigned according to the instrument stability flag (StableF). The definition was as follows:

$$W_i = \begin{cases} 1.0, & \text{StableF}_i = 1, \\ 0.7, & \text{StableF}_i \neq 1. \end{cases} \quad (1)$$

where W_i denotes the base training weight of the i-th sample. This design was intended to reduce the adverse effect of low-quality records on model training without completely excluding non-steady-state observations.

To reduce the effects of instrument status, steady-state determination, and abnormal observations on model training, QC thresholds were calculated from the training data within each LOFO fold, and samples were classified into three quality levels: A, B, and C. The QC score was defined as the total number of abnormality flags. One flag was added for each of the following conditions: `Status` $\neq$ 111,115, `StableF` $<$ 0.6666, or the presence of a negative-value flag for CO2S, C2sfc, Ci, or Ci_Pa. For h20diff, h2o_i, and ci_ca_ratio, values outside the 0.5th–99.5th percentile range estimated from the corresponding LOFO training fold, or non-finite values, were also counted as one abnormality flag. Samples with QC scores of 0, 1–2, and $\geq$3 were classified as A-, B-, and C-level samples, respectively. Three training strategies were then compared: baseline training, QC-weighted training (*qc_weighted*), and low-quality filtering (*qc_filtered*). In QC-weighted training, ordinal quality weights of 1.0, 0.7, and 0.4 were assigned to A-, B-, and C-level samples, respectively. These values were empirically specified as a monotonically decreasing weighting scheme to preserve the full contribution of high-quality observations, moderately down-weight intermediate-quality observations, and substantially reduce, rather than completely remove, the influence of low-quality observations. The weighting scheme therefore reflected the relative confidence associated with the three QC levels rather than a universal physiological threshold. The final sample weight used in `qc_weighted` training was obtained by multiplying the StableF-based base weight defined in Equation (1) by the corresponding QC-level weight. Accordingly, a zero contribution was used only in the separate `qc_filtered` strategy, in which C-level samples were excluded from training, while the

held-out test set remained complete. A key aspect of this design was that QC thresholds were calculated only from data within each training fold, without using information from the test files, thereby preventing reverse leakage during quality evaluation.

These preprocessing steps converted the heterogeneous raw Excel files into a unified analytical dataset with standardized variable names, session identifiers, quality information, and model-ready numerical inputs.

2.3. Selection of Key Factors

To reduce redundancy and unstable associations among the high-dimensional gas-exchange variables, K-means clustering and RF importance analysis were combined because they served complementary purposes. K-means was selected as an unsupervised method to identify major multivariate response regimes in the standardized predictor space without imposing a predefined linear relationship with photosynthetic rate. RF was subsequently used for supervised variable ranking because tree ensembles can represent nonlinear effects and interactions and provide a direct model-based measure of predictor importance. This two-stage design therefore combined unsupervised structure identification with supervised importance evaluation. All procedures were performed independently within each LOFO training fold, with the held-out file completely excluded. In each iteration, the held-out test file was completely excluded from K-means clustering, random-forest importance analysis, and core-variable selection.

2.3.1. K-Means Clustering

Within each LOFO iteration, predictor standardization and K-means clustering were fitted using only the corresponding training files, with the held-out file excluded from both procedures. K-means clustering was then applied to the standardized training predictor matrix to partition the samples into $k = 3$ clusters. The high-photosynthesis cluster was defined as the cluster with the highest mean photosynthetic rate (Photo) among the three clusters. K-means was implemented with $k = 3$ using standardized predictor values. Initial centroids were randomly selected from the training samples without replacement, with a maximum of 300 iterations and a base random seed of 42. Iteration stopped when cluster assignments no longer changed or the maximum number of iterations was reached. The purpose of K-means clustering was to partition the standardized training samples into distinct multivariate response regimes and thereby reduce interference from heterogeneous photosynthetic states during subsequent feature screening. Specifically, the cluster with the highest mean Photo was retained for RF-based importance evaluation, while the cluster label itself was not used as a predictor. The objective of K-means was to minimize the within-cluster sum of squared distances between data points and their assigned centroids. The objective function was expressed as follows:

$$J = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| x_{ij} - \mu_i \right\|^2, \tag{2}$$

where J is the overall target, k is the number of clusters, n_i is the number of data points in the i-th cluster, x_{ij} is the data point in the i-th cluster, and μ_i is the cluster center of the i-th cluster, defined as the mean of the cluster samples.

2.3.2. RF Feature Selection

RF was selected for feature screening because it can capture nonlinear relationships and interactions among high-dimensional gas-exchange predictors while providing a direct model-based ranking of variable importance without requiring a predefined linear relationship with Photo. Feature importance was assessed according to information gain,

with each tree splitting the dataset based on specific features. The importance of a given feature in the t-th tree was calculated from the corresponding information gain. The importance of feature x_i was calculated as follows:

$$\text{Importance}(x_i) = \frac{1}{T}\sum_{t=1}^{T} \text{FI}_t(x_i), \qquad \sum_i \text{Importance}(x_i) = 1, \tag{3}$$

where x_i denotes the i-th candidate feature, T denotes the total number of decision trees in the random forest, and $\text{FI}_t(x_i)$ denotes the importance contribution of feature x_i in tree t. In this regression task, impurity was measured by mean squared error. Therefore, $\text{FI}_t(x_i)$ represents the total MSE reduction contributed by feature x_i in tree t. The importance value of each feature was obtained by averaging its contribution across all T trees and normalizing the importance values so that their total equaled 1.

The importance grading standards were defined as follows:

1. High importance: Importance ≥ 0.10;
2. Medium importance: $0.02 \leq$ Importance < 0.10;
3. Low importance: Importance < 0.02.

RF importance was evaluated using 500 trees with `max_features='sqrt'`, bootstrap sampling, and `random_state=42`. To reduce sensitivity to a single data split, importance stability was assessed using five ShuffleSplit repetitions with a 70% training and 30% validation split. Feature importance values were averaged across the five repetitions before importance grading and core-variable determination. Parameters not explicitly specified followed the defaults of scikit-learn 1.6.1. RF importance scores and the resulting core-variable set were determined exclusively from the training files of each LOFO fold, without access to the held-out test file.

2.4. Construction of the Prediction Model

2.4.1. Model Design

Photosynthetic rate was jointly regulated by CO_2 supply status, environmental fluctuations, and nonlinear local effects. To integrate primary physiological information, supplementary environmental information, and local residual correction, a mechanism-aware stacked ensemble model was constructed. The model consisted of three base learners and one meta-learner. The first base model was built on key factors, the second further incorporated environmental variables, and the third was used to correct the residuals of the first model. The final prediction was obtained through a nonnegative linear combination. The three learners were selected to provide complementary nonlinear modeling behaviors rather than because any single algorithm was assumed to be universally optimal. CatBoost was used as the primary learner for complex nonlinear tabular relationships; ExtraTrees provided a highly randomized tree-based representation to complement the boosting model; and RBF-SVR was used for residual correction because the nonlinear kernel can represent local deviations not fully captured by the tree-based learners. Their predictions were subsequently combined by the meta-learner to exploit these complementary response structures. To avoid relying solely on the selected ensemble architecture, RF, SVR, Ridge, PLSR, LSTM, and XGBoost were additionally evaluated as benchmark models under the same validation framework.

2.4.2. Definition of the Base Models

The four core predictors were denoted as X_{core}, the environmental variables as X_{env}, and the mechanism-derived variables as X_{mech}. The three base models can therefore be expressed as follows:

$$\hat{y}_{M1} = f_1(X_{\text{core}}), \tag{4}$$

$$\hat{y}_{M2} = f_2(X_{\text{core}}, X_{\text{env}}, X_{\text{mech}}), \tag{5}$$

$$r = y - \hat{y}_{M1}, \tag{6}$$

$$\hat{y}_{M3} = \hat{y}_{M1} + f_3(X_{\text{mech}}), \tag{7}$$

where y is the true value of the sample; $\hat{y}_{M1}$, $\hat{y}_{M2}$, and $\hat{y}_{M3}$ are the predictions of models M1, M2, and M3, respectively; X_{core}, X_{env}, and X_{mech} are the core predictors, environmental variables, and mechanism-derived variables, respectively; $f_1(\cdot)$, $f_2(\cdot)$, and $f_3(\cdot)$ are the nonlinear mapping functions of CatBoost, ExtraTrees, and RBF-SVR, respectively; and r is the residual of the main model M1. Here, $f_3(X_{\text{mech}})$ was used to learn the residual correction term, and $\hat{y}_{M3}$ was the prediction after residual correction. The model hyperparameters were fixed before evaluation rather than optimized within the validation folds. M2 used the four core predictors together with the available environmental and derived variables retained after preprocessing, whereas M3 used mechanism-related variables to learn the residuals of M1. The detailed model settings are provided in Supplementary Table S2.

2.4.3. Stacked Fusion Model

The final prediction was obtained by a meta-learner, which was defined as a nonnegative constrained linear regression. The final prediction can be expressed as follows:

$$\hat{y} = w_1\hat{y}_{M1} + w_2\hat{y}_{M2} + w_3\hat{y}_{M3} + b, \qquad w_1, w_2, w_3 \geq 0, \tag{8}$$

where w_1, w_2, and w_3 denote the nonnegative weights of the sub-models, and b denotes the intercept. To avoid information leakage in stacking, the meta-learner was not trained directly on fitted values from the training set, but on out-of-fold predictions generated by nested cross-validation. This design improved the robustness of model fusion and was more consistent with the generalization requirements under strict validation settings. The inner cross-validation folds were used to generate leakage-safe out-of-fold predictions for the meta-learner rather than for hyperparameter optimization.

2.4.4. Deployable-Input Validation

To assess whether the proposed framework retained predictive utility under realistic greenhouse sensing constraints, an additional deployable-input validation was conducted. The prediction target, model family, and LOFO protocol were kept unchanged, whereas only the input feature sets were replaced with deployable greenhouse sensor variables. Detailed definitions of the inputs, evaluation settings, and the sensor-group ablation design are provided in Section 3.6, Validation under Deployable Input Constraints, and in the Supplementary Material.

2.5. Extrapolation Validation and Credibility Assessment

2.5.1. File-Level Extrapolation Validation

Because samples within the same file are strongly correlated in time, environment, and measurement operation, information overlap can easily occur between the training and test sets under ordinary random splitting. Therefore, leave-one-file-out (LOFO) validation was adopted. In each round, one complete file was first held out as the test set, and all preprocessing, feature selection, and model fitting were then performed using only the remaining training files. Specifically, K-means clustering, RF importance analysis, and core-variable selection were repeated independently within each training fold before model fitting. The held-out file was used only for final evaluation. At no stage were

observations from the held-out test file used for preprocessing, feature selection, model fitting, meta-learner training, or calibration.

One measurement session was treated as an independent test set, while the remaining sessions were used as the training and calibration sets, and extrapolation evaluation was iteratively conducted across all sessions. The LOFO extrapolation test procedure is shown in Figure 2.

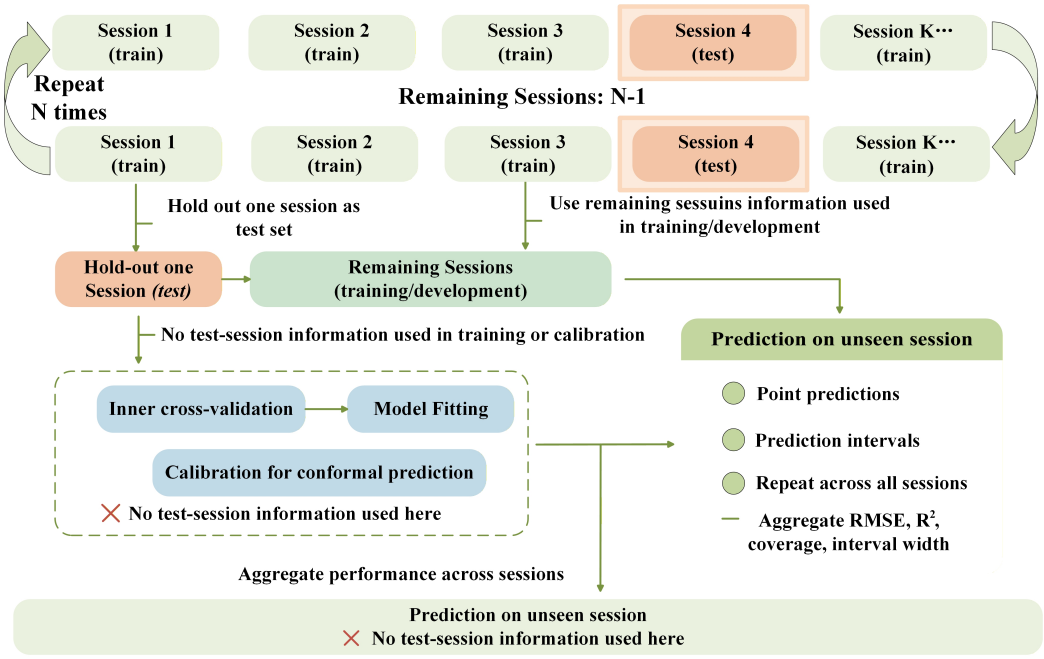

Figure 2. LOFO extrapolation test diagram.

The total number of test samples is denoted as n, and the root mean square error and coefficient of determination are defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{9}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{10}$$

where y_i and $\hat{y}_i$ denote the observed and predicted values of the i-th test sample, respectively, and $\bar{y}$ denotes the mean observed value of the test set. RMSE was used to measure prediction error, whereas R^2 was used to reflect the model's ability to explain variation. Statistical differences in session-level LOFO RMSE between the Final stack and each comparison model were evaluated using paired Wilcoxon signed-rank tests. Holm correction was applied for multiple comparisons, and an adjusted $p < 0.05$ was considered statistically significant.

2.5.2. Conformal Prediction Intervals

Point predictions alone cannot reflect model reliability on new files. Therefore, split conformal prediction was further introduced under the LOFO framework. Specifically, within each outer training fold, approximately 20% of the training files were randomly selected without replacement as the calibration subset, with at least three calibration files retained whenever possible. The remaining files formed the proper-training subset. Calibration was performed at the file level rather than by randomly splitting individual

samples, and the random-number generator was initialized with a base seed of 42. The calibration residual score was defined as

$$s_i = |y_i - \hat{y}_i|, \qquad i = 1, 2, \ldots, n. \tag{11}$$

For a given significance level α, the conformal quantile was defined as

$$q_\alpha = s_{(k)}, \qquad k = \lceil (n+1)(1-\alpha) \rceil, \tag{12}$$

where $s_{(k)}$ is the k-th residual after ascending sorting, q_α is the conformal quantile at significance level α, and α is the allowed miscoverage probability, which is complementary to the confidence level of the prediction interval. The corresponding prediction interval was written as

$$[\hat{y} - q_\alpha, \hat{y} + q_\alpha]. \tag{13}$$

In this study, 90% and 95% prediction intervals were constructed, and the interval coverage rate and mean interval width for each test file were calculated to represent the uncertainty level of the model in cross-file prediction.

2.5.3. Few-Shot Recalibration

To simulate the practical scenario of limited manual calibration followed by large-scale automated prediction, a recalibration procedure based on a small number of labeled samples from the test file was further introduced. Let the prediction of the base model at the calibration points be $\hat{y}$, and the corresponding observed value be y. The recalibrator was defined under three cases.

When $k = 0$, no recalibration was performed:

$$a = 1, \qquad b = 0. \tag{14}$$

When $k = 1$, only bias correction was applied:

$$a = 1, \qquad b = y - \hat{y}. \tag{15}$$

When $k \geq 2$, linear regression was used to fit the relationship:

$$y = a\hat{y} + b. \tag{16}$$

The recalibrated prediction was given by

$$\hat{y}_{\text{cal}} = a\hat{y} + b. \tag{17}$$

The effect of a small number of known samples on recovery of prediction performance for new files was evaluated by comparing RMSE changes under different numbers of calibration points. Of the 26 measurement files used in the primary LOFO analysis, 22 files met the sample-size requirement for the few-shot recalibration analysis. For these eligible test files, $k = 0, 3, 5, 10$, and 20 calibration samples were randomly selected without replacement, subject to sufficient remaining observations for independent performance evaluation. Each nonzero calibration setting was repeated 30 times using a reproducible random-number generator initialized with seed 42. Samples used for recalibration were excluded from the corresponding performance evaluation, so RMSE and R^2 were calculated only on the remaining test samples.

2.6. Construction of Environmental Correction and Limitation Screening

2.6.1. Environmental Correction of Photosynthetic Traits

Instantaneous environmental conditions can strongly affect the observed photosynthetic rate; direct comparison of raw photosynthetic rates across sessions is therefore easily confounded by environmental disturbance. Accordingly, an environmentally corrected photosynthetic indicator, A_{std}, was constructed by adjusting each observation to an empirically specified reference environment, denoted as E_0. The reference environment was used as a common numerical baseline for cross-session environmental correction rather than as a universally optimal physiological condition for greenhouse tomato. In this study, E_0 was specified as $PAR_i = 240.25$ μmol photons $\mathrm{m}^{-2}\,\mathrm{s}^{-1}$, $VpdL = 1.4227$ kPa, $T_{\mathrm{leaf}} = 22.38\,^{\circ}\mathrm{C}$, and $CO_2S = 371.7$ μmol $\mathrm{CO_2}$ mol^{-1}. Specifically, within each LOFO training fold, an environmental response function $g(E)$ was fitted using the training files, and predicted values for each test sample were then calculated under both the observed environment and the reference environment. The environmental response function $g(E)$ was implemented as a pipeline consisting of median imputation followed by a HistGradientBoostingRegressor with `max_depth=3`, `learning_rate=0.05`, `max_iter=400`, `l2_regularization=0`, and `random_state=42`. The environmental inputs were selected from PARi, VpdL, leaf or chamber temperature, and CO_2 variables according to data availability. The model was refitted independently using the training files of each LOFO fold.

The observed photosynthetic rate of a sample was denoted by A_{obs}, the environmental-model prediction under the current environment was denoted by $\hat{A}_{\mathrm{env}} = g(E_{\mathrm{obs}})$, and the prediction under the reference environment E_0 was denoted by $\hat{A}_{E_0} = g(E_0)$. The environmentally corrected photosynthetic indicator was then defined as

$$A_{\mathrm{std}} = A_{\mathrm{obs}} - \hat{A}_{\mathrm{env}} + \hat{A}_{E_0}. \tag{18}$$

This formulation removes the explainable component induced by the current environment from the original observation and then adds back the predicted value under the standard environment, thereby yielding an environmentally corrected indicator for cross-session comparison. Astd was not intended to represent a stable biological phenotype, because the correction may also remove part of the genuine physiological variation associated with environmental responses among sessions. The entire procedure was implemented under a leave-one-file-out strategy to avoid circular bias caused by within-file self-correction.

2.6.2. Calculation of Water-Use Traits and VPD Sensitivity

To further extend photosynthesis prediction to carbon–water coupling traits, two indicators—intrinsic water-use efficiency and transpiration efficiency—were constructed from pointwise gas-exchange data. The photosynthetic rate was denoted as A, the stomatal-conductance proxy variable as g_s, and the transpiration proxy variable as Tr. Then,

$$\mathrm{iWUE} = \frac{A}{g_s}, \tag{19}$$

$$\mathrm{TE} = \frac{A}{Tr}, \tag{20}$$

where iWUE denotes intrinsic water-use efficiency, TE denotes the amount of photosynthetic carbon assimilation per unit transpiration, A was represented by Photo, g_s by Cond, and Tr by Trmmol.

Samples were then divided into three levels—low, mid, and high—according to the 33% and 67% quantiles of global VPD, and the relative change under high VPD versus low VPD was compared at the file level. In addition, simple linear regression was used to

estimate the slopes of A, g_s, and iWUE against VPD, so that the sensitivity of each session to changes in vapor pressure deficit could be quantified. The relative change was defined as

$$\Delta X(\%) = \frac{X_{\text{high}} - X_{\text{low}}}{X_{\text{low}}} \times 100\%, \tag{21}$$

where X_{high} and X_{low} denote the mean values under high and low VPD, respectively.

2.6.3. Physiology-Informed Rule-Based Classification of Photosynthetic Limitation Types

In this study, the term limitation type refers to a physiology-informed, rule-based classification of the environmental or diffusional conditions associated with reduced photosynthetic rate. The classification was not intended to provide independent causal proof of photosynthetic limitation. Instead, it was used as a screening procedure to identify the most likely dominant constraint under the observed combinations of PAR, VPD, stomatal conductance, C_i/C_a, ambient CO_2 concentration, and leaf temperature.

To enhance the physiological interpretability of the model results, rule-based limitation classes were constructed using photosynthetic rate, PAR, VPD, stomatal conductance, C_i/C_a, C_a, and leaf temperature. All thresholds were determined automatically from data quantiles to improve adaptability across different experimental datasets. First, the following ratio was defined:

$$\frac{C_i}{C_a}. \tag{22}$$

Limitation types were then identified using quantile-based rules as follows.
Light limitation:

$$\text{PAR} \leq Q_{0.25}(\text{PAR}). \tag{23}$$

VPD-related limitation:

$$\text{VPD} \geq Q_{0.75}(\text{VPD}), \qquad g_s \leq Q_{0.25}(g_s). \tag{24}$$

Stomatal limitation:

$$g_s \leq Q_{0.25}(g_s), \qquad \frac{C_i}{C_a} \leq Q_{0.25}\left(\frac{C_i}{C_a}\right). \tag{25}$$

CO_2-supply-related limitation:

$$C_a \leq Q_{0.25}(C_a), \qquad \frac{C_i}{C_a} \leq \text{median}\left(\frac{C_i}{C_a}\right). \tag{26}$$

Temperature-related limitation:

$$[T_{\text{leaf}} \leq Q_{0.05}(T_{\text{leaf}}) \text{ or } T_{\text{leaf}} \geq Q_{0.95}(T_{\text{leaf}})], \qquad A \leq Q_{0.25}(A). \tag{27}$$

A mutually exclusive assignment was finally used to determine the dominant limitation class for each observation, with the priority order defined as light limitation > VPD-related limitation > stomatal limitation > CO_2-supply-related limitation > temperature-related limitation > other/non-limited. On this basis, the proportions of different limitation classes were summarized for the full dataset and for each file, and corresponding management interpretations were provided according to the dominant classified limitation type.

2.7. Software and Reproducibility

All data preprocessing, model development, validation, and visualization procedures were implemented in Python 3.9.7. Data handling and numerical computation were

performed using pandas 2.3.3, NumPy 2.0.2, SciPy 1.13.1, and joblib 1.5.3. Machine-learning models and evaluation procedures were implemented using scikit-learn 1.6.1, CatBoost 1.2.8, and XGBoost 2.1.4. Figures were generated using Matplotlib 3.9.4, and raw Excel files were read and processed using openpyxl 3.1.5. The raw gas-exchange Excel files were processed using customized Python scripts, which standardized column names, extracted file-level metadata, generated session labels, flagged abnormal values, calculated StableF-based sample weights, and exported the cleaned dataset as a unified CSV file.

A fixed random seed of 42 was used for procedures involving random initialization, cross-validation partitioning, and few-shot calibration sampling. The stacked prediction framework consisted of three base learners and one meta-learner. The first base learner was a CatBoostRegressor using the four selected CO_2-related predictors; the second base learner was an ExtraTreesRegressor using the four core predictors together with available environmental and derived variables; and the third base learner was an RBF-kernel support vector regression model trained to predict the residuals of the CatBoost model. The final prediction was obtained using a nonnegative linear regression meta-learner based on out-of-fold predictions generated within the nested cross-validation procedure. Missing values within the model pipelines were handled using median imputation, and standardization was applied where required for the SVR model.

For internal model evaluation, a nested cross-validation design with five outer folds and five inner folds was implemented. When a valid grouping variable was available, grouped cross-validation was used to reduce information leakage among samples from the same file, segment, date, or replicate. For extrapolation validation, leave-one-file-out validation was conducted using file name as the grouping variable. In each iteration, one complete measurement file was held out as the test set, while preprocessing, K-means clustering, RF-based feature selection, and model training were performed exclusively on the remaining files. The scripts exported fold-level predictions, model-performance summaries, meta-learner weights, feature-set definitions, session-wise LOFO metrics, and diagnostic figures for verification and reproducibility. The main hyperparameters, feature-screening settings, and validation configurations are summarized in Supplementary Table S2. The analysis scripts used for data cleaning, feature engineering, model training, nested cross-validation, LOFO validation, environmental correction, and rule-based limitation classification are available from the corresponding author upon reasonable request.

3. Results

3.1. Feature Selection and Model Evaluation

Fold-wise feature selection showed that C2sfc, CO2S, Ci, and Ci_Pa were the most consistently selected core predictors across the LOFO training folds. Physiologically, these variables collectively characterize external or chamber-level CO_2 supply and intercellular CO_2 status, which are directly associated with CO_2 diffusion and carbon assimilation in the leaf. Their selection was therefore not interpreted solely as a statistical reduction in dimensionality, but as the identification of a compact set of gas-exchange variables representing the CO_2 supply–diffusion pathway associated with photosynthetic carbon assimilation. M1, based on CatBoost [30], used only these four core factors as inputs. M2, based on ExtraTrees [31], further incorporated environmental variables on this basis. M3, an RBF-kernel support vector regression model developed from the residuals of M1, was used to capture local nonlinear deviations. To evaluate model performance under the same input conditions, XGBoost [32] and other models were also included for comparison.

Among the evaluated models, the Final stack achieved the lowest RMSE (0.806) and the highest R^2 (0.9918) under the nested cross-validation setting (Table 2). The numerical improvement over M1–M3 was modest.

Table 2. Nested cross-validation performance of different models.

Model	RMSE	R^2
M1	0.817	0.9916
M2	0.816	0.9916
M3	0.816	0.9916
Final stack	0.806	0.9918
RF	0.877	0.9034
SVR	0.968	0.8762
Ridge	0.975	0.8523
PLSR	0.947	0.8907
LSTM	1.423	0.7861
XGBoost	0.983	0.9874

To assess whether the observed performance differences were also supported under cross-session extrapolation, statistical comparisons were conducted separately using paired session-level LOFO RMSE values. The Wilcoxon signed-rank tests with the Holm correction showed that the differences between the Final stack and M1, M2, and M3 were not statistically significant (adjusted $p = 0.21$, 0.20, and 0.20, respectively). The difference between the Final stack and RF was also not statistically significant (adjusted $p = 0.25$). In contrast, the Final stack showed a significantly lower session-level LOFO RMSE than SVR ($p_{adj} = 0.003$), Ridge ($p_{adj} = 0.002$), PLSR ($p_{adj} = 0.004$), LSTM ($p_{adj} = 0.001$), and XGBoost ($p_{adj} = 0.002$).

These results indicate that the stacked ensemble provided only a modest numerical improvement over its three component models, without statistical evidence of superiority over M1–M3 or RF. Its performance advantage was more clearly supported relative to SVR, Ridge, PLSR, LSTM, and XGBoost.

3.2. Extrapolation, Uncertainty, and Few-Shot Recalibration

The validation results showed that the model maintained excellent predictive performance across all files, with an overall RMSE of 0.814 μmol m^{-2} s^{-1}, an MAE of 0.372 μmol m^{-2} s^{-1}, and an overall R^2 of 0.9917. These results indicated that, under cross-file conditions, the model still captured the main variation trend of photosynthetic rate accurately. The predicted values were highly correlated with observed photosynthetic rates, and most points were distributed near the $y = x$ reference line. Good agreement was obtained in both low- and high-photosynthesis ranges, as shown in Figure 3.

To further compare extrapolation performance across measurement sessions, the RMSE distribution for each session is shown in Figure 4. Differences in prediction error were observed among sessions, indicating that model performance was affected by variation in measurement batches and sample distributions. These results allowed sessions with relatively large extrapolation errors to be identified and provided evidence for the evaluation of cross-session robustness.

To evaluate prediction robustness and potential uncertainty, a quality-control strategy was introduced, including three training modes: baseline, *qc_weighted*, and *qc_filtered*. High-quality A-class samples accounted for about 57%, medium-quality B-class samples for about 33%, and low-quality C-class samples for about 10%. The LOFO extrapolation performance under the three QC strategies is summarized in Table 3.

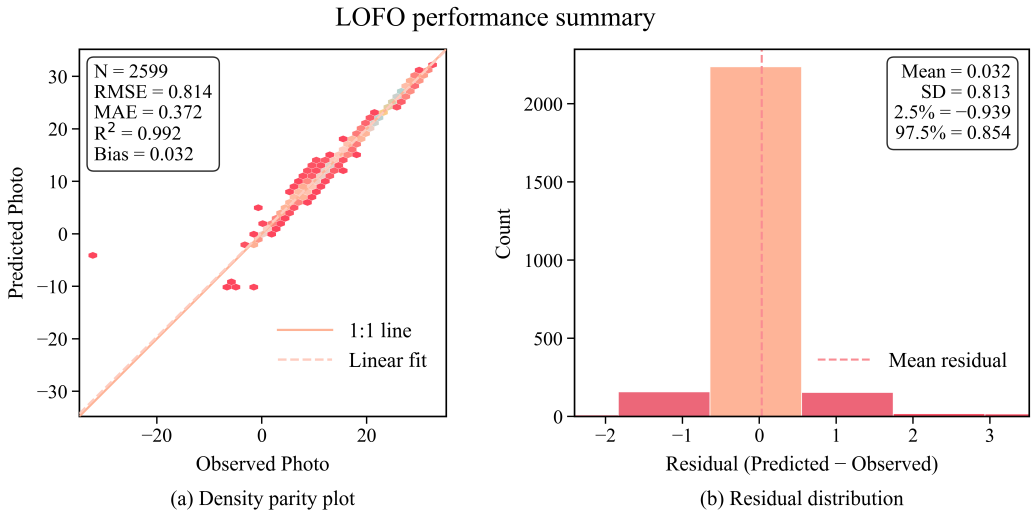

Figure 3. Overall prediction performance of the LOFO model.

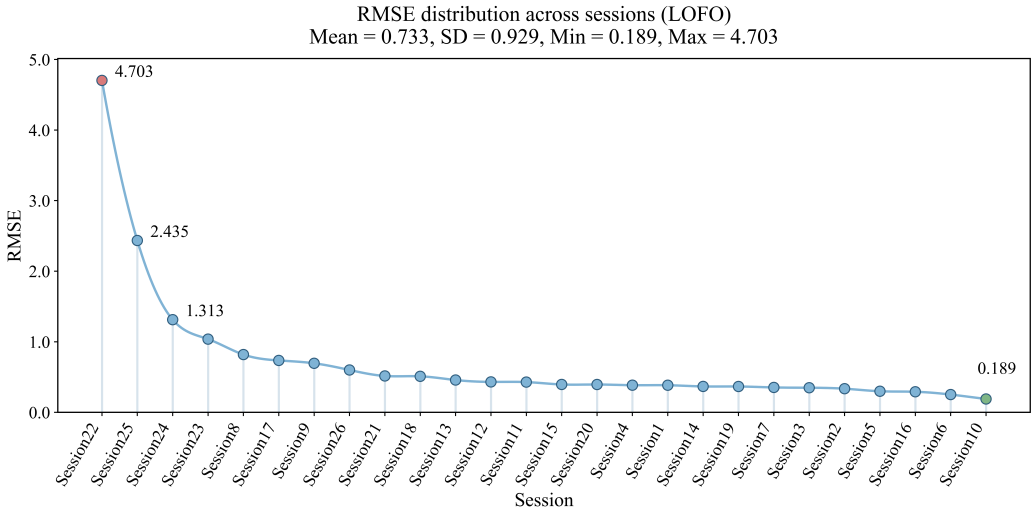

Figure 4. LOFO RMSE by session.

Table 3. Prediction performance under different QC strategies.

Training Strategy	Overall RMSE	Mean File RMSE	MAE	R^2
Baseline	0.814	0.733	0.372	0.9917
qc_weighted	0.807	0.726	0.371	0.9917
qc_filtered	0.815	0.733	0.373	0.9917

Under the LOFO extrapolation setting with conformal prediction intervals, the point predictions shown in Figure 5 yielded an RMSE of 0.848 and an R^2 of 0.991. The 90% and 95% conformal prediction intervals achieved empirical coverage rates of 0.912 and 0.953, with mean interval widths of 1.906 and 2.650, respectively. These empirical coverage rates were close to, and slightly above, their corresponding nominal levels, indicating useful uncertainty quantification under cross-session conditions, as shown in Figure 5.

To evaluate the effect of a small number of calibration samples on cross-session prediction, few-shot recalibration was performed for the 22 test files meeting the sample-size requirement. The number of calibration points (k) was set to 0, 3, 5, 10, and 20. In the overall analysis, cross-file RMSE decreased progressively as the number of calibration points increased, as shown in Figure 6, indicating that a small number of real observations could effectively correct systematic bias among files. When $k = 0$ (no recalibration), the

mean cross-file RMSE was 0.733 μmol m^{-2} s^{-1}; when $k = 3$, it decreased to 0.698 μmol m^{-2} s^{-1}; when $k = 5$, it further decreased to 0.682 μmol m^{-2} s^{-1}; when $k = 10$, the mean RMSE was 0.665 μmol m^{-2} s^{-1}; and when $k = 20$, it reached 0.658 μmol m^{-2} s^{-1}. For clarity, the upward and downward arrows in Figure 6 indicate increases and decreases in RMSE, respectively. The percentage values annotated in panel (a) indicate relative changes in RMSE, whereas the numerical values annotated in panel (c) indicate the absolute RMSE changes between $k = 0$ and $k = 10$ for the corresponding sessions.

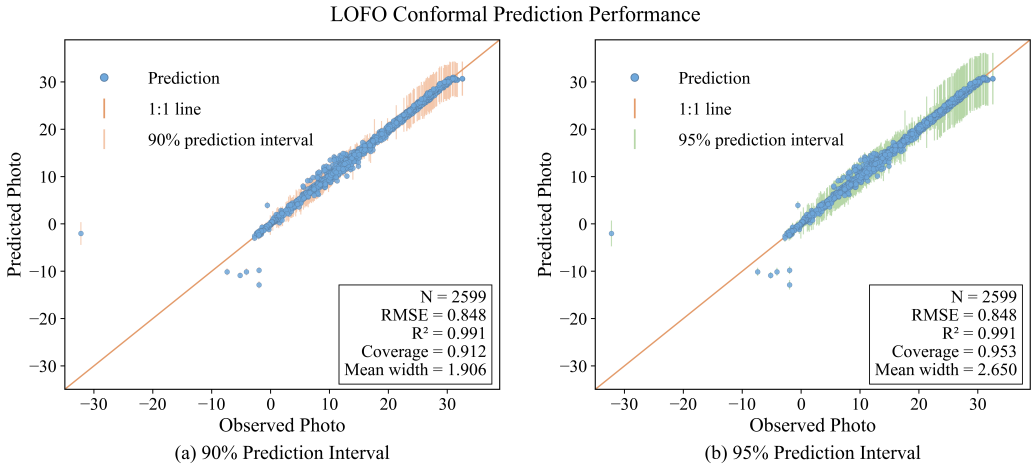

Figure 5. LOFO conformal prediction performance.

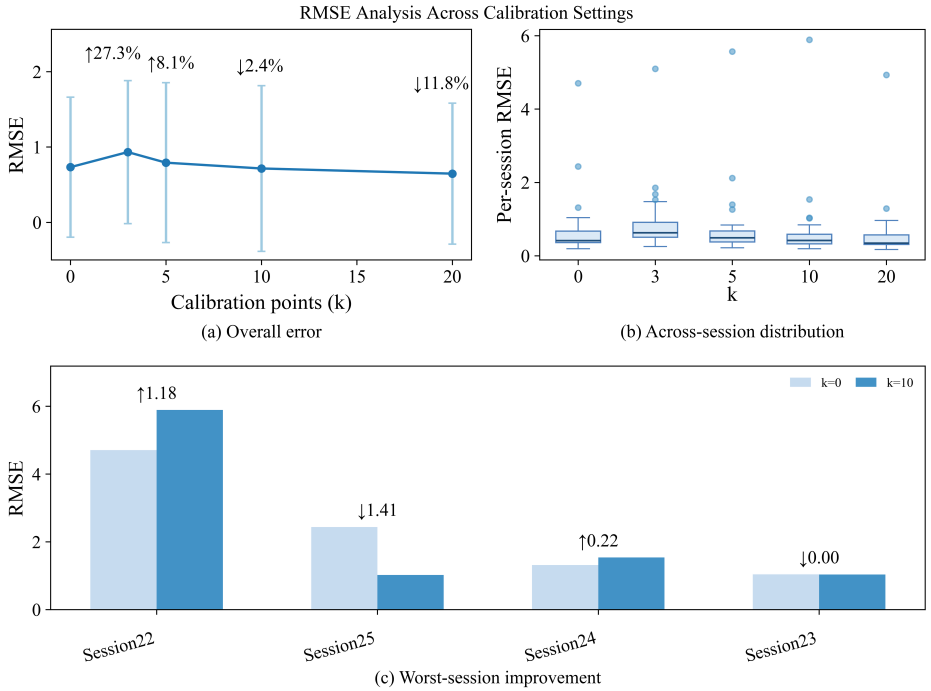

Figure 6. Few-shot recalibration results.

3.3. Environmental Correction

To reduce the influence of instantaneous environmental conditions on cross-session comparisons, each observation was adjusted to the empirically specified reference environment E_0 defined in Section 2.6.1 ($PAR_i = 240.25$ μmol photons m^{-2} s^{-1}, $VpdL = 1.4227$ kPa, $T_{leaf} = 22.38$ °C, and $CO_2S = 371.7$ μmol CO_2 mol^{-1}). The normalized value, A_{std}, was then used for cross-file and cross-batch comparison of photosynthetic rate. The coefficient of variation of the cross-file mean of the original Photo was 0.5236, whereas that of the corrected

A_{std} was 0.0208, representing a reduction of about 96.0%, as shown in Figure 7. This reduction reflects increased comparability after environmental correction and should not be interpreted as a 96% reduction in genuine biological variability.

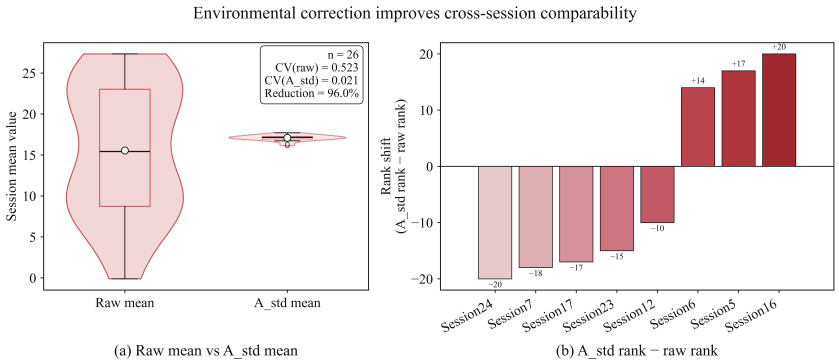

Figure 7. Environmental correction results.

3.4. Water-Use Traits and VPD Sensitivity

Across the VPD gradient, stomatal conductance (g_s) declined progressively from the low- to high-VPD class, whereas both intrinsic water-use efficiency (iWUE) and transpiration efficiency (TE) increased. Mean g_s decreased from 0.434 under low VPD to 0.196 under high VPD, while mean iWUE increased from 44.9 to 86.8; TE also increased slightly from 3.67 to 4.17. At the session level, 13 of 14 sessions showed lower g_s under high VPD, and 12 of 14 showed higher iWUE, with mean changes of −40.5% and +67.6%, respectively. Together, these results indicate that increasing VPD was associated primarily with stomatal restriction and improved water-use efficiency proxies, consistent with a predominantly water-saving response rather than a uniform increase in carbon assimilation, as shown in Figure 8.

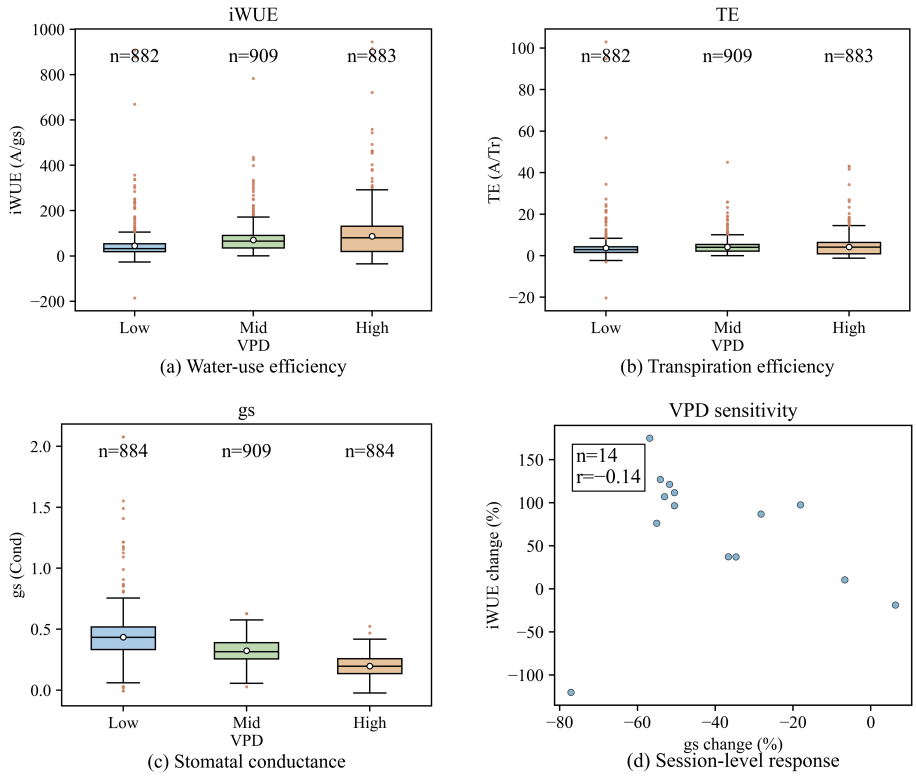

Figure 8. Overall distribution trends.

3.5. Limitation Screening

A physiology-informed rule-based limitation-screening workflow was further established to identify potential environmental or diffusional constraints associated with reduced tomato leaf photosynthetic rate under different environmental conditions, as shown in Figure 9. Photosynthetic rate, PAR, VPD, stomatal conductance (g_s), C_i/C_a, ambient CO_2 concentration, and leaf temperature were jointly considered. Stepwise classification was performed for light limitation, VPD-related limitation, stomatal limitation, CO_2-supply-related limitation, temperature-related limitation, and other/non-limited conditions.

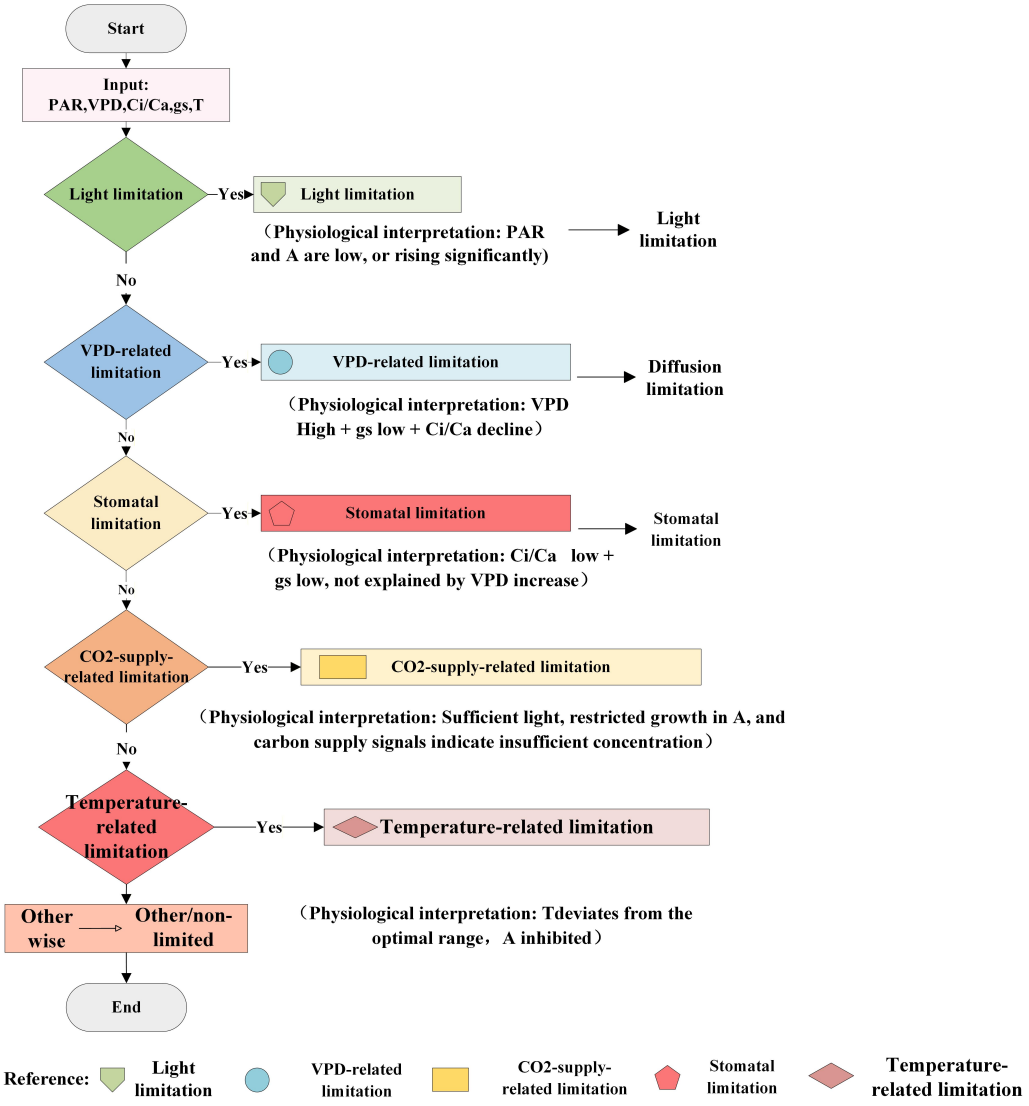

Figure 9. Photosynthetic limitation-screening flowchart.

The resulting rule-based limitation classification showed clear heterogeneity across measurement sessions, indicating that reduced tomato leaf photosynthetic rate was not associated with a single limitation class across all sessions. Instead, samples were classified into different potential constraint categories, including light limitation, VPD-related limitation, stomatal limitation, CO_2-supply-related limitation, temperature-related limitation, and other or non-limited conditions. In some sessions, CO_2-supply-related cases accounted for a relatively high proportion, suggesting that carbon supply or CO_2 diffusion status may have constrained assimilation under relatively favorable radiation conditions. In other sessions, light-limited or other/non-limited cases were dominant, indicating that insufficient incident radiation or the absence of a single strong environmental constraint better

explained the observed photosynthetic status. Temperature-related limitation accounted for only a small proportion and appeared sporadically, suggesting that temperature was not the major classified constraint within the range of this dataset. These results illustrated differences in the composition of limitation classes among sessions, as shown in Figure 10.

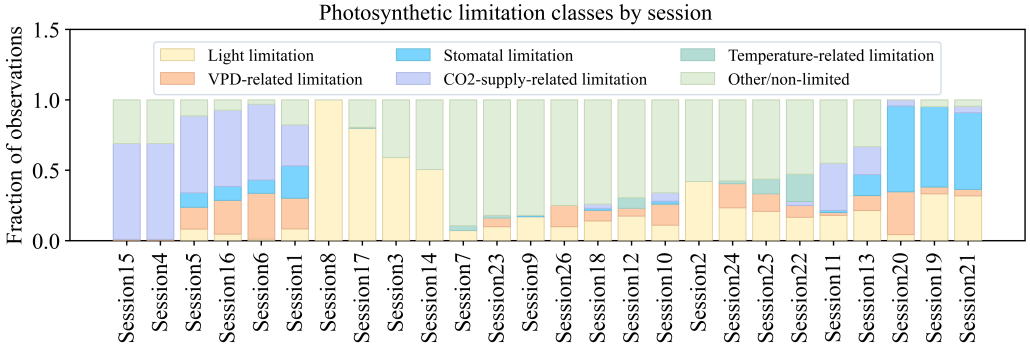

Figure 10. Composition of limitation types for each measurement session.

Evidence for separating limitation classes was further examined from the dimensions of light, carbon supply, and water cost. Under low PAR, assimilation was first associated with insufficient light drive, whereas under higher PAR, limitation more readily shifted to CO_2-supply-related or diffusional constraints, indicating that the marginal benefit of supplemental lighting was context-dependent. Meanwhile, increased VPD was often accompanied by lower C_i/C_a and low g_s, supporting the physiological chain of rising atmospheric dryness, stomatal closure, reduced intercellular CO_2 supply, and restricted assimilation, as shown in Figure 11.

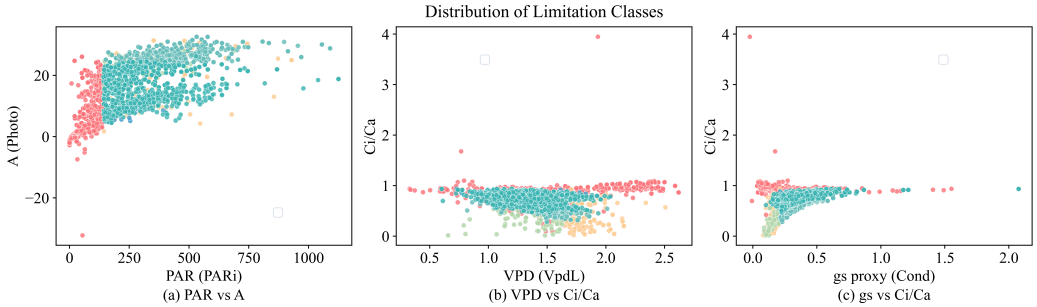

Figure 11. Distribution of restricted types.

Based on the above classification results, corresponding management directions were proposed for light limitation, VPD-related limitation, stomatal limitation, CO_2-supply-related limitation, temperature-related limitation, and other/non-limited conditions. In this way, the classification results were further translated into decision-support information for prioritizing supplemental lighting, humidity control, ventilation optimization, irrigation adjustment, and CO_2 management. A corresponding management mapping for the screened limitation classes is summarized in Figure 12.

In Figure 12, upward arrows (↑) indicate an expected increase or recovery in the corresponding monitoring indicators following management intervention, whereas downward arrows (↓) indicate an expected decrease. The arrows indicate the expected direction of change rather than the magnitude of the response.

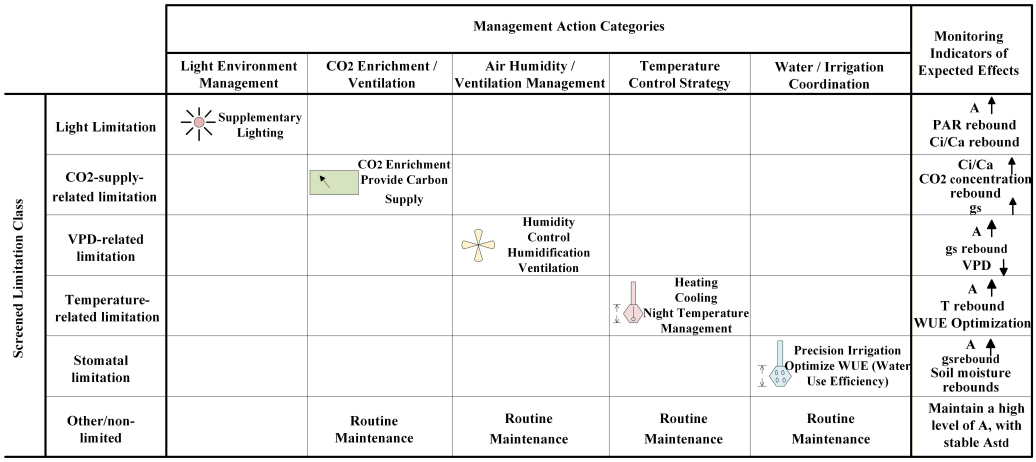

Figure 12. Diagnostic management chart.

3.6. Validation Under Deployable Input Constraints

The results showed that the baseline model using chamber/internal CO_2-related variables still achieved the highest accuracy, with an RMSE of 0.833, an MAE of 0.377, and an R^2 of 0.991, indicating that these strongly coupled variables provided a clear advantage for high-accuracy reconstruction of photosynthetic rate. In contrast, the strict deployable input set A1, including PARi, Tair, CO2R, and VPD_air reconstructed from Tair and RH_R, achieved an RMSE of 4.578, an MAE of 3.234, and an R^2 of 0.735 under the LOFO scheme. When leaf-temperature-related variables were further included, the A2 input set, including Tleaf and DeltaT_leaf_air, yielded an RMSE of 4.610, an MAE of 3.232, and an R^2 of 0.731. No improvement over A1 was achieved by the leaf-temperature-enhanced A2 setting, suggesting that, under the current data conditions, additional leaf temperature information did not stably improve cross-session extrapolation performance. These results indicated that high-accuracy photosynthetic-rate reconstruction was still more dependent on instrument-level inputs, whereas models based on routine greenhouse sensors should be regarded as approximate rather than high-accuracy substitutes for the gas-exchange-based model. A comparison of the predictive performance of the baseline and deployable-input models under LOFO cross-file extrapolation is shown in Table 4.

Table 4. Performance comparison under deployable input constraints.

Model	Input Variables	RMSE	MAE	R^2
Baseline_4 factors	C_2sfc, CO_2S, Ci, Ci_Pa	0.833	0.377	0.991
A1_strict_deployable	PARi, Tair, CO_2R, VPD_air	4.578	3.234	0.735
A2_leaftemp_enhanced	PARi, Tair, CO_2R, VPD_air, Tleaf, DeltaT_leaf_air	4.610	3.232	0.731

4. Discussion

This study developed a stacked machine-learning framework for cross-session reconstruction and environmental limitation screening of photosynthetic rate in greenhouse tomato. The results showed that a small group of CO_2-related and diffusion-related variables, including C_2sfc, CO_2S, C_i, and C_i_Pa, provided the dominant predictive information for leaf-level photosynthetic-rate reconstruction. This finding is physiologically reasonable because leaf photosynthesis is closely linked to external CO_2 supply, intercellular CO_2 status, and gas diffusion processes. The stacked ensemble provided a modest numerical reduction in RMSE relative to M1–M3, although this improvement was not statistically significant. Its performance advantage was more clearly supported against SVR, Ridge, PLSR, LSTM, and XGBoost. CatBoost and ExtraTrees captured complementary nonlinear

relationships among chamber-derived and environmental variables, whereas the RBF-SVR component provided additional residual correction.

A major strength of the proposed framework is its use of file-level extrapolation validation. In greenhouse gas-exchange datasets, samples collected within the same file or measurement session often share similar time background, environmental conditions, instrument status, and operational settings. Under such grouped data structures, ordinary random splitting may cause information overlap between training and testing samples and consequently overestimate model performance. By adopting leave-one-file-out validation, the present study evaluated the model under a stricter and more realistic cross-session setting. The high overall predictive accuracy under LOFO validation suggests that the stacked model captured stable relationships that were not confined to a single measurement batch. In addition, the use of conformal prediction intervals extended the framework from point prediction to reliability assessment, which is important for practical decision support because prediction uncertainty can help identify cases in which automatic recommendations should be interpreted with caution. Few-shot recalibration further reduced cross-session error, indicating that a small number of newly measured samples can effectively correct session-level bias.

Environmental correction provided an additional way to improve the comparability of photosynthetic traits across measurement sessions. Instantaneous photosynthetic rate is strongly influenced by light intensity, CO_2 concentration, vapor pressure deficit, and leaf temperature; therefore, direct comparison of raw photosynthetic rate across sessions can confound physiological differences with environmental variation. By adjusting observations to the specified reference environment, the environmentally corrected photosynthetic indicator reduced cross-session variability and provided a more comparable cross-session photosynthetic indicator. This does not mean that all biological differences among plants or sessions were eliminated. Rather, the environmental correction procedure was intended to reduce the component of variation explainable by short-term environmental differences, while recognizing that some genuine physiological variation may also be attenuated. For greenhouse management, this environmentally corrected indicator may be more informative than the raw instantaneous photosynthetic rate when evaluating plant status across days or measurement batches. Nevertheless, the definition of the reference environment and the environmental response function should be carefully reported, and all environmental correction procedures should be performed within the training data of each validation fold to avoid circular bias.

The analysis of water-use traits and VPD sensitivity also supported the physiological relevance of the framework. Across the VPD gradient, stomatal conductance declined, whereas intrinsic water-use efficiency and transpiration efficiency increased. This pattern suggests that higher atmospheric dryness was associated mainly with stomatal restriction and a water-saving response. Such a response is consistent with the expected physiological chain in which increasing VPD promotes stomatal closure, reduces intercellular CO_2 supply, and may constrain carbon assimilation. The limitation-screening module further translated these physiological signals into interpretable environmental constraint categories. Sessions dominated by light limitation indicate that insufficient radiation or uneven canopy light distribution may be the primary factor restricting assimilation. Sessions with high proportions of samples classified as VPD-related limitation or stomatal limitation suggest that humidity regulation, ventilation coordination, and irrigation management should be prioritized to reduce excessive transpiration demand. Sessions with high proportions of CO_2-supply-related limitation indicate that carbon supply may become limiting under relatively favorable radiation conditions, and therefore, CO_2 enrichment or ventilation-based CO_2 management may be considered. Thus, the added value of AI in this study

lies primarily in integrating, comparing, and screening complex physiological information across measurement sessions rather than in redefining the underlying crop physiological mechanisms. From a crop-science perspective, the main value of the framework is not to replace established physiological knowledge, but to organize multiple gas-exchange and environmental signals into information that can be more directly interpreted for crop management. The selected CO_2-related variables summarize carbon-supply and diffusion status, whereas the VPD, stomatal-conductance, and water-use analyses provide information on plant water regulation. By combining these physiological signals with limitation screening, the AI framework can help crop scientists identify whether reduced photosynthetic performance is more closely associated with insufficient light, atmospheric dryness and stomatal restriction, CO_2 supply, or temperature conditions, and can therefore help prioritize the corresponding management response.

The deployable-input validation clarified the practical boundary of the proposed method. The baseline model using chamber/internal CO_2-related variables achieved substantially higher accuracy than the strict deployable sensor-input models. This contrast indicates that high-accuracy photosynthetic-rate reconstruction still depends strongly on gas-exchange-derived variables, whereas routine greenhouse environmental sensors provide only an approximate representation of leaf-level photosynthetic status. Therefore, the proposed framework has two potential application levels. The first is a high-accuracy chamber-level model for gas-exchange data analysis, phenotyping, quality control, and physiology-informed limitation screening. The second is a deployable approximation model based on routine greenhouse sensors, which may support online environmental monitoring and preliminary decision support but should not be expected to reach the same accuracy as the chamber-level model. This distinction is important for avoiding overstatement of the model's operational readiness. Future work should integrate additional deployable sensing sources, such as canopy microclimate sensors, thermal imaging, chlorophyll fluorescence, hyperspectral information, or continuous CO_2 monitoring, to improve the accuracy and stability of greenhouse-level prediction.

Several biological and application-related limitations should be acknowledged. First, the study was based on a single public greenhouse-tomato gas-exchange dataset, and the physiological relationships identified here should therefore be further validated across cultivars, seasons, greenhouse types, and management regimes. Second, the analysis was conducted at the leaf level, whereas crop productivity ultimately depends on canopy photosynthesis, biomass accumulation, and yield formation; direct extrapolation from leaf-level reconstruction to whole-crop performance should therefore be avoided. Third, the dominant predictors were mainly chamber- or gas-exchange-derived CO_2 variables that are closely coupled with Photo, so their importance should be interpreted as representing gas-exchange reconstruction and carbon-supply–diffusion status rather than as independent agronomic drivers. Fourth, the environmental correction may remove part of the genuine physiological variation among sessions, and the corrected indicator should therefore not be interpreted as a stable biological phenotype. Finally, the physiology-informed limitation classes are screening indicators rather than causal diagnoses and should be validated through independent experiments involving controlled manipulation of light, VPD, CO_2, temperature, and irrigation.

For crop scientists, three practical messages emerge from this study. First, CO_2 supply and intercellular diffusion status contain strong information on leaf photosynthetic performance in gas-exchange measurements. Second, VPD-related reductions in stomatal conductance provide a physiologically interpretable signal of increasing water limitation and carbon-diffusion restriction. Third, combining photosynthetic reconstruction with physiology-informed limitation screening can help convert complex gas-exchange mea-

surements into management-oriented information for lighting, humidity and ventilation, irrigation, and CO_2 regulation. Overall, the results demonstrate that a CatBoost–ExtraTrees–RBF-SVR stacked ensemble, combined with leakage-aware cross-session validation, uncertainty quantification, environmental correction, and physiology-informed limitation screening, can provide a useful analytical framework for greenhouse tomato photosynthesis research. The framework not only improves photosynthetic-rate prediction from gas-exchange data but also helps transform numerical predictions into interpretable information for environmental management. With further validation under broader production conditions and improved integration of deployable sensors, this approach has the potential to support more robust decision-making for supplemental lighting, humidity and VPD regulation, CO_2 management, and irrigation coordination in protected tomato cultivation. For crop scientists, the principal value of the framework is therefore its ability to translate complex gas-exchange measurements into physiologically interpretable indicators and management priorities, while retaining clear boundaries between statistical reconstruction, physiological screening, and causal inference.

5. Conclusions

A gas-exchange-data-based framework for cross-session reconstruction of leaf photosynthetic rate and environmental limitation screening was established for greenhouse tomato. Cross-session stability screening reduced 50 candidate variables to four core CO_2-related predictors, and a CatBoost–ExtraTrees–RBF-SVR stacked ensemble achieved high reconstruction accuracy under nested validation and strict leave-one-file-out extrapolation, although its RMSE improvement over M1–M3 was modest and not statistically significant. However, restricting the inputs to deployable greenhouse sensors markedly reduced performance: A1 achieved an RMSE of 4.578 and an R^2 of 0.735, while A2 achieved an RMSE of 4.610 and an R^2 of 0.731, compared with an RMSE of 0.833 and an R^2 of 0.991 for the gas-exchange baseline. Quality-weighted training slightly improved cross-file robustness, and few-shot recalibration further reduced the mean cross-file RMSE from 0.733 to 0.658, corresponding to a decrease of approximately 10.23%.

Environmental correction yielded a more comparable cross-session photosynthetic indicator, reducing the coefficient of variation of session means from 0.5236 to 0.0208 (approximately 96.0%). This reduction should be interpreted as improved cross-session comparability rather than elimination of genuine biological variability. On this basis, the value of environmental limitation screening lies in identifying the potential dominant constraints associated with reduced photosynthetic rate. When light limitation dominated a session, supplemental lighting, optimization of canopy light distribution, or improvement of greenhouse light transmission should be prioritized. When VPD-related limitation or stomatal limitation accounted for a high proportion, regulation should focus on irrigation, ventilation, humidification, and the interleaf microenvironment to alleviate stomatal closure caused by excessive transpiration demand. When CO_2-supply-related limitation dominated, CO_2 fertilization or ventilation-based carbon supplementation could be further optimized. Therefore, these results supported the physiological plausibility of the limitation-screening method and showed that the method could provide decision-support information for supplemental lighting, humidity control, irrigation, and CO_2 management in greenhouse tomato production.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/agriengineering8090383/s1: Table S1: Complete Set of Candidate Predictors; Table S2: Main Hyperparameters and Validation Settings Used in This Study; Table S3: Date-Specific Sampling Information and Sample Sizes of the Greenhouse Tomato Gas-Exchange Dataset; Table S4: Overall LOFO Performance Results for Different Input Settings; Figure S1: Overall

RMSE, MAE, and R^2 of the Baseline and Deployable-Input Models under LOFO Validation; Figure S2: Session-by-Session RMSE across the 26 Measurement Sessions; Figure S3: Comparison of Observed and Predicted Photo Values for the Baseline and Deployable-Input Models; Table S5: Sensor-Group Ablation within the A1 Strict Deployable-Input Setting.

Author Contributions: Conceptualization, G.Z.; methodology, G.Z., S.Z., and H.L.; software, G.Z.; validation, L.T. and J.Z.; formal analysis, G.Z.; investigation, S.Z. and Y.Z.; data curation, S.Z., Y.Z., and H.L.; writing—original draft preparation, G.Z.; writing—review and editing, L.T. and J.Z.; visualization, G.Z. 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.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data analyzed in this study are publicly available from the dataset of photosynthesis and photosynthetic factor measurements of greenhouse tomato reported by Manjarrez-Sanchez and Martinez-Carrillo [29]. The processed data and analysis scripts generated during the current study are available from the corresponding author upon reasonable request.

Acknowledgments: The authors would like to acknowledge Shanghai Polytechnic University for providing academic support.

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

References

1. Wang, P.; Yu, A.; Huo, J.; Wang, Y.; Shang, Y.; Pang, X.; Jiang, K. Effects of nitrogen reduction with green manure incorporation on maize photosynthetic performance and yield. *J. Agric. Food Res.* **2025**, *24*, 102320. https://doi.org/10.1016/j.jafr.2025.102320.
2. Li, R.; Wang, H.; Wu, S.; Chen, J.; Zhou, Z.; Zhang, Y.; Cui, J.; Zhu, C.; Miao, C.; Chang, L.; et al. Optimizing Light Environment for Pakchoi in Plant Factories: Interactive Effects of Photoperiod and Light Intensity on Growth, Photosynthesis, and Energy-Use Efficiency. *Horticulturae* **2026**, *12*, 215. https://doi.org/10.3390/horticulturae12020215.
3. El Ouaham, W.; Sadik, M.; Ennajih, A.; Mouzouna, Y.; Orchi, H.; Elouaham, S. Smart Greenhouses in the Era of IoT and AI: A Comprehensive Review of AI Applications, Spectral Sensing, Multimodal Data Fusion, and Intelligent Systems. *Agriculture* **2026**, *16*, 761. https://doi.org/10.3390/agriculture16070761.
4. Wang, M.; Zhang, Z.; Luo, L.; Liu, F.; Bai, S.; Dong, Y.; Zhang, K.; Wan, Y. Appropriate Drip Irrigation Pipe Burial Depth Can Enhance Peanut Photosynthetic Capacity and Promote Yield Formation. *Agriculture* **2026**, *16*, 538. https://doi.org/10.3390/agriculture16050538.
5. Kowalczyk, K.; Mirgos, M.; Geszprych, A.; Przybył, J.; Sujkowska-Rybkowska, M.; Borucki, W.; Kalisz, S.; Jonczak, J.; Sobczak-Samburska, A.; Kunka, M.; et al. Effect of High Pressure Sodium and Light-Emitting Diode Lamps' Supplementary Lighting and Diffusion Glass on Growth, Yield, and Fruit Quality of Pink Tomato. *Agriculture* **2025**, *15*, 499. https://doi.org/10.3390/agriculture15050499.
6. Yu, J.; Liu, J.; Sun, C.; Wang, J.; Ci, J.; Jin, J.; Ren, N.; Zheng, W.; Wei, X. Sensing technology for greenhouse tomato production: A systematic review. *Smart Agric. Technol.* **2025**, *11*, 101020. https://doi.org/10.1016/j.atech.2025.101020.
7. Joubert, D.; Zhang, N.; Berman, S.R.; Kaiser, E.; Molenaar, J.; Stigter, J.D. A small dynamic leaf-level model predicting photosynthesis in greenhouse tomatoes. *PLoS ONE* **2023**, *18*, e0275047. https://doi.org/10.1371/journal.pone.0275047.
8. Niu, Y.; Li, Y.; Han, Y.; Zhang, M.; Li, H. Decision method for greenhouse tomato light regulation based on the concavity of photosynthesis response. *Comput. Electron. Agric.* **2024**, *222*, 109088. https://doi.org/10.1016/j.compag.2024.109088.
9. Cheng, Y.; Li, N.; Li, Z.; Zhou, A.; Li, B.; Miao, Y. Analysis of multi-environment-driven variations in net photosynthetic rate and predictive model development for tomatoes during early flowering and fruit development stages in winter solar greenhouses. *Horticulturae* **2025**, *11*, 1367. https://doi.org/10.3390/horticulturae11111367.
10. Miao, Y.; Liu, L.; Wang, M.; Zeng, Z.; Zhang, J.; Cheng, Y.; Li, B. A photosynthetic rate prediction model for cucumber based on a machine learning algorithm and multi-factor environmental analysis. *Horticulturae* **2025**, *11*, 1475. https://doi.org/10.3390/horticulturae11121475.
11. Ma, X.; Chang, J.; Li, W.; Li, R.; Jing, D.; Zhang, L.; Liu, Y.; Li, J. Limit and enhancing potential of canopy photosynthesis for greenhouse tomato: A model analysis in different climatic environments. *Plant Phenomics* **2025**, *7*, 100069. https://doi.org/10.1016/j.plaphe.2025.100069.

12. Ma, X.; Chang, J.; Chai, X.; Liu, Y.; Li, J. Evaluating the canopy light environment, photosynthesis, and fruit comprehensive performance of greenhouse tomato under different mechanized planting layouts. *Hortic. Plant J.* 2025, *in press, corrected proof.* https://doi.org/10.1016/j.hpj.2025.03.010.
13. Bua, C.; Wang, Y.; Ojo, M.O.; Giordano, S.; Zahid, A. Digital Twin-Enabled Multi-Zone Adaptive Lighting Control in Greenhouses Using Reinforcement Learning Optimization. *Smart Agric. Technol.* **2026**, *14*, 101984. https://doi.org/10.1016/j.atech.2026.101984.
14. Ding, J.; Jiao, X.; Bai, P.; Hu, Y.; Zhang, J.; Li, J. Effect of vapor pressure deficit on the photosynthesis, growth, and nutrient absorption of tomato seedlings. *Sci. Hortic.* **2022**, *293*, 110736. https://doi.org/10.1016/j.scienta.2021.110736.
15. Pan, T.; Zhang, W.; Du, W.; Fu, B.; Zhou, X.; Cao, K.; Bao, E.; Wang, Y.; Lv, G. Synergistic effects of elevated CO_2 and enhanced light intensity on growth dynamics, stomatal phenomics, leaf anatomy, and photosynthetic performance in tomato seedlings. *Horticulturae* **2025**, *11*, 760. https://doi.org/10.3390/horticulturae11070760.
16. Šalagovič, J.; Vanhees, D.; Verboven, P.; Holsteens, K.; Verlinden, B.; Huysmans, M.; Van de Poel, B.; Nicolaï, B. Microclimate monitoring in commercial tomato (*Solanum Lycopersicum* L.) greenhouse production and its effect on plant growth, yield and fruit quality. *Front. Hortic.* **2024**, *3*, 1425285. https://doi.org/10.3389/fhort.2024.1425285.
17. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A.W.; O'SUllivan, J.M. A review of feature selection methods for machine learning-based disease risk prediction. *Front. Bioinform.* **2022**, *2*, 927312. https://doi.org/10.3389/fbinf.2022.927312.
18. Sudheesh, S.G.; Beucler, T.; Tam, F.I.H.; Gomez, M.S.; Runge, J.; Gerhardus, A. Selecting robust features for machine-learning applications using multidata causal discovery. *Environ. Data Sci.* **2023**, *2*, e27. https://doi.org/10.1017/eds.2023.21.
19. Varghese, R.; Cherukuri, A.K.; Doddrell, N.H.; Doss, C.G.P.; Simkin, A.J.; Ramamoorthy, S. Machine learning in photosynthesis: Prospects on sustainable crop development. *Plant Sci.* **2023**, *335*, 111795. https://doi.org/10.1016/j.plantsci.2023.111795.
20. Deng, X.; Zhang, Z.; Hu, X.; Li, J.; Li, S.; Su, C.; Du, S.; Shi, L. Estimation of photosynthetic parameters from hyperspectral images using optimal deep learning architecture. *Comput. Electron. Agric.* **2024**, *216*, 108540. https://doi.org/10.1016/j.compag.2023.108540.
21. Wan, L.; Ma, F. Estimating leaf photosynthetic capacity using hyperspectral reflectance: Model variability and transferability. *Comput. Electron. Agric.* **2024**, *220*, 108837. https://doi.org/10.1016/j.compag.2024.108837.
22. Ji, F.; Li, F.; Hao, D.; Shiklomanov, A.N.; Yang, X.; Townsend, P.A.; Dashti, H.; Nakaji, T.; Kovach, K.R.; Liu, H.; et al. Unveiling the transferability of PLSR models for leaf trait estimation: Lessons from a comprehensive analysis with a novel global dataset. *New Phytol.* **2024**, *243*, 111–131. https://doi.org/10.1111/nph.19807.
23. Parida, P.K.; Eagan, S.; Ramanujam, K.; Sengodan, R.; Uthandi, S.; Ettiyagounder, P.; Rajagounder, R. Machine learning approaches for estimation of the fraction of absorbed photosynthetically active radiation and net photosynthesis rate of maize using multi-spectral sensor. *Heliyon* **2024**, *10*, e34117. https://doi.org/10.1016/j.heliyon.2024.e34117.
24. Li, H.; Zhang, H.; Wang, Y.; Zhao, J.; Feng, Z.; Chen, H.; Guo, X.; Xiong, T.; Xiao, J.; Li, X. Evaluation of photosynthesis estimation from machine learning-based solar-induced chlorophyll fluorescence downscaling from canopy to leaf level. *Ecol. Indic.* **2024**, *166*, 112439. https://doi.org/10.1016/j.ecolind.2024.112439.
25. John, K.; Saurette, D.D.; Heung, B. The problematic case of data leakage: A case for leave-profile-out cross-validation in 3-dimensional digital soil mapping. *Geoderma* **2025**, *455*, 117223. https://doi.org/10.1016/j.geoderma.2025.117223.
26. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. https://doi.org/10.1016/j.patter.2023.100804.
27. Zhu, J.J.; Yang, M.; Ren, Z.J. Machine learning in environmental research: Common pitfalls and best practices. *Environ. Sci. Technol.* **2023**, *57*, 17671–17689. https://doi.org/10.1021/acs.est.3c00026.
28. Lones, M.A. Avoiding common machine learning pitfalls. *Patterns* **2024**, *5*, 101046. https://doi.org/10.1016/j.patter.2024.101046.
29. Manjarrez-Sanchez, J.; Martinez-Carrillo, G. Dataset of photosynthesis and photosynthetic factors measurements of greenhouse tomato. *Data Brief* **2020**, *32*, 106274. https://doi.org/10.1016/j.dib.2020.106274.
30. Li, Y.; Liu, X.; Zhang, Z.; Lang, J. GPP estimation based on CatBoost and analysis of change driving factors in Shanxi Province, China. *Sci. Rep.* **2025**, *15*, 22346. https://doi.org/10.1038/s41598-025-08927-x.
31. Guimarães, N.; Sousa, J.J.; Couto, P.; Bento, A.; Pádua, L. Combining UAV-based multispectral and thermal infrared data with regression modeling and SHAP analysis for predicting stomatal conductance in almond orchards. *Remote Sens.* **2024**, *16*, 2467. https://doi.org/10.3390/rs16132467.
32. Liu, T.; Yang, F.; Zhang, P.; Wang, Z.; Zhang, H. Simultaneous prediction of plant photosynthetic parameters using canopy hyperspectral imaging and an explainable XGBoost-SHAP integration framework. *Measurement* **2026**, *274*, 121257. https://doi.org/10.1016/j.measurement.2026.121257.