Bound the Risk, Gate the Bias: A Mechanistic-Driven Sim-to-Real Transfer Method for Rebust Yield Prediction

1st Tiansong Wu
College of Informatics
Huazhong Agricultural University
Wuhan 430070, China
iwts1096@webmail.hzau.edu.cn

2nd Yuling Fan
College of Informatics
Huazhong Agricultural University
Wuhan 430070, China
ylfan.up@gmail.com

3rd Ning Li
College of Informatics
Huazhong Agricultural University
Wuhan 430070, China
li_ning5047@163.com

4th Dong Lin
Faculty of Electrical and Control Engineering
Liaoning Technical University
Fuxin 123000, China
donglin0705@gmail.com

***Abstract*—Precision fertilization is the core of sustainable agricultural development and relies heavily on high-precision yield prediction. Crop model-based yield prediction is crucial for sim-to-real transfer learning. However, under conditions of data scarcity, the extrapolation ability is generally limited, risks are difficult to quantify, and prediction accuracy is low. Therefore, this paper proposes a bias analysis method based on 'perfect calibration' to achieve sim-to-real transfer learning, aiming to obtain robust and reliable yield prediction. First, a crop mechanistic model is calibrated on real data to generate a large amount of simulated data as ideal yield observation data (pseudo-observation data). And the 'perfect error (PE)' is calculated under controlled conditions. Subsequently, the PE is used as a mechanistic reference lower bound (MRLB) for extrapolation error through a zero-error test and statistical analysis for prediction risk assessment. Secondly, the bias decomposition for PE is proposed, and a gated debiasing procedure is designed to obtain a low-bias dataset for pre-training the neural network. Fine-tuning is then performed on a small amount of real data to achieve mechanistic-fused transfer learning for yield prediction. Experimental results show that, under conditions of extreme data scarcity, the proposed method significantly outperforms multiple baseline methods and provides robust yield confidence intervals. This method provides a new approach for the uncertainty analysis and trustworthy prediction in agriculture under extreme data scarcity by deeply integrating mechanistic modeling with machine learning, and can be extended to the calibration and evaluation of mechanistic models under similar data scarcity conditions.**

***Index Terms*—crop yield prediction, extrapolation risk, gated debiasing, sim-to-real transfer, uncertainty quantification.**

I. Introduction

Precision fertilization is a key measure to improve crop yield and resource utilization efficiency, and is of great significance to ensuring food security and sustainable agricultural development [1]. However, its decision-making relies heavily on high-precision yield prediction. Crop mechanistic models (e.g., DSSAT, APSIM) are widely used for yield prediction and fertilization decision support due to their interpretability and composability [2], [3]. Yet high-quality field observation data is costly and time-consuming to obtain in practical production [4], leading to extreme data scarcity. Under small samples, conventional calibration methods easily overfit known scenarios, overestimating extrapolation performance [5]. In response, Researchers often use cross-validation to approximate extrapolation performance, but these methods fail to characterize extrapolation uncertainty, lacking reliable risk assessment for predictions. For instance, Yang et al. applied leave-one-out cross-validation (LOOCV) in crop phenology simulation to evaluate the uncertainty of different model structures and parameters [6]. The results show that although the calibration model performed well on LOOCV, significantly higher errors were observed in the validation set.

Moreover, crop mechanistic models inevitably have systematic errors due to the simplification and assumptions made about complex biological processes [7]. This bias reflects the sim-to-real gap between the mechanistic model and the real system. Existing research has shown that systematic bias is an important factor leading to uncertainty in yield prediction. Ignoring this gap will cause 'false confidence' in calibration models and affect the extrapolation accuracy [8], [9]. Current approaches to addressing such systematic bias during model calibration can be broadly divided into two categories: one is to mitigate systematic errors through implicit means. For example, the advantages of different models can be integrated by multi-model ensemble and Bayesian model averaging to improve the robustness of regional-scale yield prediction [10]. The other is to adopt explicit data-driven discrepancy modeling and estimate parameters and systematic errors in parallel during Bayesian calibration [11]. However, in the case of extreme data scarcity, it is often difficult to clearly isolate systematic bias and mitigate its associated risks,

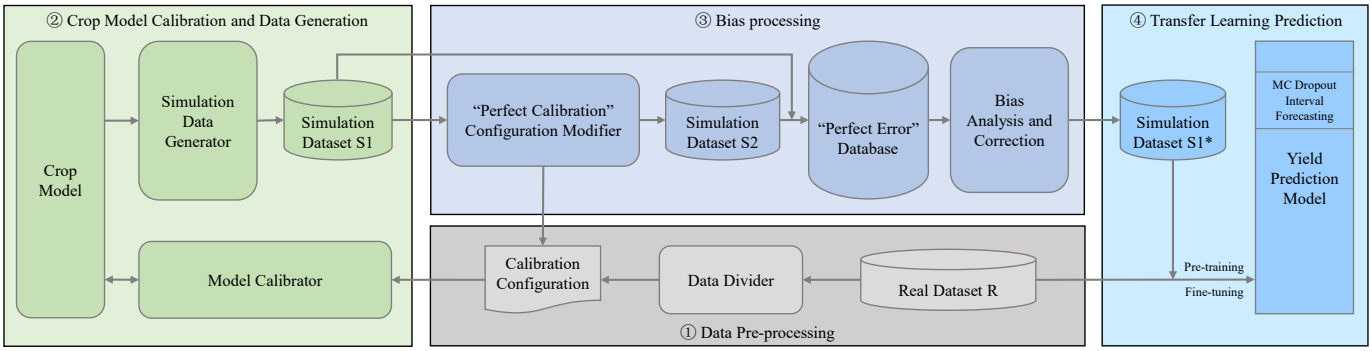

Fig. 1. Mechanism-driven transfer learning framework for crop yield prediction

making it difficult to 'prescribe the right medicine' in the calibration process, ultimately constraining the effectiveness and pertinence of subsequent debiasing interventions.

In addition, data-driven modeling is also one of the commonly used prediction methods. However, under conditions of data scarcity, pure data-driven models will also exacerbate the risk of overfitting and reduce generalization ability [12]. In contrast, hybrid modeling strategies that combine mechanistic models and data-driven models can utilize prior knowledge and statistical laws at the same time and have been recognized as an effective path to improve prediction accuracy [13]. In recent years, hybrid modeling has shown significant potential in agricultural applications. For example, Saravi et al. used DSSAT-generated data to train a deep network for maize yield prediction [14]; Tao optimized fertilization and irrigation via DSSAT and reinforcement learning [15]; Li et al. combined APSIM with machine learning for superior wheat yield prediction [16]. However, most of these studies ignore or assume unbiased simulation, inadequately addressing systematic bias and uncertainty.

To address these issues, this paper proposes a bias analysis method based on 'perfect calibration' for sim-to-real transfer learning, aiming to achieve robust yield prediction via 'risk constraint' and 'bias reduction'. The technical path is: calibration—bias analysis—gated debiasing—transfer learning—uncertainty quantification. The main contributions are as follows: (1) Introduce 'perfect calibration' and zero error test (ZET) to analyze systematic bias and its composition; propose perfect error (PE) as the mechanistic reference lower bound (MRLB) for extrapolation error, quantifying calibration uncertainty under data scarcity. (2) Propose PE bias decomposition and design a gated debiasing procedure to mitigate systematic bias, providing high-quality source-domain data for transfer learning. (3) Propose a sim-to-real transfer learning method integrating mechanistic knowledge and data-driven approaches, training on debiased simulation data and fine-tuning with limited real data; integrate MRLB with MC dropout for continuous uncertainty assessment, filling the gap in pre-transfer risk quantification.

II. Mechanistic-Driven Transfer Learning Method

To achieve robust crop yield prediction and support precision fertilization decisions, this study develops a mechanism-driven transfer learning framework (Fig. 1) that integrates crop model mechanisms, bias analysis, and data-driven approaches. The proposed framework follows a four-step workflow: (1) data preprocessing, (2) model calibration and simulation, (3) bias identification and correction, and (4) transfer-based prediction. Detailed descriptions are provided in Sections II-A, II-B, and II-C.

A. Mechanistic Model Calibration and Risk Assessment

Under data-scarce conditions, the extrapolation risk associated with model calibrated is difficult to quantify. Therefore, a perfect calibration experiment is designed and a zero-error test is conducted to deeply analyze the systematic bias and its components in the calibration process, which helps to derive and establish the MRLB, thereby providing support for extrapolation risk assessment.

1) Perfect Calibration and Zero-Error Test: The perfect calibration experiment is designed as follows. First, the real field dataset is grouped by year to form initial calibration configurations $g \in \{1, 2, \ldots, G\}$. For each configuration g, the crop model M is calibrated to derive calibrated parameters θ_1 and the mechanistic yield function:

$$y_1(x) = M(x; \theta_1) \tag{1}$$

where x denotes input variables such as fertilization (nitrogen N, phosphorus P) and meteorology. A simulated dataset S_1 is generated through the function. From S_1, a subset that matches the fertilization combinations and sample size of the original real dataset is extracted to serve as 'pseudo-observation data'. The same calibration procedure is repeated on the pseudo-observation data to obtain parameter θ_2 and the yield function:

$$y_2(x) = M(x; \theta_2) \tag{2}$$

based on which, a simulated dataset S_2 is obtained. Since the pseudo-observation data are generated by the model itself and are therefore structurally consistent with the model. This second calibration process is termed 'perfect calibration',

namely, error minimization is theoretically attainable. Thus it can be used for comparison and error analysis of real calibrations, drawing cross-domain inspiration from the ideas of 'Pseudo-observations' and 'Perfect Model Test' in climate science [17].

Under a controlled perfect calibration setting with the same scale, procedure, and objective, the errors for both calibrations are defined as:

$$\varepsilon_1(x) = |y_1(x) - y_0(x)|, \quad \varepsilon_2(x) = |y_2(x) - y_1(x)| \tag{3}$$

where ε_1 denotes the absolute error of calibration using real data (hereafter referred to as real error), ε_2 denotes the perfect calibration error, and $y_0(x)$ is the true but unobservable yield function.

Since $y_0(x)$ is unknown, direct analysis of ε_1 is infeasible. In contrast, both y_1 and y_2 are observable and deterministically generated, and ε_2 is free from observation noise. It measures whether the model can reproduce its own generated mechanistic function under identical calibration constraints and primarily reflects the limiting fitting capability of the 'calibration algorithm–mechanistic model' combination under given data and computational resources. The two errors may include: (1) Structural-side error: mismatch caused by model structural simplifications. (2) Implementation-side error: local mismatch caused by the calibration algorithm failing to reach the global optimum. (3) Combined effects of the above two.

In perfect calibration, the data structure is naturally consistent with the model structure. This allows to isolate and characterize the implementation-side error independently without interference from observation noise or structural-side error, thereby analyzing error regularities and effectively reducing the magnitude of ε_1.

To further verify this decomposition and exclude random non-systematic disturbances, a zero-error test is conducted. Instead of random initialization, the second calibration is warm-started from θ_1. Since the calibration data and the model are self-consistent, the calibration model should theoretically be able to fully reproduce $y_1(x)$. If repeated calibrations consistently satisfy $\varepsilon_2 \leq \delta_{\mathrm{t}}$ (with δ_{t} being the tolerance, e.g., 10^{-6}), this indicates that the perfect calibration error primarily reflects implementation-side error, representing the upper limit of achievable fitting accuracy and no significant random interference.

2) Mechanistic Reference Lower Bound (MRLB) for Risk Assessment: Based on the two-step calibration and the ZET, the real calibration error ε_1 consists of three components:

$$\varepsilon_1(x) = \varepsilon_{\text{impl}}(x) + \varepsilon_{\text{struct}}(x) + \varepsilon_{\text{noise}}(x) \tag{4}$$

where $\varepsilon_{\text{impl}}$, $\varepsilon_{\text{struct}}$ and $\varepsilon_{\text{noise}}$ denote implementation-side error, structural-side bias and observation/label noise, respectively. In contrast, perfect calibration eliminates observation noise and excludes structural-side error, yielding:

$$\varepsilon_2(x) = \varepsilon_{\text{impl}}(x) \tag{5}$$

Because both calibrations use the same optimization algorithm, and the mismatch between the true yield structure and the model structure must be greater than that in the perfect calibration scenario, ε_1 necessarily contains larger implementation-side error than ε_2. In addition, ε_1 contains additional error components. Therefore, the following expected lower bound can be derived under the same input distribution:

$$\mathbb{E}[\varepsilon_1(x)] \geq \mathbb{E}[\varepsilon_2(x)] =: \mathrm{LB}_{\text{mech}}^{\text{mean}} \tag{6}$$

where $\mathrm{LB}_{\text{mech}}^{\text{mean}}$ is the MRLB, which is a statistical expected lower bound rather than a pointwise absolute bound. Although reversals may occur at specific inputs, the inequality holds in expectation. MRLB transforms expert intuition about calibration precision and intrinsic model limitations into a quantitative indicator. Under extremely limited samples, it enables the mechanistic model itself to serve as a tool for uncertainty assessment, providing a theoretical uncertainty baseline that traditional methods (e.g., cross-validation) cannot offer.

B. Bias Characterization and Gated Debiasing

According to the analysis in Section II-A, implementation-side errors originate from limited samples and parameter optimization failing to reach the global optimal, thereby leading to local model mismatch. To further investigate the underlying bias patterns and implement effective interventions, multiple strategies are adopted to characterize these biases, and a gated triggering mechanism is further designed for debiasing. Structural-side errors cannot be isolated independently and are not studied in this paper.

1) Bias Characterization: Four types of robust bias patterns are analyzed on the implementation-side bias. (1) Linear bias: A linear decomposition is applied to the second calibration bias to identify the global additive shift $a_t^{\rightarrow}$ and multiplicative scaling rule $b_t^{\rightarrow}$; (2) Structural monotonicity: Spearman coefficients ρ_N, ρ_P are analyzed to describe the monotonic trends of the bias along the nitrogen N and phosphorus P gradients; (3) Year effect: For each year t, the central tendency of relative bias μ_t is defined as:

$$\mu_t = \mathrm{median}_{i \in t}\big(r_i\big) \tag{7}$$

where $r_i = b_i/(y_{1,i} + \epsilon)$ denotes the relative bias of individual samples, $b_i = y_{2,i} - y_{1,i}$ is the bias between two calibrations, and $\epsilon = 10^{-6}$ is a small constant for numerical stabilization. The variance ratio:

$$R_{\mathrm{t}}^2 = \mathrm{Var}_t(\mu_t)/\mathrm{Var}_i(r_i) \tag{8}$$

quantifies the contribution of inter-annual variability in mean bias relative to the total variance of all sample-level relative biases. Var denotes sample variance and subscripts indicate the dimension of aggregation; (4) $N \times P$ coupled bias structure: for each (N, P) grid cell, define the median relative-bias field $\delta_{NP}(N, P, t) = \mathrm{median}\big(\mathrm{RelBias}(N, P, t)\big)$, and use the interquartile range (IQR) to quantify local uncertainty of the bias within each grid cell.

2) Gated Debiasing: According to the above bias analysis, we propose task-agnostic and model-form-agnostic gating indicators and a two-stage debiasing method.

Stage 1: Global year-wise affine alignment. If the year effect R_t^2 is significant, a robust regression is fitted for year t:

$$y_2 \approx a_t^{\rightarrow} + b_t^{\rightarrow} y_1 \tag{9}$$

Then it is converted into inverse coefficients for "reverse correction" of simulated data S_1:

$$b_t = \frac{1}{b_t^{\rightarrow}}, \quad a_t = -\frac{a_t^{\rightarrow}}{b_t^{\rightarrow}} \tag{10}$$

The original simulated data $S_1(N, P, t)$ is then aligned year-wise:

$$S_{1,\mathrm{aff}}(N, P, t) = a_t + b_t \cdot S_1(N, P, t) \tag{11}$$

Stage 2: Local gated multiplicative correction. On gated grid points, the final corrected signal $S_1^*(N, P, t)$ is obtained by multiplicatively adjusting the affine-corrected signal $S_{1,\mathrm{aff}}(N, P, t)$ using a robust coupled bias field:

$$S_1^*(N, P, t) = \frac{S_{1,\mathrm{aff}}(N, P, t)}{1 + \lambda(N, P, t) \cdot \delta_{NP}(N, P, t)} \tag{12}$$

where $\lambda(N, P, t) \in [0, 1]$ denotes the gating strength. The gating strength is jointly determined by the following three decision logic indicators:

a) Gate-A Sign Consistency: For each grid cell and year, the sign-consistency ratio $\mathrm{SC}(N, P, t)$ across g independent experimental configurations is defined as:

$$\mathrm{SC}(N, P, t) = \frac{1}{G} \sum_{g=1}^{G} \mathbf{1}\Big(\operatorname{sign}\left(r^{(g)}\right) = \operatorname{sign}\left(\delta_{NP}\right) \Big) \tag{13}$$

where $r^{(g)}$ denotes the relative bias obtained under the g-th experimental configuration. A binomial test p_{SC} evaluates its deviation from random agreement. Only when multiple experimental configurations share highly consistent bias directions, it is identified that the local structure is reliable.

b) Gate-B Structural Stability: Structural stability at each grid cell is quantified by a stability weight $\mathrm{ST}(N, P, t)$, which is derived from the dispersion of long-term bias measured by $\mathrm{IQR}_{NP}(N, P)$. Larger values of ST indicate higher structural stability. The $\mathrm{clip}(\cdot)$ operator constrains the resulting weight to the range $[0, 1]$:

$$\mathrm{ST}(N, P, t) = \mathrm{clip}\left(\frac{|\delta_{NP}(N, P)|}{\mathrm{IQR}_{NP}(N, P) + \epsilon}, 0, 1 \right) \tag{14}$$

c) Gate-C Hotspot Trigger: Hotspot triggering is based on identifying a set of hotspot candidates $\mathcal{H}$, defined as grid cells that satisfy both statistical and practical significance:

$$\mathcal{H} = \Big\{ (N, P, t) : \mathrm{SC} \geq \tau_{\mathrm{ss}},\ p_{\mathrm{SC}} < \alpha,\ |\delta_{NP}| \geq \tau_{\mathrm{local}} \Big\} \tag{15}$$

where τ_{ss} denotes the sign-consistency threshold, α the significance level, and τ_{local} the local magnitude threshold. Additionally, only candidates forming connected components in the $N \times P$ neighborhood are retained to avoid isolated triggers. Final hotspot membership is encoded by a binary indicator $\mathrm{HT}(N, P, t) \in 0, 1$. The final gating strength is:

$$\lambda(N, P, t) = \mathrm{HT} \cdot \mathrm{clip}\left(\frac{\mathrm{SC} - \tau_{\mathrm{ss}}}{1 - \tau_{\mathrm{ss}}}, 0, 1 \right) \cdot \mathrm{ST} \tag{16}$$

C. Mechanistic-Driven Transfer Learning

In this section, the debiased simulation data S_1^* is used to train a neural network to extract robust mechanistic priors. The model is then fine-tuned end-to-end on real data with a small learning rate, providing point predictions and quantitative confidence intervals simultaneously.

1) Data Definition:

a) Domains: real target domain R and debiased simulation domain S_1^*.

b) Input variables: management inputs (N,P) and meteorological variables during the crop growing season (total precipitation, total radiation, and growing degree days).

c) Data split: S_1^* is split into training and validation sets with an 8:2 ratio, using year-stratified sampling to preserve consistency of N/P gradients and annual meteorological distributions. Similarly, R is split by year into training, validation, and test sets with a 3:1:1 ratio.

2) Transfer Model Architecture: The learner is a feed-forward regressor $f_\theta : \mathbb{R}^5 \to \mathbb{R}$, which captures the nonlinear environment–crop responses through a two-layer representation:

a) Representation layers: sequential mapping ($5 \to 256 \to 128$) with ReLU activations.

b) Regularization: Dropout and Layer Normalization are inserted into the representation layers to enhance training stability under small-sample conditions.

c) Regression head: maps the 128-dimensional latent representation to a scalar yield prediction.

d) Objective and optimization: Huber loss is adopted to improve robustness to outliers, and the Adam optimizer is used together with learning-rate scheduling and early stopping to constrain training.

This moderately complex architecture has sufficient fitting capacity for the highly covered simulation scenarios provided by S_1^*, while maintaining stability during fine-tuning on the target domain R through multiple regularization mechanisms, thereby avoiding degradation of generalization performance caused by excessive model complexity.

3) Probabilistic Prediction and Confidence Intervals: To quantify predictive uncertainty, Monte Carlo (MC) Dropout is employed for distributional estimation, and Bootstrap resampling is further incorporated to construct more conservative and reliable confidence intervals.

a) MC Dropout: Dropout remains active during inference. For each sample, approximately 500 stochastic forward passes are performed, yielding a set of predictions. The sample mean $\mu(x)$ serves as the point estimate, and the sample standard deviation $\sigma(x)$ measures the epistemic uncertainty. A nominal 95% interval is constructed as $[\mu(x) \pm 1.96, \sigma(x)]$.

b) Bootstrap Calibration: On the validation set, prediction interval coverage probability (PICP) and mean prediction interval width (MPIW) are evaluated. The minimum inflation factor k is searched such that the calibrated interval $[\mu(x) \pm k, \sigma(x)]$ achieves the target coverage level. This factor is then applied to the test set, yielding conservative and interpretable confidence interval outputs.

III. Experimental Results & Discussion

The experimental dataset is obtained from the National Soil Fertility and Fertilizer Effects Long-Term Monitoring Network [18]. The dataset comprises 15 maize yield records sets spanning a five-year period (1990-1994), along with corresponding meteorological and fertilization data. All data were organized following the crop model DSSAT (v4.8.5) protocol [19]. Using the year as the basic unit, the data were partitioned into 20 independent configurations following a 3:1:1 ratio for training, validation, and testing, supporting subsequent calibration, simulation, and transfer learning experiments. The performance of all experiments is evaluated using Mean Absolute Error (MAE), Mean Relative Error (MRE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2).

A. Convergence of Perfect-Calibration

Perfect-calibration experiments were conducted on each of the 10 independent configurations. For every experiment, both rounds of calibration exhibited stable convergence, and the overall performance is summarized in Table I.

The results show that all error metrics under the perfect calibration based on pseudo-observations are substantially lower than those under the initial calibration based on real observations, while R^2 is markedly higher. This indicates that perfect calibration is considerably easier than real-data calibration and confirms the objective existence of implementation-side errors. These findings provide a statistical foundation for subsequent error decomposition and lower-bound derivation.

B. Zero-Error Test and Validation of MRLB

To verify the reasonableness of the error decomposition and the correctness of the MRLB, zero-error tests are conducted, and the error expectation is calculated. The results are as follows.

All 20 configurations successfully pass the ZET, with weighted RMSE approaching zero (Table I). This demonstrates that the DSSAT model is structurally self-consistent. Combined with the findings in Section III-A, it confirms that the perfect-calibration error ε_2 mainly reflects implementation-side errors.

As shown in Table II (showing only 10 representative results due to space limitations), the expected error of perfect calibration ($\mathbb{E}_2$) is far smaller than that of initial calibration ($\mathbb{E}_1$) in the vast majority of configurations, namely, $\mathbb{E}_2 \ll \mathbb{E}_1$. This aligns with theoretical derivations and validates that MRLB can be regarded as a statistical lower bound of the initial-calibration error. Consequently, calibration uncertainty can be quantified solely using the model's own mechanism, without relying on additional external data.

TABLE I
Overall Performance and ZET Test Results of Two Calibration Processes on 20 Calibration Sets

Calibration Process	ZET	MAE	MRE	RMSE	R^2
First Calibration	pass	766.35	0.1310	987.27	0.7938
Perfect Calibration	pass	256.88	0.0452	363.73	0.9708
Improvement Δ	–	66.48%	65.50%	63.16%	22.30%

C. Bias Patterns

By analyzing the bias regularities of the simulated data generated during the perfect-calibration experiments, it identifies reproducible bias patterns in implementation-side errors. Due to space limitations, the main results are summarized in Table III.

D. Debiasing Effectiveness and Placebo Controls

To reduce simulation errors, effective combinations of debiasing gating thresholds ($\tau_{\mathrm{ss}} = 0.7$, $\alpha = 0.05$, $\tau_{\mathrm{local}} = 0.03$) were determined based on the bias patterns discovered in SectionIII-B. Four debiasing modes and corresponding placebo controls were designed:

1) Four debiasing modes: Yearwise correction only (Yc), structural correction only (Sc), yearwise followed by structural correction (YSc, our method), and structural followed by yearwise correction (SYc).

2) Placebo controls: For the first three modes (excluding the shuffled control group SYs), identity transformation (-id) and parameter shuffling (-sh) are applied to test whether performance gains originate from effective debiasing rather than random effects.

The results in Table IV (Similarly, showing only 10 representative results) show that: (1) Under the identity transformation, MAE changes are negligible, excluding accidental improvements; (2) Parameter shuffling leads to significant error increases, confirming that incorrect parameters degrade performance; (3) Under normal debiasing modes, MAE generally decreases, with YSc achieving the largest and most stable improvement. These findings verify the effectiveness of the debiasing mechanism. Occasional anomalies in the Sc-sh mode suggest that practical applications may require further scenario-specific optimization.

E. Comparison of Transfer Learning Performance

To verify the proposed transfer learning framework, comparisons were conducted against multiple baselines on 20 independent test sets in Table V (Similarly, showing only 10 representative results). The baselines are as follows: R: a neural network trained only on real observations; CM: yield predictions from DSSAT after initial calibration; S1-P / S1-F: pretrained and fine-tuned models using unbiased simulations $S1$; S1*-P / S1*-F (our method): pretrained and fine-tuned models using debiased simulations $S1*$. Overall performance follows the ranking:

$$\text{S1*-F (best)} > \text{S1-F} > \text{S1*-P} \gtrsim \text{S1-P} > \text{CM} > \text{R}.$$

TABLE II
EXPECTED STATISTICS AND IMPROVEMENT COEFFICIENTS FOR TWO CALIBRATION ERRORS

	001	002	003	004	005	006	007	008	009	010
$\mathbb{E}_1$	772.07	595.53	803.73	750.80	756.87	822.00	895.13	700.07	775.87	896.87
$\mathbb{E}_2$/MRLB	378.47	437.13	543.85	259.34	455.44	253.26	508.28	629.17	413.51	280.87

TABLE III
SUMMARY OF BIAS INDICATORS

Indicator	**Description**
Sign Consistency	Cross-group RelBias_mean shows same-sign ratio $\geq 70\%$, indicating clear systematic bias.
Structural Pattern	N-dimensional stability zone: within $N(10, 30)$, stable negative bias $(-3\%, -4.5\%)$, same-sign ratio 75–80%.
Hot Zones	Identified 9 hot zones in N-dimensions and 124 hot zones in (N, P) coupling (threshold: same-sign ratio $\geq 70\%$, $\lvert median\rvert \geq 2\%$).
Annual Effect	Additive shifts in 1990/1991/1993 ($\lvert a\rvert \geq 2\%$·mean yield); 1993 shows highest same-sign ratio ($\approx 90\%$).

TABLE IV
DEBIASING EFFECTIVENESS AND PLACEBO CONTROLS

G	**Yc-id**	**Yc-sh**	**Sc-id**	**Sc-sh**	**YSc-id**	**YSc-sh**	**Yc**	**Sc**	**YSc**	**SYc**
1	741.0 → 741.0	741.0 → 837.6	741.0 → 741.0	741.0 → 747.2	741.0 → 741.0	741.0 → 834.3	741.0 → 417.5	741.0 → 709.4	**741.0** → **397.4**	741.0 → 402.5
2	617.0 → 617.0	617.0 → 796.7	617.0 → 617.0	617.0 → 622.3	617.0 → 617.0	617.0 → 792.9	617.0 → 296.3	617.0 → 587.2	**617.0** → **278.4**	617.0 → 279.1
3	711.1 → 711.1	711.1 → 900.8	711.1 → 711.1	**711.1** → **693.8**	711.1 → 711.1	711.1 → 897.7	711.1 → 394.1	711.1 → 718.7	**711.1** → **358.7**	711.1 → 375.0
4	772.1 → 772.1	772.1 → 959.4	772.1 → 772.1	772.1 → 775.8	772.1 → 772.1	772.1 → 969.7	772.1 → 740.2	772.1 → 772.4	**772.1** → **718.7**	772.1 → 733.9
5	654.3 → 654.3	654.3 → 962.3	654.3 → 654.3	654.3 → 660.1	654.3 → 654.3	654.3 → 943.0	654.3 → 514.5	654.3 → 626.1	**654.3** → **496.3**	654.3 → 500.8
6	967.1 → 967.1	967.1 → 1175.6	967.1 → 967.1	**967.1** → **957.6**	967.1 → 967.1	967.1 → 1172.9	967.1 → 620.5	967.1 → 963.4	**967.1** → **573.6**	967.1 → 601.2
7	964.9 → 964.9	964.9 → 1031.7	964.9 → 964.9	964.9 → 970.3	964.9 → 964.9	964.9 → 1050.3	**964.9** → **526.7**	964.9 → 933.9	964.9 → 538.1	964.9 → 528.9
8	825.5 → 825.5	825.5 → 1100.1	825.5 → 825.5	825.5 → 826.3	825.5 → 825.5	825.5 → 1118.4	825.5 → 690.9	825.5 → 828.7	**825.5** → **642.1**	825.5 → 670.7
9	677.1 → 677.1	677.1 → 875.7	677.1 → 677.1	677.1 → 681.2	677.1 → 677.1	677.1 → 871.6	677.1 → 429.9	677.1 → 645.6	**677.1** → **385.2**	677.1 → 406.9
10	804.6 → 804.6	804.6 → 1027.1	804.6 → 804.6	804.6 → 809.6	804.6 → 804.6	804.6 → 1007.8	804.6 → 505.9	804.6 → 774.9	**804.6** → **461.9**	804.6 → 480.0

TABLE V
TRANSFER PERFORMANCE COMPARISON

	MAE						**MRE**						**RMSE**						$\mathbf{R^2}$					
G	R	CM	S1-P	S1-F	S1*-P	S1*-F	R	CM	S1-P	S1-F	S1*-P	S1*-F	R	CM	S1-P	S1-F	S1*-P	S1*-F	R	CM	S1-P	S1-F	S1*-P	S1*-F
1	2385.1	970.3	1043.7	737.8	1038.6	**643.4**	49.2	18.1	19.0	11.9	18.3	**10.0**	3098.4	1056.9	1151.6	746.7	1117.5	**653.2**	-1.49	0.71	0.66	0.86	0.68	**0.89**
2	2987.2	827.0	815.9	572.9	659.2	**521.1**	62.8	17.1	16.2	11.3	13.4	**10.3**	3306.0	895.9	880.6	590.7	746.1	**544.3**	-5.51	0.52	0.54	0.79	0.67	**0.82**
3	1460.6	920.3	952.1	833.1	828.9	**528.1**	24.1	15.2	15.2	13.4	13.0	**7.3**	1815.8	1163.1	1201.7	1011.2	1051.9	**598.9**	-0.39	0.43	0.39	0.57	0.53	**0.85**
4	2971.8	921.3	1008.3	785.2	965.8	**576.0**	62.5	18.6	18.9	15.2	18.1	**10.9**	3294.4	965.7	1039.3	794.7	982.9	**587.8**	-5.46	0.44	0.36	0.62	0.42	**0.79**
5	2172.9	699.0	657.3	427.5	675.8	**349.9**	34.5	14.5	13.0	8.2	13.2	**6.4**	2177.8	894.3	844.7	481.9	855.1	**382.6**	-0.23	0.79	0.81	0.94	0.81	**0.96**
6	2629.5	1160.0	1148.7	1125.0	1161.3	**1098.0**	38.3	19.3	18.9	18.3	19.2	**17.1**	2803.0	1472.9	1437.0	1433.0	1480.8	**1328.6**	-0.02	0.72	0.73	0.73	0.72	**0.77**
7	2391.7	830.0	819.9	508.8	734.0	**231.9**	51.5	18.2	16.9	9.7	15.2	**4.5**	2761.8	992.0	986.9	527.1	875.2	**251.2**	-3.54	0.41	0.42	0.83	0.54	**0.96**
8	1269.5	1188.7	1261.8	908.9	1034.3	**766.4**	23.6	17.6	18.4	12.8	14.6	**10.8**	2142.5	1230.8	1306.4	924.5	1052.1	**781.9**	-0.94	0.36	0.28	0.64	0.53	**0.74**
9	2972.1	661.7	637.9	326.8	479.2	**225.1**	62.0	15.1	14.1	6.2	11.3	**4.6**	3251.5	863.3	824.8	329.6	715.4	**239.9**	-5.29	0.56	0.60	0.94	0.70	**0.97**
10	1494.6	1086.3	1029.9	838.1	982.8	**775.6**	25.5	19.2	17.7	16.0	16.8	**15.1**	1827.4	1425.3	1298.3	1303.5	**1254.2**	1257.1	0.01	0.40	0.50	0.49	**0.53**	**0.53**

These results indicate that fine-tuning (-F) generally outperforms pretraining only (-P), demonstrating the benefit of target-domain samples. Models trained on debiased source data $S1*$ consistently outperform those trained on unde-biased data $S1$, confirming the effectiveness of the 'debias first, then transfer' strategy. Negative transfer in certain test years (e.g.,

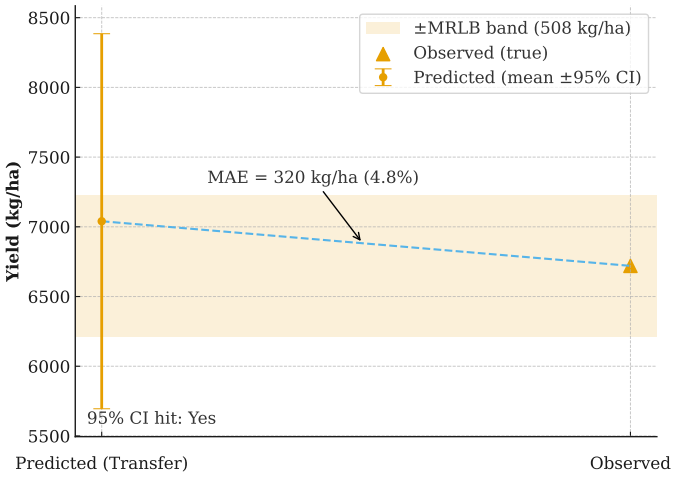

Fig. 2. Yield prediction obtained by the proposed transfer learning method.

1994) is likely due to meteorological conditions falling outside the training distribution, which degrades baseline model performance and limits transfer effectiveness.

F. Case Study

A fertilizer–yield prediction for the year 1992 is shown in Fig. 2. With nitrogen input of 165 kg/ha and phosphorus input of 82 kg/ha, the proposed transfer model predicts a yield of 7040 kg/ha, with a 95% confidence interval of [5695, 8385] kg/ha. The observed yield is 6720 kg/ha, and the MAE (320 kg/ha) is below 5% and lies within the credible interval, which is practically acceptable.

In this case study, the MRLB (508 kg/ha) quantifies the uncertainty of the source-domain data. The fact that the prediction MAE is lower than the MRLB indicates that the transfer process effectively reduces source-domain error, highlighting the value of debiasing and transfer learning. MRLB provides a critical reference for judging reliability: an MAE far below MRLB may suggest over-optimistic evaluation, whereas a large MRLB implies poor source-data quality and high uncertainty. By combining MRLB with interval prediction, the proposed method offers end-to-end risk-aware support for yield prediction and decision-making under data-scarce conditions.

IV. Conclusion

This study constructs a mechanistic-driven transfer learning framework to address crop yield prediction under extremely data-scarce conditions. We introduce the concepts of perfect-calibration experiments and a mechanistic reference lower bound (MRLB) to enable effective bias diagnosis and quantification of risk-related uncertainty. Then, a two-stage gated debiasing strategy is developed to optimize source-domain simulated data, thereby improving the quality of transfer. The debiased mechanistic priors are subsequently integrated with real-world observations through transfer learning, enabling the simultaneous generation of point predictions, credible intervals, and theoretical risk lower bounds. Results show that the proposed method improves prediction accuracy while characterizing uncertainty, providing an reliable decision-support tool for precision fertilization in data-limited settings. The framework is also generalizable to other mechanistic-driven modeling tasks constrained by data scarcity.

References

[1] S. Getahun, H. Kefale, and Y. Gelaye, "Application of precision agriculture technologies for sustainable crop production and environmental sustainability: A systematic review," The Scientific World Journal, vol. 2024, no. 1, Art. no. 2126734, 2024.

[2] A. Gobezie, D. Ademe, and L. K. Sharma, "CERES-Maize (DSSAT) model applications for maize nutrient management across agroecological zones: A systematic review," Plants, vol. 14, no. 5, Art. no. 661, 2025.

[3] B. A. Keating, "APSIM's origins and the forces shaping its first 30 years of evolution: A review and reflections," Agronomy for Sustainable Development, vol. 44, no. 3, Art. no. 24, 2024.

[4] S. A. Bhat and N.-F. Huang, "Big data and AI revolution in precision agriculture: Survey and challenges," IEEE Access, vol. 9, pp. 110209–110222, 2021.

[5] S. J. Seidel, T. Palosuo, P. Thorburn, and D. Wallach, "Towards improved calibration of crop models—Where are we now and where should we go?," European Journal of Agronomy, vol. 94, pp. 25–35, 2018.

[6] C. Yang, N. Lei, C. Menz, A. Ceglar, J. A. Torres-Matallana, and S. Li, et al., "Regional uncertainty analysis between crop phenology model structures and optimal parameters," Agricultural and Forest Meteorology, vol. 355, Art. no. 110137, 2024.

[7] B. M. Sanderson, A. G. Pendergrass, C. D. Koven, F. Brient, B. B. B. Booth, and R. A. Fisher, et al., "The potential for structural errors in emergent constraints," Earth System Dynamics, vol. 12, no. 3, pp. 899–918, 2021.

[8] X. Wei, "The performance of CERES-Wheat model in wheat planting areas and its uncertainties," J. Appl. Meteorol. Sci., vol. 20, no. 1, pp. 88–94, 2009.

[9] D. Wallach, S. Buis, D.-M. Seserman, T. Palosuo, P. J. Thorburn, and H. Mielenz, et al., "A calibration protocol for soil-crop models," Environmental Modelling & Software, vol. 180, Art. no. 106147, 2024.

[10] X. Huang, G. Huang, C. Yu, S. Ni, and L. Yu, "A multiple crop model ensemble for improving broad-scale yield prediction using Bayesian model averaging," Field Crops Research, vol. 211, pp. 114–124, 2017.

[11] T. Xu, A. J. Valocchi, M. Ye, and F. Liang, "Quantifying model structural error: Efficient Bayesian calibration of a regional groundwater flow model using surrogates and a data-driven error model," Water Resources Research, vol. 53, no. 5, pp. 4084–4105, 2017.

[12] S. Condran, M. Bewong, M. Z. Islam, L. Maphosa, and L. Zheng, "Machine learning in precision agriculture: A survey on trends, applications and evaluations over two decades," IEEE Access, vol. 10, pp. 73786–73803, 2022.

[13] Y. Chang, J. Latham, M. Licht, and L. Wang, "A data-driven crop model for maize yield prediction," Communications Biology, vol. 6, no. 1, Art. no. 439, 2023.

[14] B. Saravi, A. P. Nejadhashemi, and B. Tang, "Quantitative model of irrigation effect on maize yield by deep neural network," Neural Computing and Applications, vol. 32, no. 14, pp. 10679–10692, 2020.

[15] R. Tao, "Optimizing crop management with reinforcement learning, imitation learning, and crop simulations," Ph.D. dissertation, Univ. of Illinois at Urbana-Champaign, Urbana, IL, USA, 2022.

[16] Z. Li, Z. Nie, and G. Li, "Integrating crop modeling and machine learning for the improved prediction of dryland wheat yield," Agronomy, vol. 14, no. 4, Art. no. 777, 2024.

[17] Y. Nie and Z. Jiang, "Projection and uncertainty of eastern China summer precipitation based on the perfect model test framework," Acta Meteorologica Sinica, 2025, doi:10.11676/qxxb2025.20250019.

[18] National Ecosystem Science Data Center, "Crop harvesting period monitoring data of national soil fertility and fertilizer benefit monitoring station network (1990–2000)," Data set, 2020, doi:10.12199/nesdc.ecodb.mon.2020.dp2011.fla.003.

[19] G. Hoogenboom, C. H. Porter, V. Shelia, K. J. Boote, U. Singh, and W. Pavan, et al., "Decision Support System for Agrotechnology Transfer (DSSAT) Version 4.8.5," DSSAT Foundation, Gainesville, FL, USA, 2024. [Online]. Available: https://www.dssat.net