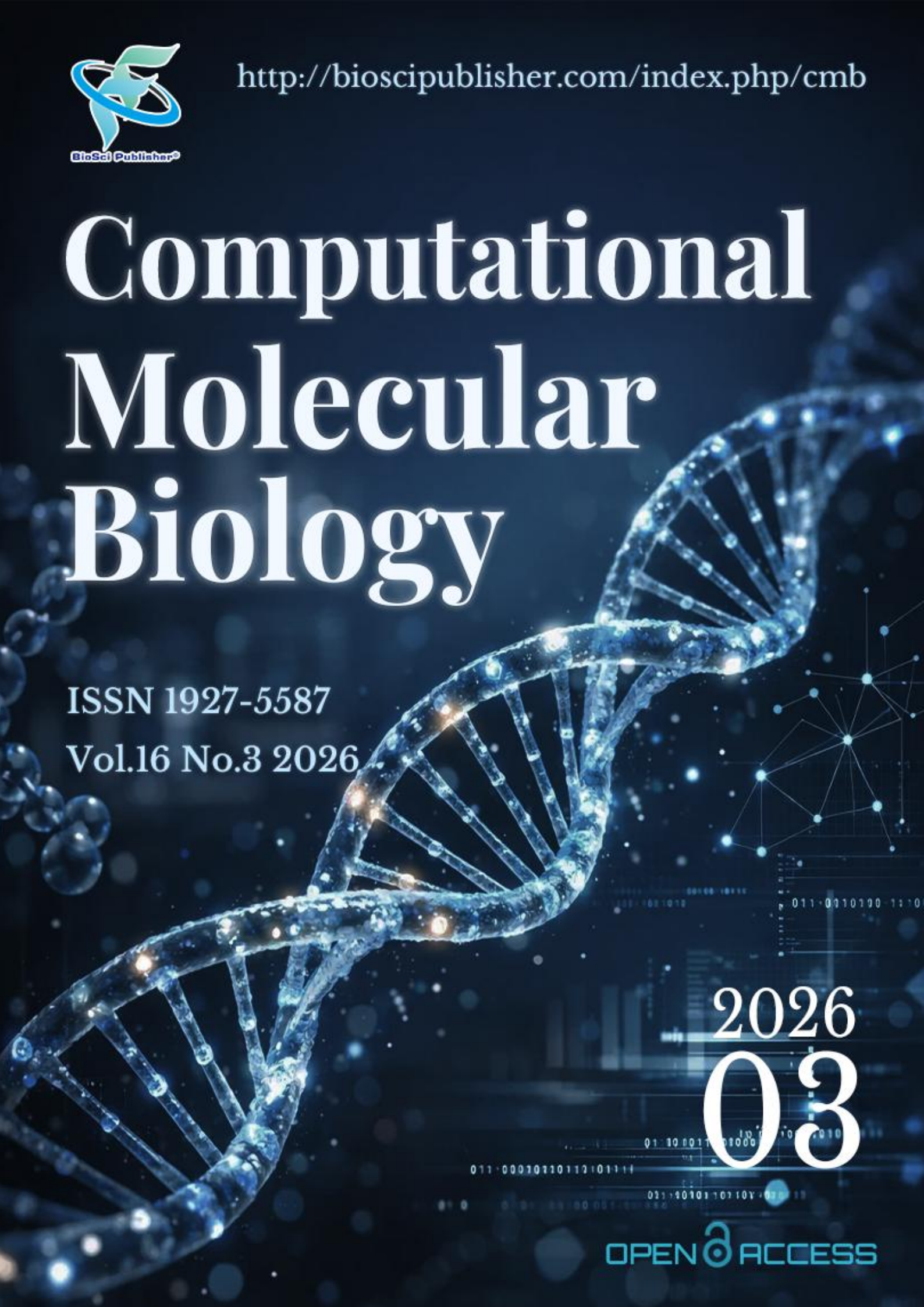
http://bioscipublisher.com/index.php/cmb

Computational Molecular Biology

ISSN 1927-5587
Vol.16 No.3 2026



2026
03

OPEN ACCESS

Publisher
BioSci Publisher

Edited by
Editorial Team of Computational Molecular Biology
Email: edit@cmb.bioscipublisher.com
Website: http://bioscipublisher.com/index.php/cmb

Address:
11388 Stevenston Hwy,
PO Box 96016,
Richmond, V7A 5J5, British Columbia
Canada

Computational Molecular Biology (ISSN 1927-5587) is an open access, peer reviewed journal published online by BioSci Publisher.

The Journal is publishing all the latest and outstanding research articles, letters, methods, and reviews in all areas of computational molecular biology, covering new discoveries in molecular biology, from genes to genomes, using statistical, mathematical, and computational methods as well as new development of computational methods and databases in molecular and genome biology. The papers published in the journal are expected to be of interests to computational scientists, biologists and teachers/students/researchers engaged in biology.

BioSci Publisher is an international Open Access publishing platform that publishes scientific journals in the field of bioscience registered at the publishing platform that is operated by Sophia Publishing Group (SPG), founded in British Columbia of Canada.

Open Access

All the articles published in Computational Molecular Biology are Open Access, and are distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

BioSci Publisher uses CrossCheck service to identify academic plagiarism through the world's leading plagiarism prevention tool, iParadigms, and to protect the original authors' copyrights.

© 2026 BioSci Publisher, registered at the publishing platform that is operated by Sophia Publishing Group, founded in British Columbia of Canada. All Rights Reserved.

Computational Molecular Biology (online), 2026, Vol. 16, No.3
ISSN 1927-5587
http://hortherbpublisher.com/index.php/cmb

Latest Content

© 2026 BioSci Publisher, registered at the publishing platform that is operated by Sophia Publishing Group, founded in British Columbia of Canada. All Rights Reserved.

Computational Molecular Biology 2026, Vol.16, No.3, 146-158
http://bioscipublisher.com/index.php/cmb

Review Article Open Access

Prediction of Eggplant Yield Based on Fertilization and Climate Variables

Guifang Li
1 Jiande Qingrun Modern Agriculture Development Co., Ltd., Jiande 311600, Zhejiang, China
2 Zhejiang Agronomist College, Hangzhou 310021, Zhejiang, China
Corresponding author: 18179387545@163.com
Computational Molecular Biology, 2026, Vol.16, No.3 doi: 10.5376/cmb.2026.16.0011
Received: 24 Mar., 2026
Accepted: 28 Apr., 2026
Published: 12 May, 2026
Copyright © 2026 Li, This is an open access article published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Preferred citation for this article:
Li G.F., 2026, Prediction of eggplant yield based on fertilization and climate variables, Computational Molecular Biology, 16(3): 146-158 (doi: 10.5376/cmb.2026.16.0011)

Abstract With the intensification of climate change and the continuous transformation of agricultural production methods, the extent to which eggplant yields are jointly influenced by fertilizer management and climatic conditions has become increasingly evident. Focusing on fertilization factors and climatic variables as the core subjects of inquiry, this study systematically analyzes the mechanisms by which temperature, precipitation, humidity, and fertilizer inputs affect eggplant yield formation, while also exploring the interactive effects between climate and fertilization. To this end, regional meteorological data, soil nutrient data, and field yield data were collected to construct an eggplant yield prediction model based on a combination of statistical analysis and machine learning techniques. The research focuses on variable selection, feature engineering, model training, and the optimization of predictive performance, while also comparing the differences in predictive accuracy and stability between regression models and machine learning algorithms. The results indicate that temperature fluctuations, soil moisture conditions, and nitrogen fertilizer inputs are critical factors influencing eggplant yields, and that the coupled effects of these multiple factors can significantly enhance the accuracy of the prediction model. A case study further validates the model's applicability within regional agricultural production contexts, providing a scientific basis for precision fertilization management, agricultural risk assessment, and smart farming decision-making. This study holds significant theoretical and practical implications for improving eggplant production efficiency, optimizing resource utilization, and fostering sustainable agricultural development.
Keywords Eggplant yield prediction; Fertilization management; Climate variables; Machine learning; Precision agriculture

1 Introduction

Eggplant (*Solanum melongena* L.) is a widely cultivated vegetable valued for its nutritional quality, including minerals, vitamins, and antioxidant phenolics that contribute to human health and dietary diversity (Başay et al., 2025). It also plays an important economic role, providing income for smallholders and contributing substantially to vegetable production in many countries, yet yields in several regions remain below global averages (Oladosu et al., 2021; Dollison and Tapas, 2024). At the same time, agriculture faces mounting pressure from climate change, with shifts in temperature and rainfall patterns already constraining productivity and threatening food and nutrition security, especially in vulnerable regions (Chioti et al., 2022; Kuradusenge et al., 2023). In this context, improving the stability and predictability of eggplant yield under varying fertilization regimes and climate conditions is critical for both farmers' livelihoods and broader food system resilience.

Fertilization management is a central lever for enhancing eggplant productivity, fruit quality, and nutritional value. Numerous studies show that optimizing macro- and micronutrient supply-through mineral NPK fertilizers, organic amendments, and foliar micronutrients-can significantly increase growth, yield components, and nutrient content of eggplant fruits and seeds (Bana et al., 2022). Integrated nutrient management approaches, combining chemical fertilizers with biofertilizers and micronutrients, have further improved yield and quality, and have been successfully modeled using data-driven techniques such as artificial neural networks to identify key nutritional predictors of yield and protein content (Thingujam et al., 2020). However, many fertilization recommendations are still static, and rarely account for interactions with variable weather, despite the fact that fertilization efficiency and crop response can be strongly modulated by temperature and moisture regimes (Gad, 2023; Chandio et al., 2025).

Climate variability and change are increasingly recognized as major drivers of year-to-year yield fluctuations across a wide range of crops. Analyses of long-term data link changes in temperature, rainfall, and the length of the rainy season to substantial variations in yields, with higher temperatures and drought often reducing productivity, while adequate or increased rainfall can partially offset these negative effects (Chioti et al., 2022). At the same time, recent work has demonstrated that combining environmental variables (such as temperature, precipitation, and evaporation) with fertilizer use data in predictive models can greatly improve crop yield forecasting performance, supporting more informed agronomic decisions (Burdett and Wellen, 2022; Krishnadoss and Ramasamy, 2024). Despite this progress, there is a notable gap regarding eggplant-specific yield prediction frameworks that jointly consider fertilization practices and climate variables, even though eggplant is sensitive to both soil fertility and temperature stress, including low-temperature constraints in certain seasons (Osman et al., 2021; Badshah et al., 2024).

This study addresses these gaps by developing a predictive framework for eggplant yield based on fertilization and climate variables, with the goal of supporting climate-smart nutrient management. Building on evidence that data-driven and machine learning models (such as random forests, ensemble approaches, and neural networks) can accurately capture complex, nonlinear relationships among weather, input use, and yields in other crops and regions, this work tailors such concepts to eggplant systems. The specific objectives are to quantify the combined and individual effects of fertilization regimes and key climate factors (e.g., temperature and rainfall) on eggplant yield, construct and evaluate predictive models that use these variables to estimate yield, and identify the most influential features governing yield variability to inform practical management guidelines. By integrating fertilization and climate information into a unified predictive approach, the study aims to contribute a scalable tool and empirical insights that can enhance fertilizer recommendations, reduce climate-related yield risks, and ultimately support more sustainable and resilient eggplant production.

2 Influence of Climate Variables on Eggplant Production

2.1 Effects of temperature variability on yield formation

Open-field work using growing degree days (GDD) shows that eggplant accessions requiring fewer accumulated heat units to first fruiting achieve higher productivity; in a Caribbean environment without temperature extremes (<15 °C or >35 °C), yields above 80 t ha^{-1} were obtained, indicating that thermally suitable sites allow full yield potential expression (Pacheco et al., 2019). Greenhouse studies reveal curvilinear temperature responses of fruit number and total yield, with lower yields when temperatures deviate from an optimum that depends on light intensity, reinforcing the non-linear nature of temperature-yield relationships.

Physiological research indicates that temperatures below about 17 °C slow growth, and near 10 °C induce metabolic disturbances, impairing membrane stability, water relations, chloroplast development, and photosynthetic efficiency, all of which ultimately reduce fruit set and yield (Shimira and Taşkın, 2022). Conversely, excessive heat accelerates development and can depress fruit set in vegetable crops, shortening the period for photoassimilate accumulation and causing yield loss, so yield prediction must account for both cold and heat stress windows around the crop's optimal growth range.

2.2 Impact of rainfall and soil moisture on crop productivity

A multi-year field trial in a moderate climate showed that eggplant yield depended strongly on both air temperature and total rainfall, with the highest yields obtained when high mean temperatures coincided with evenly distributed rainfall; periods of very low or absent rainfall shortened the harvest period and delayed first fruiting. In the Colombian Caribbean, rainfall largely met crop evapotranspiration, supplemented by irrigation to maintain soil at field capacity, and under these favorable moisture conditions no critical drought episodes occurred, supporting high yields across genotypes.

Deficit irrigation studies using field capacity (FC) as a benchmark demonstrate that, under subsurface infiltration irrigation, reducing soil moisture from 80% to 60% FC during early and mid stages can be tolerated with limited yield reduction, but deficits during the prime fruit stage markedly decrease yield and plant growth traits (Li et al., 2024). Complementary deficit drip irrigation work on sandy clay loam soils found maximum yield and irrigation

water use efficiency at about 75% FC, with both lower and higher soil moisture leading to reduced productivity, indicating an optimum soil moisture band for yield formation (Ouma et al., 2024).

2.3 Relationship between humidity conditions and plant health

Eggplant health is strongly influenced by humidity through its effects on fungal and bacterial disease development. In humid subtropical environments, high relative humidity and moderate temperatures were associated with substantial incidences of Phomopsis fruit rot and Cercospora leaf spot, with fruit rot increasing roughly tenfold over 30 days under mean temperatures around 23.7 °C and 55.5% relative humidity, and leaf spot rising fivefold when average temperature was 18.2 °C with morning humidity near 88%. Broader reviews of eggplant fungal diseases emphasize that environmental factors-particularly moisture and temperature-interact with host genetics to drive pathogenesis and yield loss, underscoring humidity as a key variable in risk-based yield prediction (Kaniyassery et al., 2022).

For Alternaria leaf spot, field monitoring across sowing dates showed disease intensity to be positively and significantly correlated with both maximum and minimum temperatures, but negatively correlated with morning and noon relative humidity; rainfall also showed a negative (though non-significant) association with disease intensity (Sharma et al., 2025). Other pathosystems, such as Verticillium wilt under greenhouse conditions, demonstrate that disease severity significantly reduces early and total yield and plant biomass, while irrigation frequency (and thus soil moisture regime) also affects plant performance, indicating that combined humidity, soil moisture, and pathogen pressure must be integrated into plant health and yield models.

3 Interaction Between Fertilization and Climate Factors in Yield Formation

3.1 Coupling effects of water and fertilizer management

Water-fertilizer coupling directly shapes crop growth environments by synchronizing soil moisture and nutrient availability, thereby affecting yield formation, resource use efficiency, and environmental impacts (Xing et al., 2024). In eggplant systems under mulched drip irrigation, factorial combinations of irrigation levels and nitrogen rates show that both water, nitrogen, and their interaction significantly alter evapotranspiration, yield, and water productivity, with mild water deficit plus moderate nitrogen achieving the highest yield and water productivity (Zhou et al., 2023). Similar coupling principles have been generalized across crops, where appropriate water-fertilizer ratios enhance soil physical structure, microbial activity, and nutrient mineralization, thus improving crop performance while reducing fertilizer loss and environmental pressure.

Studies in cold and arid oasis environments further indicate that eggplant yield, fruit quality, and water- and nitrogen-use efficiency are jointly governed by irrigation-nitrogen interactions, with mild water deficit (60%-70% field capacity) and moderate nitrogen rates outperforming both lower and higher inputs (Li et al., 2025). These results align with broader reviews of water-fertilizer coupling, which report that optimized coupling improves soil structural stability, microbial diversity, and enzyme activity, and that intelligent drip fertigation systems can enhance water use efficiency while lowering nutrient leakage and pollution risks (Xing et al., 2024). Together, this evidence highlights water-fertilizer coupling as a key mechanism through which management and climate-modulated water supply co-determine yield.

3.2 Climate-dependent fertilizer efficiency

Fertilizer efficiency is strongly modulated by climatic conditions, particularly temperature and rainfall regimes that influence nitrogen uptake pathways, losses, and crop demand. Long-term simulations for wheat-maize rotations under future climate scenarios show that, even with unchanged cultivars, warming and altered rainfall patterns reduce annual nitrogen use efficiency by about 15%, with manure-amended systems partly buffering these negative impacts by sustaining soil organic matter and nutrient supply. In rice systems, meta-analysis across climatic gradients finds that mean seasonal temperature and precipitation, along with fertilizer N rate and soil properties, jointly explain regional differences in agronomic efficiency, N recovery, and reactive nitrogen losses, underscoring that identical fertilizer rates can perform very differently under contrasting climates (Cai et al., 2022).

Experimental warming studies using 15N tracers confirm that modest temperature increases can lower fertilizer nitrogen recovery and increase nitrogen losses even when grain yield remains unchanged, indicating a hidden decline in fertilizer efficiency under warming. At a broader scale, analyses of nitrogen fertilizer use and climate interactions for maize reveal that higher temperatures and extreme heat days can diminish the yield benefits of nitrogen, while favorable growing-degree days and adequate precipitation enhance the marginal return to N, with optimal nitrogen rates shifting across climate gradients (Huang et al., 2024). These findings demonstrate that fertilizer recommendations and efficiency metrics cannot be treated as static, but must be adjusted to local and evolving climate conditions.

3.3 Synergistic effects of multi-factor agricultural inputs

Yield responses to fertilization rarely depend on nutrients alone; instead, they emerge from combined effects of climate, soil, and multiple input levels. Meta-analysis of maize fertilization across Northeast China shows that moderate NPK rates increase yield by about 20% and improve protein and fat content, but the magnitude of yield and quality gains depends on precipitation, temperature, soil pH, and soil nutrient status, with soil organic matter and available phosphorus identified as dominant drivers of fertilization benefits (Gao et al., 2025). At the process level, a global synthesis of nutrient interactions indicates that most macronutrient combinations act synergistically on yield when both are deficient, whereas certain divalent cation combinations can be antagonistic, implying that multi-nutrient strategies must be designed to exploit synergy while avoiding negative interactions.

Multi-factor management that couples irrigation, nitrogen, and delivery method can further amplify positive interactions. A large meta-analysis across Chinese cropping systems shows that drip fertigation-combining precise water and N supply-raises yield by 12%, water productivity by 26%, and nitrogen use efficiency by 34%, while reducing evapotranspiration compared with traditional irrigation and broadcasting fertilization (Li et al., 2021). Complementary analyses of irrigation-nitrogen combinations in maize and wheat demonstrate that joint application of irrigation and N typically increases yield by 9%-17% relative to controls, though the effect size varies with climate and soil, highlighting the importance of context-specific optimization of multiple inputs (Cui et al., 2024). Such evidence supports modeling approaches that integrate fertilization, water management, and climate variables when predicting yield and designing climate-resilient fertilization regimes.

4 Construction of Eggplant Yield Prediction Models

4.1 Selection of fertilization and climate variables

The selection of input variables is crucial for robust eggplant yield prediction, particularly when combining fertilization and climate information. Systematic reviews of crop-yield ML studies show that temperature, rainfall, soil type, humidity, and fertilizer-related variables are among the most frequently and successfully used features for yield estimation (Jabed and Murad, 2024; Shawon et al., 2024). Other work that jointly models environmental and chemical inputs demonstrates that precipitation, temperature, evaporation, wind speed, and chemical (fertilizer) use together can explain a large share of yield variability, supporting their inclusion in compact yet informative feature sets (Krishnadoss and Ramasamy, 2024).

At the same time, models that explicitly incorporate nutrient levels (e.g., NPK) with climatic variables such as temperature, rainfall, and humidity can generate highly accurate crop recommendations and yield responses, indicating that these variables effectively capture plant-environment-management interactions (Dey et al., 2024). Broader ML applications in agriculture reinforce that features related to soil fertility, water availability, and weather conditions (including meteorological variables and season) are central drivers of crop output and must therefore be prioritized in variable selection for eggplant yield prediction under different fertilization regimes (Figure 1) (Gupta et al., 2022; Sharma et al., 2023).

4.2 Data processing and feature engineering

Accurate prediction requires careful preprocessing to transform raw agronomic and climatic records into machine-learning-ready datasets. Studies on crop yield prediction typically perform data cleaning, normalization, and integration of heterogeneous sources (weather, inputs, yield) as early steps, sometimes engineering new targets such as yield per area from production and land area data to better reflect productivity (Iniyan et al., 2023;

Sarikonda et al., 2025). To avoid bias and overfitting, workflows also emphasize correct partitioning schemes and prevention of information leakage, along with modular feature creation from weather, soil, remote sensing, and crop-model outputs (Paudel et al., 2020; Morales and Villalobos, 2023).

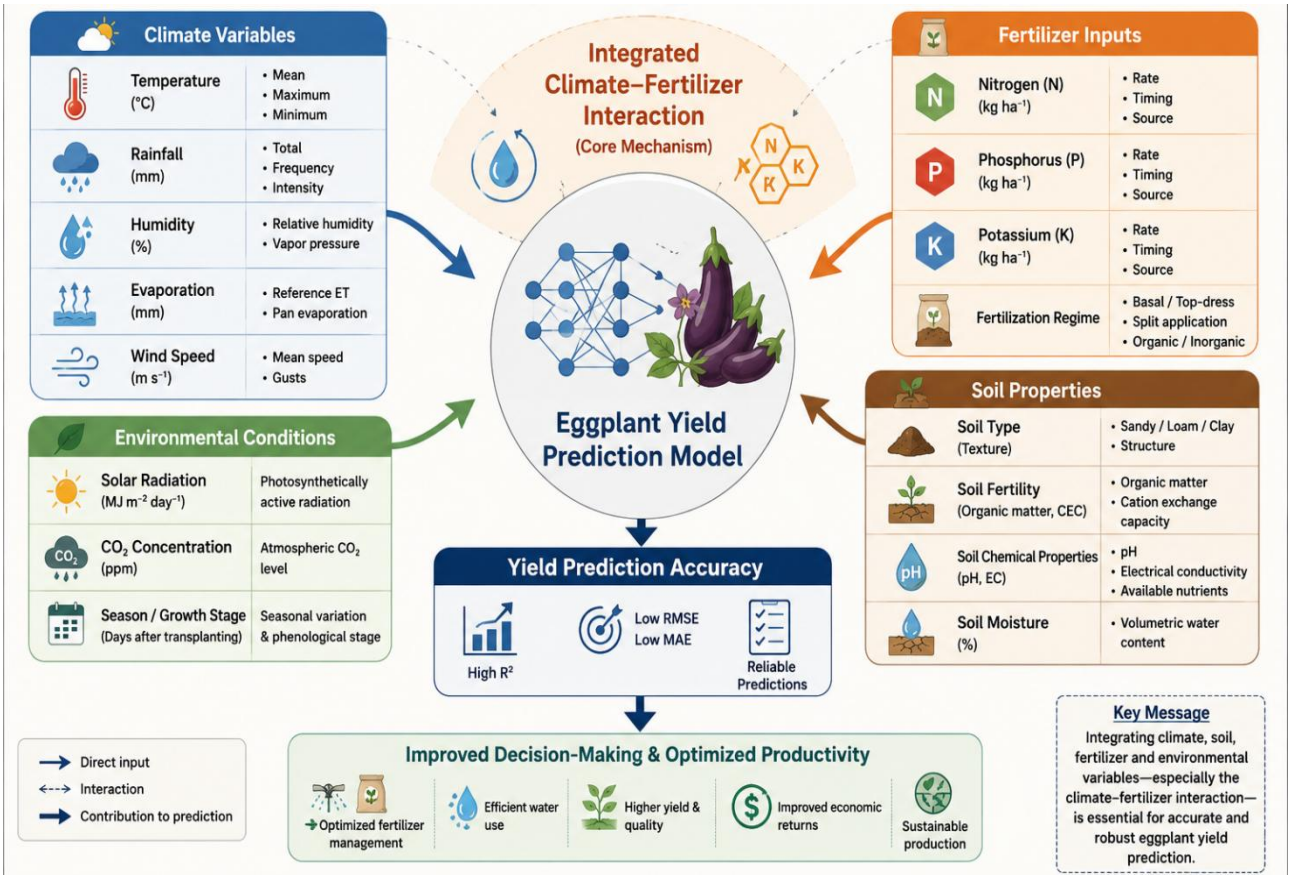


Figure 1 Conceptual framework of key input variables used in machine learning-based eggplant yield prediction models. Climate, soil, fertilizer, and environmental variables jointly influence prediction accuracy and crop productivity responses

Feature selection and extraction are key to reducing redundancy and improving generalization. Relief-based feature selection and linear discriminant analysis have been used to isolate the most discriminative predictors before training support vector machines, k-nearest neighbors, and random forests for yield classification or regression (Gupta et al., 2022). Hybrid approaches combine correlation-based filters, clustering, and recursive feature elimination to build reduced, information-rich datasets that, together with optimized support vector regressors, substantially improve prediction accuracy while lowering computational cost, illustrating the value of systematic feature engineering pipelines (Abdel-Salam et al., 2024).

4.3 Development of statistical and machine learning models

A wide range of statistical and ML algorithms has been applied to crop yield prediction, offering guidance for constructing eggplant-specific models. Linear regression, random forest, gradient boosting trees, and related methods are among the most widely used, with random forest and boosting-based techniques often achieving strong performance across diverse environments and crops (Mahesh and Soundrapandiyan, 2024; Shawon et al., 2024). Ensemble models that integrate multiple learners (e.g., Extra Trees, gradient boosting, or stacked approaches) have repeatedly reached very high R^2 and low error metrics, suggesting that ensemble strategies are promising for capturing complex fertilization-climate-yield relationships (Iniyan et al., 2023; Nossam et al., 2024).

For eggplant specifically, machine learning models using spectral vegetation indices, days after planting, and irrigation-related coefficients have successfully predicted yield; principal component analysis-based inputs combined with artificial neural networks achieved very high accuracy, indicating that nonlinear models can

effectively exploit engineered features (Taşan et al., 2022). Gradient-boosting families (CatBoost, LightGBM, XGBoost) have also shown excellent performance for general crop yield prediction and for eggplant yield based on genotype-related variables, where CatBoost provided accurate and robust forecasts, highlighting the suitability of tree-based boosting for eggplant yield modeling under varying environmental and management conditions (Islam et al., 2023; Mahesh and Soundrapandiyan, 2024).

5 Evaluation and Optimization of Yield Prediction Performance

5.1 Comparison of regression and machine learning algorithms

Crop yield prediction studies consistently show that machine learning algorithms often outperform simple regression when relationships between climate, management, and yield are nonlinear and complex. Comparative evaluations across linear regression, decision trees, random forests, support vector machines, and neural networks report that ensemble methods such as Random Forest and Gradient Boosting generally achieve higher accuracy and better generalization than traditional linear models, especially when diverse environmental and management variables are included (Kurmi and Singh, 2025). However, linear models remain competitive when relationships are close to linear, offering advantages in interpretability and lower computational cost (Nazir et al., 2025).

Broader multi-crop comparisons confirm that advanced tree-based models and k-nearest neighbors often provide lower error and higher correlation with observed yields than multiple linear regression, particularly when many climatic and soil predictors are used. Recent work further extends comparisons to deep learning (e.g., LSTM and Bi-LSTM), showing that optimized recurrent networks can substantially reduce prediction error relative to support vector regression and time-series models such as ARIMA and VAR, demonstrating the value of capturing temporal dependencies in climate and yield series (Kumar et al., 2023).

5.2 Accuracy assessment using evaluation indicators

Evaluation of yield prediction models relies on multiple complementary indicators to capture both error magnitude and explanatory power. Common metrics include root mean squared error (RMSE), mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE), and the coefficient of determination (R^2), which together provide a comprehensive view of prediction bias, dispersion, and goodness-of-fit (Kurmi and Singh, 2025; Nazir et al., 2025). Studies comparing regression and machine learning approaches typically rank models by minimizing RMSE and MAE while maximizing R^2, revealing clear performance hierarchies among algorithms under different data conditions (Pant et al., 2025).

Large-scale forecasting frameworks and ensemble systems also employ normalized RMSE (NRMSE) and additional agreement indices to compare machine-learning baselines against operational forecasting systems or process-based crop models, emphasizing reproducibility and robustness across crops, regions, and seasons (Paudel et al., 2020; Singh et al., 2025). In practice, these metrics are often computed under cross-validation or using independent test years, allowing rigorous assessment of generalization and facilitating fair comparison of alternative algorithms for integrating fertilization and climate variables in yield prediction (Sowmya and Prasad, 2024).

5.3 Optimization of model parameters and prediction stability

Model performance and stability depend strongly on appropriate hyperparameter tuning and feature selection. Grid-search and other systematic optimization methods applied to tree-based ensembles such as Random Forest and Gradient Boosting have been shown to significantly improve RMSE, MAE, and R^2 compared with default configurations, with tuned ensembles delivering more robust rice yield predictions under variable climatic conditions (Hoque et al., 2024; Sowmya and Prasad, 2024). Similarly, combining multiple tuned base learners in stacked or adaptive ensembles can further reduce prediction error relative to any single model, demonstrating the benefits of leveraging diverse algorithmic strengths (Sánchez et al., 2014).

Advanced frameworks integrate hybrid feature selection and metaheuristic optimization to enhance both accuracy and efficiency. For example, coupling clustering and correlation-based filters with feature selection methods, followed by hyperparameter optimization of support vector regression via an improved Crayfish Optimization

Algorithm, yields superior crop yield predictions compared with standard SVR and other regressors (Abdel-Salam et al., 2024). Deep learning approaches also rely on systematic hyperparameter optimization and cross-validation, where careful selection of optimizers and network configurations (e.g., Bi-LSTM with Adam) enhances prediction accuracy and reduces error variability across crops, thereby improving prediction stability over time and across environmental conditions (Kumar et al., 2023).

6 Identification of Key Determinants Affecting Eggplant Yield

6.1 Contribution analysis of fertilizer inputs

Quantifying the contribution of fertilizer inputs to yield is central for identifying leverage points in eggplant production. Pot experiments with graded nitrogen (N) and phosphorus (P_2O_5) rates showed that N applications significantly affected nearly all growth and yield components, whereas P_2O_5 influenced fewer variables; yield gains were mainly driven by fruit number and fruit weight, with optimal responses at 100-150 kg/ha of both N and P_2O_5. A separate fertigation study using factorial N and K rates found that leaf area and agronomic efficiency of N declined at higher N and K levels, indicating diminishing returns and highlighting the importance of moderate N doses and balanced K supply for efficient production.

Longer-term field experiments confirm that not only fertilizer quantity but also source and combination determine yield contributions. In a four-year eggplant trial, applying 100% recommended NPK together with farmyard manure increased fruit yield by 47% compared with mineral fertilizer alone, while also enhancing soil organic carbon and available N, P, and K, and improving agronomic efficiency and nutrient recovery (Nisar et al., 2025). In multi-crop vegetable systems on organic soils, random forest models using soil, management, and meteorological features revealed little response to added P and only null to moderate response to added N in high-P conditions, suggesting that excess P is common and that fertilizer contribution depends strongly on existing soil fertility and N-P stoichiometry (Parent, 2024).

6.2 Sensitivity analysis of climate variables

Sensitivity analyses from global and regional studies provide a framework for evaluating how climate variables modulate eggplant yield response to fertilization. Non-parametric elasticity analysis for major crops showed that yields are most sensitive to mean air temperature, with precipitation exerting a smaller but still relevant influence; the sign and magnitude of temperature elasticity varied by crop and region, with many wheat and rice systems experiencing negative yield responses to warming (Liu et al., 2020). A machine-learning study of climate extremes found that growing-season mean climate and extremes together explained up to 49% of yield anomaly variance, and that temperature-related extremes were generally more influential than precipitation-related indices, although irrigation partly mitigated heat damage.

Variance-based sensitivity analysis applied to a process-based wheat model demonstrated that yield sensitivity shifts between water-controlling factors (precipitation, soil hydraulic properties) and nitrogen-controlling factors depending on which resource is limiting under a given climate-soil-management scenario (Hao et al., 2024). In arid and semi-arid Jordan, combining machine learning with Sobol' indices showed that climate-related variables explained a large fraction of yield variance for sensitive crops like wheat, whereas more resilient crops such as barley exhibited much lower climate-driven variance, underlining the crop- and context-specific nature of climate sensitivity (Xu et al., 2025).

6.3 Identification of dominant yield-limiting factors

Disentangling dominant yield-limiting factors requires integrating fertilizer response with plant nutritional physiology and climate constraints. Nutrient-solution omission experiments in eggplant showed that withholding individual macronutrients reduced vegetative growth, dry matter, and photosynthesis, with nitrogen and calcium identified as the most growth-limiting elements despite potassium being most demanded quantitatively (Flores et al., 2015). Greenhouse studies on N and P_2O_5 rates further indicated that yield was more affected by N than by P, with excessive doses reducing performance, suggesting that sub-optimal N supply or imbalanced N:P ratios can act as primary yield constraints even when total fertilizer input is high

At larger scales and across crops, feature-importance and explainable-AI analyses consistently rank temperature, rainfall, and macronutrient levels among the most influential predictors of yield, revealing strong interactions between climate drivers and NPK supply (Meng et al., 2021; Mohan et al., 2025). In a maize yield prediction framework integrating fertilizer systems with multi-source data, random forest feature importance highlighted fertilizer variables, maximum temperature, and precipitation as key determinants, with different fertilizer systems shifting which factors were most limiting under given climatic conditions (Meng et al., 2021). Together, these results indicate that for eggplant, dominant yield-limiting factors are likely to be inadequate or poorly balanced N (and Ca), interacting with temperature and water availability, rather than single inputs considered in isolation.

7 Case Study on Regional Eggplant Yield Prediction

7.1 Overview of the selected experimental region

The experimental region represents a semi-arid to arid environment where eggplant production is constrained by high evaporative demand, limited and seasonally concentrated rainfall, and strong sensitivity of yield to microclimate modification. In cold and arid oasis conditions, such as the Hexi irrigation area of northwest China, annual precipitation is only about 183-285 mm, evaporation exceeds 1600 mm, and sunshine duration approaches 3000 h, creating a dry atmosphere where irrigation and fertilization strategies are critical to sustain eggplant productivity (Li et al., 2025). Comparable semi-arid vegetable regions, for example Carnarvon in Western Australia, face high temperatures and intense solar radiation during spring-summer, which damage crops and shorten the production season unless protective cultivation is adopted (Nguyen et al., 2022).

Within these environments, protected and controlled systems are increasingly used to create favorable microclimates for eggplant. Shade-net houses in Carnarvon, using moderate shade factors around 21%, altered light intensity and microclimatic conditions in ways that promoted taller, bushier plants and higher fruit yield compared with open-field cultivation (Figure 2) (Nguyen et al., 2022). Similarly, controlled and semicontrolled greenhouse systems in arid regions have shown that adjusting temperature, light, and nutrient sources (inorganic fertilizers, compost, and their mixtures) can strongly influence eggplant growth, yield, and water-use efficiency, providing locally specific data to calibrate yield models for such climates (Abbas et al., 2025).

7.2 Application of the prediction model to field data

Regional yield prediction relies on integrating field experiments that quantify responses to water and fertilizer regimes under real climate variability. In the Hexi oasis, split-plot experiments across two seasons with three irrigation levels (50%-60%, 60%-70%, 70%-80% field capacity) and three nitrogen rates (215, 270, 325 kg/ha) generated detailed yield, quality, and resource-use data, enabling identification of an optimal mild water deficit (60%-70% FC) with moderate nitrogen (270 kg/ha) under mulched drip irrigation (Li et al., 2025). These structured datasets, including soil properties and multi-year climate records, are well suited for training and validating regional prediction models that link fertilization and climate variables to eggplant yield.

Advanced modeling frameworks in other crops illustrate how multi-layered, multi-farm datasets can be used to forecast yield at field and regional scales. In Western Australia, yield monitor data for wheat, barley, and canola over three seasons were combined with weather and soil-related predictors to build random forest models at 100 m resolution, achieving concordance correlation coefficients of 0.89-0.92 and RMSE of 0.36-0.42 t/ha . Applying similar machine learning workflows to eggplant, using experimental and commercial field data from protected and open-field systems, allows spatially explicit yield forecasts that support regional fertilizer and irrigation decisions.

7.3 Implications for precision fertilization and farm management

Results from regional case studies highlight that optimal eggplant yield can be achieved with water- and nitrogen-saving strategies tailored to local climate, providing a basis for precision fertilization. In the cold, arid Hexi region, mild water deficit with moderate nitrogen significantly increased yield, fruit quality, and water- and nitrogen-use efficiency relative to unfertilized, fully irrigated controls, demonstrating that blanket high-input strategies are neither necessary nor efficient (Li et al., 2025). Parallel work in deficit drip irrigation on sandy clay loam soils showed that maintaining soil moisture at 75% field capacity maximized yield (≈39 t/ha) and irrigation water-use efficiency, with further increases in water supply reducing both yield and efficiency (Ouma et al., 2024).

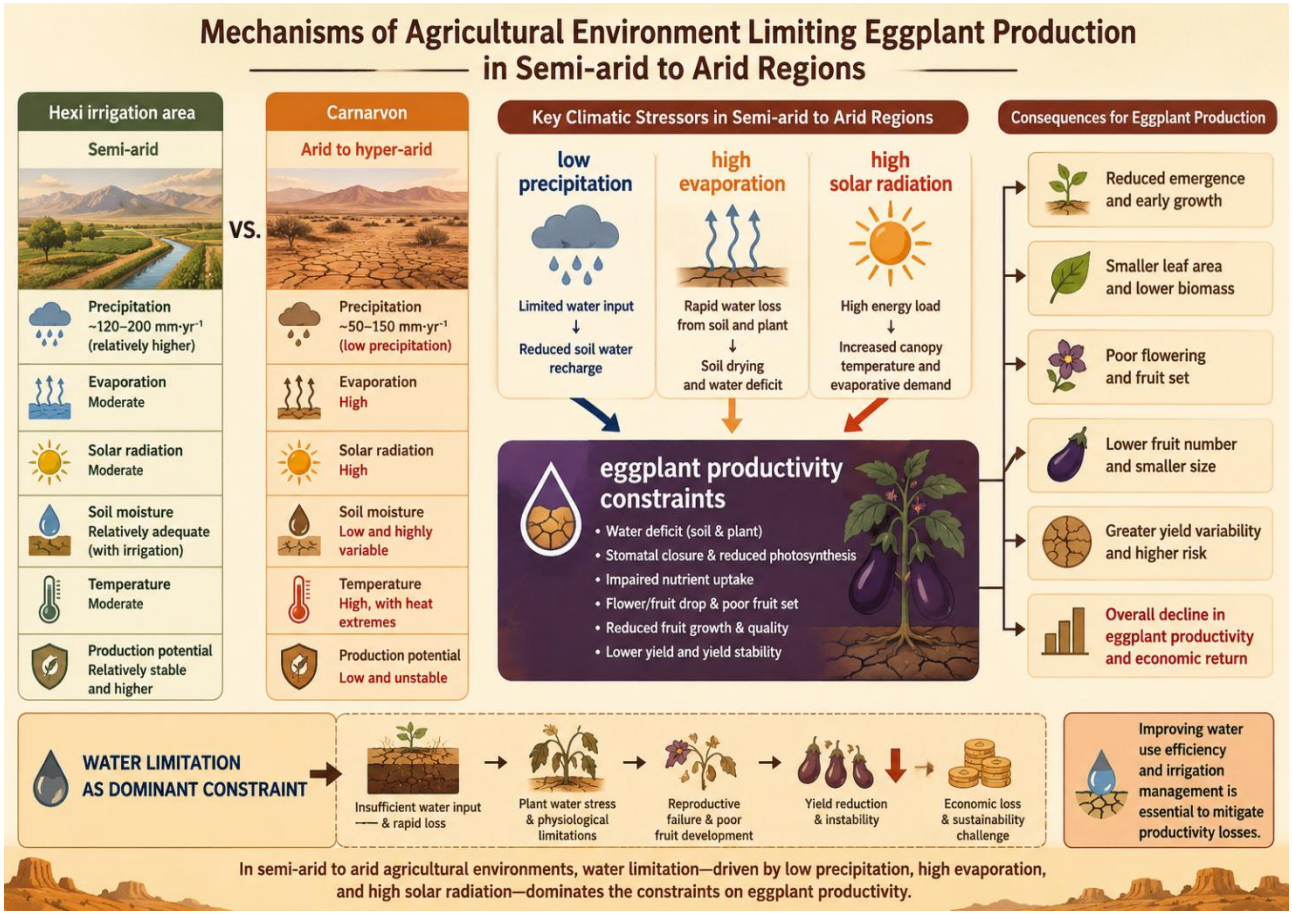


Figure 2 Schematic representation of climatic constraints on eggplant production in semi-arid to arid environments (e.g., Hexi irrigation area and Carnarvon). High evaporative demand, low precipitation, and strong solar radiation jointly limit crop productivity

These findings align with broader advances in precision water-fertilizer management. Reviews of precise water and fertilizer application technologies emphasize that integrating advanced sensors, remote sensing, and machine learning enables variable-rate fertigation and micro-irrigation that improve nutrient uptake, water-use efficiency, and environmental outcomes compared with uniform practices (Xing and Wang, 2024). Decision-support frameworks based on the Internet of Things and optimization models further show that coordinated, long-term irrigation and fertilization planning can simultaneously increase economic returns and environmental benefits compared with empirical management, indicating that regional eggplant yield prediction models can be directly embedded in smart fertigation and farm-planning systems (Lin et al., 2020).

8 Strategies for Sustainable Eggplant Production Under Climate Variability

Sustainable eggplant production under climate variability requires fertilizer strategies that enhance yield while maintaining soil health. A four-year eggplant field study showed that combining the full recommended NPK dose with farmyard manure increased yield by 47% over mineral fertilizer alone and substantially raised soil organic carbon and available N, P, and K, improving agronomic efficiency and nutrient recovery. At the broader vegetable level, a global meta-analysis found that enhanced-efficiency fertilizers (EEFs), such as nitrification inhibitors and polymer-coated urea, increased vegetable yield by about 7.5%-8.1% and improved quality while markedly reducing nitrous oxide emissions and nitrate leaching, especially when matched to soil pH and organic carbon conditions.

Optimizing nitrogen remains central, because excessive N is common in high-value vegetables and is associated with low recovery and high leaching risk. A review of nitrogen management in field vegetables emphasizes that aligning N supply with crop demand, improving synchronization via split applications, sensor-based diagnostics, and better irrigation management can simultaneously maintain yields and reduce nitrate losses below the root zone.

For eggplant in arid oasis conditions, a two-year drip-irrigation trial identified mild water deficit combined with medium N rate as the optimal strategy, significantly increasing yield, fruit quality, and water productivity compared with both higher and lower N and water levels, illustrating how fertilizer optimization must be co-designed with water management under variable climates.

Climate-smart agriculture (CSA) offers a framework to adapt eggplant systems to temperature and rainfall instability while reducing environmental impacts. A recent review highlights precision nutrient management, integrated soil fertility strategies, and regenerative practices (e.g., organic amendments, biochar, agroforestry) as key CSA options that improve soil health, raise nitrogen use efficiency, and increase carbon sequestration, thereby buffering crops against climate stress. Another synthesis of climate-change impacts on agroecosystems stresses that adaptation must address multiple risks-yield decline, water scarcity, pests, and product quality-through measures such as improved water and soil management, agronomic practices, and smart technologies tailored to local conditions.

At farm level in drought-prone regions, smallholders already employ practical adaptive strategies that are highly relevant for eggplant, including optimal water resource use, soil and water conservation, and nutrient management techniques to stabilize production under rainfall variability (Mpala & Simatele, 2024). A global scoping review of agricultural adaptation strategies further identifies crop and land-use adjustment, water and soil management, farmer training, agro-meteorological services, and early warning systems as central adaptation pillars; it emphasizes that biodiversity-based and climate-smart agriculture can simultaneously enhance resilience and productivity if supported by suitable policies and knowledge transfer.

Intelligent yield prediction systems can support sustainable eggplant production by integrating fertilization regimes, climate variables, and real-time field data to guide adaptive management. A comprehensive review of AI-based crop-yield prediction shows that machine and deep learning models using temperature, rainfall, humidity, soil type, and vegetation indices (e.g., NDVI, EVI, LAI) alongside management variables (such as irrigation and cultivation practices) substantially improve estimation accuracy and offer powerful tools for planning under environmental variability. Building on these insights, a crop yield prediction algorithm (CYPA) that combines climate, weather, yield, and chemical (including fertilizer) data demonstrated very high performance with ensemble models such as Random Forest and Extra Trees, and further enhanced efficiency via active learning to reduce labeled data needs.

For climate-resilient farming, future systems must be lightweight, deployable on edge devices, and tightly coupled with sensing infrastructures. An on-device AI framework using Random Forest on smart agricultural devices showed that integrating environmental sensor data with ML can achieve over 90% accuracy in detecting yield suitability and optimize irrigation scheduling to enhance water-use efficiency and support climate-resilient production without reliance on cloud computing. Reviews of IoT-enabled smart sensors in precision agriculture underscore that networks of soil, plant, and climate sensors, linked with AI/ML on IoT platforms, enable real-time monitoring, predictive analytics, and automated control of irrigation and fertilization, though challenges in cost, data management, and connectivity must be overcome for large-scale application in eggplant systems.

References

Abbas F., Al-Naemi S., and Al-Otoom A., 2025, Effects of controlled environment agriculture and nutrient sources on the production of eggplants (*Solanum melongena var. esculenta* L.), HortScience, 60(6): 970-980.
https://doi.org/10.21273/hortsci18550-25

Abdel-Salam M., Kumar N., and Mahajan S., 2024, A proposed framework for crop yield prediction using hybrid feature selection approach and optimized machine learning, Neural Computing and Applications, 36(33): 20723-20750.
https://doi.org/10.1007/s00521-024-10226-x

Badshah A., Alkazemi B., Din F., Zamli K., and Haris M., 2024, Crop classification and yield prediction using robust machine learning models for agricultural sustainability, IEEE Access, 12: 162799-162813.
https://doi.org/10.1109/access.2024.3486653

Bana R.S., Jat G.S., Grover M., Bamboriya S.D., Singh D., Bansal R., Choudhary A.K., Kumar V., Laing A.M., Godara S., Bana R.C., Kumar H., Kuri B.R., Yadav A., and Singh T., 2022, Foliar nutrient supplementation with micronutrient-embedded fertilizer increases biofortification, soil biological activity and productivity of eggplant, Scientific Reports, 12(1): 5146.
https://doi.org/10.1038/s41598-022-09247-0

Başay S., Dorak S., and Asik B.B., 2025, The effects of organic fertilizer applications on the nutrient elements content of eggplant seeds, Agronomy, 15(2): 439.
https://doi.org/10.3390/agronomy15020439

Burdett H., and Wellen C., 2022, Statistical and machine learning methods for crop yield prediction in the context of precision agriculture, Precision Agriculture, 23(5): 1553-1574.
https://doi.org/10.1007/s11119-022-09897-0

Cai S., Zhao X., and Yan X., 2022, Effects of climate and soil properties on regional differences in nitrogen use efficiency and reactive nitrogen losses in rice, Environmental Research Letters, 17(5): 054039.
https://doi.org/10.1088/1748-9326/ac6a6b

Chandio A.A., Ozdemir D., and Tang X., 2025, Modelling the impacts of climate change on horticultural crop production: evidence from Turkiye, Food and Energy Security, 14(1): e70040.
https://doi.org/10.1002/fes3.70040

Chioti V., Zeliou K., Bakogianni A., Papaioannou C., Biskinis A., Petropoulos C., Lamari F.N., and Papasotiropoulos V., 2022, Nutritional value of eggplant cultivars and association with sequence variation in genes coding for major phenolics, Plants: 11(17), 2267.
https://doi.org/10.3390/plants11172267

Cui J., Mak-Mensah E., Wang J.W., Li Q., Huang L., Song S., Zhi K.K., and Zhang J., 2024, Interactive effects of drip irrigation and nitrogen fertilization on wheat and maize yield: a meta-analysis, Journal of Soil Science and Plant Nutrition, 24(2): 1547-1559.
https://doi.org/10.1007/s42729-024-01650-y

Dey B., Ferdous J., and Ahmed R., 2024, Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables, Heliyon, 10(3): e25112.
https://doi.org/10.1016/j.heliyon.2024.e25112

Dollison M., and Tapas M.O., 2024, Yield components and nutritional analysis of Eggplant (*Solanum melongena* L.) under varying rates of Vermicast fertilizer, Diversitas Journal, 9(1): 316-331.
https://doi.org/10.48017/dj.v9i1.2952

Gao X.Q., Zhang L.C., An Y.L., Wang S.J., Feng G.Z., Lv J.Y., Li X.Y., and Gao Q., 2025, Synergistic effects of fertilization on maize yield and quality in northeast China: a meta-analysis, Agriculture, 15(13): 1371.
https://doi.org/10.3390/agriculture15131371

Gupta S., Geetha A., Sankaran K.S., Zamani A.S., Ritonga M., Raj R., Ray S., and Mohammed H.S., 2022, Machine learning-and feature selection-enabled framework for accurate crop yield prediction, Journal of Food Quality, 2022(1): 6293985.
https://doi.org/10.1155/2022/6293985

Hoque M.J., Islam M.S., Uddin J., Samad M.A., De Abajo B.S., Vargas D.L.R., and Ashraf I., 2024, Incorporating meteorological data and pesticide information to forecast crop yields using machine learning, IEEe Access, 12: 47768-47786.
https://doi.org/10.1109/access.2024.3383309

Huang N., Lin X., Lun F., Zeng R., Sassenrath G.F., and Pan Z., 2024, Nitrogen fertilizer use and climate interactions: Implications for maize yields in Kansas, Agricultural Systems, 220: 104079.
https://doi.org/10.1016/j.agsy.2024.104079

Iniyan S., Varma V.A., and Naidu C.T., 2023, Crop yield prediction using machine learning techniques, Advances in Engineering Software, 175: 103326.
https://doi.org/10.1016/j.advengsoft.2022.103326

Islam A., Shanto M.N.I., Rabby M.S.M., Sikder A.R., Uddin M.S., Arefin M.N., and Patwary M.J., 2023, Eggplant yield prediction utilizing 130 locally collected genotypes and machine learning model, In 2023 26th International Conference on Computer and Information Technology (ICCIT), IEEE, pp.1-6.
https://doi.org/10.1109/iccit60459.2023.10441036

Jabed M.A., and Murad M.A.A., 2024, Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability, Heliyon, 10(24): e40836.
https://doi.org/10.1016/j.heliyon.2024.e40836

Kaniyassery A., Thorat S.A., Kiran K.R., Murali T.S., and Muthusamy A., 2023, Fungal diseases of eggplant (*Solanum melongena* L.) and components of the disease triangle: a review, Journal of Crop Improvement, 37(4): 543-594.
https://doi.org/10.1080/15427528.2022.2120145

Krishnadoss N., and Ramasamy, L.K., 2024, Crop yield prediction with environmental and chemical variables using optimized ensemble predictive model in machine learning, Environmental Research Communications, 6(10): 101001.
https://doi.org/10.1088/2515-7620/ad7e81

Kiran Kumar V., Ramesh K.V., and Rakesh V., 2023, Optimizing LSTM and Bi-LSTM models for crop yield prediction and comparison of their performance with traditional machine learning techniques: V. Kiran Kumar et al, Applied Intelligence, 53(23): 28291-28309.
https://doi.org/10.1007/s10489-023-05005-5

Kuradusenge M., Hitimana E., Hanyurwimfura D., Rukundo P., Mtonga K., Mukasine A., Uwitonze C., Ngabonziza J., and Uwamahoro A., 2023, Crop yield prediction using machine learning models: Case of Irish potato and maize, Agriculture, 13(1): 225.
https://doi.org/10.3390/agriculture13010225

Li H., Mei X., Wang J., Huang F., Hao W., and Li B., 2021, Drip fertigation significantly increased crop yield, water productivity and nitrogen use efficiency with respect to traditional irrigation and fertilization practices: a meta-analysis in China, Agricultural Water Management, 244: 106534.
https://doi.org/10.1016/j.agwat.2020.106534

Li J., Zhang H., Zhou C., Teng A., Lei L., Ba Y., Yu J., and Li F., 2025, Integrated effects of water and nitrogen coupling on eggplant productivity, fruit quality, and resource use efficiency in a cold and arid environment, Plants, 14(2): 210.
https://doi.org/10.3390/plants14020210

Lin N., Wang X., Zhang Y., Hu X., and Ruan J., 2020, Fertigation management for sustainable precision agriculture based on Internet of Things, Journal of Cleaner Production, 277: 124119.
https://doi.org/10.1016/j.jclepro.2020.124119

Liu D., Mishra A., and Ray D., 2020, Sensitivity of global major crop yields to climate variables: a non-parametric elasticity analysis, Science of the Total Environment, 748: 141431.
https://doi.org/10.1016/j.scitotenv.2020.141431

Mahesh P., and Soundrapandiyan R., 2024, Yield prediction for crops by gradient-based algorithms. Plos one, 19(8): e0291928.
https://doi.org/10.1371/journal.pone.0291928

Meng L., Liu H., L. Ustin S., and Zhang X., 2021, Predicting maize yield at the plot scale of different fertilizer systems by multi-source data and machine learning methods, Remote Sensing, 13(18): 3760.
https://doi.org/10.3390/rs13183760

Mohan R.N.V., Rayanoothala P.S., and Sree R.P., 2025, Next-gen agriculture: integrating AI and XAI for precision crop yield predictions, Frontiers in Plant Science, 15: 1451607.
https://doi.org/10.3389/fpls.2024.1451607

Morales A., and Villalobos F.J., 2023, Using machine learning for crop yield prediction in the past or the future, Frontiers in Plant Science, 14: 1128388.
https://doi.org/10.3389/fpls.2023.1128388

Nguyen G.N., Lantzke N., and van Burgel A., 2022, Effects of shade nets on microclimatic conditions, growth, fruit yield, and quality of eggplant (*Solanum melongena* L.): a case study in Carnarvon, Western Australia. Horticulturae, 8(8): 696.
https://doi.org/10.3390/horticulturae8080696

Oladosu Y., Rafii M.Y., Arolu F., Chukwu S.C., Salisu M.A., Olaniyan B.A., Fagbohun L.K., and Muftaudeen T.K., 2021, Genetic diversity and utilization of cultivated eggplant germplasm in varietal improvement, Plants, 10(8): 1714.
https://doi.org/10.3390/plants10081714

Osman M.A., Onono J.O., Olaka L.A., Elhag M.M., and Abdel-Rahman E.M., 2021, Climate variability and change affect crops yield under rainfed conditions: a case study in Gedaref State, Sudan, Agronomy, 11(9): 1680.
https://doi.org/10.3390/agronomy11091680

Parent L.E., 2024, Vegetable response to added nitrogen and phosphorus using machine learning decryption and the N/P ratio, Horticulturae, 10(4): 356.
https://doi.org/10.3390/horticulturae10040356

Paudel D., Boogaard H., De Wit A., Janssen S., Osinga S., Pylianidis C., and Athanasiadis I.N., 2021, Machine learning for large-scale crop yield forecasting, Agricultural Systems, 187: 103016.
https://doi.org/10.1016/j.agsy.2020.103016

Saeed F., Chaudhry U.K., Raza A., Charagh S., Bakhsh A., Bohra A., Ali S., Chitikineni A., Saeed Y., Visser R.G.F., Siddique K.H.M., and Varshney R.K., 2023, Developing future heat-resilient vegetable crops, Functional and integrative genomics, 23(1): 47.
https://doi.org/10.1007/s10142-023-00967-8

Sharma P., Dadheech P., Aneja N., and Aneja S., 2023, Predicting agriculture yields based on machine learning using regression and deep learning, IEEe Access, 11: 111255-111264.
https://doi.org/10.1109/access.2023.3321861

Taşan S., Cemek B., Taşan M., and Cantürk A., 2022, Estimation of eggplant yield with machine learning methods using spectral vegetation indices, Computers and electronics in agriculture, 202: 107367.
https://doi.org/10.1016/j.compag.2022.107367

Thingujam U., Bhattacharyya K., Ray K., Phonglosa A., Pari A., Banerjee H., Dutta S., and Majumdar K., 2020, Integrated nutrient management for eggplant: yield and quality models through artificial neural network, Communications in Soil Science and Plant Analysis, 51(1): 70-85.
https://doi.org/10.1080/00103624.2019.1695824

Xing Y., and Wang X., 2024, Precise application of water and fertilizer to crops: challenges and opportunities, Frontiers in Plant Science, 15: 1444560.
https://doi.org/10.3389/fpls.2024.1444560

Xing Y., Zhang X., and Wang X., 2024, Enhancing soil health and crop yields through water-fertilizer coupling technology, Frontiers in Sustainable Food Systems, 8: 1494819.
https://doi.org/10.3389/fsufs.2024.1494819

Zhou C., Zhang H., Yu S., Chen X., Li F., Wang Y., and Liu L., 2023, Optimizing water and nitrogen management strategies to improve their use efficiency, eggplant yield and fruit quality, Frontiers in Plant Science, 14: 1211122.
https://doi.org/10.3389/fpls.2023.1211122

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Computational Molecular Biology 2026, Vol.16, No.3, 159-180
http://bioscipublisher.com/index.php/cmb

Research Article Open Access

Genome-wide Relationship Matrix-Based Heritability Estimation: Statistical Interpretation, Comparability, and Practical Diagnostics in the GCTA-GREML Framework

Running title: Interpreting SNP Heritability with GCTA-GREML

Xuanjun Fang
Hainan Provincial Key Laboratory of Crop Molecular Breeding, Hainan Institute of Tropical Agricultural Resources (HITAR), Sanya, 572025
Corresponding author: xuanjunfang@hitar.org

Computational Molecular Biology, 2026, Vol.16, No.3 doi: 10.5376/cmb.2026.16.0012
Received: 05 Apr., 2026
Accepted: 08 May, 2026
Published: 20 May, 2026

Copyright © 2026 Fang, This is an open access article published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preferred citation for this article:
Fang X.J., 2026, Genome-wide relationship matrix-based heritability estimation: statistical interpretation, comparability, and practical diagnostics in the GCTA-GREML framework, Computational Molecular Biology, 16(3): 159-180 (doi: 10.5376/cmb.2026.16.0012)

Abstract Heritability, as a core concept, plays a critical role in explaining trait variation and predicting selection response. Traditional heritability estimation relies on pedigree information but is limited by pedigree completeness and environmental confounding. With the development of high-throughput genotyping and genome-wide association studies, the restricted maximum likelihood method based on genomic relationship matrices (GCTA/GREML) has provided a new pathway for estimating the heritability of complex traits. This study reviews the theoretical framework and statistical assumptions of the GCTA and GREML families, elucidates their logic in variance decomposition and differences from pedigree-based models, and focuses on analyzing the sources and interpretive boundaries of the "missing heritability" problem. Further, the study explores methodological extensions such as the LOCO strategy, functional annotation partitioning, and bivariate analysis, and discusses their application value in complex trait dissection and crop breeding, supported by both simulation and empirical studies. The results indicate that GCTA/GREML not only promotes a paradigm shift in heritability research but also provides theoretical support for genomic selection and molecular breeding design. In the future, with the accumulation of sequencing data and multi-environment big data, this method is expected to more comprehensively uncover the genetic basis of complex traits.

Accordingly, this review focuses on clarifying the statistical interpretation of SNP-based heritability estimation rather than providing a general tutorial. Specifically, we (i) outline the statistical conditions required for meaningful comparisons between SNP-based and pedigree-based heritability estimates; (ii) formally define the estimand targeted by GREML and clarify its relationship to the concept of missing heritability; (iii) organize commonly used GREML extensions into a unified framework based on their inferential goals, assumptions, and diagnostic boundaries; and (iv) propose a workflow-oriented checklist to guide the interpretation of SNP heritability estimates in practice.

Keywords SNP heritability; Genome-wide relationship matrix (GRM); GCTA-GREML; Missing heritability; Statistical interpretation; Diagnostic workflow

1 Introduction

Heritability, as a central concept in quantitative genetics, has played a fundamental role in explaining the sources of trait variation and guiding genetic improvement practices since Fisher established the framework of analysis of variance. It is defined as the proportion of phenotypic variance that can be attributed to genetic differences, and serves as a cornerstone concept in both quantitative genetics and breeding science (Yang et al., 2017; Srivastava et al., 2023). In breeding, heritability not only provides a quantitative scale for evaluating the potential for trait selection, but also constitutes a key parameter for predicting selection response and optimizing population improvement strategies (Zhu and Zhou, 2020). High heritability indicates that genetic variation accounts for a large proportion of phenotypic variance, thereby leading to higher efficiency of artificial selection; conversely, low heritability suggests a dominant role of environmental variation and consequently limited selection effectiveness. Therefore, whether in the design of crop and livestock breeding strategies or in the genetic epidemiology of complex human diseases, accurate estimation of heritability remains an unavoidable core issue at both theoretical and practical levels (Yang et al., 2017).

Traditional heritability estimation primarily relies on pedigree-based variance component models, which infer additive genetic variance by comparing phenotypic similarity between related and unrelated individuals (Yang et al., 2017; Srivastava et al., 2023). However, these methods depend heavily on the completeness of pedigree information and are often constrained by simplified assumptions regarding shared environmental effects. In populations lacking detailed pedigree records or affected by environmental confounding, both their applicability and accuracy are limited (Zhu and Zhou, 2020).

With the widespread adoption of high-throughput genotyping technologies and the emergence of genome-wide association studies (GWAS), the field has undergone a methodological revolution. Yang et al. (2010; 2011) proposed the genome-wide complex trait analysis (GCTA) framework based on single nucleotide polymorphisms (SNPs), and further developed the genomic-relatedness-based restricted maximum likelihood (GREML) method based on the genome-wide relationship matrix (GRM). By constructing a GRM and leveraging SNP-derived genetic similarity among individuals after removing close relatives, this approach decomposes phenotypic variance and overcomes the limitations of traditional pedigree-based methods (Yang et al., 2011; Zhu and Zhou, 2020). Compared with pedigree models, GCTA-GREML enables direct estimation of heritability from SNP data without requiring pedigree information, and supports partitioning of genetic variance by genomic regions or functional annotations, thereby substantially expanding the scope of heritability estimation (Zhu and Zhou, 2020; Srivastava et al., 2023).

However, the introduction of the GCTA and GREML framework has also triggered extensive debate regarding the issue of "missing heritability." Classical twin and pedigree studies often yield relatively high heritability estimates, whereas SNP-based GREML estimates are typically substantially lower. This discrepancy has been interpreted as evidence that GWAS cannot fully explain the genetic variation underlying complex traits (Yang et al., 2011; 2015). Potential explanations include incomplete tagging of causal variants by SNPs, insufficient contribution from rare variants, complex genetic mechanisms such as non-additive effects and gene-environment interactions, as well as limitations of statistical modeling (Speed et al., 2016; Evans et al., 2017; Mathew et al., 2017). Furthermore, existing studies have shown that GCTA-GREML estimates are highly sensitive to factors such as GRM construction methods, sample composition, linkage disequilibrium (LD) patterns, and phenotypic measurement error, further highlighting the complexity of its application and the need for careful interpretation (Speed et al., 2012; Kumar et al., 2015; Evans et al., 2017).

Thus, the problem of missing heritability is not only a statistical challenge but also a genetic and biological one, and the associated debates have driven continuous innovation in both methodology and theory. In recent years, improvements such as LD-adjusted relationship matrices and multi-component modeling have been proposed, providing potential solutions to the limitations of the original GCTA-GREML framework (Mathew et al., 2017; Zhu and Zhou, 2020).

In crop breeding practice in China, DNA marker-assisted breeding was systematically summarized and promoted from the late 20th to early 21st century. Its core idea is to track quantitative trait loci (QTLs) or candidate genes using a limited number of molecular markers, thereby improving selection efficiency (Fang et al., 2001). This study systematically reviews the theoretical framework and statistical assumptions of GCTA and GREML relative to pedigree-based methods, clarifying their conceptual positioning and applicability boundaries in heritability estimation. The analytical framework adopted here is consistent with our previous systematic examination of the statistical continuity among linkage analysis, candidate gene strategies, and GWAS, emphasizing the continuity and division of roles among different methods in terms of statistical assumptions, signal scale, and inferential objectives (Fang and Wu, 2026). We focus on the derivation logic of the GREML method within variance component modeling, compare its estimands and interpretive scope with those of traditional pedigree models, and discuss the potential impact of model assumptions on result interpretation.

Based on the above background, this study does not aim to provide a general introductory overview, but rather focuses on the core issue of the "statistical interpretability boundaries of SNP-based heritability estimation," with the goal of constructing an operational framework for analysis and interpretation. Specifically, the study addresses

the following aspects: (1) systematically outlining the necessary conditions for comparability between SNP-based and pedigree-based heritability; (2) clarifying the statistical target of heritability estimated by GREML and the conceptual boundaries of "missing heritability"; (3) proposing a unified comparison template for common methodological extensions; and (4) providing a standardized workflow and diagnostic checklist for practical interpretation. Through theoretical derivation and empirical analysis, this study aims to offer a clearer understanding of the GCTA framework and its extensions, thereby providing a theoretical foundation and methodological reference for the application of heritability in complex trait research and crop breeding practice.

2 Basic Concepts and Classification of Heritability

Heritability is defined as a variance ratio under a specified statistical model, which depends on both the population and environmental conditions, and quantifies the proportion of phenotypic variation attributable to genetic variation. Therefore, heritability estimates are not directly comparable across different populations, environments, or modeling assumptions.

2.1 Narrow-sense and broad-sense heritability

Heritability is a core parameter in quantitative and statistical genetics, used to characterize the relative contribution of genetic factors to phenotypic variation under a given population, environment, and set of model assumptions (Vinkhuyzen et al., 2013; Yang et al., 2017). From a statistical perspective, heritability is fundamentally a variance ratio, rather than an intrinsic property of a trait or an individual.

Within the classical variance decomposition framework, heritability is typically divided into narrow-sense heritability (h^2) and broad-sense heritability (H^2).

Narrow-sense heritability is defined as the proportion of additive genetic variance (V_a) relative to total phenotypic variance (V_p):

$$h^2=\frac{V_A}{V_P}$$

where V_p represents the overall magnitude of phenotypic variation in the population. Because additive genetic effects are stably transmitted across generations and are cumulative, h^2 plays a central role in predicting the response to selection (e.g., under the Breeder's equation framework), as well as in breeding value estimation and gene mapping studies (Evans et al., 2017; Yang et al., 2017). In practical breeding, narrow-sense heritability is generally regarded as the key indicator of expected selection gain, and its practical relevance often exceeds that of broad-sense heritability (Berry, 2024).

In contrast, broad-sense heritability captures the total contribution of all genetic effects to phenotypic variation, and is defined as:

$$H^2=\frac{V_A+V_D+V_I}{V_P}$$

where V_D denotes dominance variance and V_I denotes epistatic (gene-gene interaction) variance. Although H^2 theoretically reflects the total explanatory power of genetic factors, dominance and epistatic effects depend on allele frequencies and genotype combinations, resulting in limited reproducibility and operability across generations. Therefore, H^2 is generally not suitable for directly predicting selection response (Abney et al., 2001; Zhu et al., 2015).

In most outbred or natural populations, it typically holds that $H^2 \geq h^2$, and the difference between the two reflects the presence and relative magnitude of non-additive genetic variance (Abney et al., 2001; Berry, 2024). Recent studies based on genome-wide marker data have shown that, for many complex traits, dominance variance contributes only modestly to total genetic variation, whereas rare and low-frequency variants may account for part of the "missing heritability" observed in earlier studies (Speed et al., 2012; 2016; Jang et al., 2022; Wainschtein et al., 2022; Srivastava et al., 2023). These findings help establish a consistent and interpretable framework for variance decomposition and prediction across evolutionary genetics and applied breeding (Bérénos et al., 2014; Zimmermann and Distl, 2023).

2.2 Pedigree-based vs. SNP-based heritability

Traditional heritability estimation relies primarily on pedigree information, constructing additive genetic covariance matrices among individuals based on kinship coefficients or identity-by-descent (IBD), and decomposing phenotypic variance within a linear mixed model framework (Vinkhuyzen et al., 2013; Bérénos et al., 2014). Such approaches have long played an important role in animal and plant breeding as well as in studies of natural populations. However, their estimation accuracy depends critically on the completeness and correctness of pedigree records. When shared environmental effects among related individuals are not adequately modeled, pedigree-based heritability estimates may be systematically upward biased.

With the development of high-throughput genotyping technologies and statistical genetic methods, genotype-based heritability estimation has emerged as an important complement to pedigree-based approaches. Methods represented by GCTA/GREML construct a genome-wide relationship matrix (GRM) from SNP data and estimate the additive genetic variance captured by markers within a restricted maximum likelihood (REML) framework (Speed et al., 2012; Evans et al., 2017; Yang et al., 2017).

It is important to emphasize that SNP-based heritability is not directly equivalent to the "true" heritability of a trait. Rather, it reflects the proportion of additive genetic variance that can be captured by a given set of markers under specific statistical model assumptions. Such estimates are typically obtained from samples with close relatives removed, in order to reduce confounding effects arising from shared environment and pedigree structure (Srivastava et al., 2023; Zimmermann and Distl, 2023). Therefore, differences between pedigree-based and SNP-based heritability do not necessarily imply "missing" genetic information, but more likely arise from differences in estimands, marker coverage, and modeling assumptions.

2.3 Sources of discrepancy and "missing heritability"

In numerous studies of complex traits, heritability estimates based on pedigree data are often higher than those derived from SNP-based approaches, giving rise to the so-called problem of "missing heritability" (Vinkhuyzen et al., 2013; Yang et al., 2017). From a statistical genetics perspective, this discrepancy should not be interpreted simply as a true loss of genetic information, but rather as a systematic difference arising from distinct estimands, data coverage, and modeling assumptions.

First, limited marker coverage is a major source of lower SNP-based heritability. Conventional genotyping arrays primarily capture common variants, while providing limited representation of rare variants, low-frequency variants, and structural variants. As a result, part of the genetic variance remains untagged, leading to downward-biased SNP heritability estimates (Wainschtein et al., 2019; Jang et al., 2022; Wainschtein et al., 2022). Recent analyses based on whole-genome sequencing data have demonstrated that rare variants can explain a portion of the previously "missing" heritability, further supporting this explanation.

Second, incomplete linkage disequilibrium (LD) constrains the ability of markers to capture the effects of causal variants. Even with high-density SNP data, LD between markers and true causal loci is often insufficient to fully reflect effect sizes, resulting in systematic underestimation of additive genetic variance (Speed et al., 2012; 2016; Evans et al., 2017). This issue is particularly pronounced in populations with complex LD structures or highly heterogeneous allele frequency distributions.

Third, confounding due to shared environmental effects may inflate pedigree-based heritability estimates. In pedigree studies, related individuals often share both genetic background and environmental conditions. If the model fails to adequately disentangle these contributions, environmental correlations may be incorrectly attributed to genetic variance, thereby inflating heritability estimates (Vinkhuyzen et al., 2013; Bérénos et al., 2014). In contrast, SNP-based methods are typically applied to samples with close relatives removed, reducing such confounding.

In addition, non-additive genetic effects and gene-environment interactions can further widen the gap between pedigree-based and SNP-based heritability estimates. Narrow-sense heritability and most SNP-based frameworks

primarily focus on additive genetic variance, while dominance, epistasis, and their interactions with environmental factors are often not explicitly modeled (Abney et al., 2001; Chen et al., 2015; Zhu et al., 2015). These effects may be partially absorbed into genetic variance estimates in pedigree-based analyses, but are difficult to identify in SNP-based analyses using unrelated individuals.

In summary, "missing heritability" is more appropriately understood as a difference in the identifiability of genetic variance under different statistical frameworks, rather than as an actual absence of genetic mechanisms. Pedigree-based and SNP-based heritability estimates reflect different aspects of genetic architecture; their discrepancy provides important insights into the multi-layered genetic basis of complex traits, rather than constituting mutually contradictory evidence. To facilitate a systematic comparison between traditional marker-assisted approaches and genome-wide statistical genetic methods in terms of research objectives, statistical assumptions, and application scenarios, representative methods-including linkage analysis, candidate gene approaches, and GWAS/GCTA-GREML-are summarized in Table 1.

Table 1 Comparison between traditional marker-assisted approaches and genome-wide statistical genetic methods

Comparison dimension	Traditional approaches (Linkage/Candidate gene)	Genome-wide approaches (GWAS/GCTA-GREML)
Research starting point	Hypothesis-driven candidate regions or genes	Genome-wide, hypothesis-free scanning
Primary data type	A limited number of molecular markers (e.g., RFLP, SSR)	High-density SNPs or whole-genome sequencing data
Study population	Structured populations or pedigrees	Natural populations or breeding populations
Scale of genetic signal	Single loci or local linkage intervals	Genome-wide, multi-locus signals
Core statistical assumptions	Strong prior assumptions with limited multiple testing	Explicit modeling of population structure and multiple testing
Main analytical objective	Identification of QTLs or candidate genes	Estimation of heritability and genetic architecture
Interpretation of results	Locus-specific effects and biological interpretation	Variance decomposition and predictability assessment
Suitability for complex traits	Limited power for highly polygenic traits	Well suited for highly polygenic traits
Role in breeding	Marker-assisted selection and locus validation	Guiding genomic selection and breeding strategy design
Representative methods	Linkage analysis, candidate gene analysis	GWAS、GCTA、GREMLGWAS, GCTA, GREML
Methodological limitations	Limited resolution, power depends on population design	Sample-size dependent, limited causal interpretation
Comparison dimension	Traditional approaches (Linkage/Candidate gene)	Genome-wide approaches(GWAS/GCTA-GREML)

Note: Traditional marker-assisted approaches rely mainly on linkage analysis and candidate gene strategies to identify QTLs or functional loci using a limited number of molecular markers in structured populations (Fang et al., 2001). Genome-wide methods, represented by GWAS and GCTA/GREML, use dense genome-wide markers to build statistical models for estimating heritability and dissecting the genetic architecture of complex traits. Although these approaches differ substantially in statistical assumptions and analytical scale, they are historically and conceptually connected in crop genetic improvement (Fang and Wu, 2026).

3 Principles of Constructing the Genome-wide Relationship Matrix (GRM)

3.1 Standardized genotype matrix

The construction of the genome-wide relationship matrix (GRM) is fundamentally based on a standardized genotype matrix. For each SNP locus in diploid species, genotypes are typically encoded as 0, 1, or 2, representing the number of copies of the reference allele carried by an individual. However, directly using these raw genotype encodings may introduce bias, because differences in allele frequencies across loci can lead to heterogeneity in variance (Forni et al., 2011; Wang et al., 2025).

To avoid such bias, genotype data must be standardized. Let the population frequency of the reference allele at a given locus be p. The observed genotype $x$for an individual at that locus is standardized as:

$$z=\frac{x-2p}{\sqrt{2p(1-p)}}$$

This transformation centers the genotype (by subtracting its expectation, $2p$) and scales it (by dividing by its standard deviation, $\sqrt{2p(1-p)}$), ensuring that all loci contribute comparably to the matrix calculation (Forni et al., 2011; Granato et al., 2018).

This standardization has important statistical implications. On the one hand, it removes variance heterogeneity caused by allele frequency differences across loci, making the GRM estimation more reflective of true genetic similarity (Wang et al., 2025). On the other hand, it effectively distinguishes between allele frequency differences arising from random genetic drift and those reflecting genuine shared genetic background, thereby enabling the construction of a robust relationship matrix at the genome-wide level. This approach has been widely applied in genomic prediction, heritability estimation, and association studies, and has been integrated into various molecular breeding tools (Forni et al., 2011; Granato et al., 2018).

3.2 GRM formula and intuitive interpretation

After constructing the standardized genotype matrix $\mathbf{Z}$, the GRM can be expressed as:

$$\mathbf{G}=\frac{1}{M}\mathbf{Z}\mathbf{Z}^{T}$$

where M denotes the total number of SNPs across the genome, and each matrix element G_{ij}represents the genomic similarity between individuals i and j (Forni et al., 2011; Wang et al., 2025).

Intuitively, the GRM measures the similarity between two individuals based on their standardized genotypes across all marker loci, and its values reflect their additive genetic relatedness at the population level. The diagonal elements represent self-relatedness (or inbreeding), with an expected value close to 1, while off-diagonal elements quantify pairwise relatedness between individuals. Values approaching 1 indicate high genetic similarity, whereas values close to 0 suggest near independence.

From a statistical perspective, the GRM can be interpreted as a genome-wide weighted average of identity-by-state (IBS) (Forni et al., 2011). Unlike traditional pedigree-based relationship matrices, the GRM does not rely on prior pedigree information but is constructed directly from molecular data, enabling it to capture realized genetic similarity. This property allows the GRM to be applied not only to large-scale natural populations without pedigree records, but also to more accurately characterize complex population structures and latent genetic diversity (Bilton et al., 2024; Wang et al., 2025).

3.3 Example: visualization and comparison of GRM structures in human and crop populations

In high-level human genetics studies, the GRM is often visualized using heatmaps or distributions of pairwise relatedness to intuitively illustrate additive genetic similarity among individuals (Figure 1). For example, in studies based on the UK Biobank (Yang et al., 2015; Speed et al., 2016; Hou et al., 2019), GRM heatmaps typically exhibit a highly sparse structure centered along the diagonal: diagonal elements are close to 1, reflecting the standardized genetic variance of individuals themselves, while off-diagonal elements are mostly concentrated near zero, with weak clustering patterns appearing only in the presence of subtle population structure or residual relatedness. This structural feature indicates that, after stringent quality control (QC) and removal of close relatives, the GRM can stably capture SNP-derived additive genetic similarity among unrelated individuals.

Similar structural patterns can also be observed in crop populations, but their manifestation is strongly influenced by population composition and linkage disequilibrium (LD) structure. In inbred populations such as rice or maize, where the number of chromosomes is limited, LD blocks are relatively large, and subpopulation differentiation is pronounced, GRM heatmaps often display clearer block-like structures corresponding to different genetic backgrounds or breeding origins (Granato et al., 2018). This comparison highlights that, although the statistical definition of the GRM remains consistent across species, its empirical structure is highly dependent on population history, LD architecture, and sampling design.

It is important to note that the elements of the GRM represent standardized additive genetic covariances, rather than correlation coefficients. Therefore, when the number of markers is limited and allele frequencies are estimated from the sample, diagonal elements or values for highly related individuals may slightly exceed 1.

In this context, GRM heatmaps serve not only as a visualization tool, but also as an important diagnostic instrument for understanding population structure, assessing potential confounding factors, and interpreting subsequent GREML-based heritability estimates.

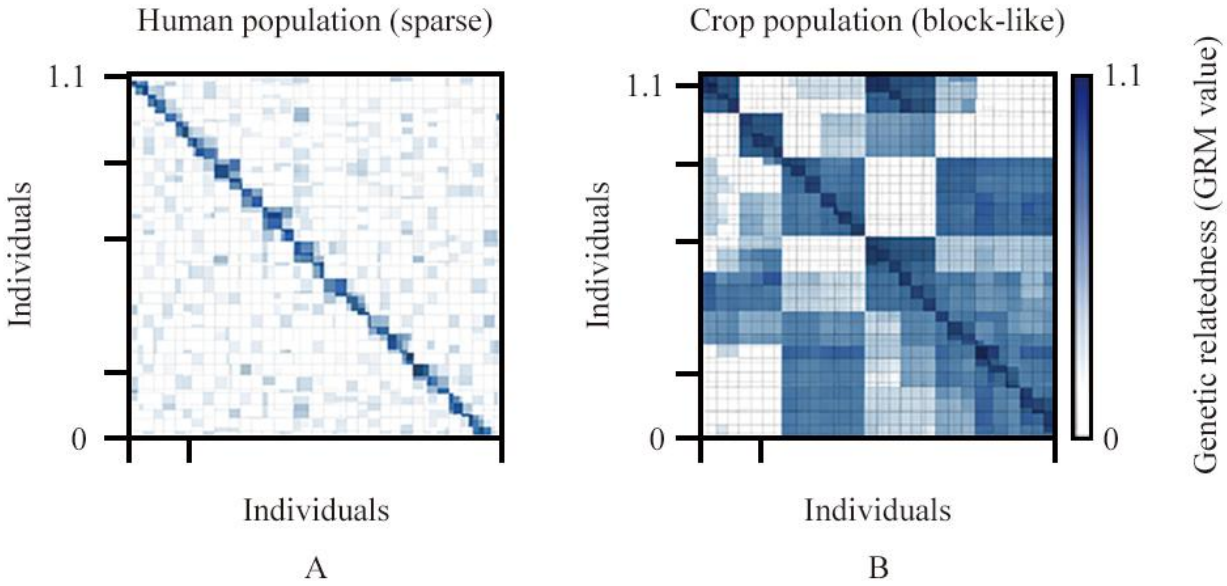


Figure 1 Illustrative comparison of GRM structures in human and crop populations

Caption: Illustrative schematic based on published studies (Yang et al., 2015; Speed et al., 2016). Panel A shows a schematic GRM heatmap representative of large human cohorts after standard quality control and removal of close relatives, as commonly observed in studies such as UK Biobank-based analyses. The matrix is characterized by strong diagonal elements (self-relatedness) and sparse off-diagonal values centered near zero, reflecting weak pairwise genetic relatedness among largely unrelated individuals. Panel B illustrates a typical GRM structure for crop populations, where pronounced block-like patterns arise due to strong population structure, limited numbers of chromosomes, extended linkage disequilibrium, and shared breeding history. These contrasting patterns highlight that, although the statistical definition of the GRM is consistent across species, its empirical structure is highly dependent on population history, LD architecture, and sampling design. The figure is schematic and intended for diagnostic illustration rather than representation of a specific dataset. Note that GRM values are not constrained to the interval [−1, 1]; diagonal elements and highly related pairs may slightly exceed 1 due to finite marker density and allele-frequency estimation.

4 GREML and REML Estimation

4.1 Model derivation

Heritability estimation based on the genome-wide relationship matrix (GRM) is typically conducted within the framework of a linear mixed model (LMM) (Da et al., 2014; Yang et al., 2016; Zhou et al., 2020). Its general form can be written as:

$$\mathbf{y}=\mathbf{X}\boldsymbol{\beta}+\mathbf{g}+\mathbf{e}$$

where $\mathbf{y}$denotes the vector of phenotypic observations, $\mathbf{X}\boldsymbol{\beta}$ represents fixed effects (e.g., population structure, environmental factors, or other covariates), $\mathbf{g}$ denotes the random additive genetic effects, and $\mathbf{e}$ is the independent residual error term. Unlike traditional heritability estimation approaches, the LMM framework allows for simultaneous control of systematic confounding and estimation of genotype-related variance components within a unified model.

In variance decomposition, the random genetic effects are assumed to follow a multivariate normal distribution with mean zero and a covariance structure proportional to the GRM:

$$\mathbf{g}\sim N(0,\sigma_g^2\mathbf{G})$$

where σ_g^2denotes the additive genetic variance and $\mathbf{G}$is the GRM. The environmental residuals are assumed to follow:

$$\mathbf{e}\sim N(0,\sigma_e^2\mathbf{I})$$

Accordingly, the variance-covariance matrix of the phenotype can be expressed as:

$$\mathrm{Var}(\mathbf{y})=\sigma_g^2\mathbf{G}+\sigma_e^2\mathbf{I}$$

Under this model, heritability is estimated as:

$$h^2=\frac{\sigma_g^2}{\sigma_g^2+\sigma_e^2}$$

This framework provides the theoretical foundation for GREML (genomic-relatedness-based restricted maximum likelihood), enabling the estimation of additive genetic variance explained by genome-wide markers through statistical inference (Da et al., 2014; Yang et al., 2016; Zhou et al., 2020). Furthermore, to better capture complex genetic architectures, extensions of the LMM have been proposed, such as models incorporating multiple random effects or covariance structures among random effects (Zhou et al., 2019; 2020).

4.2 REML estimation and the maximum likelihood framework

In terms of parameter estimation, GREML typically relies on restricted maximum likelihood (REML). Unlike conventional maximum likelihood (ML), REML eliminates fixed effects by integrating them out of the likelihood function, thereby optimizing variance parameters based on residuals. This approach effectively avoids bias in variance component estimation caused by fixed-effect estimation, and is particularly advantageous in complex models and finite-sample settings (Dao et al., 2021; Meyer, 2023).

In practical implementation, REML is carried out via numerical optimization of the log-likelihood function. The GCTA software employs the AI-REML (Average Information REML) algorithm, which iteratively updates parameters using the average information matrix and achieves efficient estimation of variance components (Yang et al., 2016; Strandén et al., 2024). BOLT-REML introduces stochastic projection and approximation techniques to substantially reduce computational complexity in large-scale datasets, making it suitable for cohorts with sample sizes on the order of hundreds of thousands to millions (Border and Becker, 2019). The GEMMA software also implements the REML framework and extends it to multivariate and Bayesian analyses, demonstrating robust convergence properties in small to medium-sized datasets (Meyer, 2023).

Recent methodological advances, including principal component-based reparameterization and stochastic optimization algorithms, have further improved the scalability and adaptability of REML estimation for large and complex datasets (Strandén et al., 2024).

4.3 Validation using simulated and empirical data

The validity of the GREML method is typically assessed through a combination of simulation studies and empirical data analyses. Simulation studies have shown that, under correct model specification and sufficiently large sample sizes, GREML can provide unbiased estimates of heritability (Da et al., 2014; Cesarani et al., 2018; Zhou et al., 2020). However, in small sample settings (e.g., hundreds of individuals), the limited information content of the GRM leads to large estimation variance, and the estimates become sensitive to assumptions regarding population structure and phenotypic distribution, potentially introducing bias (Cesarani et al., 2018; Meyer, 2023).

In contrast, in large cohorts (tens of thousands to millions of individuals), GREML is capable of more accurately capturing the genetic variation explained by genome-wide markers. Approximate methods such as BOLT-REML have been shown, in human population studies (e.g., UK Biobank), to produce heritability estimates close to true values while effectively controlling for population structure and batch effects (Nolte et al., 2017; Ni et al., 2018). In crop populations, such as maize and wheat with genome-wide data, GREML applications have revealed the heritable architecture of complex quantitative traits and provided theoretical guidance for subsequent GWAS and genomic selection. Further methodological extensions, such as CORE GREML, allow for covariance among random effects and have demonstrated improved performance over standard GREML in the presence of complex genetic architectures (Zhou et al., 2019; 2020).

5 Methodological Extensions and Variants

To facilitate comparison of different extensions in terms of statistical objectives and applicability, this study adopts a unified analytical framework for commonly used GREML-based methods (Table S2).

5.1 LOCO (leave-one-chromosome-out) strategy

When estimating heritability and performing genomic prediction within the GREML framework, sources of bias embedded in model specification are not always immediately apparent. Among these, proximal contamination represents a typical issue with important methodological implications. Specifically, when analysis focuses on a particular chromosome or a localized genomic region, if markers from that same region are simultaneously used to construct the genomic relationship matrix (GRM), the linkage disequilibrium (LD) information they carry can "feed back" into the model through the relationship structure. This feedback mechanism leads to a systematic inflation of the estimated contribution of the focal region, fundamentally reflecting a lack of identifiability in parameter decomposition and the resulting estimation bias (Yang et al., 2011; Van den Berg et al., 2019).

The LOCO (leave-one-chromosome-out) approach offers a targeted correction strategy for this problem. Rather than restructuring the model in a complex way, its core logic is to deliberately exclude all markers from the chromosome of interest when constructing the GRM used to estimate that chromosome's genetic effect. In doing so, it effectively blocks the indirect feedback pathway through which local LD information influences its own effect estimate (Yang et al., 2011). This strategy implicitly relies on the assumption that genetic contributions from different chromosomes can be treated as approximately independent in a statistical sense, such that removing information from the target chromosome does not substantially impair the modeling of the remaining genomic background. Under this condition, LOCO can mitigate endogenous bias in local effect estimation without altering the overall modeling framework.

From the perspective of genomic architecture, the advantages of LOCO become particularly evident under certain conditions. In organisms with a relatively small number of chromosomes, extended LD blocks, or phenotypic variation driven by a limited number of large-effect loci, local LD structures are more likely to generate strong signal coupling within the GRM, thereby amplifying the impact of proximal contamination. This feature is especially pronounced in many crop genomes, making the LOCO strategy highly compatible with studies in agricultural genetics and breeding. In contrast, for species with a larger number of chromosomes and rapid LD decay, the severity of proximal contamination is often reduced, and the marginal benefit of applying LOCO correspondingly diminishes.

It is important to note that LOCO is not a universal solution for bias correction. Its utility is primarily confined to addressing proximal contamination and does not extend to systematic control of population structure, long-range LD heterogeneity, or other complex confounding factors. Therefore, its application should be guided by empirical evaluation rather than assumed necessity. In practice, researchers may compare heritability estimates or marker effect sizes obtained from standard GRM-based models and LOCO-adjusted models to assess the extent of proximal contamination (Van den Berg et al., 2019). If results are highly consistent across the two settings, the additional computational burden and model partitioning introduced by LOCO may not yield substantial benefits. Conversely, pronounced discrepancies indicate that local LD "feedback" is indeed influencing parameter estimation, in which case the use of a leave-one-chromosome-out strategy is both statistically justified and practically valuable.

5.2 Partitioning heritability by functional categories

In the traditional GREML framework, all SNPs are assumed to have equal prior weights by default; that is, their contributions to the overall genetic variance are treated as statistically homogeneous. However, this assumption is often difficult to sustain for complex traits, because different functional regions of the genome vary substantially in their biological mechanisms and evolutionary constraints, which in turn leads to spatial heterogeneity in the distribution of genetic effects. Against this background, research approaches that partition heritability by functional category have gradually developed. Their core objective is to reveal how genetic variance is distributed

across different functional regions, thereby extending the question from quantifying "the magnitude of heritability" to interpreting "the structural sources of heritability." This approach can not only reduce heterogeneity-related bias in overall estimates, but also substantially improve the biological interpretability of the results, allowing heritability estimates to be more closely aligned with functional genomic information (Finucane et al., 2015; Gazal et al., 2018).

In terms of methodological implementation, such models usually rely on existing functional annotation systems, in which genome-wide SNPs are classified into categories such as coding regions, regulatory regions, and conserved sequences. A genetic relationship matrix (GRM) is then constructed separately for each category. Subsequently, within an extended multi-GRM GREML framework, multiple variance components are introduced simultaneously to jointly estimate the genetic contributions of different functional regions (Finucane et al., 2015; Wei et al., 2019). A key assumption underlying this modeling strategy is that SNPs in different functional categories differ systematically in the distribution of their effect sizes and in their relationships with linkage disequilibrium (LD) structure, and that these differences can be statistically identified and quantified through partitioned modeling.

From the perspective of data suitability, this type of method places relatively high demands on sample size and annotation quality. A larger sample size helps stabilize the estimation of multiple variance components, while high-quality functional annotation is a prerequisite for ensuring that the partitioning results have biological meaning. The number of SNPs must also be sufficient to support multi-category partitioning; otherwise, model parameters may face identification difficulties. In human and crop genetic studies, this method is particularly appropriate when the research focus shifts from a single estimate of heritability to the analysis of genetic architecture, namely when attention is directed toward the relative importance of different functional regions in contributing to a trait.

It should be noted that functional partitioning of heritability is sensitive in practice to correlations among annotations. Because different functional categories often overlap in genomic space and may exhibit highly correlated LD structures, such multicollinearity can directly affect the identifiability of variance components, leading to unstable estimates or ambiguity in interpretation. Therefore, when interpreting the results, sensitivity analyses should be incorporated to evaluate model robustness, and conclusions regarding "enrichment" in any single region should be treated with caution. Statistical association should not be equated simplistically with clear biological causality (Gazal et al., 2018).

5.3 Bivariate and cross-trait genetic correlation

Under the single-trait GREML framework, researchers can estimate the genetic variance of a single phenotype with relative robustness. However, such models essentially remain confined to variance partitioning "within a trait" and are therefore limited in addressing the more biologically meaningful question of whether different traits share a common genetic basis. Against this background, bivariate and cross-trait GREML models have gradually become important extensions in the genetic analysis of complex traits. By jointly modeling multiple phenotypes, this approach not only improves the characterization of genetic covariance structures, but also enables the quantification of correlations between traits driven by shared genetic factors (Zhou et al., 2020).

Bivariate GREML is, in essence, an extension of the variance-covariance structure of the classical linear mixed model. Within the same statistical framework, the model simultaneously estimates the genetic variance and environmental variance of two traits, as well as the genetic covariance between them, from which the genetic correlation coefficient can be derived. Its validity depends on several key assumptions. First, the genetic effects of different traits should be representable by a common genomic relationship matrix (GRM). Second, the sample data should contain sufficient information to support effective identification of the covariance structure (Figure 2).

Ideally, the two traits should be measured in the same group of individuals or in highly overlapping samples, so that genetic and environmental effects can be partitioned within a unified reference framework. Adequate sample size is also particularly important for improving the precision of covariance estimation. In crop genetic

improvement research, this method has been widely used to elucidate the intrinsic relationships among yield, stress resistance, and quality traits, and it shows particular advantages in identifying potential trade-offs between traits (Derbyshire et al., 2024).

In practical applications, bivariate GREML models are highly sensitive to data quality and model specification. On the one hand, phenotypic measurement error can directly interfere with the estimation of variance and covariance components, thereby increasing the uncertainty of genetic correlation estimates. On the other hand, insufficient sample overlap or limited information on the covariance structure may also lead to unstable parameter estimation. Therefore, when interpreting results, particular attention should be paid to the standard errors and confidence intervals of genetic correlation coefficients, so as to avoid overinterpreting genetic correlations in situations where sample size is limited or trait correlations are mainly driven by environmental factors.

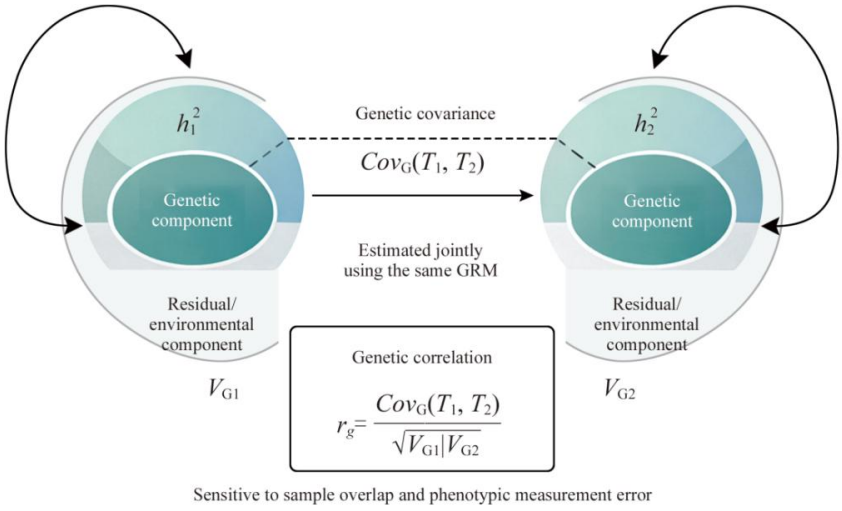


Figure 2 Schematic illustration of genetic covariance and genetic correlation in a bivariate GREML model

Note: Each trait is decomposed into a genetic component and a residual component. The bivariate GREML framework jointly estimates the additive genetic variances of Trait 1 and Trait 2, as well as their genetic covariance, using a common genome-wide relationship matrix (GRM). The genetic correlation r_g is derived from the estimated genetic covariance standardized by the square roots of the trait-specific genetic variances. This schematic emphasizes that genetic correlation reflects shared genetic architecture rather than phenotypic correlation, and its estimation is sensitive to sample overlap and phenotypic measurement error. The figure is illustrative and based on published GREML applications

6 Interpretation of Results and Common Pitfalls

6.1 Proper interpretation of "missing heritability"

This section clarifies the concept of "missing heritability" by focusing on the statistical comparability between SNP-based and pedigree-based heritability estimates. In studies based on GREML (genomic-relatedness-based restricted maximum likelihood) or SNP-derived heritability, a commonly observed phenomenon is that the proportion of phenotypic variance explained by genotyped SNPs is often substantially lower than heritability estimates derived from pedigree or twin studies (Speed et al., 2016; Evans et al., 2017; Yang et al., 2017; Wainschtein et al., 2022). This discrepancy should not be interpreted as evidence that the trait itself is weakly heritable, but rather as a reflection of differences in the identifiability of genetic variance under distinct statistical frameworks.

From a statistical genetics perspective, the systematic downward bias of SNP-based heritability primarily arises from the capturability constraints imposed by genotyping platforms, including marker coverage boundaries, heterogeneity in linkage disequilibrium (LD) structure, and the allele frequency (AF) spectrum of variants included in the analysis (Speed et al., 2016; Yang et al., 2017; Génin, 2019). Common SNP arrays provide strong tagging of common variants but have limited coverage of low-frequency, rare, and structural variants, which may contribute non-negligibly to total genetic variance (Speed et al., 2016; Wainschtein et al., 2022).

Moreover, when LD between causal variants and genotyped markers is weak, the effects of causal loci cannot be fully captured by tagging SNPs, leading to systematic underestimation of SNP-based heritability (Speed et al., 2012; 2016; Evans et al., 2017). The relationship between effect size distribution and the AF spectrum is also critical: when genetic contributions are driven primarily by rare variants or variants located in low-LD regions, the discrepancy between SNP-based and pedigree-based heritability is further amplified (Speed et al., 2016; Evans et al., 2017; Wainschtein et al., 2022).

6.1.1 Necessary conditions for comparing "heritability differences": from phenomenon to statistical framework

When discussing the discrepancy between pedigree-based heritability and SNP-based heritability, a frequently overlooked yet fundamental issue is whether such a comparison is statistically valid in the first place. These two types of estimates arise from distinct data structures and modeling frameworks; their differences are therefore not merely numerical deviations, but are embedded within their respective variance decomposition systems. Without strict alignment of underlying assumptions and conditions, the so-called "difference" often reflects only a superficial contrast between heterogeneous statistical objects, rather than an interpretable biological signal.

Consistency in phenotype definition constitutes the foundation of any meaningful comparison. A phenotype is not simply an observed variable; it directly embodies the variance structure subject to decomposition. Differences in measurement protocols, normalization procedures, or aggregation strategies across time points or traits can all alter the composition of phenotypic variance, thereby affecting both the numerator and denominator of heritability estimates. Once the phenotype definition shifts, even identical underlying genetic effects may yield systematically different estimates. As a result, comparisons lacking a unified phenotypic framework are unlikely to possess statistical interpretability.

The distribution of environmental factors and the structure of measurement error further define the reference frame for heritability estimation. Heritability is, by definition, the proportion of genetic variance relative to total phenotypic variance, and the environmental contribution to this total is highly dependent on the population context and study design. If studies differ substantially in environmental exposure, population composition, or sources of error, the decomposed variance components no longer belong to a common statistical population. Under such conditions, comparisons of heritability lose their foundation in a shared probability space.

In pedigree-based models, the treatment of shared environmental effects plays a critical role in identifying genetic variance. In twin or family studies, phenotypic similarity among related individuals arises from both genetic and shared environmental sources. If the model fails to adequately disentangle these components, part of the environmental effect may be misattributed to genetic variance, leading to systematic overestimation of pedigree heritability. This bias is structural rather than random, and often manifests as an apparent inflation of pedigree-based estimates relative to SNP-based heritability.

Finally, the coverage of the SNP marker system imposes a fundamental constraint on SNP-based heritability estimates. Estimates derived from genotyping arrays or sequencing data can only capture the genetic variation represented by observed markers and their linkage disequilibrium with causal variants. If low-frequency variants, rare variants, or structural variants are insufficiently represented, the corresponding genetic variance will be systematically missed. Therefore, even under correct model specification, SNP-based heritability cannot, in principle, reach the total level reflected by pedigree-based estimates.

6.1.2 Conceptual boundaries and interpretation of snp-based heritability

Within this analytical framework, the concept of SNP-based heritability requires a more precise definition. Rather than viewing it as a "lower-bound estimate" or a proxy for total trait heritability, it is more appropriately understood as the proportion of genetic variance captured by the observed SNP set under specific marker coverage and modeling assumptions-commonly referred to as "chip-capturable heritability." This definition highlights its conditional and tool-dependent nature, rather than treating it as a comprehensive representation of the genetic architecture of a trait.

Accordingly, interpreting SNP-based heritability derived from methods such as GREML as direct evidence of "low trait heritability" is not statistically justified. Such interpretations overlook the dependence of the estimate on marker coverage, linkage disequilibrium structure, and parametric modeling assumptions.

A more appropriate perspective is that SNP-based heritability reflects the joint explanatory capacity of multiple factors within a given analytical framework. First, the extent to which genomic markers cover true genetic variation determines the upper bound of observable genetic signal. Second, the structure of linkage disequilibrium governs whether causal variants can be effectively proxied by measured markers. Third, assumptions regarding allele frequency distributions and effect sizes further influence both the bias and variance of the estimate (Speed et al., 2016; Yang et al., 2017; Génin, 2019; Wainschtein et al., 2022).

6.2 Interpretation checklist: a standardized workflow for GREML-based SNP heritability

After obtaining an SNP-based heritability estimate within the GREML framework, a single numerical value alone does not provide sufficient explanatory power. Its statistical significance and biological interpretation both depend on the data-generating process, model specification, and stability of the estimation procedure. Therefore, a sound interpretation should not be limited to reporting the estimate itself, but should be grounded in a systematic evaluation of the entire analytical process. In other words, interpreting SNP heritability is a form of "conditional inference," whose validity depends on the consistency among data quality, model assumptions, and methodological suitability.

The foundation of result interpretation lies in the reliability of the data and the appropriateness of phenotypic modeling. The extent to which SNP markers cover genome-wide variation directly constrains the range of genetic variance that can be identified. In particular, when only common-variant genotyping array data are used, low-frequency variants, rare variants, and structural variants are not sufficiently captured. Their corresponding genetic contributions are therefore inevitably missed, leading to a systematic underestimation of SNP heritability, a phenomenon that has been clearly supported by large-scale sequencing studies (Wainschtein et al., 2019; 2022). Statistical processing of phenotypes is equally important. Phenotypes that have not been appropriately transformed or adjusted for systematic environmental factors often make effective variance decomposition difficult. In multi-environment or repeated-measure settings, if environmental heterogeneity is not explicitly modeled, part of the environmental effect may be incorrectly absorbed into the residual term, thereby weakening the ability to identify genetic variance (Evans et al., 2017; Yang et al., 2017).

The treatment of population structure and relatedness constitutes another important source of estimation bias. Systematic differences introduced by population stratification, together with correlation structures arising from cryptic relatedness, may distort heritability estimates if not adequately controlled, and the direction of such bias is not necessarily fixed. In individual-level analyses, correcting for principal components or constructing an appropriate mixed-model structure to absorb population-structure effects is a basic requirement for maintaining valid estimation. Meanwhile, the identification of close relatives and the thresholds used for their exclusion should also be subjected to sensitivity analysis, so as to avoid unstable inference caused by differences in sample structure. When individual-level data are unavailable, summary-statistics-based approaches, such as LDSC or SumHer, can serve as alternative strategies for robustly modeling population stratification and provide important references for interpreting GREML results (Ge et al., 2016; Speed et al., 2016; Speed and Balding, 2018; Speed et al., 2022).

The dependence of heritability estimation on the construction of the genomic relationship matrix (GRM) means that its interpretation must be situated within specific modeling assumptions. Because linkage disequilibrium (LD) patterns among SNPs are complex, failure to appropriately account for LD heterogeneity, or weak LD between genotyped markers and causal variants, may lead to systematic biases in different directions (Speed et al., 2012; 2016). In practice, a single standard GRM is often insufficient to fully characterize genetic architecture. Introducing LD correction or using stratified GRM models to partition SNPs by allele-frequency intervals or functional annotation categories can, to some extent, reduce model-specification bias and improve the resolution with which sources of genetic variance are interpreted. In addition, cross-checking results with frameworks that

are more sensitive to LD structure, such as SumHer, helps assess the degree to which the estimates depend on assumptions embedded in GRM construction (Speed and Balding, 2018; Speed et al., 2022).

However, even when model specification is appropriate, the statistical stability of the estimates still needs to be evaluated separately. The convergence of the REML algorithm, the magnitude of standard errors, and the width of confidence intervals are all key indicators for judging result reliability. In particular, when boundary solutions occur, such as genetic variance estimates approaching zero or reaching the upper bound of the parameter space, statistical explanations such as insufficient sample size or limited model information should be considered first, rather than assigning direct biological meaning to such results. For complex traits or studies with limited sample sizes, resampling methods such as jackknife or bootstrap can be used to evaluate estimation variability, and increasing sample size through multi-cohort joint analysis has also been shown to be an effective way to improve estimation precision (Evans et al., 2017; Wainschtein et al., 2022).

Given these multidimensional constraints, interpretation based on a single method is clearly limited. Cross-validating GREML estimates with other methods is therefore a key strategy for improving the robustness of conclusions in current research. Because different methods differ in how they capture genetic variance, their estimates for the same trait often show systematic deviations. Comparing individual-level GREML results with summary-statistics-based LDSC or SumHer estimates can help identify biases introduced by differences in data structure or model assumptions (Speed et al., 2016; Speed and Balding, 2018). Especially when SNP heritability is substantially lower than family-based heritability, the result should be interpreted comprehensively in terms of marker coverage, LD structure, non-additive genetic effects, and gene-environment interactions, rather than being simply attributed to methodological limitations or missing genetic information (Yang et al., 2017; Wainschtein et al., 2022).

In essence, SNP heritability estimated under the GREML framework is a quantitative expression of the "genetic variance identifiable under given data and model conditions." The interpretation of SNP heritability results should follow the standardized checklist shown in Supplementary Table S1. Only when data quality, model specification, statistical stability, and methodological consistency have all been adequately verified can the estimate serve as an important basis for understanding the genetic architecture of complex traits. Integrating statistical inference with the biological background of the trait and developing an interpretive pathway based on multiple lines of evidence has become a mainstream paradigm in contemporary statistical genetics.

7 Discussion

7.1 Implications of SNP-based heritability estimation for the "missing heritability" debate

The issue of "missing heritability" has long been a central debate in quantitative genetics and population genomics. Pedigree-based studies often report relatively high heritability estimates, whereas SNP-based approaches-such as GREML, which estimates genetic variance via the genome-wide relationship matrix-typically yield lower values (Evans et al., 2017; Yang et al., 2017). This discrepancy arises from multiple factors, including the limited capture of low-frequency and rare variants, incomplete linkage disequilibrium (LD) between genotyped markers and causal mutations, the omission of epistatic interactions, and the cumulative effects of numerous small-effect alleles underlying highly polygenic traits (Hou et al., 2019; Holland et al., 2020). In recent years, with the increasing use of whole-genome sequencing and the development of more refined LD-aware modeling approaches, this gap has narrowed to some extent. However, for highly polygenic traits, a portion of heritability remains unexplained (Evans et al., 2017; Hou et al., 2019).

Importantly, SNP-based heritability should not be interpreted simply as an underestimate of the true heritability, but rather as a quantitative characterization of the variance explained by the observed set of markers (Yang et al., 2017). This perspective has prompted a conceptual shift in how heritability is defined: the issue is not whether heritability is truly "missing," but whether the association between genotyped markers and causal variants is incomplete. Consequently, SNP-based heritability serves as an important indicator of the capture efficiency of genotyping platforms and provides a theoretical basis for designing higher-density genotyping strategies and improving the dissection of complex traits.

7.2 Practical implications for plant breeding

For plant breeding, SNP-based heritability estimation has substantial practical relevance. First, it provides a quantitative basis for assessing the predictability of complex quantitative traits. A high SNP-based heritability estimate suggests that the major genetic components of a trait are effectively captured by existing genotyping markers, indicating that genomic selection (GS) models are likely to achieve high predictive accuracy for that trait (Schmidt et al., 2019; Zhu and Zhou, 2020). Conversely, a low estimate implies that a significant portion of genetic variation remains unexplained, highlighting the need for increased marker density, incorporation of rare variants, or improved modeling of gene-environment interactions (Zhu and Zhou, 2020).

Second, SNP-based heritability provides valuable guidance for population design and resource allocation. In major crops such as rice, maize, and wheat, factors such as population size, genetic background, and sample representativeness significantly influence the stability of heritability estimates. By applying GREML in early-stage populations, breeders can rapidly assess whether it is necessary to increase sample size, optimize crossing schemes, or adjust selection strategies for specific traits (Schmidt et al., 2019; Holland et al., 2020). Furthermore, partitioning heritability by functional annotation or allele frequency enables breeders to identify genomic regions or variant classes that should be prioritized for improvement, thereby enhancing selection efficiency (Weissbrod et al., 2019; Zhu and Zhou, 2020).

7.3 Integration with PRS, fine-mapping, and downstream methods

Unlike early marker-assisted breeding approaches centered on QTL mapping and candidate genes (Fang et al., 2001), the GCTA/GREML framework focuses on the genome-wide proportion of genetic variance captured by markers. Its results are therefore more suitable for evaluating the predictive limits of traits, rather than directly identifying functional loci. The value of GREML lies not only in heritability estimation itself, but also in its integrative role within the broader analytical pipeline.

First, GREML is closely related to polygenic risk scores (PRS). SNP-based heritability provides a theoretical upper bound for PRS prediction accuracy. Specifically, if a trait has low SNP-based heritability, improvements in model complexity alone cannot overcome this fundamental limitation (Yang et al., 2017; Zhang et al., 2018; Wang et al., 2023). Recent studies have demonstrated that incorporating functional annotations and LD structure, as well as accounting for uncertainty in individual-level risk estimation, can substantially improve PRS performance (Weissbrod et al., 2019; Ding et al., 2021).

Second, the variance decomposition framework of GREML is highly compatible with fine-mapping approaches. By partitioning heritability across chromosomes, functional annotations, or specific gene sets, GREML can provide prioritization for identifying causal variants, thereby improving both the resolution and biological interpretability of fine-mapping results (Weissbrod et al., 2019; Gazal et al., 2022).

Taken together, GREML is not merely a tool for heritability estimation, but a methodological bridge linking variant discovery, statistical inference, and functional interpretation, thereby offering a systematic framework for future precision breeding and molecular improvement.

8 Conclusion

The development of GCTA and GREML has established a standardized framework for estimating the heritability of complex traits using genome-wide SNP data. Unlike traditional pedigree-based approaches, these methods construct a genomic relationship matrix (GRM) and decompose phenotypic variance within a linear mixed model framework, enabling robust heritability estimation in natural or breeding populations even in the absence of complete pedigree information. This framework represents a fundamental transition from classical quantitative genetics to genotype-driven modern molecular genetics, and demonstrates strong scalability and practical utility as population sizes continue to increase. Furthermore, GCTA/GREML allows heritability to be partitioned by chromosomal segments, functional annotations, or genomic regions, thereby providing a more biologically informative perspective on the genetic architecture of complex traits.

However, it is essential to recognize that GCTA/GREML estimates rely on a set of statistical assumptions and boundary conditions. First, these methods typically assume that SNP effects follow a multivariate normal distribution and primarily focus on additive genetic variance, with limited consideration of dominance and epistatic effects. Second, the heritability estimates obtained from GCTA/GREML reflect only the variance captured by genotyped or imputed markers, and are therefore influenced by marker density, allele frequency distribution, and the extent of linkage disequilibrium (LD) with causal variants. Consequently, such estimates should not be interpreted as the "true heritability" of a trait, but rather as conditional estimates based on observable genomic variation. Ignoring these underlying assumptions may lead to overinterpretation-for example, incorrectly attributing "missing heritability" to methodological limitations rather than to inherent constraints in data coverage and population characteristics.

In crop genetic improvement and molecular breeding, GCTA and GREML also demonstrate substantial practical value. On the one hand, they enable the characterization of the molecular genetic architecture of complex traits, providing a theoretical basis for quantitative trait locus (QTL) discovery and genomic selection model development. On the other hand, by comparing heritability estimates across traits or environmental conditions, it is possible to identify traits that are highly sensitive to environmental variation, thereby informing precision breeding strategies. In major crops such as rice, maize, and wheat, numerous empirical studies have demonstrated that GREML can effectively distinguish between the selectable and non-selectable components of trait variation, offering critical guidance for breeding target definition and resource allocation.

Overall, the GCTA and GREML family of methods have not only transformed the paradigm of heritability research in quantitative genetics, but have also provided practical tools for dissecting complex traits and advancing molecular breeding. Looking forward, with the continued development of population-scale sequencing, rare variant detection, and large-scale multi-environment datasets, GREML-based heritability estimation is expected to become increasingly refined and comprehensive. This progress will further enhance its role in elucidating the genetic basis of complex traits and in guiding genome-based breeding strategies.

Author Contributions

Xuanjun Fang conducted this study, including literature review, data analysis, and the drafting and revision of the manuscript. The author has read and approved the final version of the manuscript.

Acknowledgements

This work was supported by a Major Program of the National Natural Science Foundation of China (Grant No. 30490254).

References

Abney M., McPeek M., and Ober C., 2001, Broad and narrow heritabilities of quantitative traits in a founder population, American Journal of Human Genetics, 68(5): 1302-1307.
https://doi.org/10.1086/320112

Bérénos C., Ellis P., Pilkington J., and Pemberton J., 2014, Estimating quantitative genetic parameters in wild populations, Molecular Ecology, 23: 3434-3451.
https://doi.org/10.1111/mec.12827

Berry D., 2024, Many farmers want a prediction of future performance, Journal of Animal Science, 102: 51.
https://doi.org/10.1093/jas/skae234.056

Bilton T., Sharma S., Schofield M., Black M., Jacobs J., Bryan G., and Dodds K., 2024, Construction of relatedness matrices in autopolyploid populations using low-depth high-throughput sequencing data, Theoretical and Applied Genetics, 64: 137.
https://doi.org/10.1007/s00122-024-04568-2

Border R., and Becker S., 2019, Stochastic Lanczos estimation of genomic variance components for linear mixed-effects models, BMC Bioinformatics, 20: 295.
https://doi.org/10.1186/s12859-019-2978-z

Cesarani A., Pocrnić I., Macciotta N., Fragomeni B., Misztal I., and Lourenco D., 2018, Bias in heritability estimates from genomic restricted maximum likelihood methods under different genotyping strategies, Journal of Animal Breeding and Genetics, 136: 40-50.
https://doi.org/10.1111/jbg.12367

Chen X., Kuja-Halkola R., Rahman I., Arpegård J., Viktorin A., Karlsson R., Hägg S., Svensson P., Pedersen N.L., and Magnusson P.K., 2015, Dominant genetic variation and missing heritability, American Journal of Human Genetics, 97(5): 708-714.
https://doi.org/10.1016/j.ajhg.2015.10.004

Da Y., Wang C., Wang S., and Hu G., 2014, Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers, PLoS One, 9(1): e87666.
https://doi.org/10.1371/journal.pone.0087666

Dao C., Jiang J., Paul D., and Zhao H., 2021, Variance estimation and confidence intervals from genome-wide association studies through high-dimensional misspecified mixed model analysis, Journal of Statistical Planning and Inference, 220: 15-23.
https://doi.org/10.1016/j.jspi.2022.01.003

Derbyshire M.C., Newman T.E., Thomas W.J., Batley J., and Edwards D., 2024, The complex relationship between disease resistance and yield in crops, Plant Biotechnology Journal, 22: 2612-2623.
https://doi.org/10.1111/pbi.14373

Ding Y., Hou K., Burch K., Lapinska S., Privé F., Vilhjálmsson B., Sankararaman S., and Pasaniuc B., 2021, Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification, Nature Genetics, 54: 30-39.
https://doi.org/10.1038/s41588-021-00961-5

Evans L.M., Tahmasbi R., Vrieze S.I., Abecasis G.R., Das S., Gazal S., Bjelland D.W., de Candia T.R., Goddard M.E., Neale B.M., Yang J., Visscher P.M., and Keller M.C., 2017, Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits, Nature Genetics, 50: 737-745.
https://doi.org/10.1038/s41588-018-0108-x

Fang X.J., and Wu W.R., 2026, Evolution of statistical genetic paradigms: from linkage analysis and candidate gene strategies to GWAS, Molecular Plant Breeding, 24(9): 2817-2829.

Fang X.J., Wu W.R., and Tang J.L., (eds.), 2001, Crop DNA marker-assisted breeding, Science Press, Beijing, China, pp.1-84.

Finucane H.K., Bulik-Sullivan B., Gusev A., Trynka G., Reshef Y., Loh P.R., Anttila V., Xu H., Zang C.Z., Farh K., Ripke S., Day F.R., Consortium R., Schizophrenia Working Group of the Psychiatric Genomics Consortium, The RACI Consortium, Purcell S., Stahl E., Lindstrom S., Perry J.R.B., Okada Y., Raychaudhuri S., Daly M.J., Patterson N., Neale B.M., and Price A.L., 2015, Partitioning heritability by functional annotation using genome-wide association summary statistics, Nature Genetics, 47: 1228-1235.
https://doi.org/10.1038/ng.3404

Forni S., Aguilar I., and Misztal I., 2011, Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information, Genetics Selection Evolution, 43: 1.
https://doi.org/10.1186/1297-9686-43-1

Gazal S., Loh P.R., Finucane H.K., Ganna A., Schoech, A., Sunyaev, S., and Price A.L., 2018, Functional architecture of low-frequency variants highlights strength of negative selection across coding and noncoding annotations, Nature Genetics, 50: 1600-1607.
https://doi.org/10.1038/s41588-018-0231-8

Gazal S., Weissbrod O., Hormozdiari F., Dey K., Nasser J., Jagadeesh K., Weiner D., Shi H., Fulco C., O'Connor L., Pasaniuc B., Engreitz J., and Price A., 2022, Combining SNP-to-gene linking strategies to identify disease genes and assess disease omnigenicity, Nature Genetics, 54: 827-836.
https://doi.org/10.1038/s41588-022-01087-y

Ge T., Chen C., Neale B., Sabuncu M., and Smoller J., 2016. Phenome-wide heritability analysis of the UK Biobank. PLoS Genetics, 13. https://doi.org/10.1371/journal.pgen.1006711
https://doi.org/10.1371/journal.pgen.1006711

Génin E., 2019, Missing heritability of complex diseases: case solved? Human Genetics, 139: 103-113.
https://doi.org/10.1007/s00439-019-02034-4

Granato Í., Galli G., De Oliveira Couto E., Souza M., Mendonça L., and Fritsche-Neto R., 2018, snpReady: a tool to assist breeders in genomic analysis, Molecular Breeding, 38: 84.
https://doi.org/10.1007/s11032-018-0844-8

Holland D., Frei O., Fan C., Shadrin A., Smeland O., Sundar V., Thompson P., Andreassen O., and Dale A., 2020, Beyond SNP heritability: Polygenicity and discoverability of phenotypes estimated with a univariate Gaussian mixture model, PLoS Genetics, 16(5): e1008612.
https://doi.org/10.1371/journal.pgen.1008612

Hou K., Burch K., Majumdar A., Shi H., Mancuso N., Wu Y., Sankararaman S., and Pasaniuc B., 2019, Accurate estimation of SNP-heritability from biobank-scale data irrespective of genetic architecture, Nature Genetics, 51: 1244-1251.
https://doi.org/10.1038/s41588-019-0465-0

Jang S.K., Evans L., Fialkowski A., Arnett D.K., Ashley-Koch A.E., Barnes K.C., Becker D.M., Bis J.C., Blangero J., Bleecker E.R., Boorgula M.P., Bowden D.W., Brody J.A., Cade B.E., Jenkins B.W.C., Carson A.P., Chavan S., Cupples L.A., Custer B., Damrauer S.M., David S.P., de Andrade M., Dinardo C.L., Fingerlin T.E., Fornage M., Freedman B.I., Garrett M.E., Gharib S.A., Glahn D.C., Haessler J., Heckbert S.R., Hokanson J.E., Hou L.F., Hwang S.J., Hyman M.C., Judy R., Justice A.E., Kaplan R.C., Kardia S.L.R., Kelly S., Kim W., Kooperberg C., Levy D., Lloyd-Jones D.M., Loos R.J.F., Manichaikul A.W., Gladwin M.T., Martin L.W., Nouraie M., Melander O., Meyers D.A., Montgomery C.G., North K.E., Oelsner E.C., Palmer N.D., Payton M., Peljto A.L., Peyser P.A., Preuss M., Psaty B.M., Qiao D.D., Rader D.J., Rafaels N., Redline S., Reed R.M., Reiner A.P., Rich S.S., Rotter J.I., Schwartz D.A., Shadyab A.H., Silverman E.K., Smith N.L., Smith J.G., Smith A.V., Smith J.A., Tang W.H., Taylor K.D., Telen M.J., Vasan R.S., Gordeuk V.R., Wang Z., Wiggins K.L., Yanek L.R., Yang I.V., Young K.A., Young K.L., Zhang Y.Z., Liu D.J.J., Keller M.C., and Vrieze S. 2022, Rare genetic variants explain missing heritability, Nature Human Behaviour, 6: 1577-1586.
https://doi.org/10.1038/s41562-022-01408-5

Kumar K., Feldman M. W., Rehkopf D. H., and Tuljapurkar S. (2015). Limitations of GCTA as a solution to the missing heritability problem, Proceedings of the National Academy of Sciences of the United States of America, 113(1): E61-E70.
https://doi.org/10.1073/pnas.1520109113

Mathew B., Léon J., and Sillanpää M.J., 2017, A novel linkage-disequilibrium corrected genomic relationship matrix for SNP-heritability estimation and genomic prediction, Heredity, 120: 356-368.
https://doi.org/10.1038/s41437-017-0023-4

Meyer K., 2023, Reducing computational demands of restricted maximum likelihood estimation with genomic relationship matrices, Genetics Selection Evolution (GSE), 55: 19.
https://doi.org/10.1186/s12711-023-00781-7

Ni G., Moser G., Wray N., and Lee S., 2018, Estimation of genetic correlation via linkage disequilibrium score regression and genomic restricted maximum likelihood, American Journal of Human Genetics, 102(6): 1185-1194.
https://doi.org/10.1101/194019

Nolte I., Jansweijer J., Riese H., Asselbergs F., Van Der Harst P., Spector T., Pinto Y., Snieder H., and Jamshidi Y., 2017, A comparison of heritability estimates by classical twin modeling and based on genome-wide genetic relatedness for cardiac conduction traits, Twin Research and Human Genetics, 20: 489-498.
https://doi.org/10.1017/thg.2017.55

Schmidt P., Hartung J., Bennewitz J., and Piepho H., 2019, Heritability in plant breeding on a genotype-difference basis, Genetics, 212: 991-1008.
https://doi.org/10.1534/genetics.119.302134

Speed D., and Balding D., 2018, SumHer better estimates the SNP heritability of complex traits from summary statistics, Nature genetics, 51: 277-284.
https://doi.org/10.1038/s41588-018-0279-5

Speed D., Cai N., Johnson M.R., Nejentsev S., and Balding D.J., 2016, Re-evaluation of SNP heritability in complex human traits, Nature Genetics, 49: 986-992.
https://doi.org/10.1038/ng.3865

Speed D., Hemani G., Johnson M.R., and Balding D.J., 2012, Improved heritability estimation from genome-wide SNPs, American Journal of Human Genetics, 91(6): 1011-1021.
https://doi.org/10.1016/j.ajhg.2012.10.010

Speed D., Kaphle A., and Balding D., 2022, SNP-based heritability and selection analyses: Improved models and new results, BioEssays, 44(5): 2100170.
https://doi.org/10.1002/bies.202100170

Srivastava A., Williams S., and Zhang G., 2023, Heritability estimation approaches utilizing genome-wide data, Current Protocols, 3: e734.
https://doi.org/10.1002/cpz1.734

Strandén I., Mäntysaari E., Lidauer M., Thompson R., and Gao H., 2024, A computationally efficient algorithm to leverage average information REML for (co)variance component estimation in the genomic era, Genetics Selection Evolution (GSE), 56: 18.
https://doi.org/10.1186/s12711-024-00939-x

Tang M., Wang T., and Zhang X., 2022, A review of SNP heritability estimation methods, Briefings in Bioinformatics, 23(3): bbac067.
https://doi.org/10.1093/bib/bbac067

Van Den Berg S., Vandenplas J., Van Eeuwijk F., Lopes M., and Veerkamp R., 2019, Significance testing and genomic inflation factor using high-density genotypes or whole-genome sequence data, Journal of Animal Breeding and Genetics, 136: 418-429.
https://doi.org/10.1111/jbg.12419

Vinkhuyzen A.A., Wray N.R., Yang J., Goddard M.E., and Visscher P.M., 2013, Estimation and partition of heritability in human populations using whole-genome analysis methods, Annual Review of Genetics, 47(1): 75-95.
https://doi.org/10.1146/annurev-genet-111212-133258

Wainschtein P., Jain D., Zheng Z., Aslibekyan S., Becker D., Bi W., Brody J., Carlson J., Correa A., Du M., Fernández-Rhodes L., Ferrier K., Graff M., Guo X., He J., Heard-Costa N., Highland H., Hirschhorn J., Howard-Claudio C., Isasi C., Jackson R., Jiang J., Joehanes R., Justice A., Kalyani R., Kardia S., Lange E., LeBoff M., Lee S., Li X., Li Z., Lim E., Lin D., Lin X., Liu S., Lu Y., Manson J., Martin L., McHugh C., Mikulla J., Musani S., Ng M., Nickerson D., Palmer N., Perry J., Peters U., Preuss M., Qi Q., Raffield L., Rasmussen-Torvik L., Reiner A., Russell E., Sitlani C., Smith J., Spracklen C., Wang T., Wang Z., Wessel J., Xu H., Yaser M., Yoneyama S., Young K., Zhang J., Zhang X., Zhou H., Zhu X., Zoellner S., Abe N., Abecasis G., Aguet F., Almasy L., Alonso Á., Ament S., Anderson P., Anugu P., Applebaum-Bowden D., Ardlie K., Arking D., Ashley-Koch A., Assimes T., Auer P., Avramopoulos D., Ayas N., Balasubramanian A., Barnard J., Barnes K., Barr R., Barron-Casella E., Barwick L., Beaty T., Beck G., Becker L., Beer R., Beitelshees A., Benjamin E., Benos T., Bezerra M., Bielak L., Bis J., Blackwell T., Blangero J., Bowden D., Bowler R., Broeckel U., Broome J., Brown D., Bunting K., Burchard E., Bustamante C., Buth E., Cade B., Cardwell J., Carey V., Carrier J., Carson A., Carty C., Casaburi R., Romero J., Casella J., Castaldi P., Chaffin M., Chang C., Chang Y., Chavan S., Chen B., Chen W., Cho M., Choi S., Chuang L., Chung R., Clish C., Comhair S., Conomos M., Cornell E., Crandall C., Crapo J., Curran J., Curtis J., Custer B., Damcott C., Darbar D., David S., Davis C., Daya M., De Las Fuentes L., De Vries P., DeBaun M., Deka R., Demeo D., Devine S., Dinh H., Doddapaneni H., Duan Q., Dugan-Perez S., Duggirala R., Durda J., Dutcher S., Eaton C., Ekunwe L., Boueiz E., Emery L., Erzurum S., Farber C., Farek J., Fingerlin T., Flickinger M., Franceschini N., Frazar C., Fu M., Fullerton S., Fulton L., Gabriel S., Gan W., Gao S., Gao Y., Gass M., Geiger H., Gelb B., Geraci M., Germer S., Gerszten R., Ghosh A., Gibbs, R., Gignoux C., Gladwin M., Glahn D., Gogarten S., Gong D., Goring H., Graw S., Gray K., Grine D., Gross C., Gu C., Guan Y., Gupta N., Haas D., Haessler J., Hall M., Han Y., Hanly P., Harris D., Hawley N., Heavner B., Herrington D., Hersh C., Hidalgo B., Hixson J., Hobbs B., Hokanson J., Hong E., Hoth K., Hsiung C., Hu J., Hung Y., Huston H., Hwu C., Irvin M., Jaquish C., Johnsen J., Johnson A., Johnson C., Johnston R., Jones K., Kang H., Kaplan R., Kelly S., Kenny E., Kessler M., Khan A., Khan Z., Kim W., Kimoff J., Kinney G., Konkle B., Kramer H., Lange C., Lee J., Lee S., Lee W., Lefaive J., Levine D., Levy D., Lewis J., Li Y., Lin H., Lin H., Liu Y., Lunetta K., Luo

J., Magalang U., Mahaney M., Make B., Manichaikul A., Manning A., Marton M., Mathai S., May S., McArdle P., McFarland S., McGoldrick D., McNeil B., Mei H., Meigs J., Menon V., Mestroni L., Metcalf G., Meyers D., Mignot E., Min N., Minear M., Minster R., Moll M., Momin Z., Montasser M., Montgomery C., Muzny D., Mychaleckyj J., Nadkarni G., Naik R., Naseri T., Natarajan P., Nekhai S., Nelson S., Neltner B., Nessner C., Nkechinyere O., O'Connor T., Ochs-Balcom H., Okwuonu G., Pack A., Paik D., Pankow J., Papanicolaou G., Parker C., Peloso G., Peralta J., Perez M., Peyser P., Phillips L., Pleiness J., Pollin T., Post W., Becker J., Boorgula M., Qasba P., Qiao D., Qin Z., Rafaels N., Rajendran M., Rao D., Ratan A., Reed R., Reeves C., Reupena M., Rice K., Robillard R., Robine N., Roselli C., Ruczinski I., Runnels A., Russell P., Ruuska S., Ryan K., Sabino E., Saleheen D., Salimi S., Salvi S., Salzberg S., Sandow K., Sankaran V., Santibanez J., Schwander K., Schwartz D., Sciurba F., Seidman C., Seidman J., Sheehan V., Sherman S., Shetty A., Shetty A., Sheu W., Silver B., Silverman E., Skomro R., Smith A., Smith T., Smoller S., Snively B., Snyder M., Sofer T., Sotoodehnia N., Stilp A., Storm G., Streeten E., Su J., Sung Y., Sylvia J., Szpiro A., Taliun D., Tang H., Taub M., Taylor K., Taylor M., Taylor S., Telen M., Thornton T., Threlkeld M., Tinker L., Tirschwell D., Tishkoff S., Tiwari H., Tong C., Tracy R., Tsai M., Vaidya D., Van Den Berg D., Vandehaar P., Vrieze S., Walker T., Wallace R., Walts A., Wang F., Wang H., Wang J., Watson K., Watt J., Weeks D., Weinstock J., Weiss S., Weng L., Willer C., Williams K., Williams L., Wilson C., Wilson J., Winterkorn L., Wong Q., Wu J., Yang I., Yu K., Zekavat S., Zhang Y., Zhao S., Zhao W., Zody M., Cupples L., Shadyab A., McKnight B., Shoemaker B., Mitchell B., Psaty B., Kooperberg C., Liu C., Albert C., Roden D., Chasman D., Darbar D., Lloyd‐Jones D., Arnett D., Regan E., Boerwinkle E., Rotter J., O'Connell J., Yanek L., De Andrade M., Allison M., McDonald M., Chung M., Fornage M., Chami N., Smith N., Ellinor P., Vasan R., Mathias R., Loos R., Rich S., Lubitz S., Heckbert S., Redline S., Guo X., Chen Y., Laurie C., Hernandez R., McGarvey S., Goddard M., Laurie C., North K., Lange L., Weir B., Yengo L., Yang J., and Visscher P., 2022, Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data, Nature Genetics, 54: 263-273.

Wang S., Wei Y., Liu D., Zhang X., Wang Q., Pan Y., and Ma P., 2025, Impact of different genomic relationship matrix construction methods on the accuracy of genomic prediction in different species, Frontiers in Genetics, 16: 1576248.
https://doi.org/10.3389/fgene.2025.1576248

Wang X., Walker A., Revez J., Ni G., Adams M., McIntosh A., Visscher P., Wray N., Ripke S., Mattheisen M., Trzaskowski M., Byrne E., Abdellaoui A., Agerbo E., Air T., Andlauer T., Bacanu S., Bækvad-Hansen M., Beekman A., Bigdeli T., Binder E., Bryois J., Buttenschøn H., Bybjerg-Grauholm J., Cai N., Christensen J., Clarke T., Coleman J., Colodro-Conde L., Couvy-Duchesne B., Craddock N., Crawford G., Davies G., Degenhardt F., Derks E., Direk N., Dolan C., Dunn E., Eley T., Escott-Price V., Kiadeh F., Finucane H., Foo J., Forstner A., Frank J., Gaspar H., Gill M., Goes F., Gordon S., Grove J., Hall L., Hansen C., Hansen T., Herms S., Hickie I., Hoffmann P., Homuth G., Horn C., Hottenga J., Hougaard D., Howard D., Ising M., Jansen R., Jones I., Jones L., Jorgenson E., Knowles J., Kohane I., Kraft J., Kretzschmar W., Kutalik Z., Li Y., Lind P., Macintyre D., MacKinnon D., Maier R., Maier W., Marchini J., Mbarek H., McGrath P., McGuffin P., Medland S., Mehta D., Middeldorp C., Mihailov E., Milaneschi Y., Milani L., Mondimore F., Montgomery G., Mostafavi S., Mullins N., Nauck M., Ng B., Nivard M., Nyholt D., O'Reilly P., Oskarsson H., Owen M., Painter J., Pedersen B., Pedersen M., Peterson R., Peyrot W., Pistis G., Posthuma D., Quiroz J., Qvist P., Rice J., Riley B., Rivera M., Mirza S., Schoevers R., Schulte E., Shen L., Shi J., Shyn S., Sigurdsson E., Sinnamon G., Smit J., Smith D., Stefánsson H., Steinberg S., Streit F., Strohmaier J., Tansey K., Teismann H., Teumer A., Thompson W., Thomson P., Thorgeirsson T., Traylor M., Treutlein J., Trubetskoy V., Uitterlinden A., Umbricht D., Van Der Auwera S., Van Hemert A., Viktorin A., Wang Y., Webb B., Weinsheimer S., Wellmann J., Willemsen G., Witt S., Wu Y., Xi H., Yang J., Zhang F., Arolt V., Baune B., Berger K., Boomsma D., Cichon S., Dannlowski U., De Geus E., DePaulo J., Domenici E., Domschke K., Esko T., Grabe H., Hamilton S., Hayward C., Heath A., Kendler K., Kloiber S., Lewis G., Li Q., Lucae S., Madden P., Magnusson P., Martin N., Metspalu A., Mors O., Mortensen P., Müller-Myhsok B., Nordentoft M., Nöthen M., O'Donovan M., Paciga S., Pedersen N., Penninx B., Perlis R., Porteous D., Potash J., Preisig M., Rietschel M., Schaefer C., Schulze T., Smoller J., Stefansson K., Tiemeier H., Uher R., Völzke H., Weissman M., Werge T., Lewis C., Levinson D., Breen G., Børglum A., and Sullivan P., 2023, Polygenic risk prediction: why and when out-of-sample prediction R2 can exceed SNP-based heritability, American Journal of Human Genetics, 110(7): 1207-1215.
https://doi.org/10.1016/j.ajhg.2023.06.006

Wei J., Xie W., Li R., Wang S., Qu H., Ma R., Zhou X., and Jia Z., 2020, Analysis of trait heritability in functionally partitioned rice genomes, Heredity, 124(3): 485-498.
https://doi.org/10.1038/s41437-019-0244-9

Weissbrod O., Hormozdiari F., Benner C., Cui R., Ulirsch J., Gazal S., Schoech A., Van De Geijn B., Reshef Y., Márquez-Luna C., O'Connor L., Pirinen M., Finucane H., and Price A., 2019, Functionally-informed fine-mapping and polygenic localization of complex trait heritability, Nature Genetics, 52: 1355-1363.
https://doi.org/10.1038/s41588-020-00735-5

Yang J., Bakshi A., Zhu Z., Hemani G., Vinkhuyzen A.A.E., Lee S.H., Robinson M.R., Perry J.R.B., Nolte I.M., van Vliet-Ostaptchouk J.V., Snieder H., Esko T., Milani L., Mägi R., Metspalu A., Hamsten A., Magnusson P.K.E., Pedersen N.L., Ingelsson E., Soranzo N., Keller M.C., Wray N.R., Goddard M.E., Visscher P.M., 2015, Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index, Nature Genetics, 47: 1114-1120.
https://doi.org/10.1038/ng.3390

Yang J., Lee S., Wray N., Goddard M., and Visscher P., 2016, GCTA-GREML accounts for linkage disequilibrium when estimating genetic variance from genome-wide SNPs, Proceedings of the National Academy of Sciences (PNAS), 113(32): E4579-E4580.
https://doi.org/10.1073/pnas.1602743113

Yang J., Lee S.H., Goddard M.E., and Visscher P.M., 2011, GCTA: A tool for genome-wide complex trait analysis, American Journal of Human Genetics, 88(1): 76-82.
https://doi.org/10.1016/j.ajhg.2010.11.011

Yang J., Zeng J., Goddard M.E., Wray N.R., and Visscher P.M., 2017, Concepts, estimation and interpretation of SNP-based heritability, Nature Genetics, 49: 1304-1310.
https://doi.org/10.1038/ng.3941

Zhang Y., Qi G., Park J., and Chatterjee N., 2018, Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits, Nature Genetics, 50: 1318-1326.
https://doi.org/10.1038/s41588-018-0193-x

Zhou X., Im H., and Lee S., 2019, CORE GREML: Estimating covariance between random effects in linear mixed models for genomic analyses of complex traits, bioRxiv, 2019: 853515.
https://doi.org/10.1101/853515

Zhou X., Im H., and Lee S., 2020, CORE GREML for estimating covariance between random effects in linear mixed models for complex trait analyses, Nature Communications, 11: 4203.
https://doi.org/10.1038/s41467-020-18085-5

Zhu H., and Zhou X., 2020, Statistical methods for SNP heritability estimation and partition: A review, Computational and Structural Biotechnology Journal, 18: 1557-1568.
https://doi.org/10.1016/j.csbj.2020.06.011

Zhu Z., Bakshi A., Vinkhuyzen A.A., Hemani G., Lee S.H., Nolte I.M., van Vliet-Ostaptchouk J.V., Snieder H., The LifeLines Cohort Study, Esko T., Milani L., Magi R., Metspalu A., Hill W.G., Weir B.S., Goddard M.E., Visscher P.M., and Yang J., 2015, Dominance genetic variation contributes little to the missing heritability for human complex traits, The American Journal of Human Genetics, 96(3): 377-385.
https://doi.org/10.1016/j.ajhg.2015.01.001

Computational Molecular Biology 2026, Vol.16, No.3, 159-180
http://bioscipublisher.com/index.php/cmb

Appendix A. Checklist for Interpreting SNP-based Heritability Estimates under the GREML Framework

This checklist is intended to standardize the interpretation workflow of SNP-based heritability estimates derived from the GREML framework, emphasizing their dependence on data quality, model specification, and underlying statistical assumptions. Researchers may use this checklist to systematically verify each step of the analysis, thereby improving the transparency and reproducibility of inference.

Table S1 Interpretation of SNP heritability estimates followed a standardized checklist

No.	Check Dimension	Key Items to Check	Completion Status
1	Genotype quality control and variant spectrum	Whether rigorous genotype quality control has been conducted; whether the SNP set adequately covers the allele frequency spectrum, particularly low-frequency and rare variants; and whether limited coverage is expected to result in downward-biased SNP heritability estimates.	☐
2	Phenotype modeling and covariate adjustment	Whether the phenotype distribution has been examined and appropriate transformations applied; whether batch effects, environmental covariates, and other key fixed effects have been included in the model.	☐
3	Repeated measures and multi-environment structure	For phenotypes measured across multiple time points or environments, whether multi-environment or hierarchical mixed models have been adopted to avoid inflation of residual variance.	☐
4	Population structure control	Whether population stratification has been assessed and adjusted for using principal components or equivalent approaches; and whether the estimates are robust to the number of PCs included.	☐
5	Relatedness filtering and kinship threshold setting	Whether criteria for removing close relatives have been clearly defined; and whether the stability of SNP heritability estimates has been evaluated under alternative relatedness thresholds.	☐
6	GRM construction and LD sensitivity	Whether results obtained using the standard GRM have been compared with those from LD-adjusted or partitioned GRM models; and whether sensitivity to linkage disequilibrium assumptions has been assessed.	☐
7	REML convergence and boundary estimates	Whether REML optimization has converged; whether standard errors and confidence intervals have been reported; and whether boundary estimates (e.g., $h^2 = 0$or $h^2 = 1$) are interpreted as indicators of limited information rather than definitive biological conclusions.	☐
8	Estimation stability and sample size adequacy	Whether estimation stability has been evaluated using standard errors; whether resampling approaches such as jackknife or bootstrap have been applied when feasible; and whether the sample size is adequate for reliable variance component estimation.	☐
9	Cross-method validation	Whether GREML-based estimates have been compared with results from summary-statistic methods such as LDSC or SumHer.	☐
10	Integrated interpretation and boundary awareness	Whether SNP heritability estimates are interpreted in the context of marker coverage, model assumptions, and trait biology; and whether SNP heritability is not equated with the trait's "true" heritability.	☐

Appendix B. Comparative Summary of Statistical Objectives, Assumptions, and Applicability of GREML Extensions and Variants

Supplementary Table 2 summarizes the differences among commonly used extensions of the GREML framework in terms of their statistical objectives, core assumptions, applicable data structures, and key diagnostic considerations. This comparison table is intended to provide guidance for selecting appropriate extended models under different research scenarios.

Table S2 Statistical Objectives, Assumptions, and Applicability of GREML Extensions and Variants

Method extension	Primary statistical issue addressed	Core assumptions	Typical applicable data structures	Recommended diagnostics and cautions
LOCO (Leave-One-Chromosome-Out)	Proximal contamination, leading to inflation of local effects and overestimation of variance components	Genetic contributions from different chromosomes are approximately independent; excluding the target chromosome does not substantially compromise estimation of genome-wide background effects	Genomes with a limited number of chromosomes and large LD blocks; scenarios with high-effect loci; commonly applied in crop genomic datasets	Compare results from standard GRM and LOCO GRM; LOCO is designed primarily to address proximal contamination and should not be used as a general correction for population structure or LD heterogeneity
Functional partitioning of heritability (multiple GRMs)	Heterogeneous distribution of genetic variance across genomic functional regions; limited biological interpretability of aggregate heritability	SNP effect-size distributions and LD patterns differ systematically across functional annotations and can be identified through partitioned modeling	Large sample sizes; reliable functional annotations; sufficient SNP density within each category; suitable for enrichment and variance partitioning analyses	Highly sensitive to annotation correlation and multicollinearity; perform sensitivity analyses; avoid over-interpreting enrichment signals from individual categories
Bivariate / multi-trait GREML	Identifiability of genetic covariance and genetic correlation between traits; integration of information across traits	Genetic effects for multiple traits are captured by a common GRM; covariance structure is identifiable in the sample	Adequate sample size; traits measured in the same or highly overlapping samples; applications in multi-trait breeding and trade-off analyses	Sensitive to phenotypic measurement error and sample overlap; carefully inspect standard errors and confidence intervals of genetic correlations; avoid over-interpretation under limited sample sizes

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Computational Molecular Biology 2026, Vol.16, No.3, 181-193
http://bioscipublisher.com/index.php/cmb

Research Insight Open Access

Modeling the Effects of Temperature on Peach Fruit Yield and Quality

Yedan He
1 Hangzhou Fuyang Aizi Fresh Peach Professional Cooperative, Hangzhou 311404, Zhejiang, China
2 Zhejiang Agronomist College, Hangzhou 310021, Zhejiang, China
Corresponding author: 727994936@qq.com
Computational Molecular Biology, 2026, Vol.16, No.3 doi: 10.5376/cmb.2026.16.0013
Received: 09 Apr., 2026
Accepted: 14 May, 2026
Published: 29 May, 2026
Copyright © 2026 He, This is an open access article published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Preferred citation for this article:
He Y.D., 2026, Modeling the effects of temperature on peach fruit yield and quality, Computational Molecular Biology, 16(3): 181-193 (doi: 10.5376/cmb.2026.16.0013)

Abstract Peach production is highly sensitive to variations in air temperature; as a critical climatic factor, temperature plays a pivotal role in the phenological development of peach trees, yield formation, and the regulation of fruit quality. Focusing on the mechanisms by which temperature influences peach yield and quality, this paper systematically analyzes its regulatory effects across different growth stages—specifically, the temperature response characteristics observed during bud break and flowering, fruit development, and the ripening process. Building upon this foundation, and by integrating meteorological data with orchard production records, a predictive model for peach yield and quality based on temperature indicators was constructed. This model places particular emphasis on incorporating variables such as accumulated temperature, extreme heat events, and seasonal temperature fluctuations, while employing a hybrid approach that combines statistical analysis with machine learning techniques for modeling and optimization. Through model performance evaluation and sensitivity analysis, key temperature thresholds and dominant factors influencing yield and quality were identified, thereby further elucidating the mechanisms by which heat stress and low-temperature impacts contribute to yield loss and quality deterioration. Case studies demonstrate that the developed model effectively predicts regional trends in peach yield and quality, exhibiting high applicability and stability. The findings of this study provide a theoretical basis for orchard temperature management, variety selection, and disaster risk management; furthermore, they offer technical support for the advancement of precision agriculture and intelligent decision-support systems, holding significant implications for enhancing the climate adaptability and production efficiency of the peach industry.
Keywords Peach yield prediction; Temperature stress; Fruit quality modeling; Growing degree days; Precision agriculture

1 Introduction

Temperature is a primary abiotic factor shaping peach growth, fruit set, and postharvest value, and its role is becoming more critical under ongoing climate change. Experimental warming studies with early- and low-chill cultivars show that moderate increases in temperature can accelerate development and, in some cases, enhance photosynthesis and fruit mass, whereas stronger warming reduces photosynthetic performance, floral bud differentiation, and subsequent yield (Lee et al., 2022). Temperature during fruit development also alters key quality traits-such as size, sweetness, coloration, and firmness-with high temperatures often hastening maturity but compromising desirable attributes like fruit weight and soluble solids content. These responses highlight the need to understand and predict how temperature regimes across seasons and regions translate into changes in both yield and fruit quality.

Despite extensive physiological and agronomic work, quantitative models linking temperature to integrated peach yield and quality outcomes remain limited. Process-based “virtual fruit” models capture fruit mass and sugar dynamics and are sensitive to environmental drivers, yet they often treat temperature only implicitly through generic weather terms rather than explicitly parameterizing its effects on growth and compositional traits. Recent climate-driven phenological and epidemiological models have projected climate-change impacts on peach blooming, disease pressure, and yield losses at national scales, but they primarily target phenology and disease, not detailed fruit quality responses (Lee et al., 2020). As a result, growers and breeders lack predictive tools that jointly represent how intra- and inter-seasonal temperature variability influences both quantitative yield and multiple quality dimensions.

At the same time, emerging statistical and machine learning approaches demonstrate the feasibility of robust peach yield prediction when temperature and other climatic variables are explicitly incorporated. Yield models calibrated on multiyear orchard data in subtropical climates, using inputs such as chilling hours, mean temperature, and leaf nutrient status, have identified chilling and temperature as dominant predictors, with machine learning methods (e.g., Random Forest) outperforming multiple linear regression (Moura-Bueno et al., 2026). Parallel work in subtropical regions using growing degree-days (GDD) to characterize fruit development shows that thermal accumulation strongly differentiates cultivars in terms of fruit size and mass, underscoring the central role of temperature metrics for explaining variability in agronomic performance. However, these data-driven models rarely integrate detailed fruit-quality indicators, and are seldom evaluated under projected future temperature scenarios.

The present study addresses these gaps by developing and testing models that explicitly quantify the effects of temperature on peach fruit yield and quality. Building on controlled-environment and field evidence that high and low temperature regimes differentially affect photosynthesis, growth, and key quality traits, the study formulates the hypothesis that specific temperature indices (e.g., mean temperature, GDD, extreme heat indicators) are systematically associated with both yield components and multi-dimensional quality attributes, and incorporating these indices into statistical and machine learning frameworks significantly improves the prediction of combined yield-quality outcomes compared with models using generic climate covariates. By integrating physiological knowledge with modern modeling techniques, the research aims to identify temperature thresholds and response patterns critical for maintaining yield and quality, compare alternative modeling strategies for capturing these relationships, and provide a transferable framework to support cultivar selection, orchard climate adaptation, and precision management decisions in current and future temperature regimes.

2 Temperature Regulation of Peach Phenology and Fruit Development

2.1 Temperature effects on bud break and flowering dynamics

Bud break and flowering in peach are governed by the interaction of winter chilling and subsequent heat accumulation. Studies across many cultivars show that increasing chilling accumulation generally reduces the heat requirement and days to flowering, but the strength of this chilling-heat trade-off differs with the cultivar's chilling requirement; high-chill genotypes respond strongly to small increases in chilling, whereas low-chill types show weaker reductions in heat requirement (Yan et al., 2024). Classic work demonstrated an exponential decline in heat needed for floral budbreak as chilling increases, with insufficient chilling leading to extended, asymmetric budbreak and greater sensitivity to spring temperature variation.

Across wide climatic gradients, the timing of rest completion and the balance between chill and heat strongly shape bloom dates. Multi-site trials of peach and nectarine in Europe found that in colder sites, rest was completed earlier and bloom time was more tightly controlled by spring heat accumulation, while in warmer sites delayed or incomplete rest made bloom timing more sensitive to winter temperatures (Drogoudi et al., 2023). Long-term modeling shows that warming winters in major peach regions are already reducing chill accumulation, shifting the relative roles of chilling and forcing and complicating the prediction of budbreak and bloom under climate change (Yan et al., 2024).

2.2 Heat accumulation and fruit development rate

After bloom, heat accumulation is a primary driver of fruit development rate and the length of the fruit development period (FDP). Thermal-time models using growing degree hours with cultivar-specific base, optimum, and critical temperatures have been shown to predict harvest dates within 1-4 days across cultivars with FDPs ranging from 70 to 150 days, and are especially accurate when based on heat accumulated in the first 25-52 days after bloom. Analyses of spring temperature effects indicate that high early-season heat (GDH30) accelerates phenological development and shortens the period from full bloom to a reference developmental stage, but can reduce fruit size at that stage because trees cannot supply resources rapidly enough to match the higher growth potential.

Fruit growth response to temperature is stage-dependent. Under controlled conditions, peach fruit developed typical double-sigmoid growth, with higher temperatures (up to 30 °C) increasing growth rates and shortening the duration of early stages (S1-S2), thereby reducing total development time by more than two weeks compared with cooler regimes. However, the same high temperatures slowed late-stage expansion (S3) and reduced final fruit size, weight, and soluble solids, indicating that while elevated temperatures speed development and advance maturity, they can compromise key quality traits if thermal conditions exceed optimal ranges during critical growth phases.

2.3 Temperature stress impacts on reproductive success

Reproductive processes in peach are particularly vulnerable to temperature extremes around bloom. Controlled-environment studies with 'Hakuho' showed that constant temperatures of 25°C-30 °C greatly accelerated bud burst and flowering but reduced flower size, impaired embryo sac development, and markedly lowered fruit set, indicating that temperatures above about 25 °C disrupt normal reproductive organ development and fertilization success. Field and greenhouse comparisons in 'Granada' revealed that pre-bloom and bloom high temperatures advanced dormancy break and bloom but delayed female gametophyte development, induced anomalies in male gametophytes, and resulted in low pollen viability, poor synchrony of fertilization, and reduced yield.

More detailed analyses of the progamic phase show contrasting temperature sensitivities of male and female functions. Within a moderate range, increasing temperature accelerates pollen germination and pollen tube growth and increases the number of tubes reaching the style base, but simultaneously causes a sharp decline in stigmatic receptivity, first for supporting tube penetration, then germination, and finally pollen adhesion. Additional work comparing cultivars under subtropical fluctuations found that high temperatures (≥25 °C) during bloom reduced in vitro pollen germination and the proportion of normal pollen grains, with stronger negative impacts on fruit set in 'Granada' than in 'Maciel', highlighting genotype-specific vulnerability of reproductive success to heat episodes at flowering.

3 Temperature Impacts on Peach Yield Formation Mechanisms

3.1 Growing degree days and yield accumulation relationships

Growing degree-based thermal indices provide a mechanistic link between temperature, developmental timing, and yield formation in peach. In subtropical field conditions, cultivars with higher growing degree day (GDD) requirements during fruit development ("Biuti") achieved larger fruit size and mass, whereas low-GDD cultivars ("Tropical") showed smaller fruits, indicating that greater thermal accumulation supports longer growth phases and higher yield potential. Similarly, cultivar comparisons in a sub-temperate Himalayan zone showed that mid- and late-season cultivars requiring 1500-1900 GDD produced higher yields and better quality traits (TSS, sugars) than early cultivars with lower GDD, underscoring that cultivar-specific GDD thresholds structure both yield and quality outcomes.

Process-based and simulation approaches further embed growing degree metrics into yield formation. The PEACH tree growth and yield model uses growing degree hours (GDH) accumulated during the first 30 days after bloom to estimate the length of the fruit growth period; incorporating this GDH-harvest date relationship markedly improved predictions of harvest timing and yield across years and locations compared with earlier degree-day formulations. Empirical analyses of spring temperatures also show that high early GDH accumulation shortens the interval from full bloom to a reference date, increasing instantaneous growth rates but ultimately reducing reference-date fruit size when resource supply cannot keep pace with rapid phenological advancement, demonstrating that thermal time can both promote and constrain yield formation depending on seasonal context.

3.2 Heatwaves and yield reduction mechanisms

Short-term heat stress around harvest exerts distinct and sometimes counterintuitive effects on peach yield formation. A regional analysis for South Korea, using municipal yield data and thermal indicators around the 'Cheonjungdo Baekdo' harvest window, found that a higher number of hot days (>30 °C) and elevated minimum temperatures during fruit development significantly increased the probability of low-yield years, implicating

prolonged heat exposure and warm nights in yield reduction. Yet higher maximum temperatures earlier in the growth period were associated with improved productivity, and cumulative heat intensity above 30 °C around harvest showed a negative association with low yield, highlighting complex, stage-dependent responses to heat events.

Controlled-environment experiments reveal physiological mechanisms through which excessive heat depresses fruit yield and size. For the low-chill cultivar 'KU-PP2', growth at 30 °C accelerated early fruit expansion and shortened the development period by up to 18 days but substantially reduced final fruit weight and soluble solids compared with 20 °C, effects attributed to diminished photosynthetic capacity under sustained high temperature. Related work on 'Mihong' shows that moderate warming (+3.4 °C) can increase photosynthesis, stomatal density, and tree yield, while stronger warming (+5.7 °C) reduces photosynthetic rates and floral bud differentiation, thereby lowering current yield and compromising yield potential in the following year, illustrating how heatwaves that push temperatures beyond optimal thresholds can damage both current and subsequent crops (Lee et al., 2020).

3.3 Seasonal temperature variability and yield stability

Interannual variability in seasonal temperatures modulates both phenology and yield stability in peach orchards. In Moroccan Sais Valley conditions, years with higher temperatures during flowering and fruit growth showed earlier bloom and maturity but significantly lower fruit weight, suggesting that warmer seasons may compress developmental periods and reduce assimilate accumulation per fruit. Long-term observations in a warm Tunisian production area similarly indicate that exceptionally warm winters with low chill accumulation delay and desynchronize flowering, increase bud abscission, and reduce yield and fruit quality when chilling falls below cultivar-specific thresholds, demonstrating that warm-winter anomalies destabilize reproductive success and commercial output.

At broader spatial scales, process-based phenology models that couple chilling, forcing, frost risk, and growing degree days show that climate warming will shift the thermal niche of peach cultivation, with earlier bloom and easier ripening but increasing risk of insufficient winter chill in traditional warm regions. Analysis of historical low-yield events in the U.S. Midwest and Southeast further identifies "false spring" patterns-early GDD accumulation followed by hard freezes-as major drivers of regional peach crop failures, and uses surface temperature thresholds and GDD tracking to build a decision-support tool capable of predicting major yield reductions, emphasizing how seasonal temperature sequences rather than single extremes determine yield stability.

4 Temperature Regulation of Peach Fruit Quality Attributes

4.1 Temperature effects on sugar and acid metabolism

Storage and handling temperature strongly shape sweetness-acidity balance in peach by altering sugar and organic acid metabolism. Non-chilling storage around 12 °C allows normal ripening, maintaining flavor development and preventing chilling injury, whereas storage at 4 °C, although effective at slowing softening, induces off-flavors and increased bitterness linked to the accumulation of specific bitter flavonoids and related metabolites (Muto et al., 2022). Low-temperature stress can also trigger metabolic reprogramming of carbohydrates: during cold storage, sucrose and other soluble sugars often change in parallel with chilling symptoms, reflecting their dual roles as flavor components and stress protectants.

Regulation of sucrose metabolism under cold and temperature-related treatments is central for both flavor and chilling tolerance. Hot-air and methyl jasmonate treatments before storage at 5 °C increased sucrose and sorbitol contents compared with controls, associated with higher sucrose phosphate synthase activity and lower acid invertase activity, suggesting that moderate temperature stress combined with elicitors can maintain sweetness while enhancing chilling resistance (Figure 1). Similarly, salicylic acid pretreatment prior to 4 °C storage raised total soluble sugars, largely via sucrose accumulation, and modified expression of sucrose-related genes, while simultaneously activating cold-response transcription factors and reducing internal browning, indicating that temperature-driven sugar metabolism is tightly coupled to stress signaling and quality preservation.

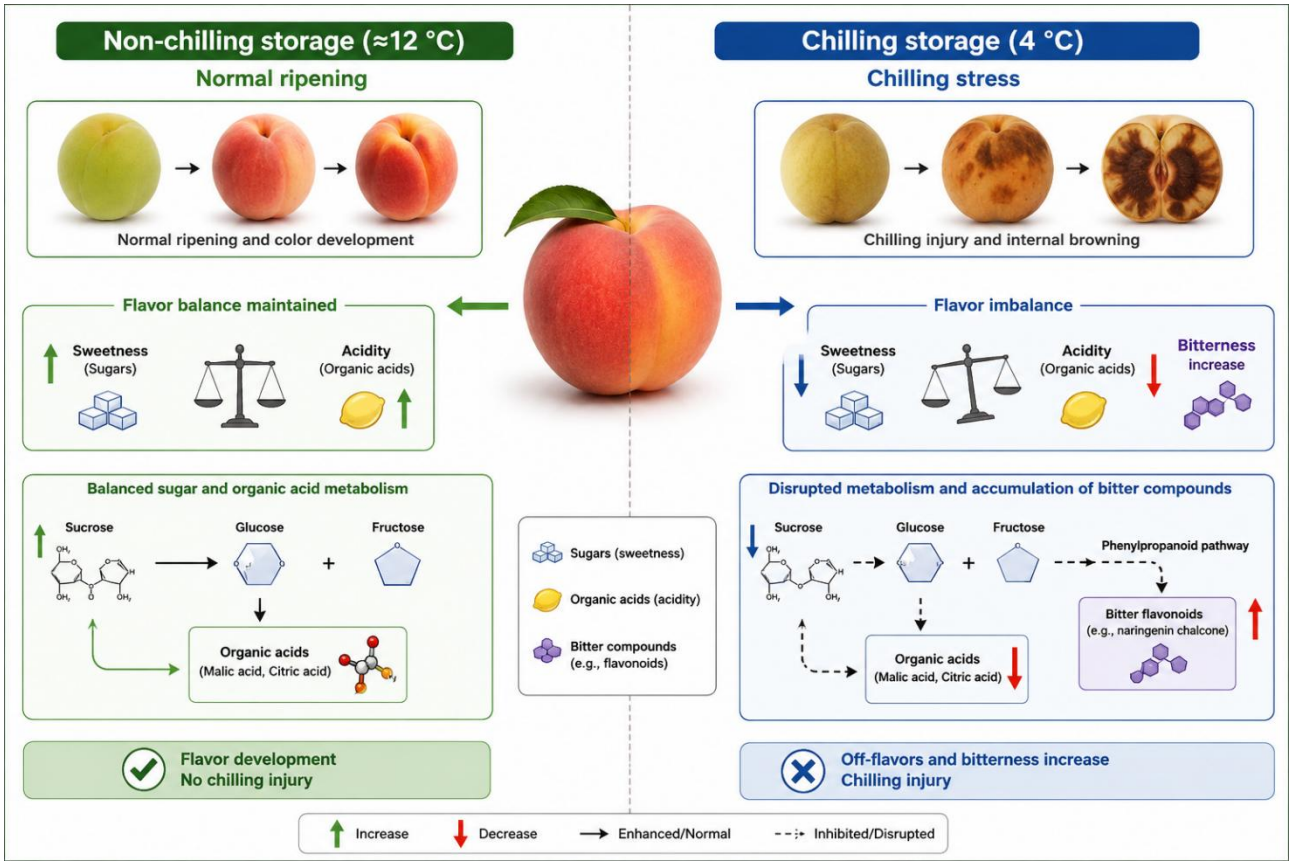


Figure 1 Effects of storage temperature on flavor balance and metabolic regulation in peach fruit

4.2 Influence on fruit size, firmness, and texture

Temperature during on-tree development controls final size and basic texture attributes. In controlled environments, increasing growth temperature from 20 °C to 30 °C accelerated early fruit expansion and shortened the development period but reduced final fruit weight, size, and sweetness, indicating that high developmental temperatures hasten maturity at the expense of key quality traits. Under projected climate-change scenarios with elevated CO_2, moderate warming (+3.4 °C) increased photosynthesis, carbohydrate content, and fruit weight, whereas stronger warming (+5.7 °C) decreased photosynthetic performance and was associated with poorer physiological status and reduced fruit quality in the subsequent year, emphasizing that beneficial temperature windows are narrow (Lee et al., 2022).

Postharvest temperature interacts with cell wall metabolism to determine firmness and textural defects. In stony-hard peaches, storage at intermediate temperatures (8°C-15 °C) induced substantial softening and strong expression of a polygalacturonase gene, whereas storage at 0 °C or 20 °C for extended periods prevented subsequent softening at 10 °C, suggesting that specific temperature ranges activate pectin-degrading machinery independently of ethylene (Tatsuki et al., 2021). Conversely, in melting-flesh peaches, low-temperature storage at 6 °C, compared with 25 °C, inhibited softening by maintaining cell wall integrity: low temperature reduced the accumulation of water- and ion-soluble pectin and suppressed activities and expression of polygalacturonase, pectate lyase, and pectin methylesterase, effects linked to cold-induced CBF transcription factors that repress pectin-degradation genes (Guo et al., 2026).

4.3 Temperature-driven changes in aroma and phytochemicals

Aroma and phytochemical profiles of peach are highly temperature-dependent during storage and ripening. Cold storage at 1 °C for 7 d significantly affected firmness, acidity, phenolics, vitamin C, and carotenoids across cultivars, with some sensory attributes (bitterness, astringency, crunchiness) increasing as firmness and acidity rose, while perceived harmony and sweetness were more closely related to °Brix, β-carotene, and specific volatiles than to simple acidity measures (Muto et al., 2022). In another study, low-temperature storage at 0.5 °C

and 5.5 °C increased aldehydes and alcohols during storage but shifted ester and lactone evolution to subsequent shelf life, with cultivar differences in chilling injury linked to differential accumulation of antioxidants and osmoprotectants such as sorbitol, putrescine, and phenolics (Brizzolara et al., 2018).

Moderately low temperatures during ripening can enhance phenolic accumulation and modify volatile pathways in the field. For a protected-origin peach, correlation analyses showed that cooler ripening conditions were associated with higher levels of phenolic compounds, particularly flavonoids and anthocyanins, while expression of a lipoxygenase gene (PpLOX1) co-varied with climate variables and LOX-derived volatiles, indicating coordinated temperature regulation of antioxidant and aroma biosynthesis. Under postharvest cold storage at 0 °C, controlled-atmosphere conditions improved sensory quality by reducing internal browning and retaining higher levels of esters and lactones; several LOX-pathway volatiles and associated biosynthetic genes were positively correlated with consumer acceptability, underscoring how temperature-atmosphere regimes modulate both aroma and perceived eating quality (Liu et al., 2022).

5 Construction of Temperature-Based Peach Yield and Quality Models

5.1 Selection of temperature indicators and feature engineering

Constructing temperature-based models for peach yield and quality begins with identifying thermal indicators that best capture phenology and fruit development. Reviews of temperature indices in temperate fruit production emphasize chill units, growing degree days (GDD), and growing degree hours (GDH) as core descriptors for dormancy release, flowering, and development, together with indices for extreme events such as frost and heat stress. In subtropical peach orchards, cultivar comparisons show that GDD accumulation during fruit development is closely linked with fruit size and mass, with higher GDD requirements associated with larger fruit, guiding the choice of development-stage-specific thermal sums as model inputs.

Feature engineering must also reflect cultivar-specific thresholds and the timing of thermal exposure. Nonlinear GDH models that incorporate base, optimum, and critical temperatures predict harvest dates within 1-4 days for cultivars with very different fruit development periods, illustrating the value of calibrated, cultivar-dependent heat-response parameters. Phenological work in sub-temperate regions further demonstrates that GDD from dormancy break to harvest differentiates early, mid, and late cultivars and is strongly associated with yield and sugar content, suggesting that cumulative heat over well-defined BBCH stages can be transformed into compact, phenology-anchored predictors for yield and quality models.

5.2 Integration of physiological and meteorological data

Accurate temperature-based models require integrating meteorological variables with physiological or structural indicators of tree status. A multi-year study on ‘Esmeralda’ peach combined meteorological indices (chilling hours, GDD, rainfall) with foliar mineral composition and previous-season yield, showing that chilling hours and GDD dominated feature rankings for yield and several quality traits, while leaf nutrients and carryover effects refined predictions (Nava et al., 2022). Similarly, a peach yield prediction study using 208 trees under subtropical climates found that hours of chilling and mean temperature, together with leaf K and N, were the most relevant predictors of yield in machine learning models, highlighting the importance of jointly representing climate and plant nutritional status (Moura-Bueno et al., 2026).

Physiological integration is also needed to capture how temperature affects photosynthesis and fruit growth potential. Controlled phytotron experiments on the early cultivar ‘Mihong’ under elevated temperatures and high CO_2 showed that moderate warming (+3.4 °C) increased photosynthetic rate, fruit weight, and carbohydrate content, whereas stronger warming (+5.7 °C) reduced photosynthesis, floral bud differentiation, and expected subsequent yield (Figure 2) (Lee et al., 2022). Temperature-controlled studies on ‘KU-PP2’ similarly demonstrated that higher growth temperatures accelerate early fruit expansion but reduce final fruit size and sweetness at 30°C, implying that model inputs should include not only simple thermal sums but also phase-specific temperature descriptors linked to physiological processes such as photosynthetic capacity and source-sink balance.

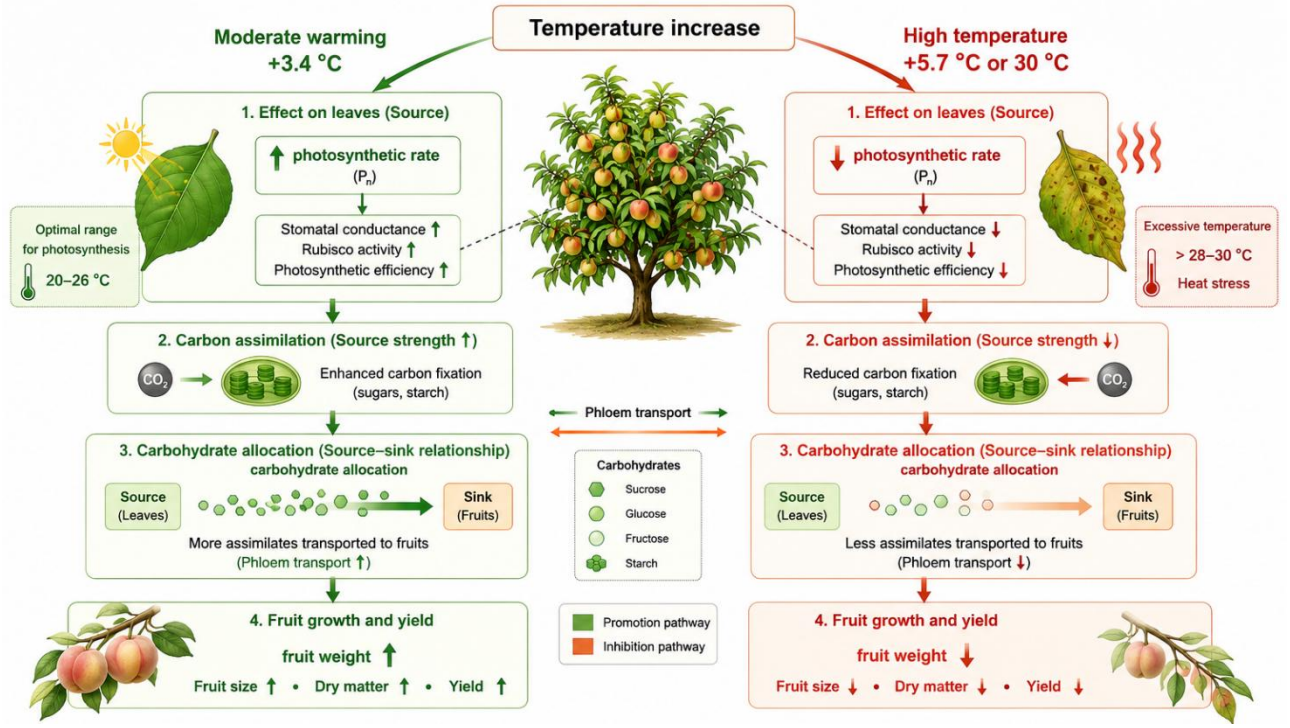


Figure 2 Effects of temperature variation on photosynthesis and source-sink dynamics in peach trees under controlled environments

5.3 Development of statistical and machine learning models

Once temperature indicators and physiological covariates are defined, they can be embedded in statistical and machine learning frameworks. For bud and bloom phenology, models combining chill accumulation (Dynamic model) with growing degree sums have predicted bud development stages across major Spanish peach regions with errors of about four days, outperforming forcing-only models and providing physiologically consistent thresholds for chill and heat. At finer scales, chill-heat models for individual cultivars have been fitted using sequential chill and GDH accumulation to estimate budbreak timing, offering simple yet robust tools for linking winter-spring temperatures to the onset of reproductive development (Cifuentes-Carvajal et al., 2023).

For yield-focused modeling, multivariate machine learning approaches appear especially promising. In 'Esmeralda' peach, k-nearest neighbors and stochastic gradient descent models trained on meteorological indices, foliar nutrients, and prior yield achieved accuracies up to 1.00 for several yield and quality indices, with chilling hours and degree-days emerging as top-ranked features (Nava et al., 2022). A broader yield prediction study comparing Random Forest, multiple linear regression, and support vector machines found that Random Forest provided the best performance, and identified hours of chilling, leaf K and N, and mean temperature as the most influential variables, confirming that nonlinear ML models can effectively learn complex temperature-nutrition-yield relationships when supported by well-engineered temperature indicators (Moura-Bueno et al., 2026).

6 Evaluation and Optimization of Temperature-Driven Prediction Models

6.1 Model performance comparison and selection

Evaluation of temperature-driven peach yield and quality models requires systematic comparison of alternative model structures and learning algorithms. For peach yield, a study comparing Random Forest, Multiple Linear Regression, and Support Vector Machine found that Random Forest achieved the highest predictive accuracy when using climatic, soil, and leaf nutrient data, with chilling hours and mean temperature among the most influential predictors (Moura-Bueno et al., 2026). Similarly, temperature-based phenology models for peach bloom (developmental rate, chill day, and new chill day models) were assessed with MAPE, R^2, and RMSE, and the new chill day model provided the best compromise between bias and precision across cultivars and sites, illustrating the value of multi-metric model comparison for temperature-driven processes.

Beyond peach, crop modeling research highlights the need to balance accuracy and interpretability when selecting prediction models. An interaction regression framework for corn and soybean yielded lower relative RMSE than state-of-the-art machine learning methods while explicitly decomposing yield into contributions from weather, soil, and management, demonstrating that carefully regularized regression with interaction selection can outperform black-box models and provide mechanistic insight on temperature effects (Ansarifar et al., 2021). Phenology-guided deep learning for soybean showed that incorporating heat-related predictors and phenological-stage windows in a Bayesian CNN architecture substantially improved yield prediction relative to benchmark models, underscoring that model choice should reflect both temperature process representation and the temporal structure of response variables (Zhang and Diao, 2023).

6.2 Error decomposition and robustness testing

For temperature-driven crop models, decomposing prediction errors helps clarify limitations in both structure and parameterization. In a grapevine phenology-yield model calibrated with a frequentist framework, the joint objective function based on normalized RMSE revealed that no single parameter vector minimized errors for both phenology and yield simultaneously, and yield RMSE exhibited much larger spread than phenology RMSE, indicating structural or parameter constraints in capturing yield responses to weather variability (Yang et al., 2024). Follow-up uncertainty analysis showed that fruit-setting parameters were the dominant contributors to yield prediction variability, illustrating how error decomposition can pinpoint biologically meaningful leverage points for improving reproductive and yield submodels.

Robustness of phenology and yield simulations to temperature extremes and calibration data coverage has been explicitly tested in multi-model rice phenology assessments. Using six model structures and leave-one-out cross-validation, regional simulations of maturity dates achieved RMSE of 2-4 days, but evaluation errors were larger than calibration errors, especially in areas with frequent high-temperature episodes, where divergent model responses increased structural uncertainty. Decomposition of total uncertainty into parameter and structural components showed that parameter variability dominated overall uncertainty in most regions, except in high-temperature zones where structural differences in temperature response functions were more important, emphasizing that robustness testing must consider both parameter and model-form uncertainties across temperature regimes.

6.3 Sensitivity and uncertainty analysis of temperature variables

Sensitivity and uncertainty analyses provide a quantitative basis for prioritizing temperature-related variables and parameters in peach yield and quality models. Global sensitivity analysis of a fertigation crop model (HORTSYST) using Sobol indices identified nine key parameters-including minimum and maximum optimal temperatures and radiation-use efficiency-as most influential on photo-thermal time, dry matter production, and transpiration, guiding calibration toward the subset of parameters that control temperature and radiation responses. In the same framework, parameters showed stage-dependent importance, with more parameters affecting outputs early in fruiting than late in the season, suggesting that temperature sensitivities should be evaluated for specific phenological windows when modeling fruit crops.

Other dynamic crop models combine variance-based sensitivity analysis with uncertainty propagation to understand climate effects on yield. For Lycium barbarum in WOFOST, Morris and extended FAST methods demonstrated that parameters related to CO_2 assimilation, leaf area expansion, and thermal time during specific periods had the largest impact on simulated yield, and sensitivity rankings were consistent across climate sites, supporting the transferability of temperature- and development-related parameter priors across regions. A similar strategy in a grapevine soil-plant-atmosphere model quantified prediction uncertainty as the spread of nRMSE across hundreds of thousands of parameter vectors, then used parameter-wise reductions in uncertainty to identify those most responsible for yield and phenology variance, providing a template for implementing global sensitivity and uncertainty analysis in peach temperature-yield-quality models.

7 Identification of Critical Temperature Thresholds in Peach Production

7.1 Optimal temperature ranges for key growth stages

Critical temperature thresholds for vegetative and reproductive development can be described using basal (minimum and maximum) temperatures and thermal sums for successive phenological phases. For 14 peach and one nectarine cultivar, minimum basal temperatures of about 8°C-10 °C were identified for pruning-sprouting and sprouting-flowering, 12°C-14 °C for flowering-fruiting, and 12°C-14 °C for ripening, while maximum basal temperatures were about 28°C-34 °C depending on the phase. These values imply that temperatures below phase-specific bases do not contribute to development, whereas temperatures above the upper limits do not further accelerate progress and may predispose to stress, providing practical bounds for "effective" temperature ranges during each stage.

Thermal-time models further refine optimal temperature concepts by combining cultivar-specific base, optimum, and critical temperatures. A non-linear growing degree hour (GDH) model using a base of 7.5 °C, an optimum of 26 °C, and a critical temperature of 38.5 °C accurately predicted harvest dates (1-4 d error) for cultivars with fruit development periods from 70 to 150 d, and an early forecast could be obtained from GDH accumulated in the first 25-52 d after bloom. Together, these results indicate that peach development proceeds most efficiently within a broad band from the low teens up to the mid-20s °C, with diminishing or saturating developmental gains as temperatures approach the upper 20s and mid-30s °C.

7.2 High-temperature stress thresholds and yield loss

Experimental warming under future-climate CO_2 has clarified high-temperature thresholds for physiological decline. For 'Mihong', a modest rise of +3.4 °C above local averages (with 700 μmol/mol CO_2) increased photosynthetic rate, carbohydrate content, and fruit weight, whereas a +5.7 °C scenario reduced photosynthesis, caused chlorophyll loss, decreased floral bud differentiation, and lowered floral bud density, leading to expected yield reduction in the following year (Lee et al., 2022). These responses suggest that warming within roughly +3-4 °C may still fall within an expanded "optimal" window, while sustained warming approaching +6 °C crosses a physiological threshold where vegetative dominance and early defoliation compromise reproductive potential.

At the orchard and regional scale, heat indicators around harvest identify damaging thresholds for yield. In South Korea, a logistic model using municipal yield data showed that a higher number of days above 30 °C and elevated minimum temperatures during fruit development significantly increased the probability of low-yield years, although higher maximum temperatures earlier in the growth period were linked to improved productivity. The positive association between counts of >30 °C days and low yield, combined with the experimental evidence of performance declines near +6 °C warming, indicates that both the intensity and persistence of temperatures above about 30 °C define critical stress thresholds for peach yield formation.

7.3 Low-temperature injury and recovery mechanisms

On the cold side, storage temperature tightly controls the onset of chilling injury (CI) symptoms and associated membrane damage. During postharvest storage, peaches kept at 4 °C rapidly developed CI, with enhanced expression of membrane lipid metabolism genes, accumulation of phosphatidic acid, and shifts in diacylglycerol and triacylglycerol profiles, whereas storage at 0 °C delayed CI by maintaining higher levels of phospholipids and promoting fatty acid desaturation and unsaturation. These findings indicate that, paradoxically, "moderate" low temperatures around 4 °C may be more injurious than near-freezing 0 °C, and that maintenance of unsaturated membrane lipids is a key protective mechanism at very low temperatures.

Pre-storage conditioning and acclimation treatments define additional functional thresholds for cold tolerance and recovery. Low temperature conditioning at 8 °C for 5 d before 0 °C storage increased ethylene production, accelerated softening, reduced internal browning, and led to higher fatty acid content, desaturation, and phospholipid levels compared with constant 0 °C storage (Song et al., 2022). Similarly, priming 'June Gold' fruit for 48 h at 20 °C before 40 d at 0 °C suppressed CI symptoms relative to fruit transferred directly to 0 °C, with distinct proteomic and metabolomic signatures indicating altered cold responses and a possible role for

branched-chain amino acids in tolerance. Collectively, these studies show that both the absolute low temperature (0°C vs 4 °C) and short exposures to intermediate "priming" temperatures (8°C-20 °C) critically determine whether cold acts as damaging stress or as a signal that triggers protective acclimation pathways.

8 Case Study on Temperature-Driven Yield and Quality Variations in Peach Orchards

8.1 Study area climate and orchard characteristics

The case study focuses on subtropical orchards in southern Brazil, where 'Maciel' and 'Chimarrita' peaches are grown under contrasting microclimates but broadly similar humid subtropical conditions with variable winter chill and warm springs. A database of 208 trees captured spatial variation in soil properties, leaf nutrient status, and localized weather, allowing climate variables such as chilling hours and mean temperature to be related to yield at tree scale (Moura-Bueno et al., 2026). In parallel, field work in Brazilian subtropical regions characterized fruit development of four cultivars across the season, using growing degree days (GDD) to describe the temperature regime governing fruit growth and size.

These subtropical environments are characterized by relatively mild winters that can constrain chill accumulation and by warm, often rapidly heating, springs and summers that drive fast GDD accumulation (Moura-Bueno et al., 2026). Under these conditions, cultivars such as 'Tropical' require lower GDD and produce smaller, lighter fruit, whereas 'Biuti' demands higher GDD and attains larger size, illustrating how local thermal regimes interact with genotype to shape orchard yield potential and quality profiles. Such climate-cultivar interactions frame the design of temperature-driven prediction models in the study area.

8.2 Application of temperature-based predictive models

In the Brazilian orchards, peach yield was modeled by combining climatic indicators with tree- and soil-level covariates. Random Forest, Multiple Linear Regression, and Support Vector Machine were trained using hours of chilling, mean temperature, and leaf and soil nutrient data; Random Forest gave the highest predictive performance, and chilling hours emerged as the single most relevant predictor of yield, followed by leaf K and N and mean temperature (Moura-Bueno et al., 2026). This structure embeds temperature both as a direct driver (chill and in-season means) and as a proxy for longer-term site suitability, while allowing nonlinear effects and interactions (Figure 3).

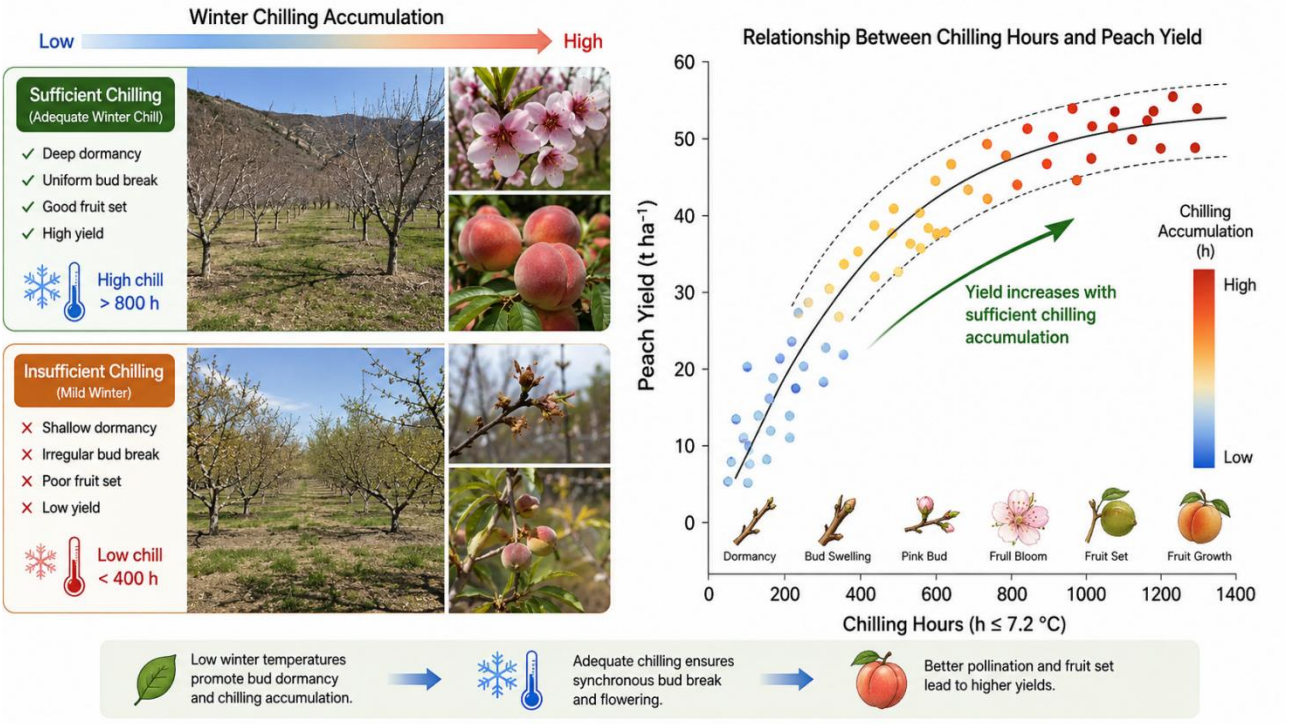


Figure 3 Relationship between winter chilling accumulation and peach yield performance in Brazilian orchards

Complementing these tree-scale models, a regional decision-support tool was developed for the U.S. Midwest and Southeast to anticipate major yield reductions from false springs, using accumulated growing degree days (GDD7.2) and minimum temperatures during freeze events. For each region, an "envelope" curve relating GDD to critical minimum temperature was derived from historical low-yield years; stations falling below this envelope in a given season were classified at risk of major peach yield loss. This approach illustrates how simple temperature-based indicators can be operationalized for risk forecasting at regional scale.

8.3 Validation of predicted yield and quality against observations

Validation of the Brazilian machine learning models used independent data from the same orchards, confirming that Random Forest calibrated with climatic, soil, and foliar variables could reproduce observed yield variation with high accuracy, while simpler linear models underperformed, especially when only a subset of predictors was used. Feature-importance analysis aligned with agronomic expectations-emphasizing chilling hours and mean temperature-supporting both statistical and physiological credibility of the fitted relationships (Moura-Bueno et al., 2026). At the quality level, comparisons among subtropical cultivars showed that modeled GDD-based development patterns were consistent with observed differences in fruit size and mass between low- and high-GDD genotypes, strengthening the case for GDD as a robust explanatory variable for quality-related traits .

For the regional false-spring tool, validation against historical high-yield years demonstrated that the GDD-minimum-temperature envelope correctly identified non-damaging seasons in all sampled years for the Midwest and in 75% of high-yield years for the Southeast, indicating strong but regionally variable skill (Chun and Changnon, 2018). Application to the 2017 false spring showed that the tool successfully anticipated widespread yield reductions in the Southeast while correctly indicating lower risk in much of the Midwest, when observed production data were later examined. Together, these evaluations show that temperature-driven models can achieve useful predictive power for both yield level and catastrophic loss when carefully calibrated to local climate and production systems.

9 Development of Climate-Resilient Peach Production Strategies

Climate-resilient orchard management increasingly focuses on mitigating insufficient winter chill and buffering trees against temperature extremes. In the southeastern United States, anthropogenic warming has already reduced winter chill, increased the probability of low-chill winters, and raised the risk of insufficient chill for moderate- and high-chill cultivars, prompting consideration of adaptive practices such as overhead irrigation for evaporative cooling, vigor control to lower chill needs, and site selection in cooler microclimates. In warm-winter regions like Israel, additional physical strategies-including shading, branch bending, and sprinkling to reduce daytime temperature-are proposed to compensate partially for chill deficits and reduce abnormal bud development and low fruit set under heat spells.

Postharvest temperature management is another critical component of climate-resilient peach systems. Poorly controlled cold chains with repeated temperature spikes to 15°C-20 °C sharply increase ethylene production, accelerate softening, and reduce phenolics, flavonoids, and antioxidant enzyme activities, whereas limiting fluctuations to around 10 °C has little impact on quality, delineating operational thresholds for transport and storage. Reviews of cold-stress physiology emphasize that careful management of storage and transport temperatures, combined with early, preferably non-destructive monitoring tools for chilling injury, is essential to safeguard fruit quality as supply chains lengthen and temperature variability increases.

Climate-resilient production depends strongly on matching cultivar chilling and heat requirements to warming agroclimates. Multi-site analyses across Tunisia and Europe show wide genotypic variation in peach chilling (≈20-63 Chill Portions) and heat requirements (≈4381-6556 GDH), with warm mean temperatures during the chilling period emerging as key drivers of flowering, providing a quantitative basis for selecting cultivars adapted to warm regions. Under mild Moroccan conditions, grouping cultivars by chill/heat needs and flowering time identified low- to medium-chill types as more suitable under climate change, while cultivars like 'Summer Lady' showed lower sensitivity to bud and fruit drop during warm autumns and chill deficits, making them strategic genetic resources (Borgini et al., 2024).

Cold-tolerance screening complements agroclimatic matching in regions facing severe winter freezes. In Gansu, China, evaluation of 28 local germplasms identified large variation in semi-lethal temperature (LT50, -28.22 to -17.22 °C), with the highly resistant 'Dingjiaba Liguang Tao' showing the lowest LT50 and strong associations between cold hardiness and soluble sugars, proteins, proline, and xylem and cork anatomy. A separate comprehensive evaluation under -5°C to -35 °C stress similarly highlighted cultivars such as 'Ziyan Ruiyang' and 'Ganlu Shumi' with low LT50, high membership scores, and good field survival, providing robust parents for breeding new cold-resistant varieties and expanding resilient cultivar portfolios.

Intelligent monitoring and control systems offer powerful tools to manage orchard microclimates under increasing thermal stress. An IoT-based "smart orchard" architecture using multi-sensors (air and soil temperature, humidity, light, rainfall, wind) and LoRa transmission demonstrated reliable environmental monitoring in peach orchards with complex terrain, enabling remote supervision and providing the data backbone for temperature-focused decision support. A related multi-parameter orchard system couples sensor data to actuators (fans, pumps, LEDs, alarms) and a cloud platform + mobile interface, allowing threshold-based, remote control of the microclimate that stabilized yields, improved fruit quality, and reduced labor costs through more precise environmental regulation.

Downstream in the supply chain, AI-based decision support can optimize temperature management for quality preservation. An artificial neural network system trained on commercial cold-room data predicts the evolution of hardness, soluble solids, and acidity as functions of storage temperature, relative humidity, and time, thereby estimating optimal commercialization windows and suggesting pre-cooling setpoints that maximize the period of peak consumer-perceived quality. Insights from virtual cold-chain experiments, which identify tolerable versus harmful temperature excursions, can be integrated into such DSS tools to define safe fluctuation ranges and reduce waste while maintaining high-quality fruit delivery.

References

Ansarifar J., Wang L., and Archontoulis S.V., 2021, An interaction regression model for crop yield prediction, Scientific Reports, 11(1): 17754
https://doi.org/10.1038/s41598-021-97221-7

Chun S.E.A., and Changnon D., 2018, Predicting major peach yield reductions in the Midwest and Southeast United States, Meteorological Applications, 26(1): 97-107.
https://doi.org/10.1002/met.1740

Cifuentes-Carvajal A., Chaves-Cordoba B., Vinson E.L., Coneva E., Chavez D.J., and Salazar-Gutierrez M.R., 2023, Modeling the budbreak in peaches: a basic approach using chill and heat accumulation, Agronomy, 13(9): 2422.
https://doi.org/10.3390/agronomy13092422

Drogoudi P., Cantín C.M., Brandi F., Butcaru A., Cos-Terrer J., Cutuli M., Foschi S., Galindo A., García-Brunton J., Luedeling E., Moreno M.A., Nari D., Pantelidis G., Reig G., Roera V., Ruesch J., Stanica F., and Giovannini D., 2023, Impact of chill and heat exposures under diverse climatic conditions on peach and nectarine flowering phenology, Plants, 12(3): 584.
https://doi.org/10.3390/plants12030584

Guo Y., Cao J., Yu W., Wang X., Zhang Y., Liu H., Zhao X., Li M., and Wang H., 2026, Chemical characterization, geographical differentiation, and climatic associations of pinggu peach based on widely targeted metabolomics, Food Chemistry, 508(Pt B): 148454.
https://doi.org/10.1016/j.foodchem.2026.148454

Lee S.K., Cho J.G., Jeong J.H., Ryu S., Han J.H., and Do G.R., 2020, Effect of the elevated temperature on the growth and physiological responses of peach 'mihong' (*Prunus persica*), Protected Horticulture and Plant Factory, 29(4): 373-380.
https://doi.org/10.12791/ksbec.2020.29.4.373

Lee S.K., Han J.H., Cho J.G., Jeong J.H., Lee K.S., Ryu S., and Choi D.G., 2022, Effect of temperature on photosynthesis and fruit quality of 'mihong' peaches under high CO_2 concentrations, Horticulturae, 8(11): 1047.
https://doi.org/10.3390/horticulturae8111047

Liu H., He H., Liu C., Song S., Wang H., Zhang H., Wang L., and Wang A., 2022, Changes of sensory quality, flavor-related metabolites and gene expression in peach fruit treated by controlled atmosphere (CA) under cold storage, International Journal of Molecular Sciences, 23(13): 7141.
https://doi.org/10.3390/ijms23137141

Moura-Bueno J.M., Betemps D.L., Marodin G.A.B., Toselli M., Natale W., and Brunetto G., 2026, Peach yield prediction models: the importance of climate variables and different machine learning, Horticulturae, 12(2): 155.
https://doi.org/10.3390/horticulturae12020155

Muto A., Christofides S.R., Sirangelo T.M., Bartella L., Muller C., Di Donna L., Muzzalupo I., Bruno L., Ferrante A., Chiappetta A.A.C., Bitonti M.B., Rogers H.J., and Spadafora N.D., 2022, Fruitomics: the importance of combining sensory and chemical analyses in assessing cold storage responses of six peach (*Prunus persica* L. Batsch) cultivars, Foods, 11(17): 2554.
https://doi.org/10.3390/foods11172554

Nava G., Reisser Júnior C., Parent L.É., Brunetto G., Moura-Bueno J.M., Navroski R., Benati J.A., and Barreto C.F., 2022, Esmeralda peach (*Prunus persica*) fruit yield and quality response to nitrogen fertilization, Plants, 11(3): 352.
https://doi.org/10.3390/plants11030352

Song C., Wang K., Xiao X., Liu Q., Yang M., Li X., Feng Y., Li S., Shi L., Chen W., and Yang Z., 2022, Membrane lipid metabolism influences chilling injury during cold storage of peach fruit, Food Research International, 157: 111249.
https://doi.org/10.1016/j.foodres.2022.111249

Tatsuki M., Sawamura Y., Yaegaki H., Suesada Y., and Nakajima N., 2021, The storage temperature affects flesh firmness and gene expression patterns of cell wall-modifying enzymes in stony hard peaches, Postharvest Biology and Technology, 181: 111658.
https://doi.org/10.1016/j.postharvbio.2021.111658

Vanalli C., Casagrandi R., Gatto M., and Bevacqua D., 2020, Shifts in thermal niche of peach under climate change, bioRxiv: 315960.
https://doi.org/10.1101/2020.09.28.315960

Yan J., Cai Z.X., Chen Z., Zhang B., Li J., Xu J., Ma R., Yu M., and Shen Z., 2024, Relationship between chilling accumulation and heat requirement for flowering in peach varieties of different chilling requirements, Agronomy, 14(8): 1637.
https://doi.org/10.3390/agronomy14081637

Yang C., Lei N., Menz C., Ceglar A., Torres-Matallana J.A., Li S., Jiang Y., Tan X., Tao L., He F., Li S., Liu B., Yang F., Fraga H., and Santos J.A., 2024, Regional uncertainty analysis between crop phenology model structures and optimal parameters, Agricultural and Forest Meteorology, 355: 110137.
https://doi.org/10.1016/j.agrformet.2024.110137

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Computational Molecular Biology 2026, Vol.16, No.3, 194-204
http://bioscipublisher.com/index.php/cmb

Review Article Open Access

Rhizosphere Microbial Diversity in Legume Cropping Systems

Weiliang Shen, Dan Luo, Xinhua Zhou
Tropical Legume Research Center, Hainan Institute of Tropical Agricultural Resources, Sanya, 572025, Hainan, China
Corresponding author: xinhua.zhou@hitar.com

Computational Molecular Biology, 2026, Vol.16, No.3 doi: 10.5376/cmb.2026.16.0014
Received: 23 Apr., 2026
Accepted: 28 May, 2026
Published: 12 Jun, 2026

Copyright © 2026 Shen et al., This is an open access article published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preferred citation for this article:
Shen W.L., Luo D., and Zhou X.H., 2026, Rhizosphere microbial diversity in legume cropping systems, Computational Molecular Biology, 16(3): 194-204 (doi: 10.5376/cmb.2026.16.0014)

Abstract Rhizospheric microorganisms associated with leguminous crops constitute a vital component in maintaining the stability of agroecosystems and promoting healthy plant growth; their diversity and functions directly influence soil nutrient cycling, nitrogen fixation efficiency, and crop stress tolerance. This study presents a comprehensive review of rhizospheric microbial diversity in leguminous crops, systematically analyzing the characteristics of the rhizosphere microenvironment, the composition of microbial communities, and their ecological functions. Particular emphasis is placed on exploring the roles of various microorganisms-including bacteria, fungi, and archaea-in plant nutrient uptake, disease suppression, and the maintenance of soil health. Furthermore, the article summarizes the primary factors influencing rhizospheric microbial diversity-such as plant genotype, tillage systems, fertilization methods, and environmental conditions-and introduces the application of modern research technologies, including high-throughput sequencing, metagenomics, and bioinformatics, in the study of rhizosphere microecology. Additionally, using soybean cropping systems as a case study, the paper analyzes variations in microbial community structure under different cultivation patterns and discusses their significance for sustainable agricultural development. Finally, this study outlines the challenges currently facing this field of research and identifies future directions-such as synthetic microbiomes, precision agriculture, and microbial engineering-with the aim of providing a theoretical foundation for the green and efficient production of leguminous crops and the effective management of agroecosystems.

Keywords Leguminous crops; Rhizospheric microorganisms; Microbial diversity; Symbiotic nitrogen fixation; Sustainable agriculture

1 Introduction

The rhizosphere, the narrow soil zone influenced by plant roots, harbors an immense and still largely unexplored diversity of microorganisms that shape plant nutrition, health, and soil functioning (Chukwuneme and Babalola, 2025). In legume-based systems, this belowground biodiversity underpins key agroecosystem services, particularly biological nitrogen fixation and improved soil fertility, making legumes central to sustainable agriculture and food security (Schaedel et al., 2021). Understanding how legumes assemble and interact with their rhizosphere microbiomes is therefore critical for designing low-input, high-efficiency cropping systems.

Research has established that the rhizosphere microbiome is a central driver of nutrient cycling, carbon sequestration, and ecosystem functioning in terrestrial systems. Microbial communities associated with plant roots form a “second genome” whose collective genes far exceed those of the host plant and are crucial for growth promotion, stress tolerance, and disease suppression. For legumes, the best-known interaction is the symbiosis with rhizobia, but it is now clear that non-rhizobial members of the rhizosphere and nodule microbiome also contribute to nodule formation, plant fitness, and broader agroecosystem benefits (Yang et al., 2024). Given the pressures of climate change and the need to reduce synthetic fertilizer inputs, harnessing this microbial diversity has major significance for sustainable intensification of legume cropping systems.

Recent advances in high-throughput sequencing and multi-omics approaches have transformed understanding of rhizosphere microbiomes by enabling cultivation-independent analysis of taxonomic and functional diversity (Chukwuneme and Babalola, 2025). Large-scale comparative studies show that legumes assemble rhizosphere communities with lower overall diversity but with strong enrichment of nitrogen-cycling taxa and nitrogen-fixing genes relative to non-legumes, revealing a pronounced functional specialization for nitrogen acquisition. At the

same time, factors such as plant species, soil type, and land-use history jointly shape microbial community structure, with bulk soil serving as the main reservoir from which rhizosphere communities are selected. These insights highlight both the selectivity of legumes in recruiting beneficial microbes and the context dependence of microbiome composition across soils and management regimes.

Despite rapid progress, important knowledge gaps remain in how rhizosphere microbial diversity in legume systems can be systematically characterized, predicted, and managed at field and farming-system scales. Most work has focused on individual symbioses or single legume species, while the broader networks of beneficial, pathogenic, and even human-pathogenic microorganisms in the legume rhizosphere remain only partially resolved. The present review aims to synthesize current knowledge on the taxonomic and functional diversity of rhizosphere microbiomes in legume cropping systems, with particular attention to nitrogen fixation, plant health, and agroecosystem services. It also seeks to integrate emerging omics-based insights with ecological theory on community assembly, and to identify opportunities to manipulate rhizosphere communities-via breeding, inoculants, and cropping system design-to enhance the sustainability and resilience of legume-based agriculture.

2 Characteristics of the Rhizosphere Microenvironment in Legume Cropping Systems

2.1 Root exudates and rhizosphere formation

Legume roots release a wide array of primary and secondary metabolites (sugars, organic acids, amino acids, flavonoids, phenolics) that both feed and signal to rhizosphere microorganisms, thereby structuring the microbial community close to the root (Chen and Liu, 2024). Temporal shifts in exudate composition during plant development generate a “chemical succession” that selects for microbes with matching substrate preferences, creating predictable patterns of community assembly along the soil-root interface (Zhou et al., 2022).

Specific exudate components, especially flavonoids and related phenolic compounds, act as key signaling molecules guiding symbioses and broader rhizomicrobiome recruitment in legumes (Chen et al., 2022; Kumar et al., 2024). These compounds mediate chemotaxis and colonization by beneficial rhizobacteria and mycorrhizal fungi, and under nutrient limitations or other stresses can be modulated to favor microbes that enhance nutrient acquisition and stress tolerance (Gong et al., 2023). In diversified or intercropped systems, changes in legume rhizodeposition can further adjust metabolite profiles and microbial functions, strengthening beneficial interactions (Qiao et al., 2024).

2.2 Soil physicochemical properties in legume rhizospheres

Legume establishment and rhizosphere activity progressively modify soil physicochemical properties, often improving pH status, organic matter, and nutrient availability. For example, legume planting in saline or degraded soils has been associated with decreased salinity and pH, and increased soil organic carbon and nitrogen pools, promoting more diverse and functionally complex bacterial networks (Liu et al., 2021; Amaya-Gómez et al., 2025). Over years of perennial or woody legume growth, rhizosphere soils can show rising organic matter and available P and K, coupled with elevated enzyme activities (e.g., urease, phosphatase) that support nutrient turnover and microbial proliferation (Ren et al., 2021; Mu et al., 2024).

Soil pH emerges as a central driver of rhizosphere bacterial diversity, composition, and function, often outweighing vegetation type or other variables (Wan et al., 2020). In acidic cropping soils, lower pH (≤5.5) is linked to reduced bacterial abundance and downregulated genes involved in C, N, P, and S cycling, which can constrain crop yield (Abd-Alla et al., 2023). Conversely, amendments such as lime, organic manure, or biochar can adjust pH and nutrient status, shifting bacterial communities toward taxa (e.g., Actinobacteria, Proteobacteria) associated with enhanced disease suppression and improved plant physiological status (Ren et al., 2021; Chen et al., 2022).

2.3 Symbiotic nitrogen fixation and nutrient cycling

Symbiotic nitrogen fixation (SNF) between legumes and rhizobia is a core process structuring rhizosphere function, converting atmospheric N_2 into plant-available ammonia in nodules and enriching soil N pools (Neda, 2021). The effectiveness of SNF varies among rhizobial strains and is strongly influenced by soil conditions and

plant demand; highly efficient symbioses can install endosphere and rhizosphere microbiomes that promote nutrient uptake beyond simple nutrient supply, including accumulation of beneficial Actinobacteria in roots (Lagunas et al., 2023). Over time, fixed nitrogen is transferred to soil through rhizodeposition, senescing roots, and residues, supporting non-legume crops and stimulating broader microbial activities in diversified systems (Qiao et al., 2024).

Excessive mineral N fertilization can suppress SNF by interfering with nodulation signaling, rhizobial chemotaxis to roots, and nitrogenase activity, thereby weakening the mutualism and altering rhizosphere microbial relationships (Abd-Alla et al., 2023). In contrast, organic inputs such as compost and vermicompost generally enhance nodulation, nodule biomass, plant growth, and yield, while improving soil biological quality and nitrogen availability (Mu et al., 2024). Free-living nitrogen-fixing bacteria in the legume rhizosphere, stimulated by legume-derived exudates (including flavonoids and coumarins), further contribute to N inputs and interact functionally with symbiotic rhizobia under intercropping or rotation schemes (Chen et al., 2022; Qiao et al., 2024).

3 Composition and Diversity of Rhizosphere Microbial Communities

3.1 Bacterial diversity in legume rhizospheres

Large-scale comparative analyses show that legume rhizospheres often exhibit lower bacterial α-diversity than non-legumes but are strongly enriched in nitrogen-cycling taxa and nitrogen-fixing genes, suggesting a specialized, function-driven bacterial assembly (Qin et al., 2025). Typical legume rhizospheres are dominated by Proteobacteria and Bacteroidetes, with notable representation of Bradyrhizobiaceae, Rhizobiaceae, and other diazotrophs, reflecting the central role of biological nitrogen fixation (Pivato et al., 2021; Yang et al., 2024). In intercropping systems, legume presence can shift bacterial communities in associated non-legume rhizospheres toward copiotrophic and nitrogen-transforming assemblages, often without major changes in overall diversity indices (Pang et al., 2022).

Cropping mode and fertilization regimes substantially modulate bacterial community richness, structure, and potential function in legume-involving systems. In hulless barley-pea mixed cropping, increasing nitrogen and phosphorus inputs caused a hump-shaped response in bacterial α-diversity, with mixed cropping supporting higher diversity than monocropping and enriching Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium relative to cereal monoculture (Fu et al., 2023; Guo et al., 2024). Sugarcane-peanut and cereal-legume intercrops similarly increased bacterial richness and the diversity of nitrogen-fixing bacteria in rhizosphere and bulk soils compared with monocultures, aligning with improved crop performance and altered soil pH and phosphorus availability (Pang et al., 2022; Yang et al., 2023). These findings indicate that legume integration and moderate nutrient inputs can promote diverse, functionally advantageous bacterial consortia in the rhizosphere.

3.2 Fungal diversity and mycorrhizal associations

Arbuscular mycorrhizal fungi (AMF) are key fungal components of legume rhizospheres, enhancing phosphorus acquisition and stress tolerance while interacting with rhizobia-dependent nitrogen fixation (Alimi et al., 2021; Pires et al., 2021). Surveys of indigenous South African legumes revealed diverse AMF communities dominated by Glomus and Acaulospora, with species richness and spore density varying markedly among hosts and being strongly structured by soil properties such as texture and nutrient status (Alimi et al., 2025). Morphological assessments across leguminous and non-leguminous crops further identified Acaulospora, *Funneliformis*, *Gigaspora*, Glomus, and *Rhizophagus* as common AMF genera, with legume hosts often supporting higher spore counts and colonization frequencies, underlining their importance as AMF reservoirs (Pires et al., 2021).

Intercropping and integrated crop-livestock systems that include legumes can enhance AMF diversity, colonization, and inoculum potential in subsequent legume phases. In maize-soybean intercropping, AMF α-diversity in soybean rhizosphere soil increased relative to monoculture at comparable nitrogen levels, and Glomus-related taxa were dominant in both soil and roots, with their abundance responding to nitrogen inputs and crop identity (Figure 1) (Zhang et al., 2020; Alimi et al., 2021). In systems where grasses are intercropped with cowpea or pigeon pea, AMF spore density, colonization, and species richness in legume-associated rhizospheres

rise, and these mycorrhizal improvements correlate positively with soybean productivity in following crops (Pires et al., 2021; Guo et al., 2024). Together, these studies show that legumes and diversified management can foster rich AMF assemblages that contribute to nutrient use efficiency and yield stability.

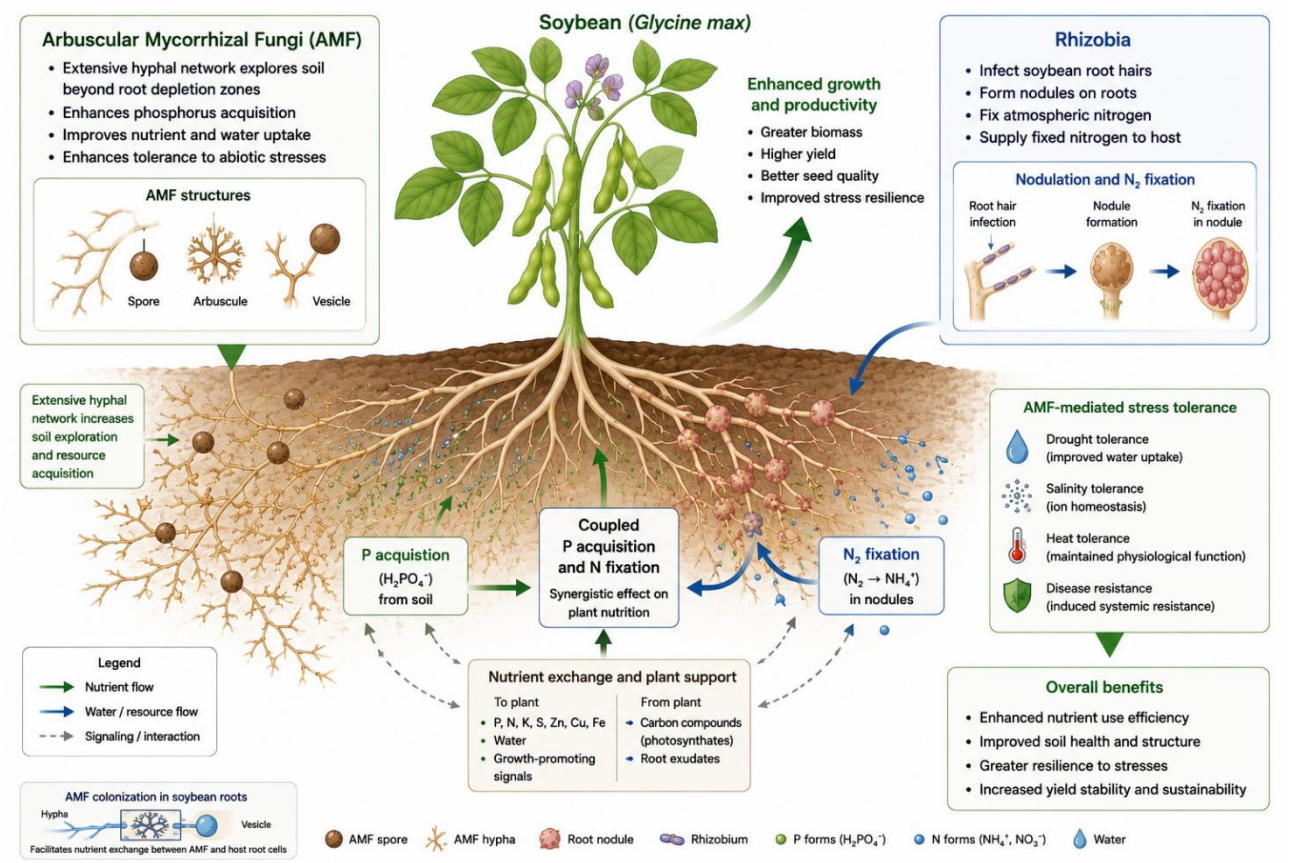


Figure 1 Illustrates the symbiotic interactions among arbuscular mycorrhizal fungi (AMF), legume roots, and rhizobia in the rhizosphere. AMF hyphae enhance phosphorus acquisition and stress tolerance, while rhizobia contribute to biological nitrogen fixation, together improving nutrient uptake and plant growth

3.3 Archaea, viruses, and other microorganisms

Beyond bacteria and fungi, legume rhizospheres host diverse archaea, phages, and other viruses whose ecological roles are only beginning to be elucidated. Conceptual and empirical work on rhizosphere “zoos” highlights that archaea, viruses, and other eukaryotes coexist with bacteria and fungi, contributing to nutrient turnover, organic matter decomposition, and plant health outcomes. Archaea, although less intensively characterized in legumes, are recognized as components of rhizosphere communities that may participate in nitrogen and carbon cycling, while protists and nematodes further shape microbial food webs and nutrient flows (Qin et al., 2025).

Recent advances in viromics demonstrate that soil and rhizosphere viral communities are taxonomically and functionally diverse, exhibiting strong spatial and temporal dynamics and exerting top-down control on bacterial hosts. Phages in rhizospheres regulate pathogen densities and suppress or exacerbate disease depending on whether they target pathogens or pathogen-suppressing bacteria, thereby influencing soil suppressiveness and plant health (Yang et al., 2023). Crop management and rotation can “prime” rhizosphere viral assemblages, altering DNA and RNA virus diversity and activity near roots and driving bacterial community succession through Kill-the-Winner dynamics (Braga et al., 2020; Muscatt et al., 2022). These findings, together with evidence that phage pressure can modify bacterial diversity and nitrogen availability, emphasize that viruses and associated microbial predators are integral but underappreciated drivers of rhizosphere community assembly and function in legume cropping systems (Wang et al., 2024).

4 Factors Influencing Rhizosphere Microbial Diversity

4.1 Plant genotype and species differences

Plant genetic variation shapes rhizosphere microbiome assembly by altering host filtering strength and the interface between rhizosphere and internal compartments. In *Medicago truncatula*, soil origin mainly structured rhizosphere communities, but plant genotype exerted strong effects in the root endosphere, indicating that different genotypes act as stronger or weaker microbial filters and influence which taxa progress inward (Brown et al., 2020). In soybean, soil type again dominated community composition, yet host genotype subtly "tuned" rhizosphere assembly and microbe-microbe interaction networks, demonstrating cooperative control by plant genetics and soil microbiome pool.

Species-level differences within legumes also generate distinct rhizosphere communities. Across five Phaseolus species, each recruited a characteristic bacterial assemblage, with notable contrasts in richness and dominant phyla; for example, *Phaseolus lunatus* showed the highest richness and an Acidobacteria-enriched rhizosphere, whereas Actinobacteria dominated several other species (Yang et al., 2023). More broadly, diversity in soil microbial community structure was greater among legume species than among grass species, and legumes generally supported higher bacterial diversity and enriched fungi, underscoring the strong niche differentiation imposed by legume species identity.

4.2 Agricultural management practices

Tillage and fertilization regimes alter soil structure, resources, and disturbance intensity, thereby reshaping rhizosphere-associated microbiomes in legume-based systems. In a long-term corn-soybean system, both tillage and fertility significantly shifted bacterial, fungal and oomycete communities, with no-till favoring ecological guilds such as arbuscular mycorrhizal fungi, mycoparasites, and nematophagous fungi, while conventional tillage promoted saprotrophs and plant pathogens. Fertilization further modified bacterial and fungal β-diversity and supported copiotrophic bacteria and Fusarium under conventional regimes, indicating that intensive inputs select for fast-growing competitors rather than mutualists (Srour et al., 2020).

Cropping sequences involving legumes also drive rhizosphere diversity and N-cycling potential. In sorghum systems, precropping with cowpea or soybean, compared with maize or no precrop, significantly altered rhizosphere bacterial α- and β-diversity and shifted key nitrogen-cycling genes (e.g., amoC, narH, gltB, glnA, ureC), with legume rotations enriching several N-transformation pathways (Enagbonma et al., 2025). In sugarcane rotations, soybean and peanut residues increased microbial biomass C, C mineralization, and nitrification capacity, although high-N soybean residues released more mineral N than low-N peanut residues, revealing crop- and residue-specific impacts on microbial functions linked to fertility (Paungfoo-Lonhienne et al., 2021).

4.3 Environmental and climatic factors

Environmental variables such as soil type, pH, and climate gradients strongly regulate rhizosphere microbial diversity and its functional consequences. Along an altitudinal and climatic gradient in mountain ecosystems, geographical and climatic factors directly and indirectly controlled rhizosphere bacterial and fungal diversity, with bacterial α-diversity and particular dominant taxa exerting strong positive or negative effects on soil multifunctionality. The balance of these effects determined net multifunctionality, and higher richness at the phylum level generally led to gains in multiple soil functions, highlighting the sensitivity of rhizosphere communities to long-term climatic contexts (Yang et al., 2023).

Among soil properties, pH is a particularly powerful predictor of rhizosphere bacterial diversity, structure, and function. In acidic crop soils, communities in pH < 5.5 versus > 5.5 clustered into distinct groups, with higher pH associated with greater bacterial abundance and diversity and more active nutrient-cycling functions (C, N, P, S) (Wan et al., 2020). In more acidic soils, bacterial interaction networks suggested reduced competition but downregulated functional genes, implying constrained ecosystem services and potentially lower crop yields when pH is not managed (Wan et al., 2020).

5 Functional Roles of Rhizosphere Microbiota

5.1 Plant growth promotion and nutrient acquisition

Legume rhizospheres are enriched in PGPR and other beneficial microbes that promote growth via biological nitrogen fixation, phosphate solubilization, siderophore production, and phytohormone synthesis (Timofeeva et al., 2023). Rhizospheric diazotrophs and phosphate-solubilizing bacteria increase plant-available N and P, while co-inoculation strategies combining these groups often outperform single strains by simultaneously enhancing nutrient supply and root development (Zeng et al., 2022). In grain legumes, such PGPR also stimulate nodulation and strengthen rhizobium-legume symbioses, further boosting nitrogen inputs and yield (Swarnalakshmi et al., 2020).

Beyond direct nutrient mobilization, PGPR and mycorrhizal fungi improve nutrient use efficiency and root architecture, enabling legumes to exploit heterogeneous soil resources (Tahat et al., 2020; Timofeeva et al., 2023). Reviews on rhizosphere-plant interactions highlight that these microbes alter root physiology, exudation, and transporter activity, thereby increasing uptake of N, P, and micronutrients while supporting growth under nutrient deficiency (Hakim et al., 2021). Harnessing these functions through microbial fertilizers and seed inoculants is increasingly proposed as a means to reduce mineral N and P inputs without compromising productivity (De Andrade et al., 2023).

5.2 Disease suppression and stress resistance

Rhizosphere microbiota contribute to legume health by forming a first line of defense against soil-borne pathogens. Beneficial bacteria and fungi protect roots via antibiosis, competition for nutrients and niches, parasitism, and induction of systemic resistance (Tahat et al., 2020; Hakim et al., 2021). In common bean, cultivars bred for resistance to *Fusarium oxysporum* harbor rhizospheres enriched in *Pseudomonadaceae*, *Bacillaceae*, and *Cytophagaceae*, along with genes for antifungal metabolites, indicating that host genetics can co-select disease-suppressive communities.

Microbiome-mediated resistance also extends to abiotic stresses such as drought, salinity, and heat. Reviews on harnessing plant-microbe interactions and rhizosphere engineering note that tailored microbial consortia and stress-resilient PGPR can improve water use efficiency, modulate stress hormones, and maintain growth under adverse conditions (Yusuf et al., 2025). Specific PGPR strains from legume rhizospheres, such as Pseudomonas chlororaphis IRHB3 in soybean, both recruit functional bacteria involved in nutrient cycling and activate jasmonate-mediated resistance, thereby simultaneously enhancing growth and suppressing root rot in the field (Kumar and Dubey, 2020; Wei et al., 2024).

5.3 Soil health and ecosystem sustainability

Soil health is tightly linked to the diversity and activity of rhizosphere microorganisms, which regulate nutrient recycling, aggregate stability, and greenhouse gas fluxes (Tahat et al., 2020; Xing et al., 2025). Beneficial rhizosphere microbes in legume systems improve soil structure and organic matter turnover, support balanced nutrient cycles, and increase resilience of soil functions to disturbance (Hakim et al., 2021; Sharma et al., 2025). Leguminous cover crops and legume-based intercropping have been shown to enhance rhizosphere microbial diversity, enrich taxa involved in nitrogen fixation and organic matter decomposition, and improve soil pH, organic carbon, and nutrient availability relative to monocultures (Jalloh et al., 2024; Pokharel et al., 2025).

Microbial-based strategies are increasingly recognized as central to sustainable agriculture, providing eco-friendly alternatives to intensive chemical inputs. Reviews on soil microbial resources and rhizosphere manipulation emphasize that bio-inoculants and management practices that favor native beneficial communities can simultaneously enhance crop productivity, soil fertility, and environmental quality (Mahmud et al., 2021; Sharma et al., 2025). Dissecting rhizosphere microbiomes into environment-dominated and plant genetic-dominated components further suggests complementary levers-agronomic management and breeding-for designing legume systems that maintain functionally robust microbiomes and deliver long-term ecosystem services (Xun et al., 2024; Xing et al., 2025).

6 Molecular and Analytical Approaches in Rhizosphere Microbial Research

6.1 High-throughput sequencing technologies

Metagenomics and marker-gene amplicon sequencing are the core high-throughput tools for rhizosphere studies, enabling cultivation-independent profiling of complex communities. Review work highlights shotgun metagenomics for unbiased recovery of genomes and functional genes, and 16S rRNA or ITS amplicon sequencing for efficient taxonomic surveys of bacteria and fungi in root-associated soils (Rajguru et al., 2024). In legume rhizospheres, such approaches reveal dominant phyla and shifts in community structure under contrasting fertilization regimes, for example in soybean grown with organic versus inorganic inputs (Babalola et al., 2025).

Recent methodological advances focus on scalability and sensitivity for plant-associated samples. A high-throughput 16S rRNA library-preparation protocol using magnetic beads for DNA extraction directly from roots and exonuclease purification before the second PCR step improves handling and detection of minor bacterial taxa, yet produces community profiles comparable to commercial kits in roots and soils (Kumaishi et al., 2022). Standardized field-to-sequencing protocols have also been developed for collecting soil, rhizosphere, and root endosphere fractions and running validated 16S pipelines, facilitating cross-study comparisons across plant species and habitats.

6.2 Bioinformatics and data analysis

Downstream of sequencing, dedicated bioinformatics pipelines convert read data into diversity metrics, taxonomic profiles, and functional inferences. A practical guide summarizes recommended workflows for amplicon and metagenomic analyses, detailing quality control, denoising or clustering, taxonomic assignment, diversity estimation, and advanced methods such as network analysis and machine learning to extract ecological meaning from microbiome datasets (Liu et al., 2020). Habitat-specific optimization is increasingly emphasized: evaluation of 35,889 microbe species and >150,000 microbiomes produced Qscore, a framework to select optimal 16S regions and strategies for different ecosystems, achieving profiling precision close to shotgun metagenomes (Zhang et al., 2023).

Pipeline choice can strongly bias apparent community structure and diversity. Comparative assessments of 16S amplicon workflows using mock communities and environmental datasets show large differences among tools such as Mothur, QIIME1, QIIME2, MEGAN, DADA2, and others; in one study, QIIME2 markedly reduced false positives and improved taxonomic and diversity estimates relative to alternatives (Straub et al., 2020). Another comparison of OTU- and ASV-based pipelines found that ASV methods such as DADA2 and USEARCH-UNOISE3 improved resolution and specificity, while some OTU workflows inflated richness and spurious taxa, underscoring the need for careful pipeline selection in rhizosphere research (Prodan et al., 2020).

6.3 Experimental models and cultivation techniques

Sequencing-based surveys are increasingly complemented by experimental models that allow mechanistic tests of plant-microbe interactions. Synthetic microbial communities (SynComs) constructed from cultured rhizosphere isolates have been systematically reviewed as tools to bridge complexity and control; SynComs ranging from a few to ~190 strains, typically dominated by Proteobacteria, Actinobacteria, and Firmicutes, are deployed on diverse plant hosts and growth systems to dissect functions such as colonization, competition, and plant growth promotion (Marín et al., 2021). A 16-member synthetic soil community derived from a single rhizosphere was further optimized for reproducibility, tunable starting composition, long-term cryopreservation, and use in standardized fabricated ecosystem devices (EcoFABs), enabling controlled plant-microbe experiments across laboratories (Coker et al., 2022).

Cultivation remains crucial for isolating functional strains and validating metagenomic predictions, but many rhizosphere microbes are recalcitrant to standard media. Improved culture-dependent strategies-such as modifying gelling agents and autoclaving steps-enhanced recovery of wheat rhizosphere bacteria from <1% to up to ~2.5% of metagenomic OTUs and yielded isolates with multiple plant growth-promoting traits (Youseif et al., 2021). Microcosm and multitrophic systems, supported by detailed manuals on soil sterilization, isolation, inoculation, and microcosm design, further allow controlled investigation of bacteria, fungi, protists, and nematodes together, better reflecting the complexity of rhizosphere food webs.

Computational Molecular Biology 2026, Vol.16, No.3, 194-204
http://bioscipublisher.com/index.php/cmb

7 Case Study: Rhizosphere Microbial Diversity in Soybean Cropping Systems

7.1 Overview of soybean rhizosphere microbiota

Soybean rhizospheres host complex, multi-kingdom microbial communities whose composition and function vary strongly across soils and regions. A metagenomic survey across 13 major soybean-producing regions in China identified over 43,000 microbial species (bacteria, archaea, fungi and viruses), with clear site-specific clustering and 556 hub taxa correlated with yield and involved in C, N, P and S cycling (Ren et al., 2025). Comparative work further shows that rhizosphere communities differ markedly from bulk soil, with enrichment of genera such as Rhizobium, Novosphingobium, Phenylobacterium, Streptomyces and Nocardioides and convergence in functional pathways linked to xenobiotic degradation, plant-microbe interactions and nutrient transport.

Soil background and plant genetics jointly modulate soybean rhizosphere assembly. Across contrasting soils and genotypes, soil type has the dominant effect, but soybean genotype subtly "tunes" recruitment and microbe-microbe networks, with wild *Glycine soja* maintaining higher rhizosphere diversity than domesticated lines (Figure 2). Other studies show that rhizocompartments (bulk soil, rhizosphere, roots, nodules) host distinct bacterial assemblages, and that rhizosphere networks include strong correlations between rhizobia and non-rhizobial taxa, which can in turn influence nodulation patterns and symbiotic efficiency (Han et al., 2020).

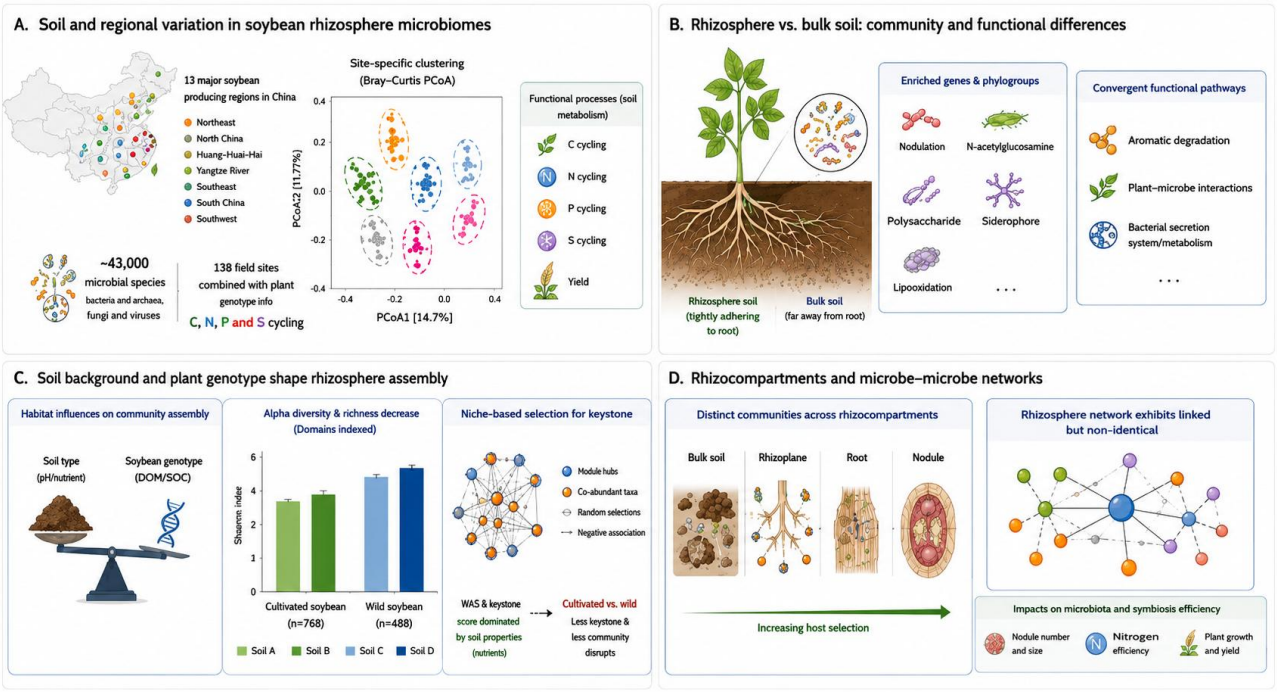


Figure 2 Multi-scale characterization of soybean rhizosphere microbiomes across contrasting soils, regions, and rhizocompartments

7.2 Effects of cropping patterns on microbial diversity

Cropping patterns substantially reshape soybean rhizosphere diversity, composition, and functional potential. In maize-soybean relay strip intercropping, soybean rhizosphere bacterial diversity increased compared with monoculture, with higher richness of Pseudomonas, Bacillus and other antagonists; several intercropping-derived strains suppressed Fusarium root rot and one *Pseudomonas chlororaphis* strain (IRHB3) promoted root growth and seedling survival under pathogen pressure (Chang et al., 2022). In coastal saline soils, soybean-corn intercropping altered soil C, N, P and salinity and significantly changed bacterial and fungal communities; intercropping increased Chao1 richness, shifted dominant phyla (Proteobacteria, Actinobacteria, Acidobacteria, Chloroflexi; Ascomycota, Mortierellomycota, Basidiomycota) and enriched taxa linked to nutrient cycling and bioremediation (Nyimbo et al., 2025).

At finer scales, belt/strip planting and intercropping layouts under field conditions also modulate multi-kingdom communities and link to yield traits. Metagenomic analysis of soybean-maize strip systems showed that bacteria

and viruses dominate inter-root communities and are more sensitive to planting mode than fungi or archaea, with shifts in Pseudomonas, rhizobia and other genera across modes that favored either soybean or maize yields (Wang et al., 2024). In maize-soybean systems compared over three years, conversion from monocropping to intercropping increased microbiome network modularity and functional diversity and enriched genes for nitrate assimilation, nitrification and dissimilatory nitrate reduction, changes that were closely related to higher yields in intercropped soybean (Shu et al., 2024).

7.3 Implications for sustainable agriculture

Evidence from soybean systems indicates that managing rhizosphere microbiota offers powerful levers for sustainable intensification. Long-term comparisons of soybean continuous monocropping versus maize-soybean rotation show that both long-term continuous soybean (13 years) and rotation can elevate soil pH and available N, P, K, and increase network complexity, while enriching beneficial Bradyrhizobium, Gemmatimonas and Mortierella and reducing pathogenic Fusarium compared with short-term continuous soybean (Liu et al., 2020). Similarly, in wheat-soybean double-cropping, introducing wheat/soybean-wheat/maize rotations improved soybean yield and a soil fertility index and shifted rhizosphere fungi toward plant growth-promoting, nematophagous and biocontrol groups, while continuous wheat/soybean favored lignocellulose degraders and pathogens (Sun et al., 2022).

Conceptual and review work suggests that such cropping-based microbiome effects can be deliberately exploited. One framework proposes dividing rhizosphere microbiota into environment-dominated and plant genetic-dominated components, with agronomic practices (e.g., rotations, intercropping, reduced tillage) used to steer the former and breeding used to enhance the latter, thereby stabilizing beneficial consortia in crops like soybean (Xun et al., 2024). More broadly, rhizosphere microbiome engineering-using indigenous consortia and designed inoculants-has been highlighted as a route to reduce synthetic inputs, enhance yield and resilience, and align soybean production with long-term soil health and environmental sustainability goals (Mahmud et al., 2021).

8 Challenges, Future Perspectives, and Conclusions

Despite rapid methodological advances, major knowledge gaps limit the application of rhizosphere microbiomes in legume-based systems. A key challenge is the complex assembly mechanisms of rhizosphere communities, where environment-dominated and plant genetic-dominated components interact in ways that are still poorly quantified, especially under realistic field conditions. This complexity hampers prediction of microbiome responses to agronomic practices or legume genotypes, and constrains efforts to design stable microbial consortia for enhanced nitrogen fixation and stress resilience.

Translating microbiome insights into reliable bioinoculants also remains difficult. Many beneficial strains perform well in controlled experiments but fail under variable soils, climates, and management, in part because interactions with native microbiota and environmental heterogeneity are insufficiently understood. Challenges specific to nitrogen-fixing systems include the ecological competitiveness of inoculant strains, context-dependent performance of symbiotic and free-living diazotrophs, and the need to match microbial partners with host genetics and local microbiomes to achieve consistent field-level benefits.

Future rhizosphere research in legumes will likely focus on predictive and integrative frameworks that connect soil factors, plant genetics, and management to microbiome structure and function. Building data-driven, high-throughput models that quantify how soil properties and agronomic practices shape environment-dominated microbiome components is a priority for precise rhizosphere regulation in real cropping systems. At the same time, identifying genes and loci controlling plant genetic-dominated microbiome fractions will support breeding of “microbiome-assisted” legumes and potentially N-self-fertilizing crops.

There is also strong momentum toward microbiome engineering and synthetic communities tailored to legume growth stages and stresses. Multi-omics meta-analyses already reveal developmental stage-specific growth-promoting marker bacteria in legumes that could guide design of multi-species inoculants. Conceptual frameworks such as microbiome-mediated smart agriculture systems emphasize combining reduced tillage,

biofertilization, increasingly complex synthetic microbiomes, and even plant genome editing to recruit beneficial microbiota and improve resilience to drought and other stresses.

Research on rhizosphere microbial diversity in legume cropping systems has demonstrated that legumes assemble functionally specialized microbiomes with strong impacts on nitrogen fixation, nutrient cycling, and stress tolerance. Reviews of legume microbiomes highlight that rhizobia operate within broader rhizosphere and nodule communities, where non-rhizobial bacteria and other microbes contribute to nodule formation, legume fitness, and agroecosystem services, including reduced fertilizer needs and pollution. Harnessing these assemblages is therefore central to strategies aiming at sustainable intensification and climate-friendly nitrogen management.

Moving forward, realizing the full potential of legume-associated rhizosphere microbiomes will require coordinated advances from microns to field scales. Sustainable agriculture perspectives stress that exploiting nitrogen-fixing rhizobacteria and other plant growth promoters depends on overcoming challenges in bioinoculant consistency, integrating omics-based discovery with agronomy, and fostering large-scale collaboration among researchers, industry, and farmers. By combining predictive microbiome management with breeding, intercropping, and reduced-chemical inputs, legume systems can become key platforms for microbiome-based solutions that support soil health, productivity, and ecosystem sustainability.

Acknowledgments

Thanks to the reviewers for providing detailed comments and guidance on the manuscript of this study. The reviewers' keen insights into the issues and attention to detail have greatly benefited the authors.

Conflict of Interest Disclosure

The authors affirm that this research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest.

References

Abd-Alla M.H., Al-Amri S.M., and El-Enany A.W.E., 2023, Enhancing Rhizobium-Legume symbiosis and reducing nitrogen fertilizer use are potential options for mitigating climate change, Agriculture, 13(11): 2092.
https://doi.org/10.3390/agriculture13112092

Alimi A.A., Adeleke R.A., and Moteetee A.N., 2021, Soil environmental factors shape the rhizosphere arbuscular mycorrhizal fungal communities in South African indigenous legumes (Fabaceae), Biodiversitas, 22(5): 2466-2476.
https://doi.org/10.13057/biodiv/d220503

Alimi A.A., Ezeokoli O.T., Adeleke R.A., and Moteetee A.N., 2025, Arbuscular mycorrhizal fungal communities and relationship with edaphic factors in the rhizospheric soil of Fabaceae in semi-arid South Africa, Scientific African, 27: e02997.
https://doi.org/10.1016/j.sciaf.2025.e02997

Amaya-Gómez C., Flórez-Martínez D., Cayuela M.L., and Tortosa G., 2025, Compost and vermicompost improve symbiotic nitrogen fixation, physiology and yield of the Rhizobium-legume symbiosis: a systematic review, Applied Soil Ecology, 210: 106051.
https://doi.org/10.1016/j.apsoil.2025.106051

Babalola O.O., Osuji I.J., and Akanmu A.O., 2025, Amplicon-based metagenomic survey of microbes associated with the organic and inorganic rhizosphere soil of *Glycine max* L., BMC Genomic Data, 26(1): 40.
https://doi.org/10.1186/s12863-025-01333-2

Braga L.P.P., Spor A., Kot W., Breuil M.C., Hansen L.H., Setubal J.C., and Philippot L., 2020, Impact of phages on soil bacterial communities and nitrogen availability under different assembly scenarios, Microbiome, 8(1): 52.
https://doi.org/10.1186/s40168-020-00822-z

Brown S.P., Grillo M.A., Podowski J.C., and Heath K.D., 2020, Soil origin and plant genotype structure distinct microbiome compartments in the model legume *Medicago truncatula*, Microbiome, 8(1): 139.
https://doi.org/10.1186/s40168-020-00915-9

Chang X., Wei D., Zeng Y., Zhao X., Hu Y., Wu X., Song C., Gong G., Chen H., Yang C., Zhang M., Liu T., Chen W., and Yang W., 2022, Maize-soybean relay strip intercropping reshapes the rhizosphere bacterial community and recruits beneficial bacteria to suppress Fusarium root rot of soybean, Frontiers in Microbiology, 13: 1009689.
https://doi.org/10.3389/fmicb.2022.1009689

Chen D., Wang X., Carrión V.J., Yin S., Yue Z., Liao Y., Dong Y., and Li X., 2022, Acidic amelioration of soil amendments improves soil health by impacting rhizosphere microbial assemblies, Soil Biology and Biochemistry, 167: 108599.
https://doi.org/10.1016/j.soilbio.2022.108599

Chen L., and Liu Y., 2024, The function of root exudates in the root colonization by beneficial soil rhizobacteria, Biology, 13(2): 95.
https://doi.org/10.3390/biology13020095

Chen Z., Wang W., Chen L., Zhang P., Liu Z., Yang X., Shao J., Ding Y., and Mi Y., 2024, Effects of pepper-maize intercropping on the physicochemical properties, microbial communities, and metabolites of rhizosphere and bulk soils, Environmental Microbiome, 19(1): 108.
https://doi.org/10.1186/s40793-024-00653-7

Chukwuneme C.F., and Babalola O.O., 2025, Microbial diversity and function in the rhizosphere microbiome: Driving forces and monitoring approaches, Agrosystems, Geosciences and Environment, 8(3): e70169.
https://doi.org/10.1002/agg2.70169

Coker J.A., Zhalnina K., Marotz C., Thiruppathy D., Tjuanta M., D'Elia G., Hailu R., Mahosky T., Rowan M., Northen T.R., and Zengler K., 2022, A reproducible and tunable synthetic soil microbial community provides new insights into microbial ecology, mSystems, 7(6): e00951-22.
https://doi.org/10.1128/msystems.00951-22

De Andrade L.A., Santos C.H.B., Frezarin E.T., Sales L.R., and Rigobelo E.C., 2023, Plant growth-promoting rhizobacteria for sustainable agricultural production, Microorganisms, 11(4): 1088.
https://doi.org/10.3390/microorganisms11041088

Enagbonma B.J., Modise D.M., and Babalola O.O., 2025, Effects of legume-cereal rotation on sorghum rhizosphere microbial community structure and nitrogen-cycling functions, MicrobiologyOpen, 14(5): e70085.
https://doi.org/10.1002/mbo3.70085

Fu X., Huang Y., Fu Q., Qiu Y., Zhao J., Li J., Wu X., Yang Y., Liu H., Yang X., and Chen H., 2023, Critical transition of soil microbial diversity and composition triggered by plant rhizosphere effects, Frontiers in Plant Science, 14: 1252821.
https://doi.org/10.3389/fpls.2023.1252821

Gong X., Feng Y., Dang K., Jiang Y., Qi H., and Feng B., 2023, Linkages of microbial community structure and root exudates: Evidence from microbial nitrogen limitation in soils of crop families, Science of the Total Environment, 881: 163536.
https://doi.org/10.1016/j.scitotenv.2023.163536

Guo T., Yao X., Wu K., Guo A., and Yao Y., 2024, Response of the rhizosphere soil microbial diversity to different nitrogen and phosphorus application rates in a hulless barley and pea mixed-cropping system, Applied Soil Ecology, 195: 105262.
https://doi.org/10.1016/j.apsoil.2023.105262

Hakim S., Naqqash T., Nawaz M.S., Laraib I., Siddique M.J., Zia R., Mirza M.S., and Imran A., 2021, Rhizosphere engineering with plant growth-promoting microorganisms for agriculture and ecological sustainability, Frontiers in Sustainable Food Systems, 5: 617157.
https://doi.org/10.3389/fsufs.2021.617157

Han Q., Ma Q., Chen Y., Tian B., Xu L., Bai Y., Chen W., and Li X., 2020, Variation in rhizosphere microbial communities and its association with the symbiotic efficiency of rhizobia in soybean, The ISME Journal, 14(8): 1915-1928.
https://doi.org/10.1038/s41396-020-0648-9

Jalloh A., Mutyambai D., Yusuf A.A., Subramanian S., and Khamis F.M., 2024, Maize edible-legumes intercropping systems for enhancing agrobiodiversity and belowground ecosystem services, Scientific Reports, 14(1): 14355.
https://doi.org/10.1038/s41598-024-64138-w

Kumaishi K., Usui E., Suzuki K., Kobori S., Sato T., Toda Y., Takanashi H., Shinozaki S., Noda M., Takakura A., Matsumoto K., Yamasaki Y., Tsujimoto H., Iwata H., and Ichihashi Y., 2022, High throughput method of 16S rRNA gene sequencing library preparation for plant root microbial community profiling, Scientific Reports, 12(1): 19289.
https://doi.org/10.1038/s41598-022-23943-x

Kumar A., and Dubey A., 2020, Rhizosphere microbiome: Engineering bacterial competitiveness for enhancing crop production, Journal of Advanced Research, 24: 337-352.
https://doi.org/10.1016/j.jare.2020.04.014

Kumar G., Kumar S., Bhardwaj R., Swapnil P., Meena M., Seth C., and Yadav A.N., 2024, Recent advancements in multifaceted roles of flavonoids in plant-rhizomicrobiome interactions, Frontiers in Plant Science, 14: 1297706.
https://doi.org/10.3389/fpls.2023.1297706

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Computational Molecular Biology 2026, Vol.16, No.3, 205-217
http://bioscipublisher.com/index.php/cmb

Feature Review Open Access

Modeling Yield Formation in Sorghum Based on Temperature and Rainfall

Mingliang Zhou
Shaoxing Yuanchu Ecological Agriculture Development Co., Ltd, Shaoxing 312000, Zhejiang, China
Corresponding author: 452589568@qq.com

Computational Molecular Biology, 2026, Vol.16, No.3 doi: 10.5376/cmb.2026.16.0015
Received: 05 May, 2026
Accepted: 07 Jun., 2026
Published: 22 Jun., 2026

Copyright © 2026 Zhou, This is an open access article published under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Preferred citation for this article:
Zhou M.L., 2026, Modeling yield formation in sorghum based on temperature and rainfall, Computational Molecular Biology, 16(3): 205-217 (doi: 10.5376/cmb.2026.16.0015)

Abstract Sorghum (*Sorghum bicolor* L.) is one of the most important cereal crops in semi-arid and drought-prone regions due to its remarkable tolerance to heat and water limitation. However, sorghum productivity remains highly dependent on climatic conditions, particularly temperature and rainfall variability. This review synthesizes current knowledge on the biological and physiological mechanisms underlying sorghum yield formation and examines how temperature, rainfall, heat stress, drought stress, and their interactions influence grain number, grain weight, and overall yield stability. The review further evaluates major approaches used in sorghum yield prediction, including empirical statistical models, process-based crop simulation models, remote sensing technologies, and machine learning methods. Case studies from semi-arid regions demonstrate that reproductive-stage heat stress, post-flowering drought, and irregular rainfall distribution are among the most critical factors limiting yield. Future climate change is expected to intensify these challenges, highlighting the need for climate-resilient cultivars, adaptive agronomic management, and integrated decision-support systems. The review concludes that combining biological understanding with advanced modeling techniques can substantially improve yield prediction accuracy and support sustainable sorghum production under changing climatic conditions.

Keywords Sorghum; Temperature; Rainfall; Yield prediction; Climate change

1 Introduction

Sorghum remains one of the most important cereals for dryland farming systems because it combines food, feed, fodder, and industrial value with a comparatively strong ability to function under water limitation and high temperature. That practical resilience explains why it is deeply embedded in semi-arid production systems across Africa and Asia, and why recent reviews increasingly frame sorghum as a strategic crop for climate adaptation rather than only a “fallback” crop for marginal lands. At the same time, this reputation should not hide the fact that sorghum productivity in many regions remains low and unstable, especially where smallholders depend on rainfed systems, shallow soils, and short, erratic wet seasons. In those environments, modest shifts in seasonal onset, dry-spell frequency, or reproductive-stage heat can have outsized effects on grain set and final harvest. The literature therefore increasingly treats sorghum not simply as a hardy crop, but as a crop whose performance is highly conditional on stage-specific weather patterns and local management. That is exactly why climate-based yield modeling has become central to sorghum research and planning (Hossain et al., 2022; Liaqat et al., 2024; Mwamahonje et al., 2024).

Temperature and rainfall influence sorghum yield through different but tightly linked pathways. Temperature controls developmental pace, especially through accumulated thermal time, and therefore shapes the timing of leaf appearance, panicle initiation, anthesis, and maturity. Rainfall, by contrast, determines whether the crop can maintain canopy expansion, transpiration, reproduction, and grain filling at those stages. In practice, yield formation depends less on either variable in isolation than on their interaction with plant development. A warm season can shorten the crop cycle, reduce the duration of grain filling, and increase atmospheric demand for water; if rainfall is poorly distributed at the same time, the combined effect can sharply reduce grain number or grain weight. Conversely, moderately warm conditions paired with timely rainfall can improve stand establishment and biomass production, especially where cold stress or delayed phenology is otherwise limiting. This stage-specific interaction is the reason recent sorghum studies rarely analyze temperature and rainfall as simple seasonal

averages; they focus instead on monthly, event-based, or growth-stage-specific conditions (Kumar et al., 2009; Prasad et al., 2021; Tolosa et al., 2023).

Climate-based yield modeling in sorghum has moved through several phases. Early work relied mainly on correlation and regression, asking which combinations of rainfall totals, rainy days, or seasonal temperatures tracked annual yield variation. That approach remains useful where data are sparse and decisions must be made quickly. More recent work has added process-based models such as APSIM, DSSAT-CERES-Sorghum, and AquaCrop, which represent phenology, biomass accumulation, water balance, and grain formation mechanistically. In parallel, remote sensing and machine learning have expanded the modeling toolbox by making it possible to estimate yield from canopy signals, spatial heterogeneity, and multi-source environmental data. The result is not a single dominant method, but an increasingly layered modeling landscape in which empirical, mechanistic, and data-driven approaches are used for different purposes. For a review aimed at a computationally oriented journal, that diversity is especially important: the field is moving toward hybrid systems that use biology to structure models and data science to improve scale, speed, and prediction accuracy (Tirfessa et al., 2023; Jabed and Murad, 2024; Gardi et al., 2025).

This study has four linked objectives. First, it summarizes the biological and physiological basis of sorghum yield formation in a form that is clear enough to support modeling decisions. Second, it examines how temperature and rainfall affect yield formation directly and through heat stress, drought stress, and their interaction. Third, it compares the main modeling approaches now used for sorghum yield prediction, not to declare a single winner, but to clarify what each method captures well and where each remains limited. Fourth, it uses published case evidence from semi-arid regions to show how climate-yield relationships work in practice and what that means for adaptation and management. The review is deliberately written as a synthesis rather than a report of new experimental data, so its contribution lies in organizing established evidence into a coherent framework that can support both research design and practical decision-making.

2 Biological and Physiological Basis of Sorghum Yield Formation

2.1 Growth and development characteristics of sorghum

Sorghum is a C4 cereal with strong adaptation to hot, high-radiation environments, and its development is usually described through discrete vegetative and reproductive stages linked to thermal time. That developmental structure matters because the crop does not respond to weather uniformly across the whole season. The timing of panicle initiation, flowering, and maturity depends on genotype, temperature, and in many materials photoperiod sensitivity; together, those factors determine whether the crop escapes or encounters stress at key points. Modeling work on diverse sorghum genotypes has shown that accurate prediction of phenology and canopy development requires explicit representation of temperature and photoperiod responses rather than broad maturity labels alone. This has two practical implications. First, cultivars that appear similar in duration can behave differently under shifting sowing dates or altered season length. Second, any serious attempt to model yield formation from climate must begin with phenology, because an error in stage timing usually propagates into errors in stress exposure, biomass partitioning, and grain yield (Tirfessa et al., 2023).

2.2 Major yield components

Sorghum grain yield is built from a small set of components, but the timing of their determination is staggered across the season. Grain number and grain weight are the dominant immediate components, while panicle size, floret fertility, seed set, and tiller contribution help explain how those two primary components are assembled. The number of kernels is largely determined during the earlier reproductive period, especially from panicle initiation through flowering and early fertilization, whereas kernel weight depends more strongly on post-flowering assimilate supply and the duration and quality of grain filling. That distinction matters for climate analysis because temperature and rainfall do not affect all components in the same way. Heat or water deficit around flowering tends to depress grain number, while post-flowering stress more often reduces individual grain weight. Genetic studies also reach the same broader conclusion from a different angle: grain size, grain number per panicle, and grain weight are central yield-related traits, but they are interconnected and subject to trade-offs.

A useful sorghum model must therefore represent not just total biomass, but how climate shapes the component pathway to yield (Baye et al., 2022; Otwani et al., 2025).

2.3 Physiological processes related to yield formation

Behind those yield components lie a set of physiological processes that climate directly modifies. Photosynthesis supplies assimilates for canopy growth, reproduction, and grain filling. Stomatal conductance and plant hydraulics regulate the trade-off between carbon uptake and water loss. Assimilate partitioning determines whether biomass supports stems, leaves, roots, or grain at a given stage. Under water stress, sorghum often maintains function better than many cereals through deep rooting, osmotic adjustment, partial stomatal control, and a capacity in some genotypes to conserve photosynthetic performance even when water becomes limiting. Recent physiological work has shown meaningful genotypic variation in intrinsic water-use efficiency and associated hydraulic traits, with improved water-use efficiency in some genotypes arising not only from stronger stomatal restriction but from better maintenance of photosynthetic capacity under stress. That is especially important for modeling because it means drought tolerance cannot be treated as a single trait or a single reduction factor. It emerges from interacting physiological controls that differ by genotype and stage (Ndlovu et al., 2021; Prasad et al., 2021; Al-Salman et al., 2024).

2.4 Environmental sensitivity across growth stages

Not all growth stages are equally vulnerable (Figure 1). The literature repeatedly shows that reproductive stages are more sensitive than vegetative stages, although early establishment can also be critical where emergence stress is severe. In sorghum, the period from panicle emergence through anthesis is especially important because it governs floret fertility, pollen development, fertilization, and embryo formation. Later, grain filling becomes the decisive stage for grain size and final grain mass. Water stress or heat stress during these windows can lower yield even when earlier biomass production looked satisfactory. The stage-specific nature of stress is one reason the same seasonal rainfall or seasonal mean temperature can produce very different outcomes in different years: what matters is where stress lands in relation to developmental timing. For modelers, that means stage-based sensitivity functions are not optional details. They are the bridge between weather time series and harvest outcomes (Prasad et al., 2015; Prasad et al., 2021; Smith et al., 2023).

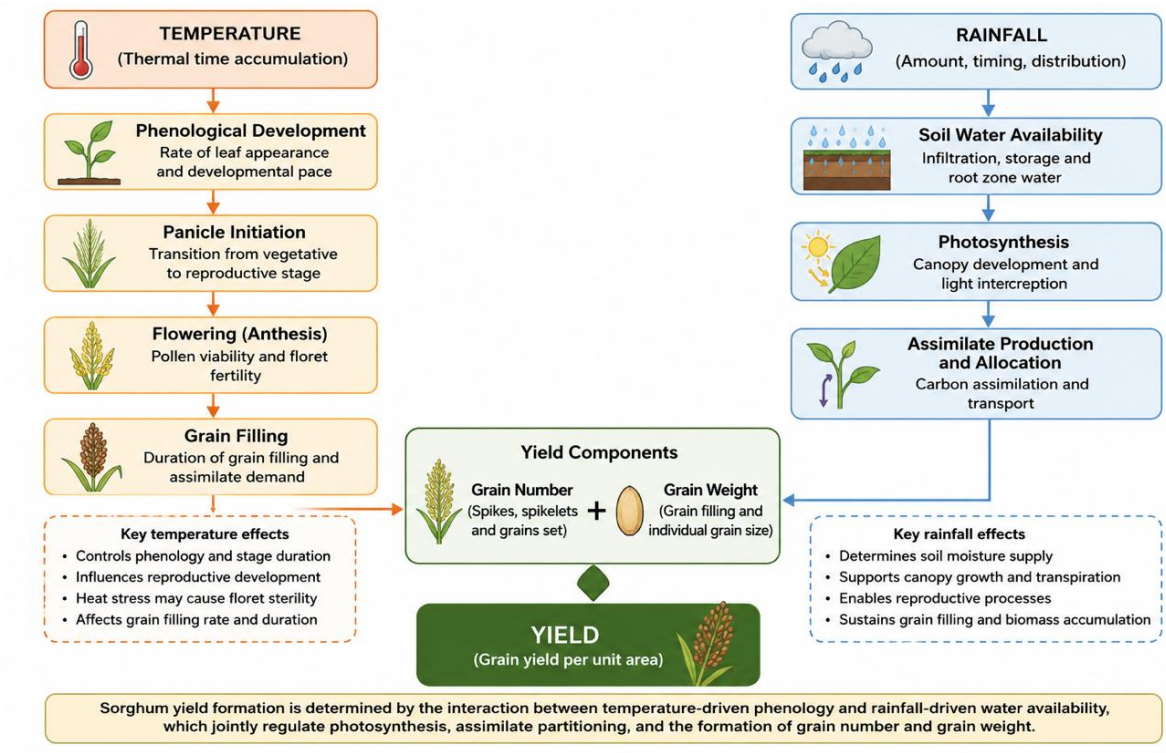


Figure 1 Biological and physiological framework of sorghum yield formation

3 Effects of Temperature and Rainfall on Sorghum Yield Formation

3.1 Temperature effects on growth and productivity

Temperature has a double role in sorghum. Within an appropriate range it accelerates development and supports rapid canopy formation, but above that range it can shorten critical phases and damage reproductive function. Growth-stage studies show that the most sensitive windows for high-temperature effects on floret fertility lie roughly from 10 to 5 days before anthesis and from 5 days before to 5 days after anthesis. In controlled and field experiments, mean daily temperatures above 25°C during panicle emergence and reproductive development reduced floret fertility sharply, with fertility reaching zero at about 37°C under one experimental setup. When heat occurs later, during grain filling, the main effect shifts from grain number to grain weight because high temperature shortens the effective filling period and limits final kernel mass. Even where sorghum is "heat adapted," these results make clear that adaptation does not equal immunity. It means the crop can perform better than alternatives under heat, not that it can ignore reproductive heat stress (Prasad et al., 2015; Smith et al., 2023).

3.2 Rainfall effects on crop development and yield

Rainfall affects sorghum yield not simply through total water supply, but through its timing, frequency, and match with soil type and plant stage. In semi-arid regions, rainfall can support germination, early canopy expansion, and reproductive development even when seasonal totals are modest, provided dry spells are short and well-positioned. Conversely, rainfall that arrives too early, too late, or in a few concentrated events can leave the crop exposed to long soil-water deficits during flowering or grain filling. Published case evidence from eastern Ethiopia illustrates this clearly: monthly rainfall amount and rainy-day number during the growing period were positively associated with sorghum yield, while temperature variables were negative. Similar conclusions also appear in dryland studies that emphasize rainfall distribution as more informative than seasonal totals. Rainfall excess can also become damaging. Waterlogging studies show that sorghum yield can decline markedly when excessive moisture occurs, particularly at early stages, because photosynthesis, enzyme activity, and panicle development are impaired. So the rainfall question is not "more or less," but "when, how often, and under what soil and stage conditions." (Tolosa et al., 2023; Zhang et al., 2023).

3.3 Heat stress, drought stress, and yield loss mechanisms

Heat stress and drought stress reduce yield through partially distinct but overlapping mechanisms (Figure 2). Heat stress during the pre-anthesis and anthesis period can disrupt tapetum development, pollen viability, pollen germination, and floret fertility, which directly lowers grain number. In one recent study, exposure to 42/32°C day/night heat at the pollen mother cell and booting stages severely disrupted male reproductive development, and 12 days of stress at the PMC stage caused almost complete loss of grain yield. Drought stress, meanwhile, acts through reduced leaf expansion, lower stomatal conductance, reduced assimilate supply, impaired reproductive success, and sometimes premature senescence. Water-deficit experiments further show that drought can alter intra-panicle grain number and depress individual grain weight, depending on timing. Under field conditions, both stresses converge on the same final logic: less successful grain set before flowering and weaker filling after flowering. The stress pathway changes, but the endpoint is the same (Adotey et al., 2021; Prasad et al., 2021; Smith et al., 2023).

3.4 Interactive effects of temperature and rainfall

The most serious yield losses often arise when high temperature and rainfall shortage occur together. Warm conditions increase vapor pressure deficit and evapotranspiration demand; if rainfall is simultaneously low or irregular, the plant faces a compounded water-energy imbalance. That interaction helps explain why a year with only moderate rainfall reduction can still perform poorly if accompanied by strong warming, especially around flowering. Reviews of combined stress in sorghum describe morphological injury, disrupted cell metabolism, lower membrane stability, reduced photosynthesis, and stronger oxidative stress under joint drought-heat exposure than under either stress alone. Modeling studies reinforce the same point. In the U.S. Great Plains, APSIM-based environment characterization identified water- and heat-stress clusters that aligned with observed yield reductions, and in the drier western sorghum belt grain-filling water stress was especially common. This is a reminder that

climate variables should not be interpreted independently in yield models when they co-determine plant demand and stress timing (Ndlovu et al., 2021; Carcedo et al., 2022).

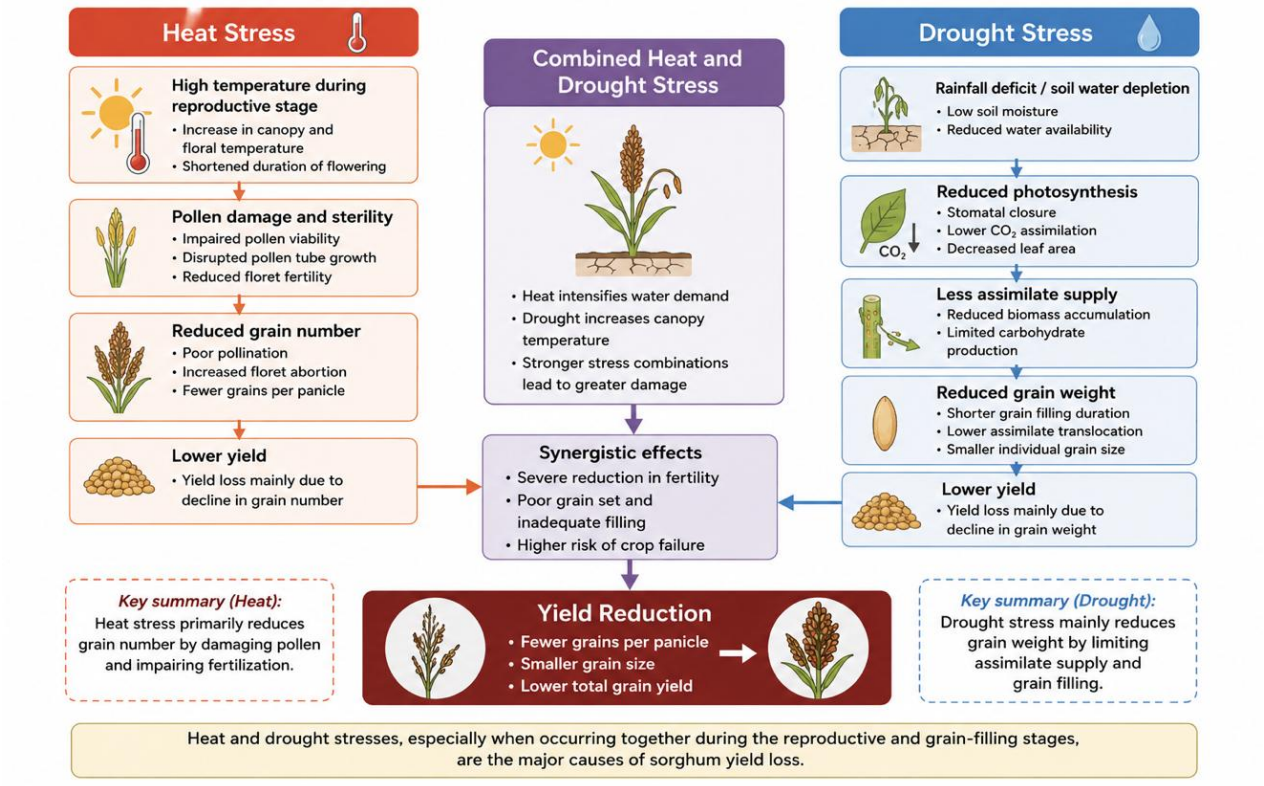


Figure 2 Mechanisms of yield loss under heat and drought stress

3.5 Climatic thresholds influencing yield stability

Thresholds in sorghum are real, but they are not universal constants. They depend on genotype, developmental stage, and stress duration. Even so, the literature gives useful working thresholds. For reproductive heat, mean daily temperatures above about 25°C during panicle emergence and early reproductive development begin to depress fertility, and sustained exposure to higher temperatures causes sharply larger losses. During grain filling, elevated temperatures mainly reduce kernel weight. For water-related thresholds, the message is less about a single rainfall total than about distribution. In Babile district, for example, August and September rainfall variability and the number of rainy days in September were among the strongest predictors of yield, whereas seasonal rainfall totals had weaker relationships. Dryland management studies likewise show that adaptation measures perform differently across rainfall bands, with some rainwater-harvesting practices helping under certain conditions but not across all semi-arid settings. In practical modeling, this means threshold thinking should focus on stage-specific heat episodes, rainfall sequence, and stress duration, not on a single whole-season cutoff (Prasad et al., 2015; Kubiku et al., 2022; Tolosa et al., 2023).

4 Modeling Approaches for Sorghum Yield Prediction

4.1 Statistical and empirical models

Statistical and empirical models remain the most straightforward entry point for sorghum yield prediction. Their usual strength is clarity: the analyst can directly test how yield covaries with rainfall totals, rainy-day frequency, monthly temperatures, or growing degree days. In data-scarce regions, that simplicity is a genuine advantage. The Babile study is a good example. Using 1995-2020 data, the authors found that a multiple regression based on monthly rainfall, rainy days, and temperature explained about 77% of the annual variation in sorghum yield, underscoring how much explanatory power can be obtained from carefully chosen climate predictors in a local rainfed system. At the same time, empirical models are only as stable as the relationships they learn from the past. They often struggle when management changes, cultivars change, or climate enters combinations outside the

historical range. They also explain association better than mechanism. So they are useful for local forecasting and first-pass diagnosis, but they rarely suffice for scenario analysis on their own (Tolosa et al., 2023).

4.2 Process-based crop simulation models

Process-based crop models are the backbone of much sorghum climate-impact research because they translate weather into plant development through explicit biological rules. APSIM and DSSAT-CERES-Sorghum are the most widely used in the literature reviewed here, with AquaCrop often used for irrigation and water-productivity questions. Their main attraction is interpretability: they can represent thermal-time driven phenology, soil-water balance, biomass production, and yield formation in a way that makes adaptation experiments possible. That is why they are especially useful for testing cultivar maturity, sowing date, fertilizer response, supplemental irrigation, and trait ideotypes under future climates. APSIM-based studies in Ethiopia and Mali, for example, have been used to characterize drought patterns, explore genotype × environment × management interactions, and identify where trait changes or sowing shifts might reduce risk. DSSAT-based studies in Ethiopia have simulated both future yield decline and the performance of adaptation packages under SSP scenarios. The cost of this power is data demand and calibration effort. A crop model can formalize biology beautifully and still perform poorly if soils, varieties, or management are mis-specified (Tirfessa et al., 2023; Diancoumba et al., 2024; Gardi et al., 2025; Ali and Kothari, 2026).

A general process-based expression of sorghum yield prediction can be written as:
$Y=f(T,R,S,G,M,\varepsilon)$

where Y is yield, T is the temperature regime, R is rainfall and soil water supply, S is soil condition, G is genotype, M is management, and ε captures unobserved variation. In empirical models, f is usually a fitted statistical relation. In crop simulation models, f is decomposed into linked sub-processes such as phenology, transpiration, biomass accumulation, and partitioning.

4.3 Remote sensing applications

Remote sensing expands sorghum yield modeling by observing the crop directly across space rather than relying only on weather and field samples. The main value of remote sensing lies in its ability to capture canopy status, vegetation indices, spatial heterogeneity, and sometimes stage-specific crop responses that are difficult to measure manually at large scale. Recent sorghum studies show that multispectral imagery from satellites or UAVs can support reasonably strong yield prediction when aligned with key phenological stages. In tropical environments, artificial neural network models built from vegetation indices and soil elevation data reached strong performance in estimating sorghum grain yield, while arid-region UAV studies found that integrating multispectral and meteorological data can predict yield with high accuracy and reveal which growth stage contributes most to predictive skill. A recurring theme is that timing matters: the best observation date is not necessarily the latest one, because some stages carry more information about final yield formation than others. Remote sensing therefore works best when it is phenology-aware, not just image-rich (Ferraz et al., 2024; Deng et al., 2025).

4.4 Machine learning and artificial intelligence approaches

Machine learning has become increasingly attractive in sorghum yield prediction because sorghum systems are shaped by non-linear interactions among climate, soils, management, and canopy signals. Algorithms such as random forests, gradient boosting, support vector machines, artificial neural networks, and stacking ensembles can absorb high-dimensional predictor sets and model interactions that are difficult to specify mechanically. Recent sorghum applications illustrate both the promise and the limits of this approach. In South Sudan, machine-learning models combining yield, climate, remote sensing, and conflict-probability data produced useful end-of-season yield predictions, with XGBoost, decision tree, and random forest performing especially well. In tropical and arid experiments, neural networks and ensemble approaches also produced strong fits. But these models can become opaque, and their success depends greatly on training data coverage and quality. When extrapolation is required, or when the user needs biological explanation rather than prediction alone, machine learning is strongest when paired with domain knowledge rather than treated as a black box (Ferraz et al., 2024; Jabed and Murad, 2024; Deng et al., 2025; Karongo et al., 2025).

4.5 Comparison of existing modeling approaches

The various modeling traditions are best understood as complementary tools rather than competing ideologies. Empirical models are attractive when transparency and low input demand matter most. Process-based models are preferable when the user needs biological realism, trait testing, or future scenario analysis. Remote sensing improves spatial monitoring and in-season updating. Machine learning is especially useful when relationships are complex and observation streams are large. The real challenge is not to choose one method forever, but to align method with question. If the purpose is local seasonal diagnosis, regression may be enough. If the purpose is trait-by-environment adaptation research, APSIM or DSSAT is more suitable. If the purpose is spatial forecasting or operational monitoring, remote sensing and machine learning become more important. Increasingly, the strongest studies combine them (Jones et al., 2003; Holzworth et al., 2014; Mihret et al., 2024).

5 Case Study: Modeling Sorghum Yield Responses to Temperature and Rainfall Variability in Semi-Arid Regions

5.1 Background and climatic characteristics

A useful published example comes from semi-arid Ethiopia, where sorghum is central to rainfed livelihoods and climate variability is already strongly visible in farm outcomes. In Babile district, eastern Ethiopia, the agro-climatic setting is semi-arid, the growing period is relatively short, rainfall is bimodal and highly erratic, and long-term average annual rainfall is about 731 mm. The area's rainfall pattern includes Belg rainfall from March to May and Kiremt rainfall from June to September, but what matters agronomically is not that there are two rainy windows; it is that their reliability is low and their intra-seasonal distribution is unstable. Published work from Kobo, Miesso, and Melkassa extends the same logic across semi-arid Ethiopia: these areas differ in rainfall response, baseline temperatures, and future vulnerability, but they share a dependence on rainfed sorghum under high interannual variability. This makes them ideal for examining how temperature and rainfall signals are translated into yield through empirical and simulation methods (Tolosa et al., 2023; Gardi et al., 2025; Ali and Kothari, 2026).

5.2 Effects of temperature variability on yield

In Babile, observed sorghum yield showed a negative relationship with both maximum and minimum temperatures during the crop-growing period, and specific monthly temperature variability explained part of the year-to-year yield fluctuation. The logic is biologically plausible: warmer conditions can accelerate development, intensify evapotranspiration, and raise the probability that reproductive processes occur under suboptimal moisture. The more recent Ethiopia modeling study reinforces this concern in forward-looking terms. Using DSSAT-CERES-Sorghum, Gardi (2025) et al. projected warming of up to about 4°C by the 2080s in semi-arid Ethiopian sites and identified Kobo as especially vulnerable where higher temperatures and reduced rainfall coincide. Taken together, the observational and simulation evidence suggests that temperature is not merely a background variable in semi-arid sorghum systems. It is an active driver of phenological compression and water-stress escalation (Tolosa et al., 2023; Gardi et al., 2025).

5.3 Effects of rainfall variability on yield

Rainfall variability in the same case-study region appears even more nuanced. In Babile, monthly rainfall and number of rainy days during the growing season were positively correlated with sorghum yield, and rainfall in August and September was more informative than crude seasonal totals. This suggests that late-season water availability supports reproductive success and grain development in that environment. Yet rainfall does not work as a simple "more is better" factor. The Zimbabwe meta-analysis on rainwater harvesting found that some adaptation practices produced neutral or even negative yield responses under specific rainfall bands and soil conditions, showing that poor alignment between technology and rainfall environment can undermine expected benefits. The lesson from these dryland case studies is that rainfall variability must be modeled at a finer temporal scale than annual or even seasonal totals. Temporal sequencing matters (Kubiku et al., 2022; Tolosa et al., 2023).

5.4 Application of yield prediction models

The case-study literature from semi-arid Ethiopia shows how different model families answer different questions. Multiple regression in Babile captured historical climate-yield relationships with useful explanatory power and highlighted specific months and rainy-day patterns. DSSAT-based regional modeling then extended the analysis into future climate scenarios, allowing more explicit testing of varietal differences and long-term adaptation options. Related APSIM work in Ethiopian drylands has gone further by examining genotype × environment × management interactions and sowing-risk trade-offs, while APSIM-based environment characterization in Mali identified drought-pattern frequencies rather than only mean conditions. This layered use of models is perhaps the most interesting lesson of the case study. Researchers did not move from a "simple bad model" to a "complex good model." They used simpler models to identify local signal and more mechanistic models to ask why the signal occurs and how it may change (Tirfessa et al., 2023; Tolosa et al., 2023; Diancoumba et al., 2024; Gardi et al., 2025).

5.5 Implications for climate adaptation and crop management

For adaptation, the published case evidence points to a clear but unspectacular conclusion: stability comes from better matching crop duration, sowing time, and water availability. In practical terms, this means cultivar choice matters, planting-date adjustment matters, and in some settings supplemental irrigation or more targeted moisture conservation may matter. It also means region-wide recommendations are risky. Even within semi-arid Ethiopia, Miesso is projected to receive larger rainfall increases than Kobo, while Kobo remains more vulnerable to heat and rainfall decline. In other words, the case study argues against generic "dryland sorghum packages" and in favor of locally parameterized decision support. Climate adaptation for sorghum is more likely to succeed when it is built from climate windows, varietal maturity, and site-specific soil-water logic than when it relies on broad labels such as drought tolerant or early maturing alone (Gardi et al., 2025; Ali and Kothari, 2026).

6 Climate Change and Future Sorghum Production

6.1 Projected changes in temperature and rainfall

The climate projections discussed across recent sorghum studies are broadly consistent even when the size of change differs by region. Temperatures are expected to continue rising across major sorghum environments, while rainfall is projected to become more variable in amount, distribution, or both. In semi-arid Ethiopia, simulation studies project warming on the order of about 2.1°C by the 2050s and around 4°C by the 2080s in some locations, with rainfall changes that vary by site rather than moving uniformly upward or downward. Similar work in India suggests that future sorghum responses may depend on whether rainfall increases offset thermal penalties, which again emphasizes that precipitation change cannot be interpreted without temperature and season type. The future climate problem for sorghum is therefore not a single trend line. It is a moving combination of faster development, stronger atmospheric water demand, altered rainy-season reliability, and more frequent extreme events (Chadalavada et al., 2022; Tolosa et al., 2023; Gardi et al., 2025).

6.2 Potential impacts on sorghum yield formation

Future sorghum yield formation is likely to be affected most where warming shifts sensitive reproductive stages into hotter and drier windows. That can reduce grain set before flowering and shorten or weaken grain filling afterward. In North Wollo, Ethiopia, future simulations suggested that rainfed grain sorghum yield would likely decline by roughly 15%-16% in mid-century and 17%-22% in late century relative to the recent baseline. Other studies, however, show that yield declines are not inevitable everywhere. In post-rainy sorghum environments in India, rising rainfall and CO2 in some scenarios were sufficient to offset part of the temperature burden and even generate simulated yield gains. The important point is not that one study is optimistic and another pessimistic. It is that future yield formation depends on how multiple climate drivers alter stage-specific stress exposure, not on warming alone (Chadalavada et al., 2022; Ali and Kothari, 2026).

6.3 Regional differences in climate vulnerability

Regional vulnerability is a recurring theme in the literature. Semi-arid production systems with high rainfall variability, shallow water storage, and low management buffering are typically more exposed than

better-resourced systems, but even within dryland sorghum regions vulnerability differs sharply. In the Ethiopian studies reviewed here, Kobo emerged as more vulnerable than Miesso in part because future warming and rainfall deficits align there more strongly with yield loss. In the Great Plains of the United States, modeled stress environments differ between the northeast and southwest sorghum belt, with grain-filling water stress dominating much of the southwest. These differences matter because they also change which traits or practices are worthwhile. A trait that helps under grain-filling drought may add little benefit in a low-stress or pre-flowering-drought environment. Vulnerability, then, is not just regional climate severity. It is the frequency of the specific stress pattern that matches the crop's sensitive stages (Carcedo et al., 2022; Gardi et al., 2025).

6.4 Future yield trends under climate scenarios

The best-supported summary of future yield trends is cautious heterogeneity. Many semi-arid sorghum systems show projected yield declines under warming, especially in rainfed and already heat-prone sites, but some environments show stable or improving simulated yields when rainfall, CO2 fertilization, or adaptation measures offset part of the thermal stress. West African APSIM simulations earlier suggested that medium-maturing material could outperform early-maturing material under some climate scenarios, while more recent Ethiopia studies show strong site and genotype contingency. This variation should not be read as contradiction. It is evidence that future yield trends are conditional on environment, cultivar duration, and management. The practical inference is simple: climate scenarios should be used to identify likely stress profiles and adaptation niches, not to produce one global verdict on sorghum (Gardi et al., 2025; Ali and Kothari, 2026).

7 Strategies for Improving Yield Stability and Prediction Accuracy

7.1 Development of climate-resilient cultivars

The breeding literature makes clear that climate-resilient sorghum will not come from one "super trait." Useful improvement will likely require trait combinations that align with the dominant stress pattern of target environments. These include reproductive heat tolerance, stable grain filling under water limitation, appropriate maturity duration, root-system traits, and physiological behaviors such as stay-green or limited transpiration in certain environments. Recent reviews on sorghum improvement emphasize that breeding progress will depend not only on identifying tolerant germplasm, but on linking physiological insight, quantitative genetics, and realistic target environments. Modeling can help here by testing the expected value of candidate traits before they are expensive to phenotype or introgress widely. In that sense, crop models do not replace breeding; they improve trait prioritization (Mwamahonje et al., 2024; Raymundo et al., 2024; Fontanet-Manzaneque et al., 2025).

7.2 Agronomic management practices

Agronomy remains the fastest route to stabilizing sorghum yield under climate variability because it can reposition the crop relative to stress without waiting for long breeding cycles. Sowing date is consistently important, especially in environments where early or delayed planting can help flowering avoid peak heat or terminal drought. Soil-water conservation, improved fertility, and context-appropriate irrigation also matter, but their value depends heavily on local rainfall regime and soil properties. The literature on rainwater-harvesting and deficit-irrigation modeling shows that management gains are not automatic: a practice that works in one rainfall band or soil class may fail in another. The most defensible management strategy is therefore adaptive rather than prescriptive. It should be based on dominant local stress patterns, cultivar duration, and short-term seasonal expectations (Kubiku et al., 2022; Fazel et al., 2023; Ali and Kothari, 2026).

7.3 Integration of climate information and decision support systems

A recurring gap in sorghum production is not the absence of climate data, but the weak translation of climate knowledge into field decisions. Decision-support systems can close that gap by combining weather history, seasonal outlooks, crop-model outputs, and local management rules. For sorghum, that might include sowing-window advisories, cultivar-duration matching, drought-risk maps, or irrigation scheduling. The strength of such systems is greatest when they integrate climate information with biologically meaningful crop thresholds rather than simply reporting rainfall probabilities. The broader crop-modeling literature also shows that operational systems become more valuable when they are iterative: they begin with pre-season planning, then

absorb in-season observations from weather and remote sensing to update expected yield and risk. For dryland sorghum systems facing increasing climate variability, this kind of staged decision support may matter as much as any single new trait. (Jabed and Murad, 2024; Mihret et al., 2024).

7.4 Improving Yield Modeling Through Emerging Technologies

The next step in sorghum yield modeling is not simply "more AI." It is better integration across scales. Emerging work already points in that direction: UAV and satellite remote sensing add frequent canopy observations, machine learning improves pattern detection, and process-based models provide biological structure. Phenotyping, explainable AI, and genotype-aware modeling can make this integration more useful rather than merely more complex. Particularly promising are hybrid frameworks in which a crop model provides stage structure and water-balance logic, while data-driven methods update parameters or correct prediction error using contemporary observations. This may be the most realistic way to improve sorghum yield prediction under rapidly changing climates, because it preserves interpretability while benefiting from rich data streams. The most valuable future systems will likely be those that can explain why a yield prediction changed, not only produce a more accurate number (Figure 3) (Jabed and Murad, 2024; Deng et al., 2025; Karongo et al., 2025).

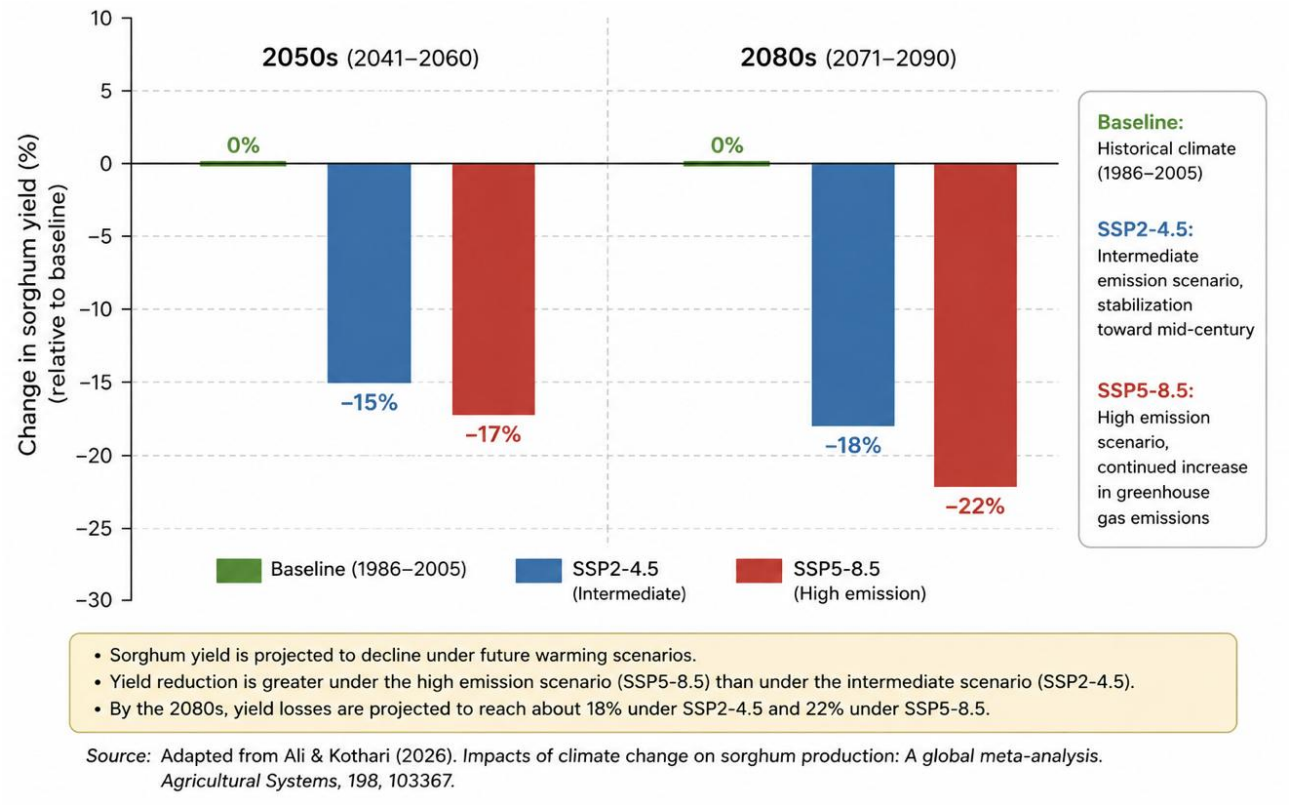


Figure 3 Integrated framework for next-generation sorghum yield prediction

8 Conclusions and Future Perspectives

This study argues that sorghum yield formation cannot be understood, or modeled well, by treating temperature and rainfall as broad background variables. Their impact depends on developmental timing, stress duration, and interaction. Temperature drives phenology and can damage reproduction directly, while rainfall determines whether the crop can sustain the physiological processes that support grain set and grain filling. Reproductive-stage heat, post-flowering drought, and poorly distributed rainfall are repeatedly identified as the most consequential threats to stable yield. At the modeling level, empirical, process-based, remote-sensing, and machine-learning approaches all contribute something important, but their real value is highest when they are combined rather than isolated.

Several gaps remain. First, many studies still rely on seasonal climate summaries that are too coarse to represent the biological reality of stage-specific stress. Second, genotype differences are often acknowledged but insufficiently parameterized in operational models. Third, interactions among heat, drought, soil constraints, and excess rainfall remain under-modeled in many sorghum systems. Fourth, strong local case studies exist, but transferability across regions is still limited. Finally, predictive accuracy is improving faster than interpretability in some data-driven studies, which risks producing models that are useful technically but harder to trust agronomically.

Future work should move toward integrated sorghum modeling systems that connect phenology, plant physiology, remote sensing, and climate analytics in the same framework. More attention is needed on stress timing around flowering and grain filling, on genotype-specific calibration of water-use and heat-response traits, and on decision tools that translate model output into locally actionable advice. For both researchers and practitioners, the most productive perspective may be to treat sorghum neither as a miracle crop nor as a victim crop, but as a biologically understandable crop whose yield can be better stabilized when climate signals are interpreted through the lens of development, physiology, and carefully chosen models.

Acknowledgments

I am deeply grateful to Professor R. Cai for his multiple reviews of this paper and for his constructive revision suggestions.

References

Adotey R.E., Patrignani A., Bergkamp B., Kluitenberg G., Prasad P.V.V., and Jagadish S.V.K., 2021, Water-deficit stress alters intra-panicle grain number in sorghum, Crop Science, 61(4): 2680-2695.
https://doi.org/10.1002/csc2.20532

Ali K.H., and Kothari K., 2026, Assessing future climate change impacts and adaptation strategies for sorghum yield in North Wollo, Ethiopia, Theoretical and Applied Climatology, 157(1): 36.
https://doi.org/10.1007/s00704-025-05986-y

Al-Salman Y., Cano F.J., Mace E., Jordan D., Groszmann M., and Ghannoum O., 2024, High water use efficiency due to maintenance of photosynthetic capacity in sorghum under water stress, Journal of Experimental Botany, 75(21): 6778-6795.
https://doi.org/10.1093/jxb/erae418

Baye W., Xie Q., and Xie P., 2022, Genetic architecture of grain yield-related traits in sorghum and maize, International Journal of Molecular Sciences, 23(5): 2405.
https://doi.org/10.3390/ijms23052405

Carcedo A.J.P., Mayor L., Demarco P., Morris G.P., Lingenfelser J., Messina C.D., and Ciampitti I.A., 2022, Environment characterization in sorghum (Sorghum bicolor L.) by modeling water-deficit and heat patterns in the Great Plains region, United States, Frontiers in Plant Science, 13: 768610.
https://doi.org/10.3389/fpls.2022.768610

Chadalavada K., Gummadi S., Kundeti R.K., Kadiyala D.M., Deevi K.C., Dakhore K.K., Diana R.K.B., and Thiruppathi S.K., 2022, Simulating potential impacts of future climate change on post-rainy season sorghum in India using CERES-Sorghum model, Sustainability, 14(1): 334.
https://doi.org/10.3390/su14010334

Deng L.Q., Li Y.Y., Liu X.F., Zhang Z.M., Mu J.J., Jia S.J., Yan Y.Q., and Zhang W.P., 2025, Sorghum yield prediction using UAV multispectral imaging and stacking ensemble learning in arid regions, Frontiers in Plant Science, 16: 1636015.
https://doi.org/10.3389/fpls.2025.1636015

Diancoumba M., Kholová J., Adam M., Famanta M., Clerget B., Traore P.C.S., Weltzien E., Vacksmann M., McLean G., Hammer G.L., van Oosterom E.J., and Vadez V., 2024, APSIM-based modeling approach to understand sorghum production environments in Mali, Agronomy for Sustainable Development, 44(3): 25.
https://doi.org/10.1007/s13593-023-00909-5

Fazel F., Ansari H., and Aguilar J., 2023, Determination of the most efficient forage sorghum irrigation scheduling strategies in the U.S. Central High Plains using the AquaCrop model and field experiments, Agronomy, 13(10): 2446.
https://doi.org/10.3390/agronomy13102446

Ferraz M.A.J., Barboza T.O.C., Piza M.R., Von Pinho R.G., and dos Santos A.F., 2024, Sorghum grain yield estimation based on multispectral images and neural network in tropical environments, Smart Agricultural Technology, 9: 100661.
https://doi.org/10.1016/j.atech.2024.100661

Fontanet-Manzaneque J.B., Hernández D.M., Giordano A., and Caño-Delgado A.I., 2025, Sorghum as a monocot model for drought research, Frontiers in Plant Science, 16: 1665967.
https://doi.org/10.3389/fpls.2025.1665967

Gardi M.W., Zewdu E., and Sida T.S., 2025, Modeling sorghum yield response to climate change in the semi-arid environment of Ethiopia, Journal of Agriculture and Food Research, 22: 102143.
https://doi.org/10.1016/j.jafr.2025.102143

Holzworth D.P., Huth N.I., deVoil P.G., Zurcher E.J., Herrmann N.I., McLean G., Chenu K., van Oosterom E.J., Snow V., Murphy C., Moore A.D., Brown H., Whish J.P.M., Verrall S., Fainges J., Bell L.W., Peake A.S., Poulton P.L., Hochman Z., Thorburn P.J., Gaydon D.S., Dalgliesh N.P., Rodriguez D., Cox H., Chapman S., Doherty A., Teixeira E., Sharp J., Cichota R., Vogeler I., Li F.Y., Wang E., Hammer G.L., Robertson M.J., Dimes J.P., Whitbread A.M., Hunt J., van Rees H., McClelland T., Carberry P.S., Hargreaves J.N.G., MacLeod N., McDonald C., Harsdorf J., Wedgwood S., and Keating B.A., 2014, APSIM-Evolution towards a new generation of agricultural systems simulation, Environmental Modelling & Software, 62: 327-350.
https://doi.org/10.1016/j.envsoft.2014.07.009

Hossain M.S., Islam M.N., Rahman M.M., Mostofa M.G., and Rahman Khan M.A., 2022, Sorghum: A prospective crop for climatic vulnerability, food and nutritional security, Journal of Agriculture and Food Research, 8: 100300.
https://doi.org/10.1016/j.jafr.2022.100300

Jabed M.A., and Murad M.A.A., 2024, Crop yield prediction in agriculture: A comprehensive review of machine learning and deep learning approaches, with insights for future research and sustainability, Heliyon, 10(24): e40836.
https://doi.org/10.1016/j.heliyon.2024.e40836

Jones J.W., Hoogenboom G., Porter C.H., Boote K.J., Batchelor W.D., Hunt L.A., Wilkens P.W., Singh U., Gijsman A.J., and Ritchie J.T., 2003, The DSSAT cropping system model, European Journal of Agronomy, 18(3-4): 235-265.
https://doi.org/10.1016/S1161-0301(02)00107-7

Karongo J., Mwaniki J.I., Ndiritu J., and Mokaya V., 2025, Sorghum yield prediction based on remote sensing and machine learning in conflict affected South Sudan, Scientific Reports, 15(1): 4469.
https://doi.org/10.1038/s41598-025-89030-z

Kubiku F.N.M., Mandumbu R., Nyamangara J., and Nyamadzawo G., 2022, Sorghum (Sorghum bicolor L.) yield response to rainwater harvesting practices in the semi-arid farming environments of Zimbabwe: A meta-analysis, Heliyon, 8(3): e09164.
https://doi.org/10.1016/j.heliyon.2022.e09164

Kumar S.R., Hammer G.L., Broad I.J., Harland P., and McLean G., 2009, Modelling environmental effects on phenology and canopy development of diverse sorghum genotypes, Field Crops Research, 111(1-2): 157-165.
https://doi.org/10.1016/j.fcr.2008.11.010

Liaqat W., Altaf M.T., Barutçular C., Mohamed H.I., Ahmad H., Jan M.F., and Khan E.H., 2024, Sorghum: A star crop to combat abiotic stresses, food insecurity, and hunger under a changing climate: A review, Journal of Soil Science and Plant Nutrition, 24(1): 74-101.
https://doi.org/10.1007/s42729-023-01607-7

Mihret Y.C., Ketsela G.M., and Mintesinot S.M., 2024, Implementation and application of APSIM for crop modelling in Ethiopia: A comprehensive review, Heliyon, 10(10): e31612.
https://doi.org/10.1016/j.heliyon.2024.e31612

Mwamahonje A., Mdindikasi Z., Mchau D., Mwenda E., Sanga D., Garcia-Oliveira A.L., and Ojiewo C.O., 2024, Advances in sorghum improvement for climate resilience in the global arid and semi-arid tropics: A review, Agronomy, 14(12): 3025.
https://doi.org/10.3390/agronomy14123025

Ndlovu E., van Staden J., and Maphosa M., 2021, Morpho-physiological effects of moisture, heat and combined stresses on Sorghum bicolor [Moench (L.)] and its acclimation mechanisms, Plant Stress, 2: 100018.
https://doi.org/10.1016/j.stress.2021.100018

Otwani D., McLean G., Hammer G., Cruickshank A., Hunt C., Tao Y., Koltunow A., Mace E., and Jordan D., 2025, Extended grain filling has potential to improve yield in grain sorghum, Journal of Experimental Botany, 76(10): 2763-2774.
https://doi.org/10.1093/jxb/eraf117

Prasad P.V.V., Djanaguiraman M., Perumal R., and Ciampitti I.A., 2015, Impact of high temperature stress on floret fertility and individual grain weight of grain sorghum: Sensitive stages and thresholds for temperature and duration, Frontiers in Plant Science, 6: 820.
https://doi.org/10.3389/fpls.2015.00820

Prasad V.B.R., Govindaraj M., Djanaguiraman M., Djalovic I., Shailani A., Rawat N., Singla-Pareek S.L., Pareek A., and Prasad P.V.V., 2021, Drought and high temperature stress in sorghum: Physiological, genetic, and molecular insights and breeding approaches, International Journal of Molecular Sciences, 22(18): 9826.
https://doi.org/10.3390/ijms22189826

Raymundo R., McLean G., Sexton-Bowser S., Lipka A.E., and Morris G.P., 2024, Crop modeling suggests limited transpiration would increase yield of sorghum across drought-prone regions of the United States, Frontiers in Plant Science, 14: 1283339.
https://doi.org/10.3389/fpls.2023.1283339

Smith A., Gentile B.R., Xin Z., and Zhao D., 2023, The effects of heat stress on male reproduction and tillering in Sorghum bicolor, Food and Energy Security, 12(6): e510.
https://doi.org/10.1002/fes3.510

Tirfessa A., Getachew F., McLean G., van Oosterom E., Jordan D., and Hammer G., 2023, Modeling adaptation of sorghum in Ethiopia with APSIM-opportunities with G×E×M, Agronomy for Sustainable Development, 43(1): 15.
https://doi.org/10.1007/s13593-023-00869-w

Tolosa A.A., Dadi D.K., Mirkena L.W., Erena Z.B., and Liban F.M., 2023, Impacts of climate variability and change on sorghum crop yield in the Babile district of eastern Ethiopia, Climate, 11(5): 99.
https://doi.org/10.3390/cli11050099

Zhang R.D., Yue Z.X., Chen X.F., Huang R.D., Zhou Y.F., and Cao X., 2023, Effects of waterlogging at different growth stages on the photosynthetic characteristics and grain yield of sorghum (Sorghum bicolor L.), Scientific Reports, 13(1): 7212.
https://doi.org/10.1038/s41598-023-32478-8

Disclaimer/Publisher's Note

The statements, opinions, and data contained in all publications are solely those of the individual authors and contributors and do not represent the views of the publishing house and/or its editors. The publisher and/or its editors disclaim all responsibility for any harm or damage to persons or property that may result from the application of ideas, methods, instructions, or products discussed in the content. Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

BioSci Publisher®

Reasons to publish in BioSci Publisher

An Online Publishing Platform

★Peer review quickly and professionally
☆Publish online immediately upon acceptance
★Deposit permanently and track easily
☆Access free and open around the world
★Disseminate multilingual available

Submit your manuscript at: **http://bioscipublisher.com/**