

## Feature Review

## Open Access

# Integrating Genomic Selection and Machine Learning for Predicting Maize Yield Under Drought

Weichang Wu ✉

Biotechnology Research Center, Cuixi Academy of Biotechnology, Zhuji, 311800, China

✉ Corresponding author: [weichang.wu@cuixi.org](mailto:weichang.wu@cuixi.org)Maize Genomics and Genetics, 2025, Vol.16, No.5 doi: [10.5376/mgg.2025.16.0021](https://doi.org/10.5376/mgg.2025.16.0021)

Received: 05 Jul., 2025

Accepted: 22 Aug., 2025

Published: 07 Sep., 2025

**Copyright** © 2025 Wu, 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:**

Wu W.C., 2025, Integrating genomic selection and machine learning for predicting maize yield under drought, Maize Genomics and Genetics, 16(5): 239-250 (doi: [10.5376/mgg.2025.16.0021](https://doi.org/10.5376/mgg.2025.16.0021))

**Abstract** Drought stress severely constrains maize yields, posing a significant challenge to global food security. This study explores the integration of genomic selection (GS) and machine learning (ML) methods to improve the accuracy of maize yield prediction under drought conditions. First, we outline the principles of GS, highlighting its advantages over traditional breeding methods and its growing application in drought-tolerant breeding. Next, we explore the application of various ML algorithms (such as random forests, support vector machines, and deep learning) for crop yield prediction, along with their strengths and limitations in the context of genomics. We then propose strategies for integrating GS with ML, including hybrid modeling frameworks and context-specific optimization, and discuss recent trends and research advances. Particular emphasis is placed on drought-specific modeling approaches that incorporate stress-responsive traits and evaluate their predictive accuracy under water-deficit environments. A case study from sub-Saharan Africa illustrates the practical application of an integrated GS-ML prediction system and its implications for climate-resilient maize breeding. Despite this promising outlook, challenges remain, including data heterogeneity, model interpretability, and implementation barriers. This study summarizes the future prospects of advancing the integration of genomic selection and machine learning (GS-ML) through technological innovation and its potential to support global climate-smart maize breeding.

**Keywords** Genomic selection; Machine learning; Drought tolerance; Maize yield prediction; Climate-smart breeding

## 1 Introduction

No one would object to the fact that corn plays a significant role in the global food supply issue. But the problems are not small either, especially when there is a drought. During some critical growth stages, such as pumping or filling, an untimely drought can cause a direct drop in yield. Moreover, in recent years, the weather has become increasingly unreasonable (Zhang and Xu, 2024). Droughts not only occur more frequently but also cause much greater damage. The traditional drought-resistant breeding methods have long been struggling. It's not that the technology is lacking, but rather that the trait of drought resistance itself is troublesome, involving too many genes and having significant environmental disturbances (Zhang et al., 2022). The result is that the same set of breeding strategies may perform completely differently in various places. Low efficiency and slow pace will eventually make it impossible to keep up with the express train of climate change. What should I do? Take another route. To ensure that corn yields are not controlled by weather or mood, a faster and more accurate screening mechanism is needed to pick out in advance those genotypes that can truly "withstand the pressure".

At this point, genomic selection (GS) becomes a powerful tool. It does not focus on a few key genes but takes the entire genome together, using molecular markers to predict the potential of each variety. For the complex trait of drought resistance, the applicability of GS is quite high. After all, it can handle all kinds of genetic effects, big and small, without distinguishing between primary and secondary (Yuan et al., 2019). Of course, it's not a divine skill either. When there is too much data and too many variables, especially when genotypes interact with the environment, the model is prone to "overload" (Wang et al., 2025). At this point, relying solely on GS is indeed a bit inadequate. So, it's time for machine learning (ML) to come into play. Methods like random forests, neural networks, and support vector machines are particularly adept at handling complex data and nonlinear relationships. Once they are combined with GS, the accuracy of prediction reaches a higher level (Saimon et al., 2023; Azrai et

al., 2024). Ultimately, the combination of GS and ML is not for showing off skills, but to truly find a way to deal with yield prediction in complex environments-especially under the increasingly normalized stress condition of drought (Wu et al., 2024; He et al., 2025).

This study is not intended to propose a new method, but rather to systematically review the existing achievements: This includes the challenges encountered in predicting corn yields under drought conditions, the application logic and development progress of GS and ML in breeding, several real cases verifying the effectiveness of their combination, as well as future directions worth paying attention to, such as how to build data infrastructure and how policies can support to truly promote the implementation of these technologies in agricultural practice.

## **2 Genomic Selection in Maize Breeding**

### **2.1 Principles of genomic selection (GS)**

On the surface, GS is a modern breeding tool that "predicts traits using whole-genome molecular markers". It sounds grand and sophisticated, but in essence, it's about using all the genotype information you have at hand, regardless of whether it has a significant or minor impact, and then putting it all into the model for training-the aim is to estimate the potential of each corn material in a certain trait, such as drought resistance. This is quite different from the previous practice of selecting seeds by relying on a few principal QTLs. In the past, methods focused on key points, but GS adheres to the principle of "leaving no one unchecked". Not only potent sites but also those with minor effects are included. In this way, we can obtain the so-called genomic estimated breeding value (GEBV), and then we do not have to rely entirely on phenotypes when making subsequent seed selection decisions. Of course, there are many types of models, such as ridge regression, Bayes A/B, and random forest methods, all of which are used to train prediction accuracy (Shikha et al., 2017; Nepolean et al., 2018). However, which model to choose actually depends on the characteristics of the data. Sometimes, there is no "universal answer".

### **2.2 Advantages over conventional breeding**

To be honest, when it comes to drought-resistant breeding, traditional methods are indeed not very easy. For decades, breeding experts have been struggling with the old problems of "low heritability, significant environmental interference and slow progress", especially when it comes to complex traits, they find it even more difficult to take steps forward. But the emergence of GS has torn a new knot in this deadlock. It doesn't wait for the corn to grow out to observe its performance, but rather determines in advance-at the seed stage or even earlier-whether this plant is worth continuing to cultivate through genetic information (Chen, 2024). The benefits of this "early judgment" are obvious: it saves time, reduces experimentation, and also avoids wasting resources on materials with little potential. For those who are in a hurry to shorten the breeding cycle, GS is more like a speed-up tool. However, GS is not just about being fast. It can also be "multi-functional"-not only considering drought resistance, but also taking into account other traits simultaneously, such as yield, plant type, and even quality. This kind of multi-objective improvement is almost impossible to achieve in traditional methods, or rather, it is extremely inefficient. Some studies simply rely on figures: under drought conditions, using the GS method can increase the yield of corn by approximately 7.3% (Vivek et al., 2017; Das et al., 2021). This extent, when placed in actual production, is already a considerable gain-not a theoretical improvement, but a real increase in money.

### **2.3 Application to drought tolerance**

When it comes to application, the performance of GS in drought-resistant breeding of corn is obvious to all. Especially in those regions with intense climate change and frequent droughts, GS has significantly improved the efficiency of seed selection. The efficiency of trying one piece at a time as in the past is far from sufficient. Nowadays, researchers have been able to predict in advance the performance of certain corn strains under drought by modeling whole-genome SNP data. Not only that, these models can also identify key genetic factors related to the drought resistance response mechanism, such as root development, stomatal regulation, and even hormone signals (Figure 1) (Liu and Qin, 2021; Sheoran et al., 2022). More detailed approaches also include the introduction of multi-environmental test data, taking into account both additive and dominant effects to adapt to

the differences in drought among different production areas. Some people have combined GS with high-throughput phenotypes and initiated a rapid breeding cycle. The results also proved that in an environment prone to drought, this set of combined measures can indeed bring visible genetic gains (Dias et al., 2018), laying the foundation for the promotion of climate-adapted corn varieties.

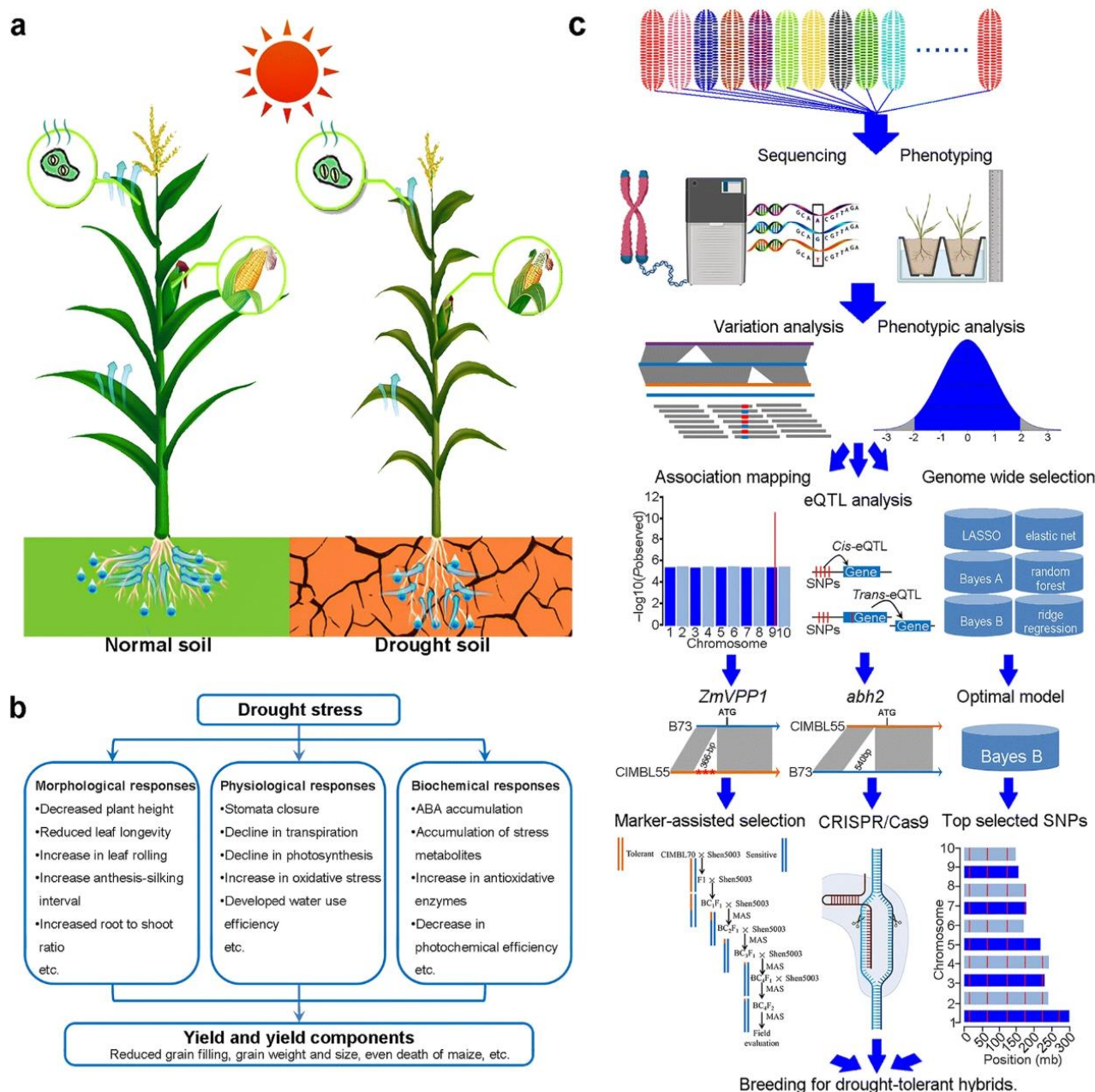


Figure 1 Effect of drought stress on maize growth and development and the research strategy for the trait improvement. a. An illustration describing the morphological changes that occur in plants in response to drought stress. b. The physiological and cellular responses that occur in maize in response water-deficit conditions and lead to reductions in growth and yield. c Schematic of the research strategy employed in genetic dissection of maize drought resistance for trait enhancement (Adopted from Liu and Qin, 2021)

### 3 Machine Learning in Maize Crop Yield Prediction

#### 3.1 Types of ML algorithms used

The number of machine learning methods currently used to predict corn yields is not an exaggeration. From the earliest linear regression model to the current situation where neural networks and ensemble algorithms are all at work. Traditional methods like Lasso and ridge regression, although "old", still have their place in some simple

problems. In recent years, however, people's focus has clearly shifted towards tree-based ensemble methods such as Random Forest (RF) and XGBoost, as well as classic nonlinear models like Support Vector Machine (SVM). Deep learning has not been idle either. Architectures such as CNN and LSTM have gradually emerged in breeding research, especially in those involving time series or image processing (Kang et al., 2020; Cheng et al., 2022). However, whether an algorithm is new and cool or not has nothing to do with whether it is "easy to use" or not. Some studies have instead found that XGBoost and RF are more reliable in terms of prediction accuracy and stability than certain deep learning models (Tahi et al., 2024). So, when it comes to different data and problem scenarios, no algorithm can be a one-size-fits-all solution. It's still necessary to "select the type as needed".

### **3.2 Advantages of ML in breeding contexts**

In fact, it's not that no one uses traditional methods; it's just that in some situations they really can't handle them. For example, what factors affect the output? Genotype, climate, management methods... These factors are intertwined, and the relationship is complex and non-linear, which conventional statistical methods simply cannot sort out. Machine learning, however, has no such concern. It doesn't care whether your underlying logic is clear or not. As long as there is sufficient data, it can learn something from it. There is another practical problem: the data is too diverse. Genomic information, remote sensing images, weather records, soil parameters... These data have different sources and structures, but ML can still swallow them all up to run the model. This ability is precisely one of its greatest advantages over traditional regression (Guo et al., 2023; Miao et al., 2024). For breeders and farmers, it never hurts to know more about the situation earlier. ML models can provide predictions in the early stage of crops and also use feature analysis to identify which factor has the greatest impact on yield (Wu et al., 2024). This not only helps in selecting materials, but also facilitates subsequent planting management and resource allocation.

### **3.3 Challenges in ML application to genomics**

Although it sounds like machine learning has a promising future in the application of genomics, there are also many problems in actual operation. First of all, you need to have a large amount of high-quality data and clear labels; otherwise, how can the model learn? Besides, genomic data itself has too high a dimension, and it is easy to overfit if one is not careful. The interpretability of deep models is another troublesome point. The trained accuracy rate may be very high, but if you ask it why it makes such a prediction, it is hard for it to explain clearly (Shahhosseini et al., 2019; Abbasi et al., 2025). Especially when you are dealing with a multi-environment and multi-group data scenario, a model that performs well in one place does not necessarily mean it can be smoothly transferred to another for use. Furthermore, the formats of genomic, phenotypic, and environmental data are inherently inconsistent. To integrate them into a single model is not only a matter of computing power but also technically challenging (Van Klompenburg et al., 2020). If these obstacles are not overcome, ML will also find it difficult to fully realize its true potential in corn breeding.

## **4 Integrating Genomic Selection and Machine Learning**

### **4.1 Rationale for integration**

In fact, putting GS and ML together is not out of the pursuit of some kind of "theoretical integration", but rather a result of pragmatism. After all, in the field of breeding, there are too many complex variables. Relying solely on one set of methods often leads to neglecting one aspect for another. GS has a set of methods for processing genomic data, which can provide genomic breeding values for each material. ML is more flexible and adept at handling nonlinear and multi-dimensional data structures, such as environmental factors and phenotypic information. It can also capture these. However, if we talk about the greatest use of the combination of the two, it is still to improve the accuracy of prediction in a variable environment-especially for traits like drought resistance that are influenced by both genotype and environment (Varshney, 2021). In the current situation where climate change is becoming increasingly uncontrollable, how to stabilize the output of crops such as corn has become even more urgent.



## 4.2 Integration strategies

Nowadays, integrating GS and ML is not uncommon, and the practices are also increasing. Some research studies choose to process the data uniformly first, and then feed the genomic, phenotypic and environmental data together using models such as random forests or deep neural networks. This approach is also often referred to as "multimodal fusion", with the aim of enhancing the overall prediction performance. Some teams also adopt the approach of transfer learning, first training their models on datasets related to output and then fine-tuning them to the target data-the TrG2P framework does it this way. The advantage of this strategy is that it can utilize data that are not direct production indicators, helping the model learn useful information more quickly. In addition, some people are more sensitive to the "explanatory power of models" and tend to use deep learning models with attention mechanisms. This way, they can see which variables have the greatest impact (Togninalli et al., 2023). Another approach is simply to start with variable screening. Through feature selection techniques, SNPS or environmental factors with large amounts of information can be picked out in advance, which not only reduces the dimension but also lowers the risk of overfitting (Bayer et al., 2021; Sirsat et al., 2022). Of course, which features are useful varies from data to data and there is no absolute answer.

## 4.3 Current research trends

At present, many studies are attempting to use GS in combination with ML to enhance the accuracy of yield prediction for different crops in various environments. Compared with traditional methods, multi-omics and multi-modal machine learning models perform more stably overall, especially when considering high-throughput phenotypic and environmental data simultaneously, their advantages are more obvious. Methods like transfer learning and deep frameworks are receiving increasing attention because they can utilize data with complex structures and related traits. It has recently been reported that the accuracy rate of predicting corn yield through these methods has increased by approximately 6.8% (Li et al., 2024), which is not a small figure. But don't overlook some hidden issues either: for instance, although the model is accurate, can it still work well in other environments? How can different types of data be unified? And whether breeders feel "handy" when using it-all these require corresponding tools and platforms to be addressed. Not to mention the issues of infrastructure such as policies, data platforms and hardware. If these links are not kept up, if this set of combined measures is to be applied to large-scale breeding projects, many detours may still be needed.

# 5 Drought-Specific Modeling Approaches

## 5.1 Incorporating drought response traits

Not all yield prediction models take drought resistance traits into account, but this has become increasingly common now. Especially those physiological or agronomic indicators that can directly reflect the plant's response to water stress, such as SPAD (relative chlorophyll content), LAI (leaf area index), flowering and silk production time, as well as stress resistance index (STI, DTI), etc., are often regarded as important variables and incorporated into statistical or machine learning models. The combination of the values SPAD and LAI works quite well. Some studies have observed that during the VT stage of corn, the correlation between them and yield is the most obvious (Szeles et al., 2023). But this does not mean that other stages are unimportant; it's just that this correlation may change. In breeding practice, some studies have gone further by introducing the multitrait index and the calculation method of "distance between genotype and ideal type" to pick out more drought-resistant materials (Kumar et al., 2022). In addition, incorporating the mark-trait association information related to drought resistance into genomic prediction models has indeed improved the accuracy of yield prediction under stress conditions, especially for those complex polygenic control traits.

## 5.2 Environment-specific modeling

Drought is not the same every year or everywhere, which poses a challenge to modeling. Environment-specific modeling aims to address this issue. It hopes to make the prediction framework more detailed by taking into account the temporal and spatial variations in drought occurrence. Remote sensing or process indicators such as solar-induced chlorophyll fluorescence (SIF), soil moisture simulation values, and cumulative drought index (CDI) have all been widely used as tools to describe drought conditions in recent years. If these variables are added to the model, the prediction accuracy can be significantly improved-especially in years with particularly severe

drought (Vergopalan et al., 2020; Wang et al., 2023). Furthermore, some hybrid methods are also being used, such as combining biophysical crop models with ML, training the models with data across years and regions, and the resulting predictions often hold water in different drought scenarios (Attia et al., 2022; Wang et al., 2022; Li et al., 2023).

### 5.3 Evaluation of prediction performance under drought

It is not enough to merely judge whether a model runs accurately in a certain year or a certain place. A truly reliable model needs to be able to stand the test of many years, multiple regions, and various drought intensities. This is also why many current studies are conducting cross-environment model evaluations. Some models use genetic algorithms or feature selection to optimize their structure, and the prediction results are indeed quite impressive. For instance, the predicted  $R^2$  for production can reach 0.92, and the resilience index (such as STI) also has a level of 0.82, which is already quite good. The introduction of drought-specific factors such as SIF, CDI, and soil moisture has also helped the model maintain high stability in extreme years (Shuai and Basso, 2022; Luo et al., 2024). However, to be fair, not all models can handle extreme situations perfectly. Some models still have the problem of "underestimating production loss due to extreme drought", which is quite likely to be exposed in practice (Amiri et al., 2022; Bueechi et al., 2023). This indicates that we still need to continue to strengthen the mechanism construction of the model in terms of drought response.

## 6 Case Study: Integrated Prediction System for Drought-Tolerant Maize in Sub-Saharan Africa

### 6.1 Background and objectives

Sub-saharan Africa (SSA) is not short of sunlight or arable land, but the food problem has never been solved. The reasons are very complex. Among them, drought is the most direct and common limiting factor affecting corn yield. Especially in some areas where water resources are already tight, poor harvests for several consecutive seasons are the norm. Therefore, drought-resistant corn varieties, along with a reliable yield prediction system, become particularly crucial. The SSA region is promoting an integrated prediction framework, with a clear goal: to integrate genomic, remote sensing, meteorological and environmental data and use machine learning to predict corn yields under drought conditions (Ndlovu et al., 2024). This not only serves breeding projects but also aims to provide some more practical reference information for farmers and policymakers.

### 6.2 Model development and integration

Not all models can be truly implemented. Many solutions run very fast in the laboratory, but once they are put into practical application, they fail to adapt to the local environment. But what the SSA team did this time was a bit different. They did not start from the top-level design but from the local demands, prioritizing the issue of "locality" first. One of the core data sources of the model is Earth Observation (EO) data. Such as rainfall, water availability, extreme temperatures, number of drought days... All these indicators have been integrated into the system, and the time scale has been refined to the sub-month level. Don't underestimate this detail. The significant improvement in prediction performance is largely supported by it. As for the genetic aspect, they used the RR-BLUP model in the multi-environment corn experiment and identified many key quantitative trait nucleotides (QTN) and candidate genes related to drought resistance. Furthermore, they also incorporated the results of GWAS for verification, which helps improve prediction accuracy and makes the model output more biologically explanatory (Amadu et al., 2025). It is worth mentioning that they did not stop at making predictions. Instead, they used big data platforms and spatial modeling to precisely "deliver" these drought-resistant materials to high-risk drought areas. In this way, the materials cultivated will not be "selected well but not put to use".

### 6.3 Outcomes and implications

Ultimately, whether a system is worth promoting depends on whether it can truly be "put into use". At least in terms of on-site performance, this system of the SSA team has passed the test. Just in terms of yield prediction, the model's accuracy during the corn growing season is quite good, with a Nash-Sutcliffe efficiency value exceeding 0.6 and an average relative error controlled within 20% (Lee et al., 2022). This level of accuracy is not only sufficient for research but can also be used for early warning and making decisions on grain dispatching. In terms

of breeding, they have also screened out multiple potential drought-tolerant haplotypes and candidate genes through the system. This is of direct significance for the subsequent selection of more stable yield strains. But what is truly impressive is the geospatial analysis they conducted. In areas with a high incidence of drought, after planting these drought-resistant materials, the yield can be 5% to 40% higher than that of commercial varieties. This range sounds like a large fluctuation, but in reality, even a 10% gap is very realistic for farmers' income (Tesfaye et al., 2016). Of course, the issues of scalability and infrastructure have not been completely resolved yet—such problems are not uncommon in Africa. But taking a step back, this case at least shows that combining genomic data, remote sensing information and machine learning technology is a promising path, especially for regions like SSA that are greatly affected by climate change, which may indeed bring about some changes.

## **7 Challenges and Limitations**

### **7.1 Data-related constraints**

Not to mention how the model is built, many problems actually get stuck at the "data" step. Methods like genomic selection and machine learning rely particularly on a large amount of high-quality data when it comes to predicting corn yields under drought conditions. But what about the reality? Although high-throughput genotyping and phenotypic analysis generate a large amount of data, they are often incomplete, noisy, or have an unbalanced distribution of variables. Especially the data related to drought, the degree of standardization is generally not high (Tong and Nikoloski, 2020). Often, data from different environments are difficult to concatenate, with a lack of vertical information and incomplete annotations. This makes model training very challenging, not to mention generalization ability. Moreover, when there is a large amount of omics data, it is prone to dimensional explosion. If feature screening or dimensionality reduction is not done well, overfitting is very likely to occur. The model may seem very accurate on the surface, but in reality, it is unstable.

### **7.2 Model interpretability and biological relevance**

Some models do make good predictions, but if you really ask them "Why do you predict like this?", they can't explain it clearly either. This kind of "black box" problem is particularly common in complex models such as deep learning and ensemble methods (Maløy et al., 2021). Moreover, no matter how accurate the model is, breeders are more concerned about: which genes and which environmental variables are truly effective? If this issue is not clarified, actual breeding decisions will be hesitant. Although some methods, such as attention mechanisms and feature importance analysis, are attempting to make models "speak human language", they are far from enough to turn these results into truly actionable and verifiable biological knowledge (Shook et al., 2020). So a core question remains unsolved: Can the model capture the interactions between those real and biologically significant genotypes and the environment? If not, even if the prediction is extremely accurate, it will be very difficult to be truly put into use.

### **7.3 Practical implementation barriers**

No matter how beautiful a model is built, if it is not used in the end, it is just for show. The integration technology of GS and ML does have potential, but in practical operation, problems keep emerging one after another. First of all, it has relatively high requirements for resources. No matter how powerful a model is, there must be someone who can run it. To have sufficient computing power, there must be someone to manage the data and a team that can understand the model and adjust the parameters. These conditions are not available in many regions, especially in areas with weak breeding foundations, where they are simply "unusable". Another problem that is not easy to solve is poor mobility. A model that works well in location A may not adapt to the local environment in location B. When the variety changes or the environment changes, the effect is compromised and re-training and re-validation are required (McBreen et al., 2025). This means that you want to rely on a universal model to cover all breeding scenarios? Ideals may be full and rich, but reality may not buy them. And there is another point that is often overlooked: people. Whether the policies can support it and how well the digital infrastructure is built are one aspect. More importantly, have breeders and data scientists been properly trained? If people can't keep up, this entire system will eventually be hard to be truly implemented. Ultimately, the integration approach of GS-ML is not without prospects; it's just that the road ahead has not yet been paved. No matter how advanced the technology is, if no one uses it or uses it poorly, it can only remain at the stage of being "beautiful on paper".

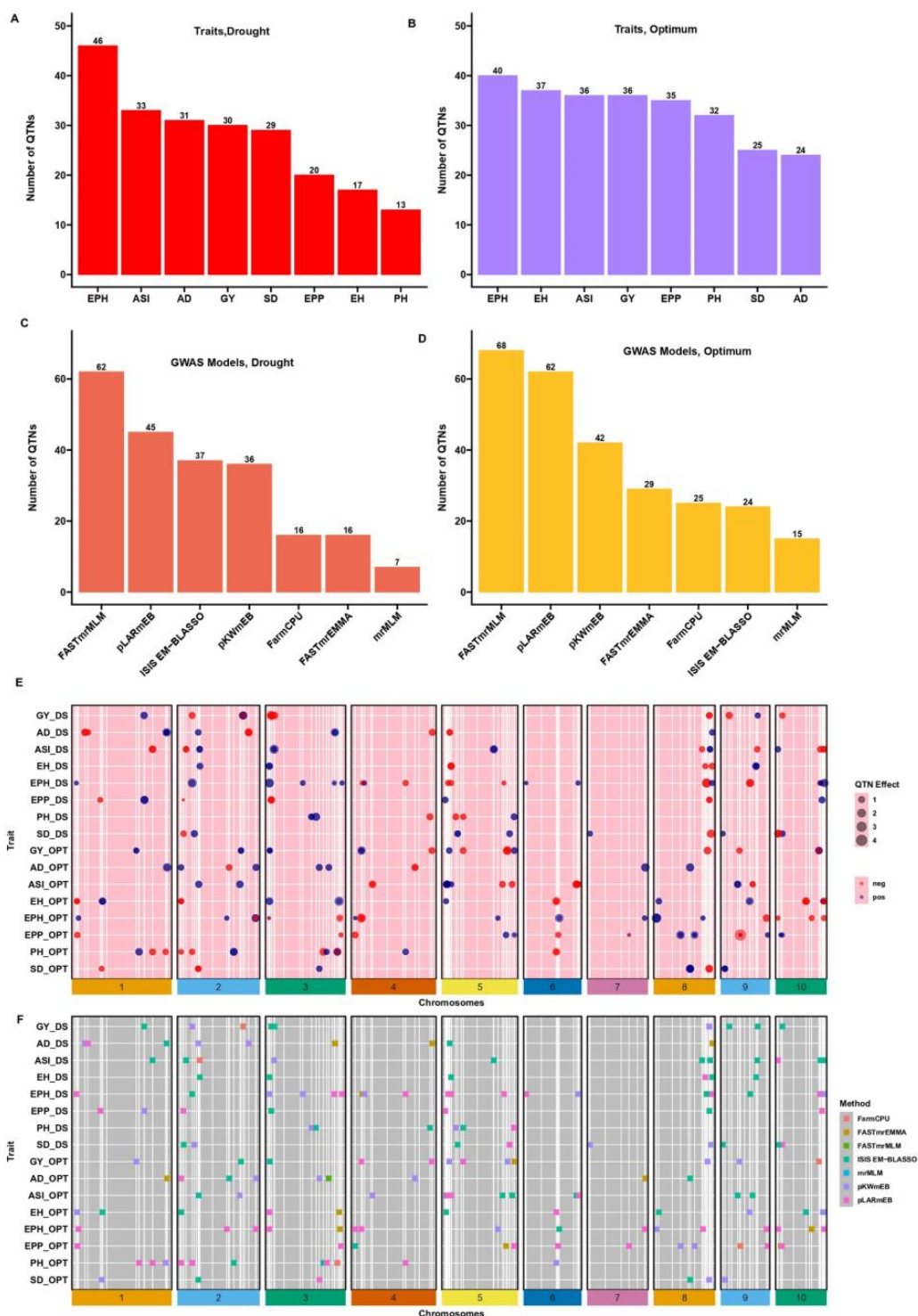


Figure 2 Number of significantly associated QTNs detected by each of the seven GWAS models implemented in this study. A Number of QTNs/SNP detected for grain yield (GY); days to 50% anthesis (AD); days to 50% silking (SD); Anthesis-Silking interval (ASI); plant height (PH), eight height (EH), ear-plant height ratio (EPH) and ear per plant (EPP) under drought condition (B). Number of QTNs/SNP detected for grain yield (GY); days to 50% anthesis (AD); days to 50% silking (SD); Anthesis-Silking interval (ASI); plant height (PH), eight height (EH), ear-plant height ratio (EPH) and ear per plant (EPP) under optimum condition C. Number of QTNs detected by the eight GWAS models under drought condition (D) Number of QTNs detected by the eight GWAS models under Optimum condition (E) Chromosomal Distribution of QTN effects. The circle diameter is proportional to the absolute value of the QTN effect. The colors indicate the direction of the effects: red indicates negative QTN effect, and blue indicates positive QTN effect. F Chromosomal distribution of QTNs based on seven GWAS methods. The x-axis indicates genomic locations by chromosomal order, and the significant QTNs are plotted against genome location. Each row represents one QTN identified by a different method (Adopted from Amadu et al., 2025)



## 8 Future Perspectives and Concluding Remarks

In recent years, the pace of the breeding field has indeed changed, so fast that it has caught people off guard. In the past, it would take at least seven or eight years or even longer for a variety to go from breeding to promotion. But what about now? With high-throughput genotyping, multi-omics data and an automated phenotypic platform, the efficiency of many links has been improved by more than a little. But these alone are not enough. It is not uncommon for genetic effects to be handled. The key point is that models are now daring to deal with things that were previously difficult to handle, such as non-additive genetic effects, the interaction between genotypes and the environment, and the indescribable interrelationships between traits. Machine learning has indeed come in handy in this regard, especially as model structures have become increasingly complex. However, no matter how powerful the technology is, it is impossible to master it all by oneself. Things like deep learning, pan-genomics, and AI algorithms, which sound very trendy, can only remain at the level of academic papers in the end without platform support and tool matching. Fortunately, in recent years, many integrated platforms and open-source tools have emerged, at least giving us a glimmer of hope-these models may really have the chance to move from the laboratory to the fields.

How can breeding respond to climate change? No one can give a universal answer. But at least one thing is clear: the future breeding process needs to be more "climate-smart". That is to say, not only should high yields be pursued, but also adaptability and stress resistance should be taken into consideration. To achieve this, relying solely on traditional methods is far from enough. The integration of GS-ML is precisely part of this new path. It can help breeders identify materials that can truly "withstand" extreme weather more quickly. However, to make it run, it requires far more than one model-cross-regional data integration, environmental variable access, and multi-trait collaborative modeling-all of which are indispensable. And there are also quite a few practical problems. From data sharing and standard setting among institutions, to the construction of digital infrastructure, and then to the connection and collaboration between breeders and data scientists, every step requires people to do it, money to invest, and a willingness to cooperate. Without solving these problems, the value of technology will also be difficult to be fully released.

But then again, don't be dazzled by the "coolness" of the technology itself. The integration of GS and ML is not merely about achieving a few more percentage points of accuracy, but rather offers the opportunity to redefine the breeding process itself, especially in the face of extreme conditions like drought. Of course, there are still quite a few problems. The data is still lacking, the interpretability of the model has not kept up, and the threshold for implementation is also high. These are all realities blocking the way and cannot be resolved in a year or two. But if we can truly make some breakthroughs in these difficult areas, especially by facilitating cross-disciplinary cooperation and integrating what each party excels at, those changes that once seemed distant might come in the blink of an eye. The future is hard to predict, but it's not completely unprepared. Against the backdrop of increasingly unstable climate, these technologies might just become one of the key supports for the food system.

## Acknowledgments

I would like to express my heartfelt gratitude to Ms. Xuan for reviewing the draft of this paper and providing suggestions for improvement, which made the structure and content of the paper clearer and more rigorous. I would also like to thank the two anonymous peer reviewers for their professional reviews and insightful comments, which have helped me further enhance the quality and academic rigor of this paper.

## Conflict of Interest Disclosure

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

## References

- Abbasi M., Váz P., Silva J., and Martins P., 2025, Machine learning approaches for predicting maize biomass yield: leveraging feature engineering and comprehensive data integration, *Sustainability*, 17(1): 256.  
<https://doi.org/10.3390/su17010256>

- Amadu M., Beyene Y., Chaikam V., Tongoona P., Danquah E., Ifie B., Burgueño J., Prasanna B., and Gowda M., 2025, Genome-wide association mapping and genomic prediction analyses reveal the genetic architecture of grain yield and agronomic traits under drought and optimum conditions in maize, *BMC Plant Biology*, 25: 135.  
<https://doi.org/10.1186/s12870-025-06135-3>
- Amiri E., Irmak S., and Araj H., 2022, Assessment of CERES-Maize model in simulating maize growth, yield and soil water content under rainfed, limited and full irrigation, *Agricultural Water Management*, 259: 107271.  
<https://doi.org/10.1016/j.agwat.2021.107271>
- Attia A., Govind A., Qureshi A., Feike T., Rizk M., Shabana M., and Kheir A., 2022, Coupling process-based models and machine learning algorithms for predicting yield and evapotranspiration of maize in arid environments, *Water*, 14(22): 3647.  
<https://doi.org/10.3390/w14223647>
- Azrai M., Aqil M., Andayani N., Efendi R., Suarni, Suwardi, Jihad M., Zainuddin B., Salim, Bahtiar, Muliadi A., Yasin M., Hannan M., Rahman, and Syam A., 2024, Optimizing ensembles machine learning, genetic algorithms, and multivariate modeling for enhanced prediction of maize yield and stress tolerance index, *Frontiers in Sustainable Food Systems*, 8: 1334421.  
<https://doi.org/10.3389/fsufs.2024.1334421>
- Bayer P., Peterit J., Danilevicius M., Anderson R., Batley J., and Edwards D., 2021, The application of pangenomics and machine learning in genomic selection in plants, *The Plant Genome*, 14(3): e20112.  
<https://doi.org/10.1002/tpg2.20112>
- Bueechi E., Fischer M., Crocetti L., Trnka M., Grlj A., Zappa L., and Dorigo W., 2023, Crop yield anomaly forecasting in the Pannonian basin using gradient boosting and its performance in years of severe drought, *Agricultural and Forest Meteorology*, 340: 109596.  
<https://doi.org/10.1016/j.agrformet.2023.109596>
- Chen I., 2024, Genome-wide association studies of disease resistance genes in maize, *Genomics and Applied Biology*, 15(1): 12-21.  
<https://doi.org/10.5376/gab.2024.15.0003>
- Cheng M., Peñuelas J., McCabe M., Atzberger C., Jiao X., Wu W., and Jin X., 2022, Combining multi-indicators with machine-learning algorithms for maize yield early prediction at the county-level in China, *Agricultural and Forest Meteorology*, 323: 109057.  
<https://doi.org/10.1016/j.agrformet.2022.109057>
- Das R., Vinayan M., Seetharam K., Patel M., Kumar R., Singh S., Shahi J., Sarma A., Barua N., Babu R., and Zaidi P., 2021, Genetic gains with genomic versus phenotypic selection for drought and waterlogging tolerance in tropical maize (*Zea mays* L.), *The Crop Journal*, 9(6): 1438-1448.  
<https://doi.org/10.1016/J.CJ.2021.03.012>
- Dias K., Gezan S., Guimarães C., Nazarian A., Da Costa E Silva L., Parentoni S., De Oliveira Guimarães P., De Oliveira Anoní C., Pádua J., De Oliveira Pinto M., Noda R., Ribeiro C., De Magalhães J., Garcia A., De Souza J., Guimarães L., and Pastina M., 2018, Improving accuracies of genomic predictions for drought tolerance in maize by joint modeling of additive and dominance effects in multi-environment trials, *Heredity*, 121: 24-37.  
<https://doi.org/10.1038/s41437-018-0053-6>
- Guo Y., Xiao Y., Hao F., Zhang X., Chen J., De Beurs K., He Y., and Fu Y., 2023, Comparison of different machine learning algorithms for predicting maize grain yield using UAV-based hyperspectral images, *International Journal of Applied Earth Observation and Geoinformation*, 124: 103528.  
<https://doi.org/10.1016/j.jag.2023.103528>
- Hársányi E., Bashir B., Arshad S., Owca A., Vad A., Alsalmán A., Bácskai I., Rátonyi T., Hijazi O., Széles A., and Mohammed S., 2023, Data mining and machine learning algorithms for optimizing maize yield forecasting in Central Europe, *Agronomy*, 13(5): 1297.  
<https://doi.org/10.3390/agronomy13051297>
- He K., Yu T., Gao S., Chen S., Li L., Zhang X., Huang C., Xu Y., Wang J., Prasanna B., Hearne S., Li X., and Li H., 2025, Leveraging automated machine learning for environmental data-driven genetic analysis and genomic prediction in maize hybrids, *Advanced Science*, 12(17): 2412423.  
<https://doi.org/10.1002/advs.202412423>
- Kang Y., Ozdogan M., Zhu X., Ye Z., Hain C., and Anderson M., 2020, Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest, *Environmental Research Letters*, 15: 064005.  
<https://doi.org/10.1088/1748-9326/ab7df9>
- Kumar A., Singh V., Saran B., Al-Ansari N., Singh V., Adhikari S., Joshi A., Singh N., and Vishwakarma D., 2022, Development of novel hybrid models for prediction of drought- and stress-tolerance indices in teosinte introgressed maize lines using artificial intelligence techniques, *Sustainability*, 14(4): 2287.  
<https://doi.org/10.3390/su14042287>
- Lee D., Davenport F., Shukla S., Husak G., Funk C., Harrison L., McNally A., Rowland J., Budde M., and Verdin J., 2022, Maize yield forecasts for Sub-Saharan Africa using Earth Observation data and machine learning, *Global Food Security*, 33: 100643.  
<https://doi.org/10.1016/j.gfs.2022.100643>
- Li J., Li G., Wang L., Li D., Li H., Gao C., Zhuang M., Zhuang J., Zhou H., Xu S., Hu Z., and Wang E., 2023, Predicting maize yield in Northeast China by a hybrid approach combining biophysical modelling and machine learning, *Field Crops Research*, 302: 109102.  
<https://doi.org/10.1016/j.fcr.2023.109102>
- Li J., Zhang D., Yang F., Zhang Q., Pan S., Zhao X., Zhang Q., Han Y., Yang J., Wang K., and Zhao C., 2024, TrG2P: a transfer-learning-based tool integrating multi-trait data for accurate prediction of crop yield, *Plant Communications*, 5(7): 100975.  
<https://doi.org/10.1016/j.xplc.2024.100975>
- Liu S., and Qin F., 2021, Genetic dissection of maize drought tolerance for trait improvement, *Molecular Breeding*, 41: 8.  
<https://doi.org/10.1007/s11032-020-01194-w>

- Luo Y., Wang H., Cao J., Li J., Tian Q., Leng G., and Niyogi D., 2024, Evaluation of machine learning-dynamical hybrid method incorporating remote sensing data for in-season maize yield prediction under drought, *Precision Agriculture*, 25: 1982-2006.  
<https://doi.org/10.1007/s11119-024-10149-6>
- Måløy H., Windju S., Bergersen S., Alsheikh M., and Downing K., 2021, Multimodal performers for genomic selection and crop yield prediction, *Smart Agricultural Technology*, 1: 100017.  
<https://doi.org/10.1016/j.atech.2021.100017>
- McBreen J., Babar M., Jarquin D., Ampatzidis Y., Khan N., Kunwar S., Acharya J., Adewale S., and Brown-Guedira G., 2025, Enhancing genomic-based forward prediction accuracy in wheat by integrating UAV-derived hyperspectral and environmental data with machine learning under heat-stressed environments, *The Plant Genome*, 18(1): e20554.  
<https://doi.org/10.1002/tpg2.20554>
- Miao L., Zou Y., Cui X., Kattel G., Shang Y., and Zhu J., 2024, Predicting China's maize yield using multi-source datasets and machine learning algorithms, *Remote Sensing*, 16(13): 2417.  
<https://doi.org/10.3390/rs16132417>
- Ndlovu N., Gowda M., Beyene Y., Chaikam V., Nzuve F., Makumbi D., McKeown P., Spillane C., and Prasanna B., 2024, Genomic loci associated with grain yield under well-watered and water-stressed conditions in multiple bi-parental maize populations, *Frontiers in Sustainable Food Systems*, 8: 1391989.  
<https://doi.org/10.3389/fsufs.2024.1391989>
- Nepolean T., Kaul J., Mukri G., and Mittal S., 2018, Genomics-enabled next-generation breeding approaches for developing system-specific drought tolerant hybrids in maize, *Frontiers in Plant Science*, 9: 361.  
<https://doi.org/10.3389/fpls.2018.00361>
- Saimon M., Moniruzzaman M., Islam M., Ahmed M., Rahaman M., Hossain S., and Manik T., 2023, Integrating genomic selection and machine learning: a data-driven approach to enhance corn yield resilience under climate change, *Journal of Environmental and Agricultural Studies*, 4(2): 20-27.  
<https://doi.org/10.32996/jeas.2023.4.2.6>
- Shahhosseini M., Martinez-Feria R., Hu G., and Archontoulis S., 2019, Maize yield and nitrate loss prediction with machine learning algorithms, *Environmental Research Letters*, 14: 124026.  
<https://doi.org/10.1088/1748-9326/ab5268>
- Sheoran S., Kaur Y., Kumar S., Shukla S., Rakshit S., and Kumar R., 2022, Recent advances for drought stress tolerance in maize (*Zea mays* L.): present status and future prospects, *Frontiers in Plant Science*, 13: 872566.  
<https://doi.org/10.3389/fpls.2022.872566>
- Shikha M., Kanika A., Rao A., Mallikarjuna M., Gupta H., and Nepolean T., 2017, Genomic selection for drought tolerance using genome-wide SNPs in maize, *Frontiers in Plant Science*, 8: 550.  
<https://doi.org/10.3389/fpls.2017.00550>
- Shook J., Gangopadhyay T., Wu L., Ganapathysubramanian B., Sarkar S., and Singh A., 2020, Crop yield prediction integrating genotype and weather variables using deep learning, *PLoS ONE*, 16(6): e0252402.  
<https://doi.org/10.1371/journal.pone.0252402>
- Shuai G., and Basso B., 2022, Subfield maize yield prediction improves when in-season crop water deficit is included in remote sensing imagery-based models, *Remote Sensing of Environment*, 272: 112938.  
<https://doi.org/10.1016/j.rse.2022.112938>
- Sirsat M., Oblessuc P., and Ramiro R., 2022, Genomic prediction of wheat grain yield using machine learning, *Agriculture*, 12(9): 1406.  
<https://doi.org/10.3390/agriculture12091406>
- Széles A., Horváth É., Simon K., Zagyi P., and Huzsvai L., 2023, Maize production under drought stress: nutrient supply, yield prediction, *Plants*, 12(18): 3301.  
<https://doi.org/10.3390/plants12183301>
- Tahi S., Hounmenou C., Houndji V., and Kakaï R., 2024, An experimental analysis of traditional machine learning algorithms for maize yield prediction, *Contemporary Mathematics*, 5(4): 6208-6224.  
<https://doi.org/10.37256/cm.5420244481>
- Tesfaye K., Sonder K., Cairns J., Magorokosho C., Tarekegn A., Kassie G., Getaneh F., Abdoulaye T., Abate T., and Erenstein O., 2016, Targeting drought-tolerant maize varieties in Southern Africa: a geospatial crop modeling approach using big data, *The International Food and Agribusiness Management Review*, 19: 1-18.
- Togninalli M., Wang X., Kucera T., Shrestha S., Juliana P., Mondal S., Pinto F., Govindan V., Crespo-Herrera L., Huerta-Espino J., Singh R., Borgwardt K., and Poland J., 2023, Multi-modal deep learning improves grain yield prediction in wheat breeding by fusing genomics and phenomics, *Bioinformatics*, 39(6): btad336.  
<https://doi.org/10.1093/bioinformatics/btad336>
- Tong H., and Nikoloski Z., 2020, Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data, *Journal of Plant Physiology*, 257: 153354.  
<https://doi.org/10.1016/j.jplph.2020.153354>
- Van Klompenburg T., Kassahun A., and Catal C., 2020, Crop yield prediction using machine learning: a systematic literature review, *Computers and Electronics in Agriculture*, 177: 105709.  
<https://doi.org/10.1016/j.compag.2020.105709>

- Varshney R., 2021, The *Plant Genome* special issue: advances in genomic selection and application of machine learning in genomic prediction for crop improvement, *The Plant Genome*, 14(3): e20178.  
<https://doi.org/10.1002/tpg2.20178>
- Vergopolan N., Xiong S., Estes L., Wanders N., Chaney N., Wood E., Konar M., Caylor K., Beck H., Gatti N., Evans T., and Sheffield J., 2020, Field-scale soil moisture bridges the spatial-scale gap between drought monitoring and agricultural yields, *Hydrology and Earth System Sciences*, 25: 1827-1847.  
<https://doi.org/10.5194/hess-2020-364-supplement>
- Vivek B., Krishna G., Vengadessan V., Babu R., Zaidi P., Kha L., Mandal S., Grudloyma P., Takalkar S., Krothapalli K., Singh I., Ocampo E., Xingming F., Burgueño J., Azrai M., Singh R., and Crossa J., 2017, Use of genomic estimated breeding values results in rapid genetic gains for drought tolerance in maize, *The Plant Genome*, 10(1): 70.  
<https://doi.org/10.3835/plantgenome2016.07.0070>
- Wang J., Liu L., He K., Gebrewahid T., Gao S., Tian Q., Li Z., Song Y., Guo Y., Li Y., Cui Q., Zhang L., Wang J., Huang C., Li L., Guo T., and Li H., 2025, Accurate genomic prediction for grain yield and grain moisture content of maize hybrids using multi-environment data, *Journal of Integrative Plant Biology*, 67(5): 1379-1394.  
<https://doi.org/10.1111/jipb.13857>
- Wang Y., Leng P., Shang G., Zhang X., and Li Z., 2023, Sun-induced chlorophyll fluorescence is superior to satellite vegetation indices for predicting summer maize yield under drought conditions, *Computers and Electronics in Agriculture*, 205: 107615.  
<https://doi.org/10.1016/j.compag.2023.107615>
- Wang Y., Lv J., Sun H., Zuo H., Gao H., Qu Y., Su Z., Yang X., and Yin J., 2022, Dynamic agricultural drought risk assessment for maize using weather generator and APSIM crop models, *Natural Hazards*, 114: 3083-3100.  
<https://doi.org/10.1007/s11069-022-05506-5>
- Wu C., Luo J., and Xiao Y., 2024, Multi-omics assists genomic prediction of maize yield with machine learning approaches, *Molecular Breeding*, 44: 1-17.  
<https://doi.org/10.1007/s11032-024-01454-z>
- Yuan Y., Cairns J., Babu R., Gowda M., Makumbi D., Magorokosho C., Zhang A., Liu Y., Wang N., Hao Z., Vicente S., Olsen M., Prasanna B., Lu Y., and Zhang X., 2019, Genome-wide association mapping and genomic prediction analyses reveal the genetic architecture of grain yield and flowering time under drought and heat stress conditions in maize, *Frontiers in Plant Science*, 9: 1919.  
<https://doi.org/10.3389/fpls.2018.01919>
- Zhang X., and Xu M.L., 2024, Adaptation of maize to various climatic conditions: genetic underpinnings, *Bioscience Evidence*, 14(3): 122-130.  
<https://doi.org/10.5376/be.2024.14.0014>
- Zhang A., Chen S., Cui Z., Liu Y., Guan Y., Yang S., Qu J., Nie J., Dang D., Li C., Dong X., Fan J., Zhu Y., Zhang X., Crossa J., Cao H., Ruan Y., and Zheng H., 2022, Genomic prediction of drought tolerance during seedling stage in maize using low-cost molecular markers, *Euphytica*, 218: 154.  
<https://doi.org/10.1007/s10681-022-03103-y>



---

#### 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.

---