

One Earth

Perspective Interdisciplinary strategies to enable data-driven plant breeding in a changing climate

Aaron Kusmec,¹ Zihao Zheng,¹ Sotirios Archontoulis,^{1,4} Baskar Ganapathysubramanian,^{2,4} Guiping Hu,^{3,4} Lizhi Wang,^{3,4} Jianming Yu,^{1,4} and Patrick S. Schnable^{1,4,*}

¹Department of Agronomy, Iowa State University, Ames, IA 50011, USA

²Department of Mechanical Engineering, Iowa State University, Ames, IA 50011, USA

³Department of Industrial and Manufacturing Systems Engineering, Iowa State University, Ames, IA 50011, USA

⁴Plant Sciences Institute, Iowa State University, Ames, IA 50011, USA

*Correspondence: schnable@iastate.edu

https://doi.org/10.1016/j.oneear.2021.02.005

SUMMARY

This perspective lays out a framework to enable the breeding of crops that can meet worldwide demand under the challenges of global climate change. Past work in various fields has produced multiple prediction methods to contribute to different plant breeding objectives. Our proposed framework focuses on the integration of these methods into decision-support tools that quantify the effects on multiple objectives of decisions made throughout the plant breeding pipeline. We discuss the complementarities among these methods with an emphasis on integration into tools that utilize operations research and systems approaches to help plant breeders rapidly and optimally design new cultivars under extant time, cost, and environmental constraints. In illustrating this potential, we demonstrate the interconnectedness and probabilistic nature of plant breeding objectives and highlight research opportunities to refine and combine knowledge across multiple disciplines. Such a framework can help plant breeders more efficiently breed for future environments, including so-called minor crops, leading to an overall increase in the resiliency of global food production systems.

INTRODUCTION

To ensure food security crop cultivars must not only produce high yields, but also do so across a range of management and environmental conditions. Global climate change challenges food security by increasing the intensity and frequency of environmental perturbations,¹ such as extreme precipitation events,² which threaten yield stability. In subsistence agriculture, crop stability is of paramount importance for local food security.³ In high-input systems, management decisions play an important role in mediating environmental variability to reduce farmers' risk; however, in the absence of correspondingly well-adapted cultivars, management decisions alone cannot ameliorate the detrimental effects of climate change.

Global climate change poses a challenge to plant breeders in part because cultivars are developed on a time lag. For example, in hybrid maize breeding, development of a cultivar can take 5–7 years, but these cultivars have a half-life of only 4 years, with complete turnover in 7 years.⁴ Variety development in soybean takes a comparable 6 years.⁵ Thus, breeders must develop cultivars for certain and uncertain changes in environments of a decade into the future based on field trials conducted under current weather conditions.⁶ The value of an early selection decision made under current conditions depends on the uncertainties in the responses of the selected cultivars to environmental factors and in predictions of decadal-scale weather patterns. Early selection decisions for traits with significant genotype-environment interactions (GEIs) under such uncertainty can lead to the loss of useful genetic variation, decreasing genetic gain in the target environments. The effects of GEIs can be partially controlled by the use of managed-stress environments and adequate environmental characterization, although the application of these will depend on the resources available to individual breeding programs. Low GEIs for grain yield and the use of these strategies have contributed to the relative success of some breeding programs, for example, maize in the central US Corn Belt.⁷

Precision agriculture has been proposed as an important component of a solution to global food security.^{8,9} Recent advances in remote-sensing technologies, cloud-based computing, and farm machinery enable the collection of realtime, fine-grained data that can be used to make agile in-season management decisions to preserve and enhance crop yields. These technologies promise to shrink the yield gap (the difference between yield under optimal conditions and realized yield in production systems) by improving farmers' ability to respond appropriately to environmental conditions. However, as retrospective analyses show, precision agriculture (or new agronomic practices in general) must be complemented by well-adapted cultivars that maintain or improve yield potential as climates change.^{10,11} Recent work by Cooper et al.⁷ has demonstrated the utility of a prediction framework that integrates traditional







Figure 1. Gains in hybrid maize yields require larger R&D investments over time (in constant dollars)

(A) Average US maize yields (in bushels/acre and kilograms/hectare). Yields are categorized by predominant breeding practice according to Troyer.²⁵ Best-fit linear regression lines are fit within each category, and the average rate of gain in bushels per acre per year is reported. Approximately 50%–60% of increased yields are due to genetic gain.¹⁰

(B) Public and private sector spending on agricultural research and development in billions of constant dollars (2013 US\$).

plant breeding measures of genetic gain with agronomic gap analysis techniques to target breeding efforts on those biophysical constraints with the highest impact on realized yields. Development of cultivars with reduced yield gaps will be especially important in the developing world, where the application of precision agriculture would first require massive investments in infrastructure¹² and education.¹³

Historically, plant breeding has relied on massive networks of multienvironment trials (METs) to identify candidate cultivars worthy of commercial deployment.¹⁴ If interactions between a cultivar's genotype and environment/management practices are important to performance, the scale of METs required to test these interactions can quickly become prohibitively costly and/or inefficient. Although the use of METs has been enormously successful in multiple crops,15 as a consequence of the law of diminishing returns, maintenance or acceleration of current rates of genetic gain requires ever greater investment¹⁶ (Figure 1). Hence, this traditional strategy is not economically sustainable. Global climate change threatens agriculture by, for example, increasing the yield gap of existing cultivars, threatening further genetic gain,¹⁷ and challenging the intensification of agricultural systems.¹⁸ Although there is evidence that existing germplasm will contribute to adaptation to climate change (e.g., in maize¹⁹), the combination of uncertainty in long-term climate predictions and uncertainty in crop responses to likely stresses²⁰ promises to greatly increase the difficulty of breeding for future climates. Some plant scientists and breeders envision furthering genetic gain by targeting physiological traits such as radiationuse efficiency and transpiration efficiency^{21,22} or populationlevel characteristics.²³ However, regardless of the targeted traits, due to the time and cost required for traditional breeding programs, innovation through in planta and field-based experiments is costly and time consuming, suggesting an urgent need for efficient, complementary approaches to improve plant breeding processes.^{14,24}

Until recently, data from METs have generally been used to identify regionally adapted cultivars¹⁴ without a long-term and large-scale design, with breeding decisions made on the basis of theory and experimental design derived from quantitative genetics, statistics, biophysical models, and best practices developed over the course of the past century.²⁶ Advances in crop growth models (CGMs), high-throughput genotyping, high-throughput phenotyping (HTP), and environmental sensing and

modeling have transformed different parts of plant breeding pipelines, but largely without the integration that offers the possibility of revolutionizing the process of plant breeding. These individual components are used to make decisions whose impacts may not be realized for several years.

In this perspective, we argue that by synthesizing these emerging technologies within flexible decision-support tools, plant breeding can be radically transformed to be able to rapidly, accurately, and cost-effectively evaluate cultivars and development pipelines *in silico* to more optimally use limited *in planta* trials. Building on previous work in various fields,^{24,27–30} we propose a framework that would facilitate the rapid prototyping of different breeding and management strategies, allowing breeders to increase the rate of genetic gain, improve yield stability, decrease the lengths of breeding cycles, and develop long-term breeding plans that incorporate both the nearly certain and the less certain effects of global climate change.

SOURCES OF UNCERTAINTY IN PLANT BREEDING

Plant breeders make probabilistic decisions in the context of environmental and genetic uncertainties and under constraints imposed by time and cost. These decisions often involve trade-offs between competing objectives such as genetic gain for two different target traits. Twentieth-century plant breeders made extensive use of different predictive models (discussed below) to increase the probability of making beneficial, low-risk decisions. Central to any attempt to develop predictive models is a thorough enumeration and understanding of the uncertainties, constraints, and trade-offs within the modeled system. In this section, we briefly review the major sources of uncertainty, constraints, and trade-offs navigated by plant breeders during cultivar development.

First, breeding decisions are inherently constrained by genetic stochasticity and reproductive biology. Breeders rely on genetic segregation and recombination events to combine favorable alleles from multiple parents into candidate cultivars. These events are predictable³¹ in aggregate, but not yet controllable.³² Because recombination between relevant loci can be rare, it may be necessary to genotype and/or field evaluate thousands of offspring to identify a single desired event. Advances in genome editing techniques hold promise for large-scale, precise genetic changes,³³ although further development is required to



increase their efficiency, throughput, and generalizability to the diversity of crop species.

Second, breeders must also plan within the constraints imposed by the environments in which cultivars are to be grown, which requires breeding cultivars for environments sometimes a decade in the future while evaluating them in present environments.¹⁴ Large-scale, long-term climate predictions of sufficient accuracy exist but are still subject to uncertainty,³⁴ which increases as models reach regional scales.³⁵ For example, this has negative impacts on the accuracy of prediction of precipitation,³⁶ a key input factor in crop yields. Finer-scaled predictions at the level of individual fields or parts of fields are more uncertain. It should also be recognized that there are additional factors, such as soil fertility dynamics and pest and disease damage, that are not incorporated into current models yet affect crop performance.³⁷

Third, uncertainty in weather prediction interacts with uncertainty in the responses of crops to different environmental factors to determine the overall level of prediction uncertainty.²⁰ Quantification of the relative uncertainties from these factors can be used to direct future research/breeding efforts and should be included in models to provide more realistic assessments. For example, incorporation of moderately accurate weather forecasts has been shown to improve farmer profitability when used to optimize crop mixtures and management decisions,^{38,39} which could be easily extended to the choice of cultivar(s) to plant.

Fourth, breeding decisions are constrained by cost and time. Fundamental time constraints are imposed by the reproductive biology of different crops and genetics, although some methods to relax this constraint have been developed for some crops, including speed breeding,⁴⁰ doubled-haploid technology,⁴¹ offseason nurseries, and tissue-culture-based embryo rescue that can reduce the length of the crop life cycle. The number and scale of yield tests that are performed also substantially influence the time and cost of a breeding program¹⁴ and affect the precision with which differences between candidate and current cultivars can be estimated.⁴² The length of the development cycle also influences the amount of uncertainty in environmental predictions because longer development cycles require the prediction of environmental conditions further into the future.

Finally, breeding decisions and these constraints interact to produce trade-offs. Many commercial cultivars have had traits introgressed from non-commercial cultivars or carry packages of traits introduced via genetic modification (GM) technologies. Introducing these traits into new cultivars adds time and cost to the breeding pipeline.⁴³ In addition, trait introgression is constrained by the biology of the crop, and, in the case of GM traits, is subject to a patchwork of regulatory regimes depending on target markets. An additional constraint comes from the necessary erosion of genetic diversity over the course of the breeding pipeline.⁴⁴ Although this is an intrinsic effect of the process of breeding superior cultivars, the maintenance and introduction of genetic variation are required for continued genetic gain.45 Breeders must balance the need for short-term gain to produce new cultivars with the consideration for long-term gain alongside maintaining and introducing favorable alleles for multiple traits in their populations without yield penalties. Although genomic prediction has the capacity to increase the rate of genetic gain in the

One Earth Perspective

short term, it also rapidly erodes genetic variance.⁴⁴ Methods that preserve genetic variance have been proposed,^{46–49} potentially at the cost of reduced rate of genetic gain in the short term.

Explicit consideration of these constraints by breeders in the context of robust management options available to farmers is critical to improving the ability of breeders to make data-driven, probability-based decisions by integrating diverse datasets and enabling stochastic modeling of outcomes. When extended beyond the breeding pipeline for a single cultivar, this approach can help breeders plan long term to produce crops that can handle the demands of global climate change and global food security. However, these constraints can be handled only to the extent that their effects are faithfully reproduced and accurately and precisely predicted by the models used to make plant breeding decisions.

PREDICTIVE MODELS FOR PLANT BREEDING

Accurate yield predictions help famers make informed economic and management decisions and can support strategies for plant breeding programs of various purposes.²⁴ Historically, yield predictions have been generated by statistical models based on quantitative genetics theory,⁵⁰ CGMs⁵¹, and machine learning⁵² (ML), which have complementary strengths and limitations. The application of any particular modeling approach depends on the objective of the prediction and the relevant constraints. In this section, we provide a brief overview of these three modeling paradigms in the context of plant breeding and then argue for the value of their combination in a nested modeling framework. Combining these approaches offers the potential to improve prediction accuracies for decision-making in breeding while providing insights into underlying biological processes.^{53,54}

Historical models for yield prediction

Prior to the advances in computing power that have enabled large-scale application of ML and CGMs, trait prediction and breeding methods relied heavily on quantitative genetics theory. The pioneering work of Charles Henderson on the use of pedigrees and mixed linear models for predictions of genetic merit in animal breeding was widely adopted in both animal and plant breeding.⁵⁵ The development of molecular markers in the 1980s allowed breeders to make predictions on the basis of realized genetic relationships rather than expected statistical relationships. However, the full power of molecular markers was not realized until the development of genomic prediction methods that use all available markers in regularized statistical models.⁵⁶ The further development of high-throughput genotyping and doubled-haploid technologies substantially improved breeding programs.⁵⁷ Much subsequent research has focused on the application of genomic prediction to breeding populations⁵⁸⁻⁶⁰ and the elaboration of more flexible models that also include GEIs.^{61–65} These models are, however, fundamentally simple, modeling the genotype-phenotype relationship as the sum of additive genetic effects. Increasing complexity by increasing the number of markers, accounting for epistasis, allowing GEIs, and considering non-linear relationships is a challenging statistical and computational problem.⁶⁶

CGMs were originally developed to assist with agronomic management decisions.⁶⁷ CGMs are biophysical models based

on physiological parameters and non-linear relationships between phenotypes and environment.⁶⁸ Although several applications of CGMs in major crops have been successful,⁶⁹ links between CGMs and plant breeding are still in their infancy.^{7,70–73} Because CGMs depend on physiological parameters that may vary genetically, experimental estimation of these parameters limits the application of CGMs at the scale of breeding programs. Recent work has demonstrated some success with *in silico* estimation of these parameters from experimental data,^{74,75} but challenges remain.⁷⁶ Alternatively, yield predictions can be improved using models that are agnostic to the actual values of the CGM parameters.^{73,77}

ML has gained popularity in human and plant genetics in recent years.78,79 In the context of trait prediction, ML models treat phenotype as an unknown function of genotype, environment, and their interactions and attempt to approximate their complex, often non-linear relationships by learning from large sets of training data. Many methods, such as neural networks, can approximate any complex function, although identifying the optimal hyperparameters is often challenging.^{80,81} Other methods are scalable to large datasets and have reasonably high prediction accuracies.⁸² However, due to the black-box nature of many ML models, it can prove difficult to explain why predictions are accurate or not, and they do not reveal the underlying effects that can form the basis of further biological research to improve physiological models and refine breeding objectives. Endowing ML models with interpretability mechanisms and principled approaches, such as adding domain constraints into ML models, is a promising avenue that the ML community is exploring to broaden their applicability.83

Nested models for predictive plant breeding

The strengths and weaknesses of these three modeling approaches are complementary; thus, combining them has the potential to improve prediction of yield and other traits. We propose a nested modeling strategy (Figure 3B) in which, by directly modeling physiological processes, CGMs integrate the outputs of quantitative genetic and ML models as biologically interpretable patterns of variation²⁶ for key adaptive traits and other traits of interest that are segregating in the breeding population.

For example, parallel advances in phenomics and ML have enabled the measurement and extraction both of traditional phenotypes at expanding scales and of novel phenotypes. Incorporation of these data into breeding programs is an open research question.⁸⁴ From a basic science perspective, they provide the opportunity to increase the surveyed genetic variation in the construction of CGMs,⁸⁵ potentially leading to refinement of models through expansion of phenotyping for physiological processes. This can also benefit traditional quantitative genetics characterization of context-dependent allelic effects in networks of METs.²⁷ Identification of key alleles for adaptive traits may then directly inform breeding decisions or influence the construction of CGMs. In this nested framework, the strengths of ML models for extracting traits from HTP data and quantitative genetic models for modeling the genetic basis of phenotypic variation are complemented by the incorporation of physiological processes and environmental variation from CGMs to produce probabilistic evaluations of likely yield outcomes under various climate change scenarios.



CGMs, therefore, playing a key role in linking the different classes of models used in the various disciplines that contribute to plant breeding, and their improvement are of vital importance to the success of various proposals for the future of predictive plant breeding.^{14,27,73,86} This integrative role, then, depends on the construction of CGMs with appropriate structures that operate on appropriate data.

CROP GROWTH MODELS

Structurally, a CGM can be thought of as a system of differential equations whose functional forms encapsulate the relationships between abiotic and biotic factors, crop physiology, and yield determined from retrospective modeling of experimental data with concomitant levels of accuracy and precision. For predictive breeding, the coefficients of these equations may vary genetically, introducing another layer of uncertainty.

Considering the demonstrated and expected contributions of CGMs to predictive plant breeding, it is important to consider in what ways CGMs might be improved as models of crop physiology in general and for their utility in simulating the outcomes of breeding decisions. Below, we discuss five general areas in which CGMs can be improved by refining the modeled relationships between environment and physiology (the first and second areas), updating the modeled environmental factors with current knowledge (the third and fourth), and incorporating genetic variation for the modeled physiological relationships (the fifth). We note that the following discussion is necessarily general and reflects the authors' judgment; further in-depth discussion of these and related domains can be found in related literature (e.g., Muller and Matre²⁸ and references therein).

Improve modeling of physiological processes

Much progress in modeling dynamic, biophysical processes has been made since the earliest CGMs were proposed. However, substantial knowledge gaps in fundamental processes remain, and new research has highlighted the importance of previously unmodeled processes.⁸⁷ Over the past decade improvements have been made in the simulation of maize nitrogen dynamics⁸⁸, maize grain growth dynamics⁸⁹, simulation of root depth by considering excess water stress⁹⁰, improvement of maize phenology prediction⁹¹, and new routines to simulate biomass partitioning in crops.⁹² These are all promising avenues of research, but they are often made in different crop species and CGMs. Coordination of physiological modeling progress across models and species will be required to realize the full potential of this research.

Introduce precise measurements of current and novel physiological parameters by high-throughput phenotyping

Reynolds et al.⁸⁴ reviewed numerous HTP methods and how the traits they measure relate to current breeding practices. Phenotyping is needed not only for selection decisions but also for genomic studies and translational research, among others. The specific-use case and cost-benefit ratio will be driving factors in the adoption of these technologies for various breeding purposes. HTP has the potential to reduce the cost and time necessary to measure key CGM parameters, improve the quality of



CGM simulations for individual cultivars, and enable prediction of CGM parameters for unphenotyped germplasm. In addition to the accuracy of the trait measurements, emphasis should also be placed on collecting accurate management inputs for the CGM. For example, Archontoulis et al.⁵¹ reported that the accuracy of management inputs accounted for one-quarter of the variability in maize yield prediction.

Improve modeling of extreme weather events

The majority of CGMs have been calibrated and tested using a narrow set of observations that do not include extreme weather events. Extreme weather events are predicted to increase in frequency and severity^{1,2}, and their impacts on crop yields are often complex, non-linear, and dependent on other stresses.93,94 For example, uncertainty in crop responses to temperature has a relatively high contribution to uncertainty in the effects of climate change on agriculture.²⁰ High temperatures decrease crop yields in a non-linear fashion, with variable effects when combined with drought stress and/or increasing CO₂ concentrations,^{19,93} and these interactions are species dependent. CGMs that accounted for the effects of such events would be better able to inform breeders about the impact of environmental change and uncertainty on breeding decisions. Weather events can also damage plants, leading to changes in trait values that have an impact on yield, such as hail damage that reduces leaf area index.⁹⁵ Thus, in addition to physiological processes, future improvements should incorporate functions to simulate plant damage.

Improve modeling of edaphic factors

A recent study reported that simulation of soil processes is less accurate than simulation of crop processes.⁵¹ This suggests that more emphasis needs to be placed on modeling dynamic environmental factors such as the microbiome, soil, water, and nutrients that determine the magnitude of plant stress and therefore the resulting yield gap.

Integrate genetic variation for key crop model input parameters

The difficulty of measuring the physiological parameters required by CGMs means that these models generally incorporate only a small sample of available genetic variation at best (see, for example, Padilla and Otegui⁸⁵). Promising results for expansion of this genetic sample have been obtained using combinations of HTP, statistical models, ML, and CGMs,^{72,74,75} although challenges remain due to the large number of combinations of CGM parameters that can lead to the same output.⁷⁶ To a certain extent, these all represent partial statistical fixes. A gold-standard CGM would incorporate direct measurements of physiological parameters, likely measured by HTP, and functional forms based on knowledge about the biological networks underlying the modeled processes at an appropriate level of abstraction. Although previous work has demonstrated that it is possible to improve the prediction accuracy of maize grain yield even without accurate knowledge of genetic variation for CGM parameters,^{73,77} the effect on prediction accuracy of integrating the measurements discussed here remains an open research question.

One Earth Perspective

DATA TO ENABLE PREDICTIVE PLANT BREEDING

The success of predictive models depends not only on their structures but also on the data on which they are trained. Plant breeding models depend on three broad classes of data: genetic, phenotypic, and environmental (Figure 4). Advances in molecular, remote-sensing, computational, and other technologies have greatly expanded the volume and variety of data available to plant breeders. Equally important to the integration of different modeling paradigms will be the principled integration of these data classes into plant breeding efforts (see Reynolds et al.⁸⁴ for a recent example). In this section, we briefly review salient advances and promising research directions in genotyping, phenotyping, and envirotyping.

Genotyping

Next-generation sequencing technologies such as mRNA sequencingenabled high-throughput discovery of molecular markers^{96,97} and characterization of molecular intermediaries between DNA and phenotype.^{98,99} The quantitative importance to phenotypic variation of other sequence-level characteristics, such as epigenetic marks, non-coding variation, and micro-RNAs, is an area of ongoing investigation.¹⁰⁰ In parallel, investments in computational infrastructure to store and analyze these data have been, and will continue to be, made.¹⁰¹⁻¹⁰⁴ Recent research on cryptographic protocols for genomic data sharing and analysis^{105,106} may further collaboration between the public and the private sectors to leverage the extensive databases of genotypic and phenotypic data warehoused by industry while preserving proprietary information. Because a limited number of candidate cultivars can be field tested, genotypes serve as a critical link between field-tested materials and untested materials held by breeding companies and in international gene banks.¹⁰⁷ We expect that further advances in high-throughput genotyping will continue to stimulate research into efficient and robust methods for identifying genotype-phenotype associations and prediction that will increase biological insight into physiological processes that can be used to improve CGMs and allow breeders to utilize larger pools of genetic diversity in breeding decisions.

Phenotyping

HTP has stimulated research into novel plant phenotypes,¹⁰⁸ ML,¹⁰⁹ artificial intelligence,¹¹⁰ and computer vision.¹¹¹ For example, HTP research has expanded beyond traditional RGB (red, green, blue) imaging to include hyperspectral, 3D, and other spectroscopic modalities.¹¹² The proximal challenge for HTP is its integration into plant breeding, where the definition and utility of different phenotypes may differ between the two fields.⁸⁴ Depending on the stage of cultivar development, crop, and phenotype, accurate and precise phenotypes may be needed for individual plants or rows to make selection decisions. Therefore, HTP requires rigorous evaluation to identify and mitigate sources of error^{113,114} (Y. Zhou, A.K., and P.S.S., unpublished data). Numerous avenues for improving the accuracy of HTP methods exist, including data augmentation,¹¹⁵ algorithmic developments, objective redesign, and multisensor modalities.^{116,117} Moreover, HTP also enables the measurement of novel phenotypes through multimodal, time-series-aware sensing technologies¹¹⁸ that can

CellPress OPEN ACCESS

Decisions:

- Which inbreds to cross?
- Test crosses to make?
- # of DHs to generate?
- Scale of per se testing?
- Introgression?
- Introgression?
- GM trait introduction?
- Implementation of GP?

Decisions:

- Selection of training population and/or environments?
 Which traits to select and when?
- when?
- Scale of tests?
- Location of tests?
- Design to supplement GP?
- How to advance hybrids?
- When to test untested hybrids?
- Implementation of GP?



help identify previously uncharacterized phenotypic hierarchies.¹¹⁹ It has long been acknowledged that HTP can be used to improve and train various predictive models. Combinations of these technologies can help breeders identify early yield and yield stability signatures and enable inference of difficult-to-measure traits from more easily measured traits using appropriate modeling strategies.⁸⁶ Finally, HTP can provide directions and feedback on the development of CGMs and plant ideotypes that will be beneficial in future environments.

Envirotyping

Environments include both the impacts of weather and climate and the management decisions made by farmers. They are the context in which cultivars develop and express phenotypes. The challenges posed by the interaction of genotype and environment have long been recognized.¹²⁰ Proper integration of environmental information for agriculture includes both short-term predictions used by farmers to make management decisions and long-term predictions used by breeders to develop new cultivars. Optimal decision-making in plant breeding will rely on the proper propagation of uncertainty from climate models at relevant spatiotemporal scales into CGMs and simulation platforms. Recent work on characterizing environments using multidimensional indices²⁹ and simulation platforms³⁰ represent promising advances in this direction. Continued improvement of environmental-sensing technologies complements HTP methods¹¹⁶ by enabling field-level monitoring for management decisions and linkage of plant-level phenotypes to microenvironments. Concurrently, research should also be focused on linking data from controlled and managed stress environments with performance in field environments^{14,113}

Figure 2. An idealized hybrid maize breeding program

Separate male and female heterotic pools are maintained and improved by crossing elite inbred lines within pools, generating doubled-haploid lines (DHs), and retaining candidates that perform well in *per* se field evaluations. New DHs are crossed with appropriate DHs (testers and elite lines) from the opposite heterotic pool and subjected to successive yield trials at increasing scale, leading to the release of 0–2 new hybrids. Examples of required breeding decisions are noted to illustrate the diversity, complexity, and number of decisions leading to the production of a single hybrid. Modified from Bernardo¹²² and Cooper et al.¹⁴ GM, genetically modified; GP, genomic prediction.

and the construction of facilities for phenotyping in extreme environments that are important to future conditions but unlikely to occur at present.¹²¹

INTEGRATING DECISION-SUPPORT TOOLS/PLATFORMS AND PLANT BREEDING

Plant breeders make numerous and diverse decisions during cultivar development. These include decisions driven by biology (e.g., choice of existing cultivars

to mate, advancement decisions, locations and scales of yield tests) and decisions driven by economics (e.g., budget constraints, development cycles, target markets, and the discount rate) (Figure 2). For example, Syngenta estimated that its soybean breeders make approximately 200 binary decisions over the course of 6 years to develop a single cultivar.⁵ These decisions interact with one another and with various stochastic factors (e.g., environmental conditions and genetic recombination) to affect the probability of success. The consequences of these decisions may not manifest until years later and be attributable to specific decisions only with low confidence, if at all.

Thus, it is critical that plant breeders be able to integrate data and models from multiple disciplines so they can make more informed decisions to balance risk and profitability subject to biological and economic constraints. Decision-support tools/ platforms (DSTs) that leverage advances in operations research, computational intelligence, and simulation models are a promising avenue to achieve this integration goal.^{5,123} These tools and platforms must be flexible and efficient, allowing breeders to simulate and evaluate the effects of different decisions on the probability of successful cultivar development in the context of environmental uncertainty caused by global climate change.

A key component of operations research approaches is proper model formulation that reflects the breeding system decision dynamics under uncertainty. Some aspects of these models will depend on the details of the reproductive biology of different crops and the strategies and technologies available to breeders. In this section, we provide example applications of DSTs to plant breeding, integrating the uncertainties, decision points, models, and data classes discussed in the previous sections (Figure 4).



CellPress OPEN ACCESS

Figure 3. Schematic of model integration for improved yield predictions

(A) Diverse, high-dimensional data sources generated by high-throughput phenotyping are used to train machine learning (ML) trait extraction models to produce numerous (semi-)automated phenotypes (\tilde{t}) at scale. Extraction models are calibrated on a smaller set of manually collected phenotypes (T). (B) Traits from high-throughput phenotyping (\tilde{t}), genotypes (G), and environmental factors (E) serve as inputs to ML and quantitative genetics-based models for phenotype prediction (\hat{T}) of unphenotyped candidates. Predicted phenotypes and environmental factors serve as input to crop growth models, which produce yield predictions (\hat{Y}). Biological insights derived from quantitative genetics-based models and comparison of predicted yields to actual yields (Y) can be used to stimulate basic science research leading to refinement of crop growth models.

Throughout this section, we use hybrid maize breeding as an example (Figure 2).

The first potential application is to the definition of the target population of environments (TPE) and target traits in the planning stage of a cultivar development pipeline. The TPE defines the environmental and management conditions under which a cultivar is expected to perform. This helps define targeted adaptive traits and provides a benchmark for genetic gain through comparison with the observed performance of cultivars currently grown in the current TPE. A TPE needs to be narrow enough that a reasonable suite of adaptive traits can be defined, that achievable goals for genetic gain can be determined, and that stable performance within the TPE is highly probable. At the same time, a TPE has to be wide enough to recoup the development costs of new cultivars. Climate modeling combined with digital gap analysis⁷ could be used to identify where yield gaps are likely to occur by the end of a development pipeline to guide cultivar design decisions. The same climate simulations could also be used to assess likely patterns of future environmental variation that may require redefinition of the TPE.¹⁹ Physiological modeling could then be used to identify adaptive traits that contribute to increased yield in the TPE.⁶ A range of possible ideal cultivars could be evaluated for feasibility using estimates of predicted genetic gain for the identified traits that incorporate, among other factors, phenotyping costs, available genetic variation, and

Perspective

One Earth

accuracy of genomic prediction models. This may lead to an iterative process that further refines the TPE and target adaptive traits based on expected development costs and time.

A second potential application is recurrent population improvement (e.g., Figure 2, part 1). Generally, recurrent population improvement seeks to produce genetic gain for one or more key traits while meeting two conflicting objectives. The first objective is to realize enough genetic gain that the population can be sampled periodically to produce improved cultivars (e.g., hybrids in Figure 2, part 2). The second objective is to maintain genetic variation generally in the population to enable longterm genetic gain and selection on traits that may increase in importance in the future. Due to correlations between multiple traits and the negative impact of selection on genetic variation, these objectives are often conflicting. Recent work⁴⁹ has demonstrated how such multitrait objectives could be addressed by proposing flexible multitrait desired gains indices that are evaluated through stochastic genetic simulation of proposed crosses subject to time and resource constraints. The proposed approach was shown to have a higher probability of achieving the desired genetic gains while maintaining genetic variation for target and non-target traits.

A third potential application is the design of METs. METs narrow down the set of candidate cultivars by testing them at an everwider network of trial locations that sample the TPE (Figure 2, part 2). The scale of testing depends on the expected variability within the TPE and is constrained by per-trial costs. It is critical that the trial locations be a representative sample of the TPE's diversity because the expected genetic gain in the TPE depends on the similarity between the TPE and the environments where selection is performed (i.e., the MET). Climate modeling and stochastic weather generators can be used to quantify the likely characteristics of a future TPE. These forecasts can then help with several decisions. By providing an estimate of the relevant dimensions of environmental variability, they can be used to determine which characteristics of trial locations should be prioritized and the sampling density of those environmental variables. Another application is to cross-reference the forecasts with historical and expected conditions at current trial locations to determine their suitability for making selection decisions. This may lead to the consideration of new trial locations, which may be undesirable due to economic constraints, and this could be weighed against probabilistic estimates of genetic gain under different location combinations. Possible reductions in the number of trial locations could also lead to the possibility of evaluating more candidate cultivars, improving sampling of the genetic and GEI variability.

CONCLUSIONS

This perspective lays out a vision for a framework to integrate CGMs, high-throughput genotyping, HTP, and environmental sensing and modeling with new advances in statistical modeling, ML, and operations research in flexible and efficient decision-support tools (Figure 4). These tools would facilitate the rapid design, development, evaluation, and modeling of new crop cultivars under genetic and environmental uncertainty. To that end, we have emphasized the complementarity of emerging, data-driven techniques and traditional statistical and biological models. We also recognize that there is often significant lag





Figure 4. Concept map illustrating the relationships among decision-support tools, models, genotypes, phenotypes, and environments Genotype, phenotype, and environment are incorporated into the nested model of Figure 3B, which produces yield and other trait predictions. These predictions along with genotype, phenotype, and environment are evaluated in decision-support tools that can direct further research.

time between the development of new methods and technologies and their widespread adoption. Prospects for rapid deployment of these tools must consider as development objectives consistency with users' end needs, materials for education and training, and development of necessary operational infrastructure.

The data-driven nature of these tools makes them applicable to three further cases. First, approximately 7.4 million plant accessions are held in gene banks along with vast libraries of insufficiently field-tested materials in public and private breeding programs that are underutilized in current breeding schemes due to resource constraints. The approaches described in this perspective can be used to leverage these accessions for further crop improvement, bringing in much needed sources of genetic variation.¹⁰⁷

Second, these tools are applicable to so-called minor crops that play outsized roles in human nutrition (e.g., casava and banana, among others), which have thus far received less attention from plant breeders, despite the fact that some of them contribute substantially to the human food supply.¹²⁴ The lack of genetic, phenotypic, and physiological resources for these crops hampers traditional breeding efforts. The tools we have described could be used to accelerate the accretion of better knowledge bases in these crops for future breeding efforts and development of predictive models. Investment in these crops may provide outsized benefits compared with the return achieved from investments in established breeding programs for major crops, including by diversifying global food systems.¹²⁵

Third, persistent hunger¹²⁶ and famine¹²⁷ remain obstacles to global food security in the 21st century. Ending global hunger, improving food security, and developing sustainable agricultural practices are components of the United Nations Sustainable Development Goals for 2030 (https://sdgs.un.org/goals). These

efforts would be assisted through the diversification of global food systems discussed above. In addition, the tools and integration we describe can contribute to the realization of these goals and serve as foundations for broader simulations and predictions of the vulnerabilities in complex agricultural systems to predict impending famines with sufficient advance notice to potentially mitigate deleterious health and social effects.

Due to population growth and changing consumer preferences, maintenance of global food security requires more than maintaining current rates of genetic gain, particularly in the context of the diminishing returns of R&D investments in breeding major crops (Figure 1). The challenge of breeding to meet these demands would exist even in the absence of global climate change. However, global climate change challenges the fundamentally prospective nature of plant breeding by shifting the climate and increasing weather variability, introducing more uncertainty into the evaluation of new cultivars in current environments. Therefore, improved cultivars must also demonstrate improved performance and stability with respect to a wider range of environmental perturbations based on predictions of future environments. The data-driven approaches we describe and their integration into decision-support tools have the potential to revolutionize 21st century plant breeding by enabling breeders to rapidly and cost-effectively design and develop appropriate cultivars for these future environments.

EXPERIMENTAL PROCEDURES

Resource availability Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the corresponding author, Patrick S. Schnable (schnable@ iastate.edu).



Materials availability

This study did not generate new unique materials. **Data and code availability**

Data and code to reproduce Figure 1 can be obtained from GitHub (https:// github.com/amkusmec/OneEarthPerspective). Data for Figure 1A were retrieved from the US Department of Agriculture-National Agricultural Statistics Service (USDA-NASS) website (https://quickstats.nass.usda.gov/) on August 20, 2020. The following options were selected:

- Select Commodity: SURVEY > CROPS > FIELD CROPS > CORN > YIELD > CORN, GRAIN - YIELD, MEASURED IN BU/ACRE > TOTAL
- Select Location: NATIONAL > US TOTAL
- Select Time: 1866–2019 > ANNUAL > YEAR

Data for Figure 1B were retrieved from the USDA-Economic Research Service (USDA-ERS) website (https://www.ers.usda.gov/data-products/ agricultural-research-funding-in-the-public-and-private-sectors/) accessed on August 20, 2020.

Regression analysis

Following Troyer²⁵, average yield for each year was assigned to one of four categories: "open-pollinated" (1866–1929), "double cross" (1930–59), "single cross" (1960–94), and "biotech/GMO" (1995–2019). Linear regression of yield on year within each category was used to estimate the average rate of gain in bushels per acre per year. Best-fit lines and average rates of gain are reported in Figure 1.

ACKNOWLEDGMENTS

The authors would like to thank the three anonymous reviewers whose comments and suggestions greatly improved the manuscript. This material is based upon work supported by the National Science Foundation under grant 1842097 (to L.W., P.S.S., B.G., and S.A.) of the EAGER program and grant 1830478 (to L.W., S.A., and G.H.) of the LEAP HI and GOALI programs. This work is also supported by Agriculture and Food Research Initiative grant 2017-67007-26175/project accession no. 1011702 (to P.S.S., G.H., and L.W.) and by Plant Health and Production and Plant Products: Plant Breeding for Agricultural Production grant 2020-68013-30934/project accession no. 1022368 from the USDA National Institute of Food and Agriculture (to P.S.S.), and by Hatch project no. IOW04714 from the National Institute of Food and Agriculture (to P.S.S.). Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the US Department of Agriculture. Z.Z. is supported by the Department of Energy Advanced Research Projects Agency-Energy (grant DE-AR0000826 to P.S.S. and J.Y.). Additional support is received from Iowa State University's Raymond F. Baker Center for Plant Breeding (to J.Y.) and Iowa State University's Plant Sciences Institute Scholars' Program (to S.A., B.G., G.H., L.W., J.Y., and P.S.S.). Cheng-Ting Yeh (Schnable Lab, Iowa State University) prepared the figures in consultation with A.K., Z.Z., and P.S.S.

AUTHOR CONTRIBUTIONS

All authors contributed to conceptualization of the manuscript. A.K., Z.Z., and P.S.S. wrote the original draft with the support and review of the other coauthors.

DECLARATION OF INTERESTS

A.K., Z.Z., S.A., B.G., and J.Y. declare no competing interests. G.H. and L.W. are co-founders of LookAhead Breeding, LLC. P.S.S. is a Changjiang Scholar at China Agriculture University. He is a co-founder and managing partner of Data2Bio, LLC; Dryland Genetics, LLC; and EnGeniousAg, LLC. He is a member of the scientific advisory board and a shareholder of Hi-Fidelity Genetics, Inc., and a member of the scientific advisory boards of Kemin Industries and Centro de Tecnologia Canavieira.

REFERENCES

 Hartmann, D.L., Klein Tank, A.M.G., Rusticucci, M., Alexander, L.V., Bronninmann, S., Charabi, Y.A.-R., et al. (2013). Observations: atmosphere and surface. In Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, T.F. Stocker, D. Qin, G.-K.



Plattner, M. Tignor, S.K. Allen, and J. Boschung, et al., eds. (Cambridge University Press), pp. 159–254.

- Myhre, G., Alterskjær, K., Stjern, C.W., Hodnebrog, Ø., Marelle, L., Samset, B.H., Sillmann, J., Schaller, N., Fischer, E., Schulz, M., et al. (2019). Frequency of extreme precipitation increases extensively with event rareness under global warming. Sci. Rep. 9, 16063.
- Kruseman, G., Bairagi, S., Komarek, A.M., Milan, A.M., Nedumaran, S., Petsakos, A., Prager, S., and Yigezu, Y.A. (2020). CGIAR modeling approaches for resource-constrained scenarios: II. Models for analyzing socioeconomic factors to improve policy recommendations. Crop Sci. 60, 568–581.
- 4. Edgerton, M.D. (2009). Increasing crop productivity to meet global needs for feed, food, and fuel. Plant Physiol. *149*, 7–13.
- Byrum, J., Davis, C., Doonan, G., Doubler, T., Foster, D., Luzzi, B., Mowers, R., Zinselmeier, C., Kloeber, J., Culhane, D., et al. (2016). Advanced analytics for agricultural product development. Interfaces (Providence) 46, 5–17.
- Hammer, G.L., McLean, G., van Oosterom, E., Chapman, S., Zheng, B., Wu, A., Doherty, A., and Jordan, D. (2020). Designing crops for adaptation to the drought and high-temperature risks anticipated in future climates. Crop Sci. 60, 605–621.
- Cooper, M., Messina, C., Tang, T., Gho, C., Hart, T., and Hammer, G. (2020). Integrating genetic gain and gap analysis to predict improvements in crop productivity. Crop Sci. 60, 582–604.
- Cassman, K.G. (1999). Ecological intensification of cereal production systems: yield potential, soil quality, and precision agriculture. Proc. Natl. Acad. Sci. U S A 96, 5952–5959.
- Alexandratos, N., and Bruinsma, J. (2012). World Agriculture towards 2015/2030: The 2012 Revision (Food and Agriculture Organization of the United Nations).
- Duvick, D.N. (2005). Genetic progress in yield of United States maize (Zea mays L.). Maydica 50, 193–202.
- Challinor, A.J., Koehler, A.-K., Whitfield, S., and Das, B. (2016). Current warming will reduce yields unless maize breeding and seed systems adapt immediately. Nat. Clim. Chang. 6, 954–958.
- Lowenberg-DeBoer, J., and Erickson, B. (2019). Setting the record straight on precision agriculture adoption. Agron. J. 111, 1552–1569.
- Kitchen, N.R., Snyder, C.J., Franzen, D.W., and Wiebold, W.J. (2002). Educational needs of precision agriculture. Precis. Agric. 3, 341–351.
- Cooper, M., Messina, C.D., Podlich, D., Totir, L.R., Baumgarten, A., Hausmann, N.J., Wright, D., and Graham, G. (2014). Predicting the future of plant breeding: complementing empirical evaluation with genetic prediction. Crop Pasture Sci. 65, 311.
- Xu, Y., Li, P., Zou, C., Lu, Y., Xie, C., Zhang, X., Prasanna, B.M., and Olsen, M.S. (2017). Enhancing genetic gain in the era of molecular breeding. J. Exp. Bot. 68, 2641–2666.
- Ruttan, V.W. (1982). Agricultural Research Policy (University of Minnesota Press).
- Pardey, P.G., Beddow, J.M., Hurley, T.M., Beatty, T.K.M., and Eidman, V.R. (2014). A bounds analysis of world food futures: global agriculture through to 2050. Aust. J. Agric. Resour. Econ. 58, 571–589.
- Fischer, R.A., Byerlee, D., and Edmeades, G.O. (2014). Crop yields and global food security: will yield increase continue to feed the world?, ACIAR Monogr. No. 158 (Australian Centre for International Agricultural Research).
- Harrison, M.T., Tardieu, F., Dong, Z., Messina, C.D., and Hammer, G.L. (2014). Characterizing drought stress and trait influence on maize yield under current and future conditions. Glob. Chang. Biol. 20, 867–878.
- Lobell, D.B., and Burke, M.B. (2008). Why are agricultural impacts of climate change so uncertain? The importance of temperature relative to precipitation. Environ. Res. Lett. *3*, 034007.
- Fischer, R.A., and Edmeades, G.O. (2010). Breeding and cereal yield progress. Crop Sci. 50, 85–98.
- Hall, A.J., and Richards, R.A. (2013). Prognosis for genetic improvement of yield potential and water-limited yield of major grain crops. Field Crops Res. 143, 18–33.
- Weiner, J. (2019). Looking in the wrong direction for higher-yielding crop genotypes. Trends Plant Sci. 24, 927–933.
- Messina, C.D., Cooper, M., Reynolds, M., and Hammer, G.L. (2020). Crop science: a foundation for advancing predictive agriculture. Crop Sci. 60, 1–3.
- Troyer, A.F. (2006). Adaptedness and heterosis in corn and mule hybrids. Crop Sci. 46, 528–543.

One Earth

Perspective

- Hammer, G., Cooper, M., Tardieu, F., Welch, S., Walsh, B., van Eeuwijk, F., Chapman, S., and Podlich, D. (2006). Models for navigating biological complexity in breeding improved crop plants. Trends Plant Sci. 11, 587–593.
- Tardieu, F., Simonneau, T., and Muller, B. (2018). The physiological basis of drought tolerance in crop plants: a scenario-dependent probabilistic approach. Annu. Rev. Plant Biol. 69, 733–759.
- Muller, B., and Matre, P. (2019). Plant and crop simulation models: powerful tools to link physiology, genetics, and phenomics. J. Exp. Bot. 70, 2339–2344.
- Resende, R.T., Piepho, H.-P., Rosa, G.J.M., Silva-Junior, O.B., e Silva, F.F., de Resende, M.D.V., and Grattapaglia, D. (2020). Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. Theor. Appl. Genet. 134, 95–112.
- **30.** de los Campos, G., Perez-Rodriguez, P., Bogard, M., Gouache, D., and Crossa, J. (2020). A data-driven simulation platform to predict cultivars' performances under uncertain weather conditions. Nat. Commun. *11*, 4876.
- Rodgers-Melnick, E., Bradbury, P.J., Elshire, R.J., Glaubitz, J.C., Acharya, C.B., Mitchell, S.E., Li, C., Li, Y., and Buckler, E.S. (2015). Recombination in diverse maize is stable, predictable, and associated with genetic load. Proc. Natl. Acad. Sci. U S A *112*, 3823–3828.
- Bernardo, R. (2017). Prospective targeted recombination and genetic gains for quantitative traits in maize. Plant Genome 10, https://doi.org/ 10.3835/plantgenome2016.11.0118.
- 33. Jenko, J., Gorjanc, G., Cleveland, M.A., Varshney, R.K., Whitelaw, C.B.A., Woolliams, J.A., and Hickey, J.M. (2015). Potential of promotion of alleles by genome editing to improve quantitative traits in livestock breeding programs. Genet. Sel. Evol. 47, 55.
- 34. Marotzke, J. (2019). Quantifying the irreducible uncertainty in near-term climate projections. WIREs Clim. Chang. *10*, e563.
- Shepherd, T.G. (2014). Atmospheric circulation as a source of uncertainty in climate change projections. Nat. Geosci. 7, 703–708.
- Hawkins, E., and Sutton, R. (2011). The potential to narrow uncertainty in projections of regional precipitation change. Clim. Dyn. 37, 407–418.
- 37. Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., et al. (2017). Toward a new generation of agricultural system data, models, and knowledge products: State of agricultural systems science. Agric. Syst. 155, 269–288.
- Messina, C.D., Hansen, J.W., and Hall, A.J. (1999). Land allocation conditioned on El Nino-Southern Oscillation phases in the Pampas of Argentina. Agric. Syst. 60, 197–212.
- Jones, J.W., Hansen, J.W., Royce, F.S., and Messina, C.D. (2000). Potential benefits of climate forecasting to agriculture. Agric. Ecosyst. Environ. 82, 169–184.
- Watson, A., Ghosh, S., Williams, M.J., Cuddy, W.S., Simmonds, J., Rey, M.-D., Hatta, M.A.M., Hinchliffe, A., Steed, A., Reynolds, D., et al. (2018). Speed breeding: a powerful tool to accelerate crop research and breeding. Nat. Plants 4, 23–29.
- Andorf, C., Beavis, W.D., Hufford, M., Smith, S., Suza, W.P., Wang, K., Woodhouse, M., Yu, J., and Lübberstedt, T. (2019). Technological Advances in Maize Breeding: Past, Present and Future (Springer Berlin Heidelberg).
- Schmidt, P., Hartung, J., Bennewitz, J., and Piepho, H.-P. (2019). Heritability in plant breeding on a genotype-difference basis. Genetics 212, 991–1008.
- Cameron, J.N., Han, Y., Wang, L., and Beavis, W.D. (2017). Systematic design for trait introgression projects. Theor. Appl. Genet. 130, 1993–2004.
- Jannink, J.-L. (2010). Dynamics of long-term genomic selection. Genet. Sel. Evol. 42, 35.
- Gorjanc, G., Jenko, J., Hearne, S.J., and Hickey, J.M. (2016). Initiating maize pre-breeding programs using genomic selection to harness polygenic variation from landrace populations. BMC Genomics 17, 30.
- 46. Daetwyler, H.D., Hayden, M.J., Spangenberg, G.C., and Hayes, B.J. (2015). Selection on optimal haploid value increases genetic gain and preserves more genetic diversity relative to genomic selection. Genetics 200, 1341–1348.
- Goiffon, M., Kusmec, A., Wang, L., Hu, G., and Schnable, P.S. (2017). Improving response in genomic selection with a population-based selection strategy: optimal population value selection. Genetics 206, 1675–1682.
- Akdemir, D., and Sánchez, J.I. (2016). Efficient breeding by genomic mating. Front. Genet. 7, 210.

- Moeinizade, S., Kusmec, A., Hu, G., Wang, L., and Schnable, P.S. (2020). Multi-trait genomic selection methods for crop improvement. Genetics 215, 931–945.
- Xavier, A., Muir, W.M., Craig, B., and Rainey, K.M. (2016). Walking through the statistical black boxes of plant breeding. Theor. Appl. Genet. 129, 1933–1949.
- Archontoulis, S.V., Castellano, M.J., Licht, M.A., Nichols, V., Baum, M., Huber, I., Martinez-Feria, R., Puntel, L., Ordóñez, R.A., Iqbal, J., et al. (2020). Predicting crop yields and soil-plant nitrogen dynamics in the US Corn Belt. Crop Sci. 60, 721–738.
- Shahhosseini, M., Martinez-Feria, R.A., Hu, G., and Archontoulis, S.V. (2019). Maize yield and nitrate loss prediction with machine learning algorithms. Environ. Res. Lett. 14, 124026.
- Washburn, J.D., Burch, M.B., and Franco, J.A.V. (2020). Predictive breeding for maize: making use of molecular phenotypes, machine learning, and physiological crop models. Crop Sci. 60, 622–638.
- McCormick, R.F., Truong, S.K., Rotundo, J., Gaspar, A.P., Kyle, D., van Eeuwijk, F., and Messina, C.D. (2020). Intercontinental prediction of soybean phenology via hybrid ensemble of knowledge-based and datadriven models. bioRxiv. https://doi.org/10.1101/2020.09.22.306506.
- Henderson, C.R. (1984). In Applications of Linear Models in Animal Breeding Models, Third Edition (University of Guelph).
- Meuwissen, T.H.E., Hayes, B.J., and Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.
- Bernardo, R., and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. Crop Sci. 47, 1082–1090.
- Heffner, E.L., Sorrells, M.E., and Jannink, J.-L. (2009). Genomic selection for crop improvement. Crop Sci. 49, 1–12.
- 59. Jannink, J.-L., Lorenz, A.J., and Iwata, H. (2010). Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genomics 9, 166–177.
- Voss-Fels, K.P., Cooper, M., and Hayes, B.J. (2018). Accelerating crop genetic gains with genomic selection. Theor. Appl. Genet. 132, 669–686.
- Heslot, N., Akdemir, D., Sorrells, M.E., and Jannink, J.-L. (2014). Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theor. Appl. Genet. 127, 463–480.
- Morota, G., and Gianola, D. (2014). Kernel-based whole-genome prediction of complex traits: a review. Front. Genet. 5, 363.
- 63. Jarquín, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., Piraux, F., Guerreiro, L., Pérez, P., Calus, M., et al. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. Theor. Appl. Genet. 127, 595–607.
- 64. Ly, D., Huet, S., Gauffreteau, A., Rincent, R., Touzy, G., Mini, A., Jannink, J.-L., Cormier, F., Paux, E., Lafarge, S., et al. (2018). Whole-genome prediction of reaction norms to environmental stress in bread wheat (Triticum aestivum L.) by genomic random regression. Field Crops Res. 216, 32–41.
- 65. Elias, A.A., Rabbi, I., Kulakow, P., and Jannink, J.-L. (2018). Improving genomic prediction in cassava field experiments by accounting for interplot competition. G3 (Bethesda) 8, 933–944.
- 66. Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., Campos, G.de los, Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y., et al. (2017). Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 22, 961–975.
- Jones, J.W., Antle, J.M., Basso, B., Boote, K.J., Conant, R.T., Foster, I., Godfray, H.C.J., Herrero, M., Howitt, R.E., Janssen, S., et al. (2016). Brief history of agricultural systems modeling. Agric. Syst. 155, 240–254.
- 68. Keating, B.A., Carberry, P.S., Hammer, G.L., Probert, M.E., Robertson, M.J., Holzworth, D., Huth, N.I., Hargreaves, J.N.G., Meinke, H., Hochman, Z., et al. (2003). An overview of APSIM, a model designed for farming systems simulation. Eur. J. Agron. 18, 267–288.
- 69. Di Paola, A., Valentini, R., and Santini, M. (2016). An overview of available crop growth and yield models for studies and assessments in agriculture. J. Sci. Food Agric. *96*, 709–714.
- Chapman, S., Cooper, M., Podlich, D., and Hammer, G. (2003). Evaluating plant breeding strategies by simulating gene action and dryland environment effects. Agron. J. 95, 99–113.
- Chenu, K., Chapman, S.C., Tardieu, F., McLean, G., Welcker, C., and Hammer, G.L. (2009). Simulating the yield impaces of organ-level quantitative trait loci associated with drought response in maize: a "gene-tophenotype" modeling approach. Genetics 183, 1507–1523.







- Messina, C.D., Technow, F., Tang, T., Totir, R., Gho, C., and Cooper, M. (2018). Leveraging biological insight and environmental variation to improve phenotypic prediction: integrating crop growth models (CGM) with whole genome prediction (WGP). Eur. J. Agron. 100, 151–162.
- 74. Lamsal, A., Welch, S.M., Jones, J.W., Boote, K.J., Asebedo, A., Crain, J., Wang, X., Boyer, W., Giri, A., Frink, E., et al. (2017). Efficient crop model parameter estimation and site characterization using large breeding trial data sets. Agric. Syst. 157, 170–184.
- Casadebaig, P., Debaeke, P., and Wallach, D. (2020). A new approach to crop model calibration: phenotyping plus post-processing. Crop Sci. 60, 709–720.
- Lamsal, A., Welch, S.M., White, J.W., Thorp, K.R., and Bello, N.M. (2018). Estimating parametric phenotypes that determine anthesis date in Zea mays: challenges in combining ecophysiological models with genetics. PLoS One 13, e0195841.
- Technow, F., Messina, C.D., Totir, L.R., and Cooper, M. (2015). Integrating crop growth models with whole genome prediction through approximate Bayesian computation. PLoS One 10, e0130855.
- 78. Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nat. Rev. Genet. *16*, 321–332.
- Ma, C., Zhang, H.H., and Wang, X. (2014). Machine learning for Big Data analytics in plants. Trends Plant Sci. 19, 798–808.
- Hornik, K., Stinchcombe, M., and White, H. (1990). Universal approximation of an unknown mapping and its derivatives using multilalyer feedforward networks. Neural Netw. 3, 551–560.
- Bellot, P., de los Campos, G., and Pérez-Enciso, M. (2018). Can deep learning improve genomic prediction of complex human traits? Genetics 210, 809–819.
- Khaki, S., and Wang, L. (2019). Crop yield prediction using deep neural networks. Front. Plant Sci. 10, 621.
- Washburn, J.D., Mejia-guerra, M.K., Ramstein, G.P., Kremling, K.A.G., Valluru, R., Buckler, E.S., and Wang, H. (2019). Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proc. Natl. Acad. Sci. U S A *116*, 5542–5549.
- Reynolds, M., Chapman, S., Crespo-Herrera, L., Molero, G., Mondal, S., Pequeno, D.N.L., Pinto, F., Pinera-Chavez, F.J., Poland, J., Rivera-Amado, C., et al. (2020). Breeder friendly phenotyping. Plant Sci. 295, 110396.
- Padilla, J.M., and Otegui, M.E. (2005). Co-ordination between leaf initiation and leaf appearance in field-grown maize (Zea mays): genotypic differences in response of rates to temperature. Ann. Bot. 96, 997–1007.
- Eeuwijk, F.A. Van, Bustos-Korts, D.V., Millet, E.J., Boer, M.P., Kruijer, W., Thompson, A., Malosetti, M., Iwata, H., Quiroz, R., Kuppe, C., et al. (2018). Modelling strategies for assessing and increasing the effectiveness of new phenotyping techniques in plant breeding. Plant Sci. 282, 23–39.
- Boote, K.J., Jones, J.W., White, J.W., Asseng, S., and Lizaso, J.I. (2013). Putting mechanisms into crop production models. Plant Cell Environ. 36, 1658–1672.
- Soufizadeh, S., Munaro, E., McLean, G., Massignam, A., van Oosterom, E.J., Chapman, S.C., Messina, C., Cooper, M., and Hammer, G.L. (2018). Modelling the nitrogen dynamics of maize crops – Enhancing the APSIM maize model. Eur. J. Agron. 100, 118–131.
- Messina, C.D., Hammer, G.L., McLean, G., Cooper, M., van Oosterom, E.J., Tardieu, F., Chapman, S.C., Doherty, A., and Gho, C. (2019). On the dynamic determinants of reproductive failure under drought in maize. In Silico Plants 1, diz003.
- Ebrahimi-Mollabashi, E., Huth, N.I., Holzwoth, D.P., Ordóñez, R.A., Hatfield, J.L., Huber, I., Castellano, M.J., and Archontoulis, S.V.((2019). Enhancing APSIM to simulate excessive moisture effects on root growth. Field Crops Res. 236, 58–67.
- 91. Tollenaar, M., Dzotsi, K., Kumudini, S., Boote, K., Chen, K., Hatfield, J., Jones, J.W., Lizaso, J.I., Nielsen, R.L., Thomison, P., et al. (2017). Modeling the effects of genotypic and environmental variation on maize phenology: the phenology subroutine of the AgMaize crop model. In Agroclimatology: Linking Agriculture to Climate Agronomy Monograph, J.L. Hatfield, M.V.K. Sivakumar, and J.H. Prueger, eds., pp. 173–200.
- Brown, H.E., Huth, N.I., Holzworth, D.P., Teixeira, E.I., Wang, E., Zyskowski, R.F., and Zheng, B. (2019). A generic approach to modelling, allocation and redistribution of biomass to and from plant organs. In Silico Plants 1, diy004.

- Schauberger, B., Archontoulis, S.V., Arneth, A., Balkovic, J., Ciais, P., Deryng, D., Elliott, J., Folberth, C., Khabarov, N., Muller, C., et al. (2017). Consistent negative response of US crops to high temperatures in observations and crop models. Nat. Commun. 8, 13931.
- Lesk, C., Coffel, E., and Horton, R. (2020). Net benefits to US soy and maize yields from intensifying hourly rainfall. Nat. Clim. Chang. 10, 819–822.
- Conley, S.P., Abendroth, L., Elmore, R., Christmas, E.P., and Zarnstorff, M. (2008). Soybean seed yield and composition response to stand reduction at vegetative and reproductive stages. Agron. J. 100, 1666–1669.
- 96. Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A robust, simple genotypingby-sequencing (GBS) approach for high diversity species. PLoS One 6, e19379.
- Ott, A., Liu, S., Schnable, J.C., Yeh, C.-T.E., Wang, K.-S., and Schnable, P.S. (2017). tGBS[®] genotyping-by-sequencing enables reliable genotyping of heterozygous loci. Nucleic Acids Res. 45, e178.
- Lin, H.-Y., Liu, Q., Li, X., Yang, J., Liu, S., Huang, Y., Scanlon, M.J., Nettleton, D., and Schnable, P.S. (2017). Substantial contribution of genetic variation in the expression of transcription factors to phenotypic variation revealed by eRD-GWAS. Genome Biol. 18, 192.
- Kremling, K.A.G., Chen, S.-Y., Su, M.-H., Lepak, N.K., Romay, M.C., Swarts, K.L., Lu, F., Lorant, A., Bradbury, P.J., and Buckler, E.S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. Nature 555, 520–523.
- 100. Xu, G., Lyu, J., Li, Q., Liu, H., Wang, D., Zhang, M., Springer, N.M., Ross-Ibarra, J., and Yang, J. (2020). Evolutionary and functional genomics of DNA methylation in maize domestication and improvement. Nat. Commun. 11, 5539.
- Nekrutenko, A., and Taylor, J. (2012). Next-generation sequencing data interpretation: Enhancing reproducibility and accessibility. Nat. Rev. Genet. 13, 667–672.
- 102. Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E.S. (2007). TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23, 2633–2635.
- 103. Lipka, A.E., Tian, F., Wang, Q., Peiffer, J.A., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E.S., and Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. Bioinformatics 28, 2397–2399.
- 104. Paten, B., Novak, A.M., Eizenga, J.M., and Garrison, E. (2017). Genome graphs and the evolution of genome inference. Genome Res. 27, 665–676.
- 105. Blatt, M., Gusev, A., Polyakov, Y., and Goldwasser, S. (2020). Secure large-scale genome-wide association studies using homomorphic encryption. Proc. Natl. Acad. Sci. U S A 117, 11608–11613.
- 106. Mott, R., Fischer, C., Prins, P., and Davies, R.W. (2020). Private genomes and public SNPs: homomorphic encryption of genotypes and phenotypes for shared quantitative genetics. Genetics 215, 359–372.
- 107. Yu, X., Li, X., Guo, T., Zhu, C., Wu, Y., Mitchell, S.E., Roozeboom, K.L., Wang, D., Wang, M.L., Pederson, G.A., et al. (2016). Genomic prediction contributing to a promising global strategy to turbocharge gene banks. Nat. Plants 2, 16150.
- 108. Zheng, Z., Hey, S., Jubery, T., Liu, H., Yang, Y., Coffey, L., Miao, C., Sigmon, B., Schnable, J.C., Hochholdinger, F., et al. (2020). Shared genetic control of root system architecture between Zea mays and Sorghum bicolor. Plant Physiol. 182, 977–991.
- 109. Miao, C., Pages, A., Xu, Z., Rodene, E., Yang, J., and Schnable, J.C. (2020). Semantic segmentation of sorghum using hyperspectral data identifies genetic associations. Plant Phenomics, 4216373.
- Singh, A., Ganapathysubramanian, B., Singh, A.K., and Sarkar, S. (2016). Machine learning for high-throughput stress phenotyping in plants. Trends Plant Sci. 21, 110–124.
- 111. Singh, A.K., Ganapathysubramanian, B., Sarkar, S., and Singh, A. (2018). Deep learning for plant stress phenotyping: Trends and future perspectives. Trends Plant Sci. 23, 883–898.
- 112. Araus, J.L., Kefauver, S.C., Zaman-Allah, M., Olsen, M.S., and Cairns, J.E. (2018). Translating high-throughput phenotyping into genetic gain. Trends Plant Sci. 23, 451–466.
- 113. Liang, Z., Pandey, P., Stoerger, V., Xu, Y., Qiu, Y., Ge, Y., and Schnable, J.C. (2018). Conventional and hyperspectral time-series imaging of maize lines widely used in field trials. Gigascience 7, 1–11.
- 114. Gage, J.L., de Leon, N., and Clayton, M.K. (2018). Comparing genomewide association study results from different measurements of an underlying phenotype. G3 (Bethesda) 8, 3715–3722.
- 115. Perez, L., and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. arXiv, 1712.04621.



- 116. Araus, J.L., and Cairns, J.E. (2014). Field high-throughput phenotyping: the new crop breeding frontier. Trends Plant Sci. *19*, 52–61.
- 117. Jiang, Y., and Li, C. (2020). Convolutional neural networks for imagebased high-throughput plant phenotyping: a review. Plant Phenomics 2020, 4152816.
- 118. Miao, C., Xu, Y., Liu, S., Schnable, P.S., and Schnable, J.C. (2020). Increased power and accuracy of causal locus identification in time series genome-wide association in sorghum. Plant Physiol. 183, 1898–1909.
- 119. Yang, C.J., Samayoa, L.F., Bradbury, P.J., Olukolu, B.A., Xue, W., York, A.M., Tuholski, M.R., Wang, W., Daskalska, L.L., Neumeyer, M.A., et al. (2019). The genetic architecture of teosinte catalyzed and constrained maize domestication. Proc. Natl. Acad. Sci. U S A *116*, 5643–5652.
- Allard, R.W., and Bradshaw, A.D. (1964). Implications of genotype-environmental interactions in applied plant breeding. Crop Sci. 4, 503–508.
- 121. Bao, Y., Zarecor, S., Shah, D., Tuel, T., Campbell, D.A., Chapman, A.V.E., Imberti, D., Kiekhaefer, D., Imberti, H., Lübberstedt, T., et al. (2019). Assessing plant performance in the Enviratron. Plant Methods 15, 117.

- **122.** Bernardo, R. (2010). Breeding for Quantitative Traits in Plants, Second Edition (Stemma Press).
- 123. Varshney, R.K., Singh, V.K., Hickey, J.M., Xun, X., Marshall, D.F., Wang, J., Edwards, D., and Ribaut, J.-M. (2016). Analytical and decision support tools for genomics-assisted breeding. Trends Plant Sci. 21, 354–363.
- 124. Khoury, C.K., Bjorkman, A.D., Dempewolf, H., Ramirez-Villegas, J., Guarino, L., Jarvis, A., Rieseberg, L.H., and Struik, P.C. (2014). Increasing homogeneity in global food supplies and the implications for food security. Proc. Natl. Acad. Sci. U S A *111*, 4001–4006.
- 125. Renard, D., and Tilman, D. (2019). National food production stabilized by crop diversity. Nature 571, 257–260.
- 126. Baro, M., and Deubel, T.F. (2006). Persistent hunger: perspectives on vulnerability, famine, and food security in Sub-Saharan Africa. Annu. Rev. Anthropol. 35, 521–538.
- 127. Maxwell, D., Khalif, A., Hailey, P., and Checchi, F. (2020). Determining famine: multi-dimensional analysis for the twenty-first century. Food Policy 92, 101832.