

1 **Problems with Estimating Anthesis Phenology Parameters in *Zea mays*: Consequences for Combining**
2 **Ecophysiological Models with Genetics**

3 Abhishes Lamsal^a, Stephen M. Welch^a, Jeffrey W. White^b, Kelly R. Thorp^b, and Nora Bello^c

4 ^aDepartment of Agronomy, Kansas State University, 2104 Throckmorton Plant Science Center, Manhattan, Ks
5 66502, USA

6 ^bUSDA-ARS Arid-Land Agricultural Research Center, 21881 North Cardon Ln, Maricopa, Az 85138, USA

7 ^cDepartment of Statistics, Kansas State University, 002 Dickens hall, Manhattan, Ks, 66502, USA

8 Corresponding Author: Abhishes Lamsal, abhilam@ksu.edu

9

10 **Abstract**

11 Ecophysiological crop models encode intra-species behaviors using constant parameters that are presumed to
12 summarize genotypic properties. Accurate estimation of these parameters is crucial because much recent work
13 has sought to link them to genotypes. The original goal of this study was to fit the anthesis date component of
14 the CERES-Maize model to 5266 genetic lines grown at 11 site-years and genetically map the resulting
15 parameter estimates. Although the resulting estimates had high predictive quality, numerous artifacts emerged
16 during estimation. The first arose in situations where the model was unable to express the observed data for
17 many lines, which ended up sharing the same parameter value. In the second (2254 lines), the model
18 reproduced the data but there were often many parameter sets that did so equally well (equifinality). These
19 artifacts made genetic mapping impossible, thus, revealing cautionary insights regarding a major current
20 paradigm for linking process based models to genetics.

21 **Introduction**

22 In the opening sentences of the 1968 book, *The Population Bomb*, Paul Ehrlich (and his wife Anne,
23 uncredited at publisher behest) wrote, “The battle to feed all of humanity is over. In the 1970s hundreds of
24 millions of people will starve to death in spite of any crash programs embarked upon now” and, in a subsequent
25 chapter, “I don't see how India could possibly feed two hundred million more people by 1980.” Fortunately,
26 research started in Mexico, India and elsewhere by Norman Borlaug before 1968 created high yielding dwarf
27 wheat varieties that, worldwide, are credited with averting one billion deaths from famine. India also
28 introduced IR8, the so-called “miracle rice” developed at the International Rice Research Institute in the
29 Philippines and the predicted human catastrophe was averted.

30 Nearly 50 years later, the specter of global disruption is again upon us. The challenges today are not
31 only increasing human population (which has doubled since 1970) but emerging concerns like climate change

32 and declining water resources. The confluence of these manifold trends makes finding ways to feed nine billion
33 people by 2050 one of the most pressing issues of our time [1]. However, the annual percentage increase rates
34 for crop yields are only half those required to meet that goal [2].

35 Beginning over 20 years ago, a paradigm has emerged offering the promise of dramatically accelerating
36 breeding programs via improved phenotype prediction of prospective crop genotypes in novel, time-varying
37 environments subject to sophisticated management practices [3–7]. The basic notion has two parts. The first is
38 to exploit ecophysiological crop models (ECM's) to describe the intricate, dynamic, and environmentally
39 responsive biological mechanisms that determine crop growth and development on daily or even hourly time
40 scales. The aim is to use highly detailed, nonlinear simulation models to predict the phenotypes of interest
41 within a subsample of possible environments and in-field management options. ECMs, whose origin is often
42 credited to Wit. (1965), encode intra-species behavioral differences in terms of parameters that are intended to
43 summarize genotypic properties. On the strength of that presumption, the constants are termed *genotype-*
44 *specific parameters* (GSP's).

45 The second part of the paradigm is to use quantitative genetic methods such as genomic prediction [9]
46 to relate the GSP's to genotypic markers [3]. Next, the outcomes of crosses are estimated by (1) calculating the
47 GSP values that would arise from possible offspring genotypes. These values are then (2) used in
48 ecophysiological model runs to predict the phenotypes in the target population of environments (for which
49 detailed descriptive data must be available). In simplified instances, this approach has seen remarkable success
50 (e.g., [10]).

51 Composed of large coupled sets of continuous-time differential equations, ecophysiological models
52 simulate many interacting processes [11,12] operating in the soil-plant-atmosphere continuum. These
53 processes include physiology (e.g., photosynthesis, respiration, resource partitioning to various plant parts, and
54 growth), phenology (leaf emergent timing, the date of vegetative-to-reproductive development, etc.), as well as

55 chemistry and physics (soil water flows, chemical transformations, energy fluxes, gas exchange, etc.). During
56 simulation runs, model formulas compute instantaneous process rates based on plant status and environmental
57 conditions at each time point. These rates are integrated (*sensu* calculus) to output time series of dozens of
58 plant variables. The models typically have 10 to 20 GSP's whose estimates are read from input files at the start
59 of model execution. Numerous other inputs (e.g. soil water holding capacities by layer; measured daily solar
60 radiation, rainfall, maximum and minimum temperatures; etc.) further quantify the physical environment.

61 The lynchpin of the two-step paradigm is the accurate estimation of the GSP's so that these can be
62 related to allelic states of the individual lines. Unfortunately, the direct measurement of GSP's is so time- and
63 resource-demanding as to be infeasible for large numbers of lines. Indirect GSP estimation via model inversion
64 is also challenging because easily-measured plant phenotypes exhibit strong interactions with the environment
65 [13] thus increasing data requirements by necessitating trait measurement in multiple settings [14]. Even so,
66 ecophysiological crop models enjoy extensive global use in areas ranging from global climate change, policy
67 analysis, crop management, etc. Indeed, a Google search on the abbreviations of just two major model systems
68 [namely "DSSAT" [15] and "APSIM" [16]] returned 134,000 hits. Not surprisingly, there is an extensive literature
69 (reviewed briefly below) on ecophysiological model parameter estimation.

70 Initially, the authors' intent was to apply the two-step method to anthesis date using data from over
71 5000 lines comprising the maize nested association mapping population (NAM) [17], which was developed
72 specifically to enable high-resolution studies of trait genetic architectures. Not only is anthesis date a
73 phenotype of major biological significance, but it was also studied in this same panel using conventional
74 statistical genetic methods [18,19]. Our hypothesis was that applying the proposed 2-step paradigm would
75 demonstrate its merit in the specific context of the large data sets increasingly used in crop breeding programs
76 to interrelate genotypes and phenotypes. Contrasting the results of the standard and ecophysiological
77 approaches was expected to be interesting and informative. Granted, the model fitting methods to be used

78 were not novel, but we expected that a further demonstration of their value with data sets much larger than
79 ever used before would have utility.

80 However, something quite different happened. We discovered modeling issues and estimation artifacts
81 that are of sufficient severity and generality that, if not addressed, are likely to imperil the breeding
82 acceleration paradigm. Therefore, the objectives of this paper were 1) to describe these problems and the
83 methods that revealed them (which can be applied as detection tools in studies of other traits) and 2) to discuss
84 research directions that might ameliorate the problems.

85 **Background**

86 Numerous optimization methods have been used to estimate parameters for ECM's. Surprisingly,
87 perhaps the most common approach has been that of trial and error [20], wherein different parameters values
88 are manually tested until an acceptable match between simulated and observed data is found. This approach,
89 of course, becomes highly inefficient as the number of model parameter increases. Thus, numerous off-the-
90 shelf, automated optimization techniques have been developed. Examples include the simplex method [21],
91 simulated annealing [22,23], sequential search software (GENCALC) [24], Uniform Covering by Probabilistic
92 Region (UCPR) [25], particle swarm optimization (PSO) [26], and generalized likelihood uncertainty estimation
93 (GLUE) [27]. While these traditional optimization techniques have advantages, they can be inefficient in terms
94 of runtime and are highly dependent on optimization settings when thousands of combinations of line \times
95 planting site-years are involved – a situation that is becoming common in the era of massive genetic mapping
96 populations. The fundamental issue is that, as the number of lines and environments increases, estimating
97 GSP's for each line independently usually involves highly redundant simulation. To this end, we adapted an
98 algorithm pioneered by Welch et al. (2000) and Irmak et al. (2000), as described in methods section. The
99 approach exhibits particular efficiencies when individual plantings incorporate large numbers of lines and,
100 serendipitously, supports a close examination of the estimation process, itself.

101 The vast majority of prior ECM parameter estimation studies have been conducted in non-genetic
102 contexts. Against these backgrounds, the sole merit criterion has been the predictive skill demonstrated by the
103 GSP estimates obtained. However, the current setting, however, is markedly different. GSP's are not just inputs
104 to ecophysiological crop models; GSP's simultaneously function as the outputs (i.e. dependent) variables of
105 genetic prediction models. As such, GSP's are at least as closely related to tangible biochemical processes at
106 the molecular level as they are summative of physiological properties (e.g. maximum photosynthetic rates) in
107 higher organizational realms. Therefore, a deeper inspection of their estimation is warranted and two concepts
108 are helpful in achieving the enhanced discernment now required.

109 We employ the term “expressivity” (and the adjective “expressive”) to describe a model's innate ability
110 to reproduce a set of observations independent of particular parameter values. An expressive model may fail
111 to replicate data because an unskilled optimizer cannot find a meritorious combination of parameter values. In
112 contrast, a model with low expressivity will fail to fully mimic actual data irrespective of what (biologically or
113 physically reasonable) values are assigned to its parameters. In cases where the latter behavior is detected,
114 remedies will be vigorously sought. However, as shown below, however, systematic gaps in expressivity can
115 coexist even within an overall framework of predictively skilled model performance.

116 Another model property that has received little attention in previous estimation studies is equifinality.
117 Equifinality describes a situation in which multiple sets of parameter values generate identical model
118 predictions. In statistics, a synonym for “equifinality” is “parameter non-identifiability” [30]31. When the
119 only concern is prediction quality and that seems “good enough”, it is easy to consider equifinality a non-
120 problem. However, when parameters are intermediaries rather than just inputs and equifinality exists, it begs
121 the question as to what relationship, if any, putative GSP estimates might bear to allelic states across the
122 genotype? A moment's reflection shows that equifinality and expressivity are different model properties. The
123 former relates to how many different estimates yield identical predictions; the latter refers to the possible
124 existence of systematic failures of those predictions to mimic observed data.

125 In this paper, we explore these issues in modeling and estimation using the anthesis phenology
126 component of the CERES-Maize ECM [32–34] and observed dates from multiple plantings of three maize
127 genetics panels totaling nearly 5300 lines. Anthesis initiates the period of grain development and is therefore a
128 critical milestone toward grain yield. As such, it mediates the adaptation of the crop to its environment by
129 determining the relative length of the vegetative and reproductive growth phases and is a key target of
130 breeding programs [18]. (Although at the apical meristem, floral initiation precedes the visible morphological
131 change of anthesis, the linkage between the two is tight enough that we follow common modeling practice and
132 consider them as effectively synonymous.) The genetics of flowering time has been intensively studied in the
133 model plant *Arabidopsis thaliana* where well over 100 influential genes are now known [35]. Indeed, gene
134 expression models of flowering time of *A. thaliana* based on differential equations have been developed [36],
135 and genetically-informed approaches have established the relationships between network-level function and
136 common ecophysiological time formulations [37]. In maize, our understanding of the genetic control on
137 flowering time is more limited but has been advancing in recent years. More than 30 genes have been
138 described and conservation of key features from *A. thaliana* seems apparent (Table 1 in [38]). A quantitative
139 gene network model based on a number of these loci has been published [38].

140 The general desire within applied quantitative genetics to probe genetic architectures has led to the
141 construction of ever-larger and/or special purpose mapping populations [18]. The maize NAM panel [17] was
142 constructed by making bi-parental crosses between one common parent, B73, and each of a set of 25 other
143 inbreds that collectively encompassed a wide range of maize diversity. Approximately 200 offspring from each
144 of these 25 crosses were then inbred for a number of generations to ensure, to the greatest degree feasible,
145 that the influence of each locus on any trait of interest reflected the contribution of one parent only. Individual
146 plant genotypes produced in this fashion are called “recombinant inbred lines” (RIL’s). Buckler et al. (2009)
147 reported a seminal study of maize anthesis dates using this NAM panel. Demonstrating the power of these
148 lines to finely dissect genetic contributions to traits of interest, they identified 36-39 QTL, where the exact

149 number depended on the analysis method used. Most of loci had small effects but collectively, they explained
150 89% of total variation in anthesis date.

151 For the reasons outlined above, accurate prediction of anthesis date is a major target for
152 ecophysiological crop models [25]. However, few studies exist have used large data sets for ECM calibration.
153 Mavromatis et al. (2002) reported 5,109 site-year-line-parameter combinations and Welch et al. (2002)
154 estimated 4,620 site-year-line-parameters. The effort presented herein encompassed 197,964 site-year-line-
155 parameter combinations – to our knowledge, the largest such study ever reported. As the following sections
156 document, it was the sheer scale of this data set and the resulting scatterplots depicting thousands of lines that
157 revealed worrisome issues of equifinality and expressivity that might be overlooked in studies of smaller scale.

158 **Materials and Methods**

159 **Experimental data**

160 Observations collected on anthesis date for a total of 5266 maize lines were obtained from the Panzea
161 data repository (<http://www.panzea.org>). The lines used were members of three genetic panels. In particular,
162 4785 lines were from the 25 RIL panels comprising the maize NAM set described above. Also included were an
163 additional 200 RIL lines commonly referred to as the IBM panel because they originated by Intermating B73 ×
164 Mo17 [40]. Finally, a maize diversity panel [41] contributed data on 281 additional lines. Various combinations
165 of these lines were grown at six US sites: New York (NY), North Carolina (NC), Illinois (IL), Missouri (MO), Florida
166 (FL) and Puerto Rico (PR), during 2006 and 2007 for a total of eleven site-years. In what follows “NY6” denotes
167 the 2006 planting in New York, respectively by state abbreviation and year for other site-years. Table 1 gives
168 the exact locations of the experimental sites, and the respective sowing dates. The “Total Lines” row of the
169 table gives the number of lines from the three panels that were present in each study. The “Lines with data”
170 row lists the number of lines with available observations on anthesis date. Data on daily maximum and
171 minimum temperatures for each site were provided by the maize NAM collaborators (H. Hung, personal

172 communication, 2010) and did not included metadata on position of the weather stations to the field plots,
173 types and calibration of sensors or types of radiation shields used.

174 **Table 1. Sowing dates, geographical coordinates, total number of lines planted and number of lines for which**
 175 **anthesis dates were observed for all site-year combinations used in this study.**

	NY6	NY7	NC6	NC7	MO6	MO7	IL6	IL7	FL6	FL7	PR6
Sowing Date (DOY)	128	135	122	120	137	138	128	137	265	280	314
Latitude (deg)	42.73	42.73	35.67	35.67	38.89	38.89	40.08	40.08	25.51	25.51	18.00
Longitude (deg)	-76.66	-76.66	-78.49	-78.49	-92.23	-92.23	-88.2	-88.2	-80.49	-80.49	-66.51
Number of total lines sown	5478	5478	5478	5478	5478	5478	5478	5478	5026	3753	5131
Number of lines with data	4743	5236	5236	5160	3261	2555	5036	5178	4943	3742	4401

176 **CERES-Maize model**

177 The Crop Estimation through Resource and Environment Synthesis (CERES)-Maize model is one of the
 178 oldest, most widely used ecophysiological crop models for maize [42]. We used the CERES-Maize version
 179 incorporated in CSM (Cropping System Model) 4.5 ([11,15]. The CERES-Maize simulation of development
 180 toward anthesis is controlled by a set of GSP's and environmental inputs [33,34]. Specifically, the GSP's studied
 181 herein were thermal time from emergence to juvenile phase (P1), critical photoperiod (P2O), sensitivity to
 182 photoperiods longer than P2O (P2), and the phyllochron interval (PHINT) as measured in thermal time. The
 183 duration of Stage 1, the interval from emergence through the end of the juvenile phase, is calculated by

184 accumulating daily thermal time until P1 is reached. Stage 2 follows immediately and lasts until tassel initiation.
185 Stage 2 lasts a minimum of four days when the photoperiod (including civil twilight) is less than P20. P2
186 specifies the number of extra days required for every hour by which the photoperiod exceeds P20. The model
187 continues to accumulate thermal time through Stage 2. The model assumes that (1) there are five embryonic
188 leaves; (2) two new leaves initiate during each phyllochron interval; and (3) that anthesis date, which
189 terminates Stage 3, occurs when all leaves present at the end of Stage 2 (i.e., total leaf number, TOLN) are fully
190 expanded. The date on which this happens is when the ongoing thermal time accumulation reaches $TOLN \times$
191 PHINT.

192 Thermal time is calculated from inputs of daily maximum and minimum temperatures. Sowing dates
193 (Table 1) determined the time series of weather data that control simulated plant growth and development.
194 The model calculated daily photoperiods from geographic position. Other required model inputs did not affect
195 predicted anthesis dates and were not considered here. For example, the soil water and nutrient balance
196 components of the model do not affect simulated anthesis date in the CERES-Maize model and therefore were
197 not used in this study. The model also requires row spacing and planting depth, which were set to 0.5 m and
198 2.5 cm, respectively. No tillage, pest, or disease effects were simulated.

199 **Parameter estimation**

200 **Search strategy**

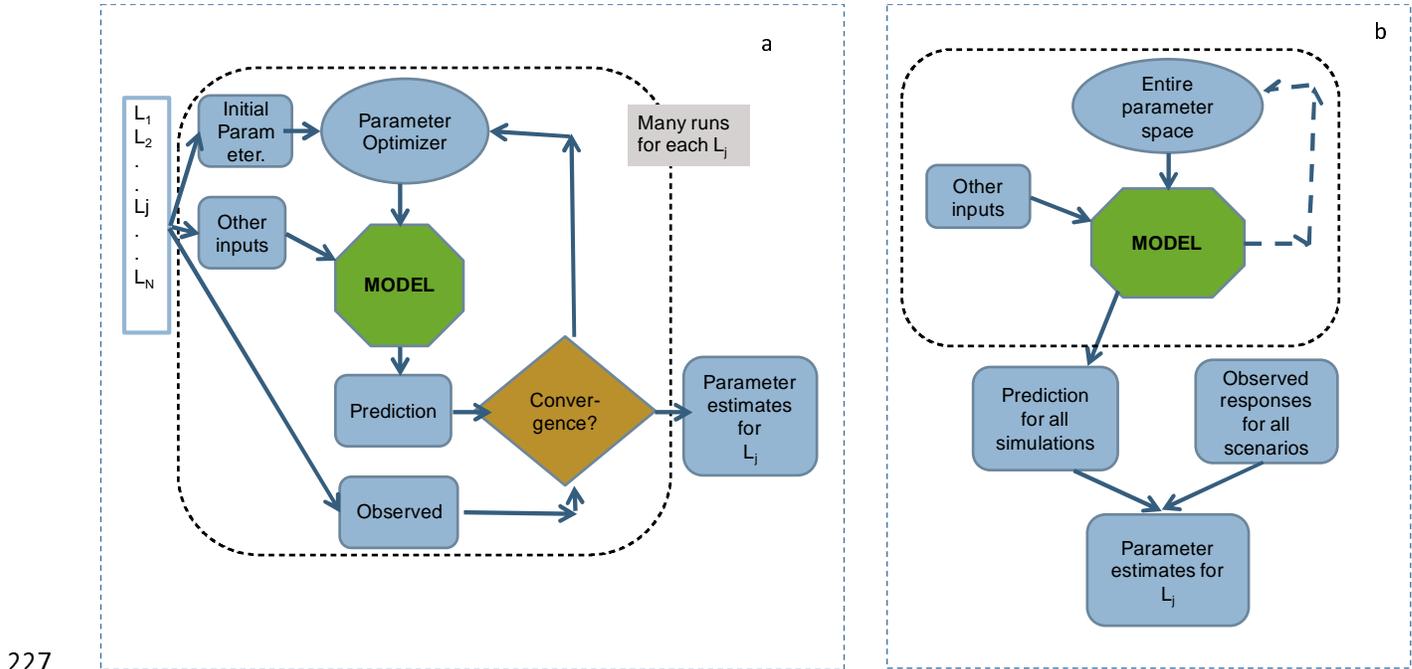
201 In the conventional approach to parameter estimation (Fig 1a), an optimizer iterates through a series of
202 trial solutions for which model predictions are generated in each environment. The entire process is repeated
203 for each line. This approach becomes inefficient when many lines are planted together in large experiments
204 and are therefore exposed to identical environments. This is because estimates approaching optimal goodness-
205 of-fit will only emerge in the latter stages of an iterative optimization run. Therefore, the majority of early

206 iterations for each line entail the repeated evaluation of estimates with mediocre predictive ability in the same
207 environment.

208 To overcome this problem, we adapted an approach described by Irmak et al. (2000) and Welch et al.
209 (2002, 2000). In their scheme (Fig 1b), model simulations were conducted for each planting across a
210 multidimensional grid of parameter value combinations. The resulting predictions were stored in a database.
211 As a second step, for each line the root mean square error objective function (RMSE; [43]) between observed
212 and predicted anthesis day of year was evaluated with respect to all combinations of parameter values across
213 all site-years. That is, for line l ,

$$214 \quad RMSE_l = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_p - Y_o)^2} \quad (1)$$

215 where, n is the number of observations for that line (consisting of one observation per site-year
216 combination), and Y_p (Y_o) is the predicted (observed) anthesis date. The optimizer goal was to minimize the
217 RMSE for each line. If a unique minimum existed, it defined the combination of GSP values that best fit each
218 line. Total computational time was reduced because time-consuming model simulations for each combination
219 of GSP parameter values were only performed once, but those outputs were reused many times in the much
220 faster RMSE calculations. Another benefit is that a combination of GSP values that yielded poor predictability
221 for one variety might perform better for a different line. Additionally, this process ensured that identical
222 parameter combinations were tested for each line, which can aid in comparing the results achieved. Finally,
223 simply by retabulating the database, any number of different optimizations could be performed using different
224 observations, alternative subsets of site-years plantings or combinations of parameter values. The use of
225 alternative objective functions is also possible without requiring additional simulations. Because of the central
226 role played by the database of simulation outputs, we will refer to this scheme as the *database method*.

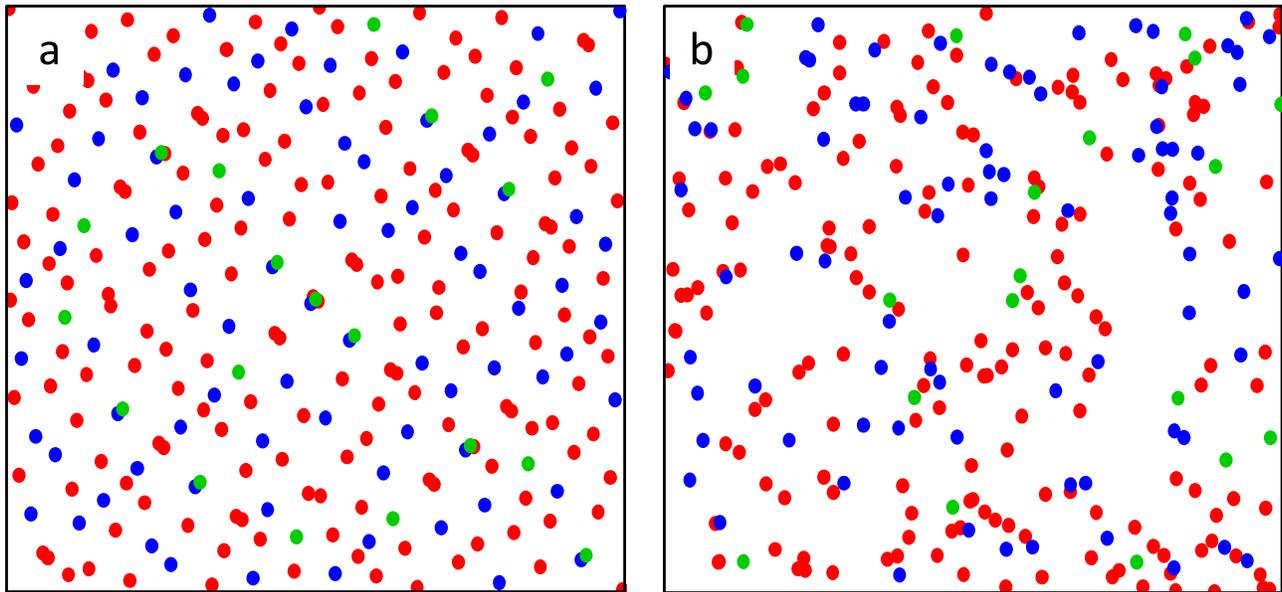


228 **Fig 1. Parameter search strategies.** (a) Conventional method (b) Database method. $L_{1...N}$ is the number of lines.

229 Sampling the model parameter space with sobol sequences

230 Unlike Irmak et al. (2000) and Welch et al. (2002, 2000) who sampled the parameter space with a
231 rectilinear grid, we employed Sobol sequences so as to avoid the combinatorial explosion in computational
232 requirements that accompany increasing dimensionality. Sobol sequences belong to a family of quasi-random
233 processes designed to generate samples of multiple parameters dispersed as uniformly as possible over the
234 multi-dimensional parameter space [44]. Sobol sequences are specifically designed to generate samples with
235 low discrepancy – that is, a minimal deviation from equal spacing. Unlike random numbers, quasi-random
236 algorithms can effectively identify the position of previously sampled points and fill the gaps between them
237 [45], thus avoiding the formation of clusters. Further, Sobol sequences offer reduced spatial variation compared
238 to other sampling methods (e.g., random, stratified, Latin hypercube; see Fig 2a vs. 2b), make this method more
239 robust [46]. We used a Python-based algorithm to generate a Sobol sequence of quasi-random numbers for
240 calculating 32,400,070 sets of the four CERES-Maize GSP's, leading to a uniformly-sampled four-dimensional

241 parameter space for P1, P2, P2O, and PHINT. To construct the database, CERES-Maize calculated anthesis date
242 for each GSP combination in each of the 11 site-years – a total of 356,400,770 model runs. Table 2 describes the
243 upper and lower bounds and the number of distinct values obtained for each parameter.



244

245 **Fig 2. Quasi-random points from two-dimensional Sobol sequence.** (a) The first 275 quasi-random
246 points from a two-dimensional Sobol sequence. (b) The first 275 points produced by the commonly used
247 Mersenne twister pseudo-random number generator [47]. The Sobol sequence covers the space more
248 evenly. The first 20 points are green, the next 80 are blue, and the final 175 are red, thus demonstrating
249 Sobol gap filling.

250

251 **Table 2. Parameter ranges used in generating sobol sequence.**

Parameter	Definition	Unit	Min	Max	No. of unique values
P1	Thermal time from seedling emergence to end of juvenile phase	GDD (°C)	150	450	30,001
P20	Critical photoperiod hour	hrs.	10	14	401
P2	Days of anthesis date delay for each hour by which the day length exceeds P20	rate	0	2	20,001
PHINT	Phylochron interval (Interval between successive leaf tip appearances)	GDD (°C)	25	70	45001

252

253 High performance computing

254 The number of model runs was too large for lab-scale computing facilities, so we used the “Stampede”
 255 supercomputer at the Texas Advanced Computing Center (TACC) [46]. *In toto*, the CERES-Maize runs required
 256 63,372 CPU-hours, which equates to ca. 176 simulations per second distributed across 112 processors. The
 257 predicted anthesis dates were collated and transferred to the “BeoCat” computing cluster at Kansas State
 258 University (https://support.beocat.ksu.edu/BeocatDocs/index.php/Compute_Nodes). There, RMSE values were
 259 tabulated for each line × parameter value combination across all site-years in which anthesis date was
 260 observed. As combinations of GSP values were found that had progressively lower RMSE values, they were

261 recorded by the computer. This process required ca. 15 minutes of wall clock time per line so the total
262 estimation process was completed in ca. 7 h on 200 Xeon E5-2690 cores.

263 **Assessing estimate properties**

264 **Equifinality**

265 Equifinality occurs when multiple combinations of parameter estimates generate the same minimal
266 RMSE value, often because they generate identical model predictions [30], in this case identical integer DOY
267 values for anthesis dates. We quantified "equifinality" by defining "number of ties" as the number of Sobol sets
268 of parameter combinations that produced the same optimal RMSE values, minus one. No equifinality is present
269 in a line if there is only one combination of parameter values that minimizes the RMSE. That is, there are zero
270 ties among its estimates. To illustrate the magnitude of the problem and our motivation to study it more
271 closely, we note that 2254 (43%) of the 5266 lines available in the data exhibited equifinality. The worst case
272 was represented by a line that had 1,043,933 distinct combinations of GSP values that produced identical
273 anthesis date predictions, and thus the same RMSE, thereby yielding 1,043,932 ties.

274 During the database tabulation phase, the values of the "best combination of parameter estimates seen
275 so far" was updated only if its RMSE value was strictly better than all previously evaluated ones. So, when
276 equifinality was present, the final GSP estimate was the first combination of parameter values encountered that
277 had a minimal RMSE value. As a result, some of the analyses described below are sensitive to equifinality,
278 illustrating the fact that subtle optimizer algorithm idiosyncrasies can have marked impacts on the overall
279 results. Such cases are noted explicitly along with the procedures used to mitigate the effects.

280 **Interrelationships between parameter estimates**

281 Correlations and other relations among parameter estimates are highly important to breeding programs
282 and related simulation studies. When correlations between parameter estimates are present, opportunities
283 exist to select on one plant trait by selecting on a related phenotype instead. Additionally, there have been a
284 number of *in silico* studies where CERES models were used to design crop ideotypes [48,49]. Such efforts find
285 combinations of model parameter values that predict phenotypes well suited to the target population of
286 environments. Once identified, lines with those values become breeding targets. However, a potential pitfall
287 arises if realizing the desired genotype involves changing parameter values in directions contrary to the
288 correlations that exist between them.

289 For this reason, we explored the pairwise correlation structure of the GSP parameter estimates and
290 generated pairwise scatter plots of their line-specific values. However, the latter revealed a bizarre pattern, the
291 diagnosis of which ultimately led us to the second problem alluded to in the introduction – the inability of the
292 model to reproduce certain observational combinations – and to the methods presented next.

293 **Model expressivity**

294 A common graphical method to assess the quality of model fit is to plot the predicted vs. observed
295 values (e.g., Fig 3). Such scatterplots can be informative in detecting areas of mismatch between observed and
296 predicted values, thus providing specific characterization of the model's lack of fit. By definition, each point in
297 the scatterplot corresponds to a prediction that a model is able to make given an optimized set of parameter
298 values. However, an entirely different question is whether there are observations that a given model cannot
299 reproduce using *any* reasonable combination of parameter values? That is, one might seek to assess whether a
300 given model has the requisite expressivity to reproduce the data.

301 The database approach allows such a question to be addressed using what we term *phenotype space*
302 scatter plots. In such plots, each axis corresponds to a different site-year. The coordinates along the axes
303 represent the observed or predicted anthesis dates for each site-year. Model expressivity is then assessed by

304 comparing the scatter of predicted anthesis date generated from a wide range of GSP value combinations to
305 the scatter of observed values in large data sets. Because equifinality does not affect predictions, this method
306 of evaluating model expressivity is independent of the order in which an optimizer locates points that minimize
307 RMSE values (see the second paragraph in section “Equifinality”).

308 **Testing for parameter stability across environments**

309 In order for the two-step paradigm outlined in the Introduction to work, the estimates of GSP’s should
310 not vary across the set of environments used to estimate them, a property called “stability” [4]. If GSP
311 estimates did vary across environments, there would be no way to tell what GSP values to input to the
312 ecophysiological model to predict traits whenever daily weather time series or soils differed from those used in
313 the paradigm’s first step. This might seem an insuperable barrier to readers for whom G×E interactions are
314 virtually ubiquitous among quantitative plant phenotypes, but it is not. This is because the *raison d’être* of
315 models like CERES-Maize is to explain crop variety × environment interactions mechanistically based on
316 physiological principles.

317 Many GSP’s, including the ones in this study, explicitly relate plant behaviors (e.g., development toward
318 anthesis) to environmental variables (e.g., temperature and photoperiod in the current case). Modelers assert
319 that GSP’s are properties of the individual lines (i.e., stable) and, therefore, by implication, have a genetic basis
320 because genotypes do not change with the environment. Over time, it is thus expected that research will
321 mechanistically link at least some GSP’s to molecular genetic processes. For example, both short (P2O) and long
322 day critical photoperiods are determined by the dynamics of the *CONSTANS* protein in a range of plants
323 including *Arabidopsis* [50] and a number of grasses [4], albeit not maize [51]. In rice (*Oryza sativa*), critical short
324 day length has even been successfully predicted from a differential equation model of the diurnal expression
325 patterns of the *CONSTANS* ortholog [52].

326 Because stability is both important and reasonable to expect given the goals of ecophysiological
327 modeling, it has been argued [5] that finding a putative GSP to be unstable is *prima facie* evidence of a problem.
328 Possible causes of instability include: (1) the model incompletely or incorrectly disentangles $G \times E$; (2) a stable
329 answer exists but the optimizer is insufficiently skilled to find it; (3) undiscovered equifinality is present, and the
330 solutions found depend on low-level algorithmic idiosyncrasies of the optimizer (e.g. method section
331 “equifinality”); and (4) unique best GSP estimates exist that the optimizer can find, but because the model is
332 over-parameterized, the values obtained reflect noise signals that differ between environments.

333 All sources of instability, whether these or others, are detrimental to the two-step ecophysiological
334 genetic approach to phenotype prediction. Thus, it is critical to know when parameter instability is present, so
335 herein we developed a statistical approach to detect and test for it. The specific question asked was "Do the
336 GSP estimates depend on the particular set of environments used to construct them?" A conceptually simple
337 way to answer this might be to (1) obtain a combination of parameter estimates from one subset of site-years,
338 (2) repeat the estimation with a different subset, and (3) test whether the two sets of parameter estimates
339 differ according to an appropriate statistical test.

340 A more general and robust approach, however, might be to obtain parameter estimates from many
341 site-year subsets chosen according to a principled method. Preliminary tabulations of the Sobol database
342 revealed that equifinality increased dramatically when fewer than seven site-years were used for estimation
343 (see Results). Therefore, the subset size was set to seven site-years. One method for selection of site-year
344 subsets might be to resample site-years with replacement. However, as shown by analogy in Fig 2b,
345 randomization adds a source of variability to the results that could be of concern given that sampling by
346 replacement would have $P_7^{11} = 39,916,800$ possible site-year subsets. Therefore, analogous to Fig 2b, we used
347 a combinatorics-based sampling pattern leading to more uniformly-distributed site-year subsets by taking all
348 combinations of 11 site-years 7 at a time, of which there are $C_7^{11} = 330$ possibilities. To maximize the amount

349 of data available for each line in any subset, we focused on the 539 lines for which observation were available in
350 all 11 site-years.

351 We then conducted 177,870 four-dimensional optimizations to obtain GSP parameter estimates for
352 each of the 539 line \times 330 site-year set combinations. These optimizations involved only Sobol database
353 retabulations rather than new model runs, again illustrating the computational efficiency of the database
354 approach. When forced to generate a single result, the database search returned the combination of GSP
355 estimates yielding a minimal RMSE that it happened to encounter first. To focus on the subset that lacked this
356 element of optimizer arbitrariness, we first dropped the 114,314 line \times site-year combinations that had ties (i.e.
357 more than one set of GSP estimates yielding the same RMSE). Because our primary interest was in the
358 variability that different site-year combinations might contribute to GSP estimates, we further restricted our
359 attention to the 297 site-year subsets that had at least 100 lines remaining after ties were removed. Each of
360 the 539 lines was present in at least 28 site-year subsets, which was deemed adequate for GSP estimation.
361 These actions left a total of 60,834 estimates for each of the four GSP's in the study. This became our base
362 group for analysis. We acknowledge that the estimates dropped share a common property (i.e., ties) that might
363 have systematic effects influencing the results. So, in addition to the base group just described above, we also
364 examined the set of (1) all 177,870 GSP sets and (2) the 114,314 results for which ties existed. In both cases we
365 used the optimizer-selected values

366 We then specified a statistical model to test for stability in parameter estimates across environmental
367 subsets, as follows:

$$368 \quad \rho_{l,e} = \mu_{\rho} + \alpha_l + \beta_e + \varepsilon_{l,e} \quad (2)$$

369 where $\rho_{l,e}$ represents an estimate of the GSP ρ (i.e. either P1, P2, P2O, or PHINT) for the l^{th} line ($l =$
370 1,2,... 539) obtained from the e^{th} site-year set ($e = 1,2,\dots 297$), μ is the intercept parameter, acting as an overall

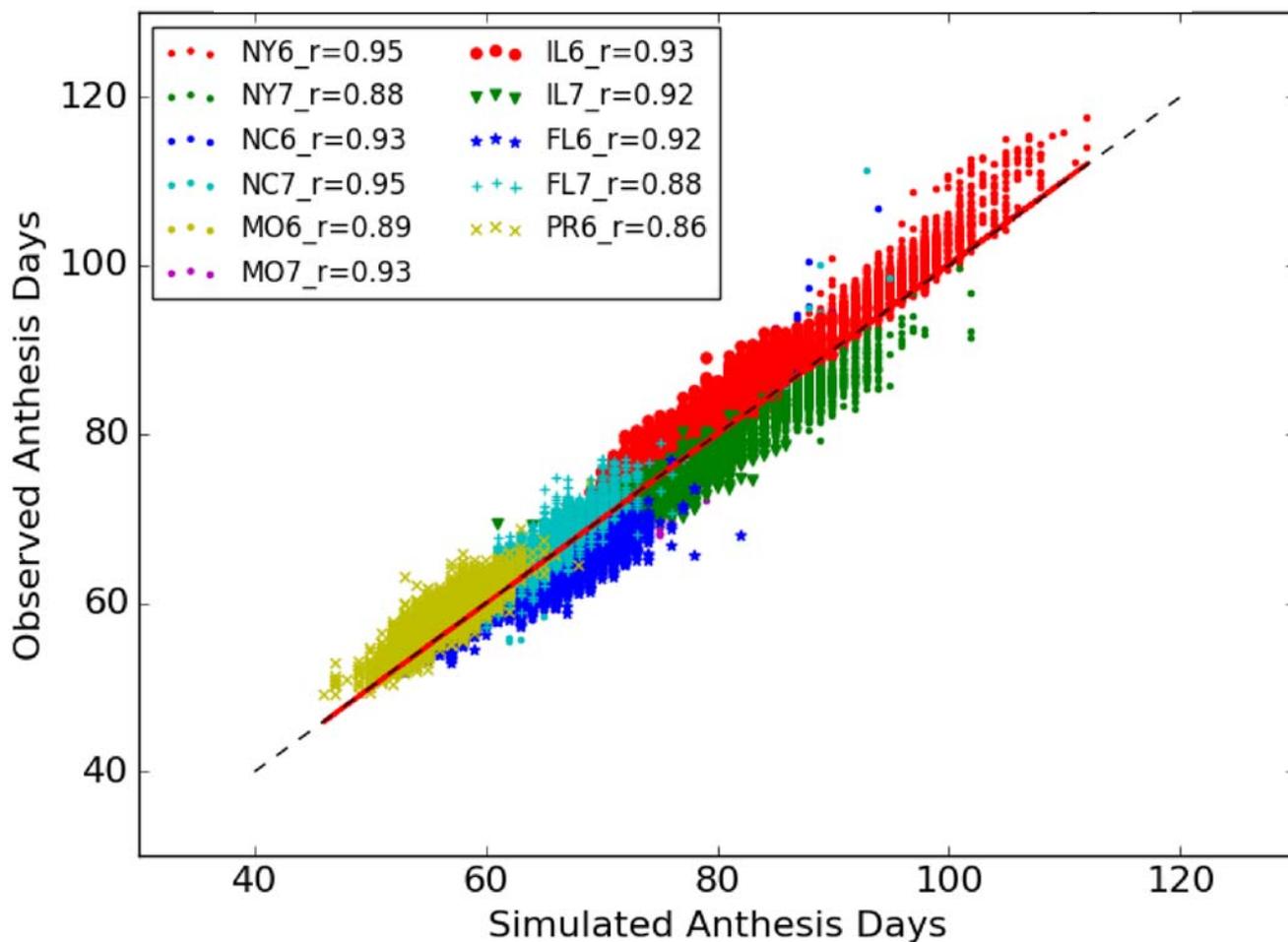
371 mean of GSP ρ across all lines and site-year subsets; α_l is the differential random effect of line l , assumed to
372 be distributed $\alpha_l \sim N(0, \sigma_l^2)$; β_e is the differential random effect of the e^{th} set of site-years, assumed to be
373 distributed $\beta_e \sim N(0, \sigma_e^2)$; and $\varepsilon_{l,e}$ is the left-over residual unique to the l, e^{th} observed GSP estimate and
374 assumed $\varepsilon_{l,e} \sim NIID(0, \sigma_\varepsilon^2)$. The differential line effects α_l are considered to be random, as is common in
375 field studies of plant population biology. Further, the differential effects of site-year sets, β_e , were treated as
376 random because the corresponding environmental sets are combinations of 7 out of 11 plantings considered to
377 be a representative, if not random, sample of the population of possible site-years to which we are interested in
378 inferring.

379 If the estimation of any GSP parameter ρ were stable across the site-year subsets, one would expect
380 the variance of β_e , namely σ_e^2 , to be zero; alternatively, if estimation is unstable, one would expect $\sigma_e^2 > 0$.
381 To test this hypothesis set, we fit two competing versions of the statistical model in equation (1), one with and
382 one without the random effect of site-year subsets β_e for each of the GSP's $\rho = \text{P1, P2, P2O, and PHINT}$.
383 For each GSP, we then compared the two competing models using a likelihood ratio test statistic against a
384 central chi-square distribution with half a degree of freedom to account for the fact that the test is being
385 conducted on the boundary of the parameter space. Statistical models were fitted using the liner mixed-
386 effects model package lmer in R [53] with optimization based on the log-likelihood option. The lmer package
387 also calculated the Akaike and Bayesian Information Criteria [AIC [54] and BIC [55], respectively], which allowed
388 for an additional assessment of fit for statistical models that included or excluded the random effects of site-
389 year subsets.

390 Results

391 **Observations vs. Predictions**

392 Fig 3 shows a color-coded scatterplot of observed vs. predicted days to anthesis for 49,491 line \times site-
393 year combinations; the cloud of points is concentrated along the identity line, therefore suggesting accurate
394 prediction; the overall estimated RMSE is 2.39 days. Also, there seem to be considerable differences between
395 sites on anthesis days, whereby Florida and Puerto-Rico show very short vegetative durations (ca. 50 d), which
396 are more than doubled in New York (120 d). Empirical correlation coefficients (\hat{r}) were high across site-years
397 and ranged from 0.86 to 0.95, thus indicating an overall responsiveness across lines to the range of site-year
398 conditions on anthesis dates. The standard deviations of the predicted values and their corresponding
399 observations are 10.336 and 10.639, respectively, which, with the overall empirical correlation coefficient of
400 0.974, account for a close to 1-to-1 estimated regression slope of observations vs. predictions [i.e. $1.002 = ($
401 $10.639 / 10.336) * 0.974$], as per the established statistical identity between these four sample quantities [56].



402

403 **Fig 3. Predicted and Observed anthesis days of all 5,266 lines from 11 site-year combinations.** The graph has
404 49,491 points and an overall RMSE of 2.39 days.

405 Equifinality

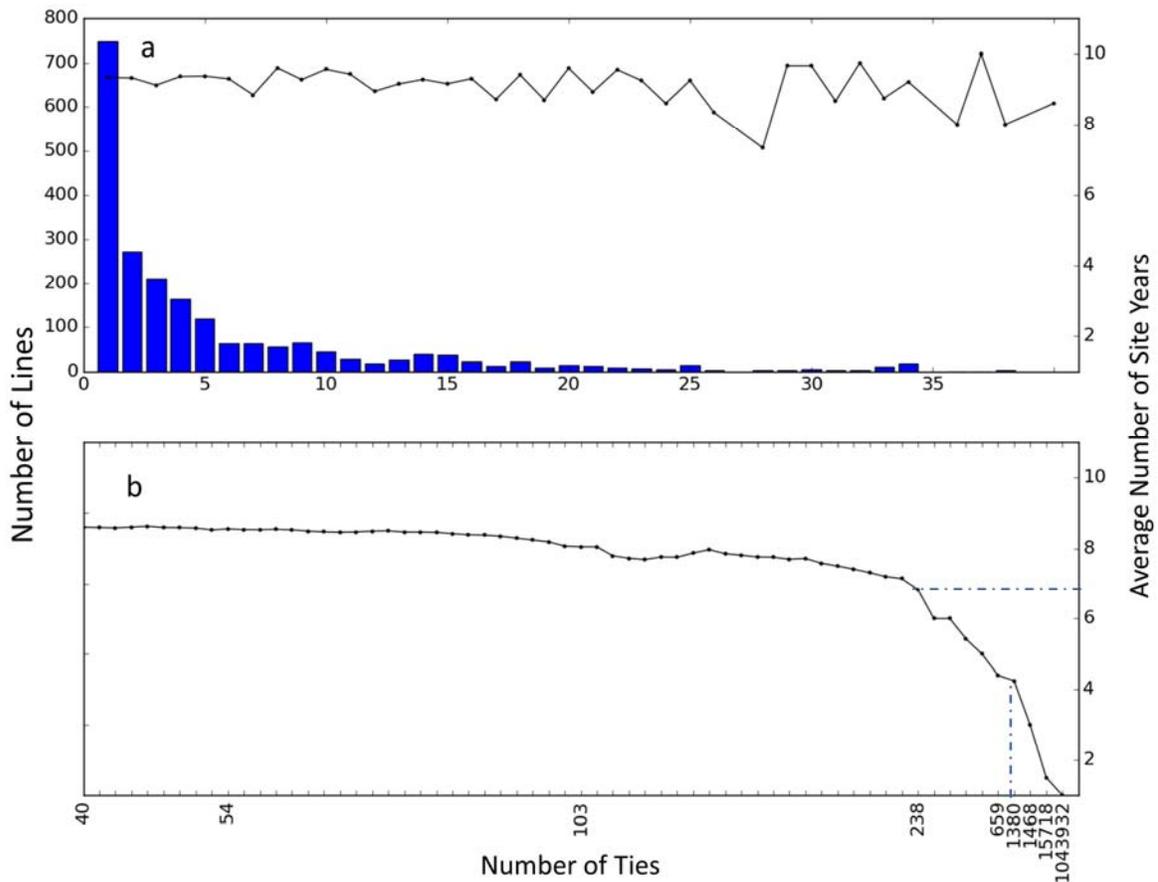
406 A more complex picture emerges when the prevalence of equifinality is considered. As noted in 3.4.1,
407 for the 2,254 lines exhibiting equifinality, the number of ties can exceed 1M. The histogram in Fig 4a tabulates
408 the frequency of ties across lines. There are 2,153 lines with fewer than or equal to 40 ties. The line trace along
409 the upper portion of the top and bottom panels shows the average number of site-years in each bin.

410 In Fig 4a, the empirical distribution of ties was right skewed, thereby indicating that a relatively large
411 number of maize lines had few ties and thus low levels of equifinality. This is particularly true when parameter
412 estimates were computed using data from 7 to 11 site-years (right axis of Fig 4b). Further, the distribution of
413 ties appears to have a very long tail to the right, whereby the number of lines with increasing amounts of
414 equifinality declines very slowly while the number of site-year combinations used for estimation seems to
415 plateau (Fig 4a). This pattern continues into Fig 4b, which shows the 101 lines with more than 40 ties. (No bars
416 are shown in Fig 4b due to scale of the y-axis, as each bin generally contains one to three lines.) Interestingly,
417 the number of ties, and thus equifinality, seems to increase precipitously for the 56 out of 5,266 lines that have
418 fewer than seven site-years of data (Fig 4b).

419 As the number of ties increases, one can expect that the range of indistinguishable estimates for any
420 GSP will widen. To illustrate this phenomenon, a set of GSP estimates were obtained using just two illustrative
421 site-years (NY6 and NY7) so as to artificially inflate equifinality. Fig 5 shows scatterplots of coordinate pairs of
422 either predicted (a) or observed (b) values for anthesis days from NY6 (horizontal axes) and NY7 (vertical axes).
423 Points in each scatterplot are color-coded to represent the number (on a \log_{10} scale) of tied GSP combinations.
424 Each tied GSP combination, when simulated using the weather data for NY6 and NY7, predicts the same
425 anthesis dates that form the point's coordinates. Dark red indicates 235,976 ties and blue indicates 1 tie. It is
426 reasonable to expect that as the number of ties increases, the range (max minus min) of the equifinal estimates
427 will increase. The size of each circle indicates the range of tied P1 estimates expressed as a percentage of the
428 mean. These percentages extend from 0.36% to 65.68%. The association of redder colors with larger circles
429 indicates that estimate ranges do, indeed, increase with the level of equifinality.

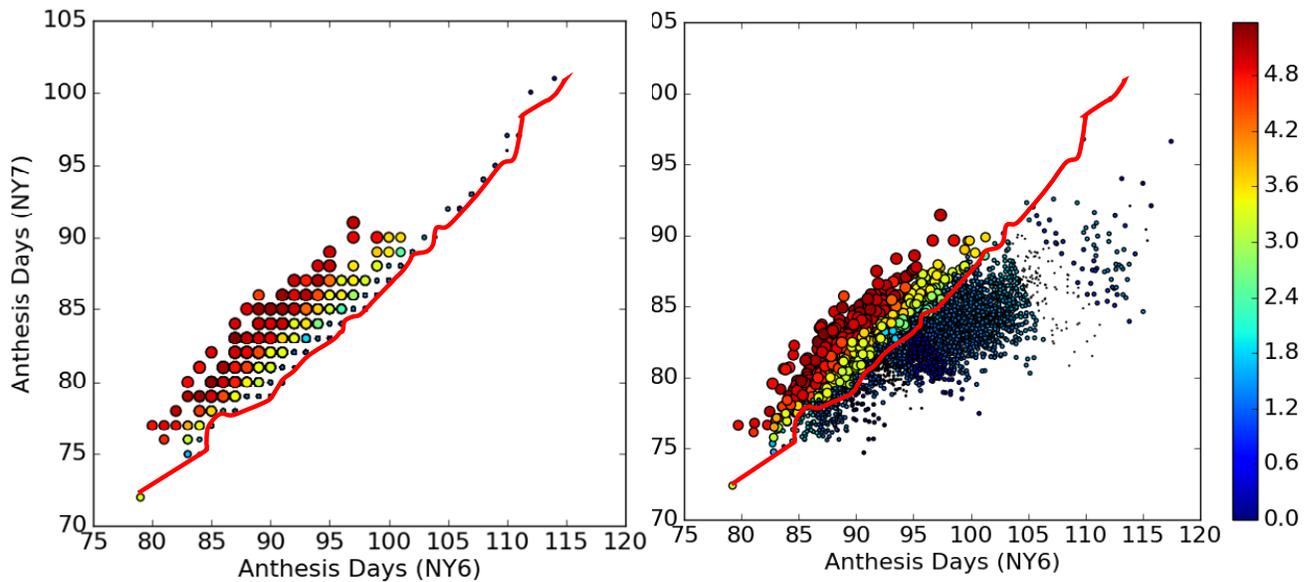
430 This is an example of a phenotype space plot that can be used to show how properties of interest (e.g.
431 number of ties and estimate ranges in this case) are distributed across the range of predictions made by the
432 model given the weather in a pair of site-years. Notice that (1) the cloud of observed points (Fig 5b) is more
433 dispersed than that of the predicted points (Fig 5a), suggesting that model responses to the environment were

434 less plastic than those of real plants and (2), as indicated by the red lines, the lowest numbers of ties in Fig 5b
435 (blue points) appear to fall in empty regions of Fig 5a where predictions are lacking. This pattern has important
436 consequences to be explained later in section “model expressivity”.



437

438 **Fig 4. Histogram depicting the frequency distribution of number of ties for 2,254 lines, used here to**
439 **characterize equifinality.** (a) Histogram of number of ties for 2153 lines with fewer than or equal to 40 ties. (b)
440 Continuation of the histogram tail from the upper panel figure representing frequency of ties for the 101 lines
441 with more than 40 ties. The trace at the top of each panel represents the average number of site-year
442 combinations (right axis) used as data for parameter estimation.



443

444 **Fig 5. Phenotype space plots of predicted (a) and observed (b) values of anthesis dates for site-years NY6 and**

445 **NY7. The marker sizes and colors respectively express the levels of equifinality based on number of ties for P1**

446 **(log₁₀ scale) and the relative ranges of its tied values. The red line is explained in the text.**

447 **Interrelationships between parameter estimates**

448 Fig 6 presents a combined plot depicting histograms of GSP parameter estimates based on all 5,266

449 lines along the main diagonal and corresponding pairwise GSP scatterplots in the upper right panels. The GSP

450 estimates were obtained using all site-years. The lower left panels in Fig 6 show the estimated Pearson

451 correlation coefficients (\hat{r}), estimated regression slopes (\hat{b}), and corresponding p -values for each mirrored

452 scatterplot. Two immediately apparent features on the scatterplots are to be noted, which might readily escape

453 notice in data sets with fewer lines. The first is the pronounced banding pattern appearing in all plots except,

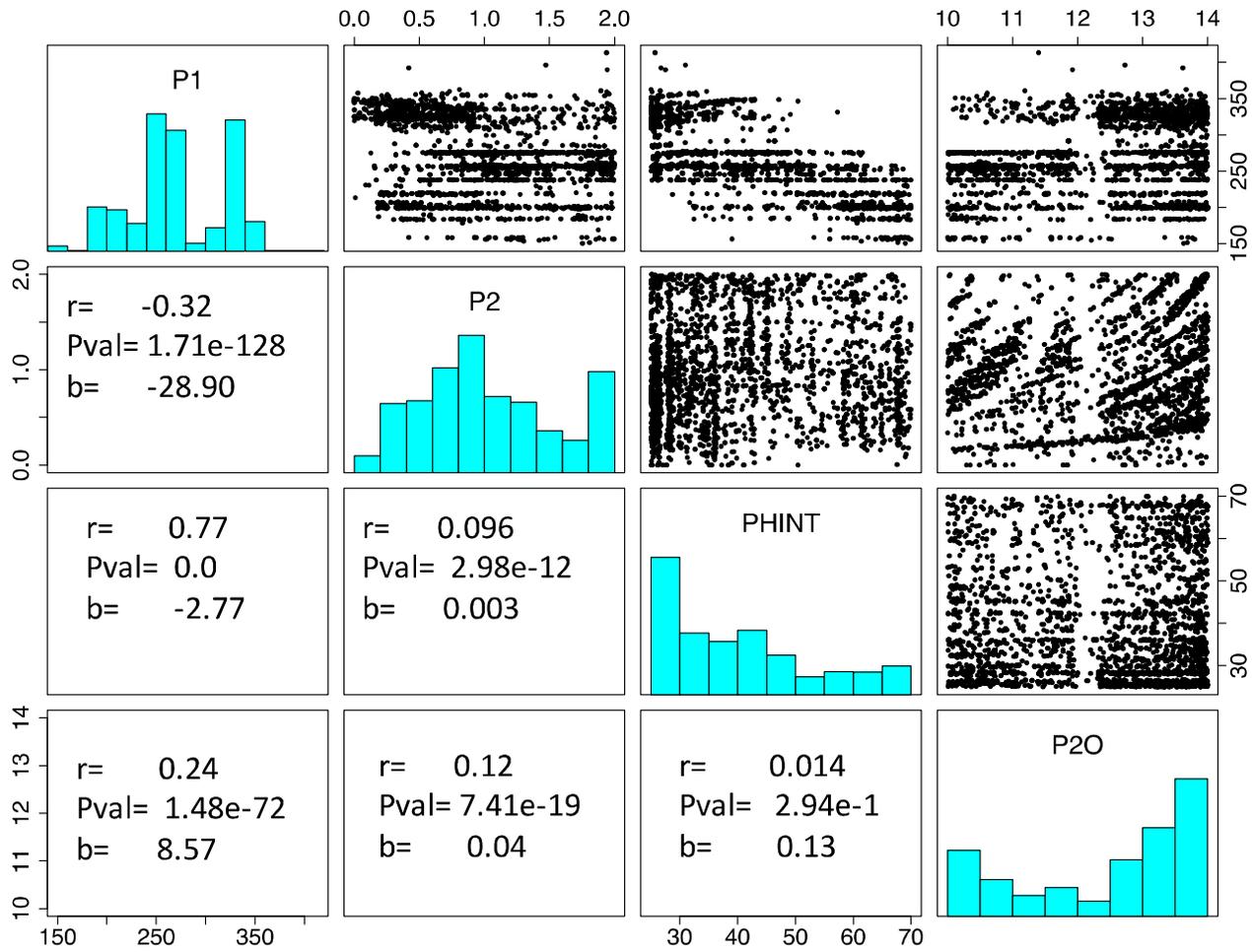
454 perhaps, P2O vs. PHINT. Most bands seem to be linear except for those on the scatterplot of P2O and P2 plot,

455 which exhibits curvilinearity. The second is the pronounced vertical gap in all P2O scatterplots. In an attempt to

456 understand the reasons for such patterns, the authors explored multiple seemingly plausible hypotheses,

457 ranging from genetics to input file coding quirks (e.g., unintended rounding of parameter values) and many

458 more, all of which were tested and discarded. Ultimately, the results presented in the following sections
 459 provided the explanations.

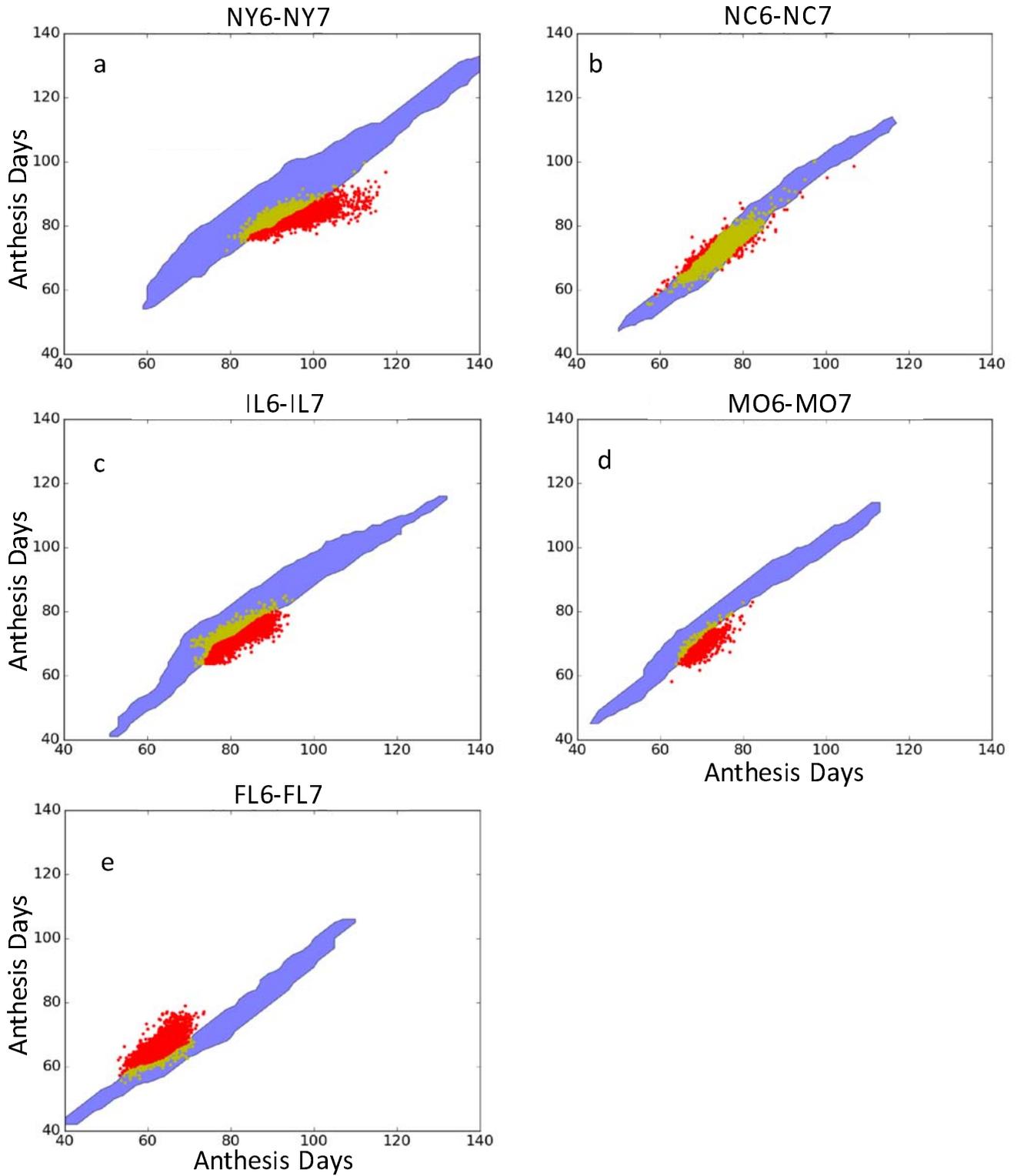


460

461 **Fig 6. Empirical distribution of selected GSP parameter estimates (main diagonal), pairwise scatterplots**
 462 **(upper right triangle) and empirical estimates of Pearson correlation coefficients, regression coefficients and**
 463 **p-values (Lower left triangle).** Each dot in the scatter plots represents a pair of GSP estimates from a single
 464 line.

465 **Model expressivity**

466 The first clue to the cause of the banding pattern emerges from the phenotype space plots in Fig 7.
467 Each plot corresponds to an independent fit to just one particular pair of site years. The blue regions in each
468 panel of Fig 7 outline predicted anthesis date pairs for two consecutive years in a given site, where model
469 prediction are constrained by the bounds imposed on the range of values allowed for each of the four GSP's
470 (Table 2). Also, for each panel in Fig 7, a dot depicts an observed anthesis date pair for a line present in a given
471 site in both 2006 and 2007. Yellow (red) dots represent observed anthesis date pairs that the model was able
472 (unable) to reproduce. We characterize each observation corresponding to a yellow (red) dot as “expressible”
473 (“inexpressible”). Except for the two North Carolina site-years, there were many lines (Table 3) for which
474 observations on anthesis date could not be predicted despite: (1) the seeming breadth of GSP values allowed by
475 Table 2; and (2) the fact that the model was only being asked to match two data points, which would seem to
476 greatly relax the constraints on GSP estimates.



477

478 **Fig 7. Phenotype space plots for predicted and observed anthesis dates.** Each panel corresponds to a pair of
 479 site-years for which fits were done. Regional color codes are described in the text.

480 **Table 3. Numbers of model expressible and inexpressible observations for selected site-year pairs.**

Lines that are ^a :	NY6/NY7	NC6/NC7	IL6/IL7	MO6/MO7	FL6/FL7
Expressible	2189	4964	2024	146	193
Inexpressible	2542	168	2946	637	3339

481 ^a These numbers refer to lines with data in both years of each pair and therefore do not precisely align with Table 1.

482 This begs the question as to what would happen to model expressivity if an even broader range of GSP
 483 values were allowed. In an attempt to investigate in a computationally efficient way how the outputs of a more
 484 conventional optimizer might appear when viewed in phenotype space, the CERES-Maize anthesis date routine
 485 was ported to Python and fit to NY6/NY7 via Differential Evolution (DE) [57]. DE is a well-established (63K
 486 Google Scholar hits on “Differential Evolution” as of October 21, 2016) and highly effective evolutionary
 487 algorithm that embodies mechanisms reminiscent of techniques ranging from the Nelder-Mead Simplex [58]
 488 method to Particle Swarm Optimization [59]. Among the algorithm’s initiating inputs is the range of parameter
 489 values within which to search, which were set as shown in Table 4. These ranges are greatly broadened from
 490 that used in the database search (Table 2); in fact, the values in Table 4 are intentionally broader than biological
 491 experience would suggest as reasonable.

492 **Table 4. Extended range of parameter values used for DE search.**

Parameter	Definition	Unit	Min	Max	Percent of Sobol Range

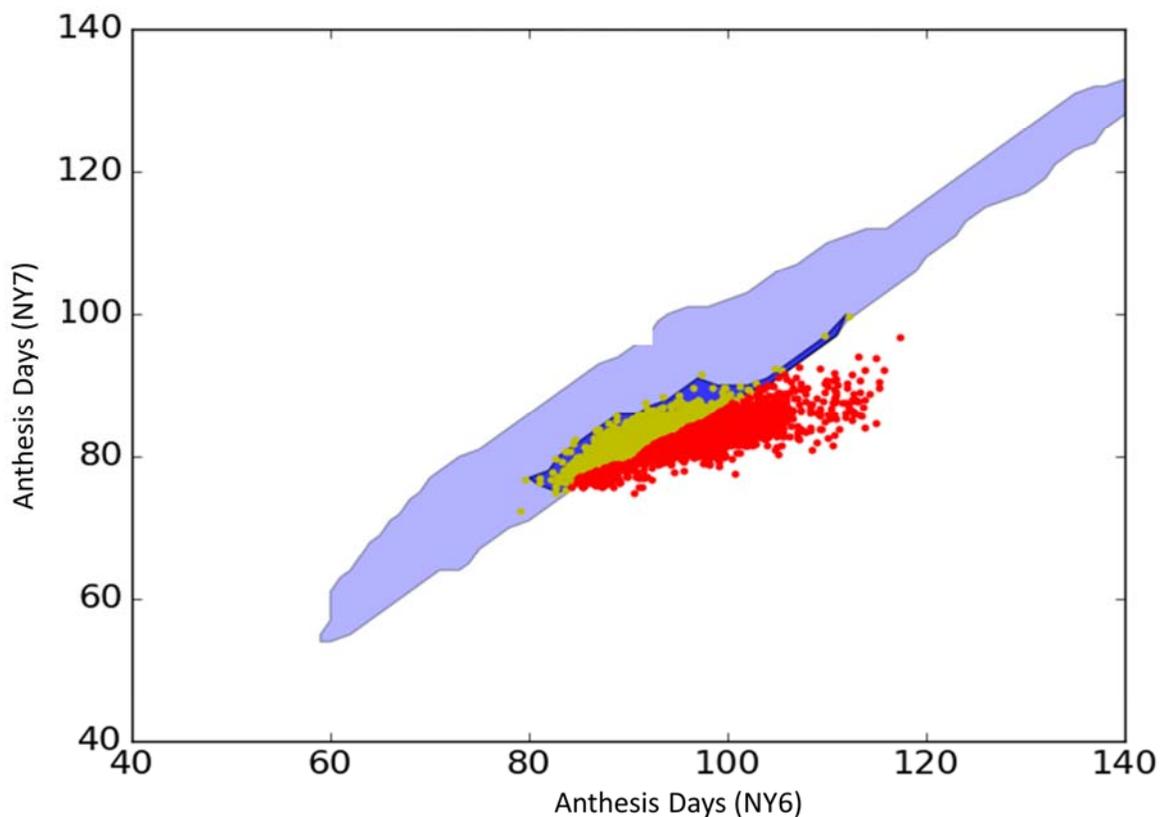
P1	Thermal time from seedling emergence to end of juvenile phase	GDD (°C)	75	600	175%
P20	Critical photoperiod hour	hrs.	6	21	300%
P2	Days of anthesis date delay for each hour by which the day length exceeds P20	rate	0	6	375%
PHINT	Phylochron interval (Interval between successive leaf tip appearances)	GDD (°C)	20	110	200%

493

494 Fig 8 shows overlapping predictions based on the database search under the range of parameters in
 495 Table 2 and on the DE search under the extended range of parameter values (Table 4). Specifically, the light
 496 blue area represents the anthesis date region that was reachable through predictions based on the database
 497 search. In contrast, the dark blue area is the predicted anthesis date region within which the DE algorithm
 498 converged. Note the almost perfect overlap of the lower edges of the light blue (i.e. database search) and dark
 499 blue (i.e. DE search) areas, indicating that, despite its much larger starting parameter search space, DE did not
 500 extend model predictions. This suggests limitations in model expressivity that go beyond the method of
 501 parameter estimation or the initial parameter space used for the search.

502 As a corollary, it is worth noting that more site-years of data of similar quality are unlikely to improve
 503 model expressivity, as illustrated by the following thought experiment. Suppose a community has developed
 504 the univariate deterministic model $y = \arctan(\theta)$, where θ is a parameter, with $0 \leq \theta \leq 10$ by solid prior
 505 knowledge and y is some dependent variable of interest. Assume that this is viewed as a very complex model
 506 requiring simulation to solve. The community understands that no model is perfect, but no specific flaws of this

507 one are known. Extant data for y ranges from 1.31 to 1.61 and yields the point estimate $\hat{\theta} = 5.79$ (RMSE =
508 0.12). Due to its complexity, no one has noticed that the model cannot reproduce any $y > \arctan(10) = 1.47$
509 or, for any θ , a $y > \pi/2 \approx 1.57$. Now suppose that: a very large set of new y data is collected. Depending on
510 the distribution of the new data either: (1) a new $\hat{\theta} < 10$ will be found or (2) $\hat{\theta}$ will rise significantly above 10,
511 leading to a rejection of the model. However, what will *not* happen is that the increase in data will enable
512 observations > 1.57 to be reproduced. The model simply lacks the expressivity to do so. Analogously, increasing
513 the amount of anthesis date data may narrow GSP estimate confidence limits, but the reachable region of
514 predicted phenotype space is unlikely to extend beyond the edges of the light blue regions. Therefore, any
515 improvement in the ability to predict the large numbers of red points in Figs 7 and 8 is unlikely.

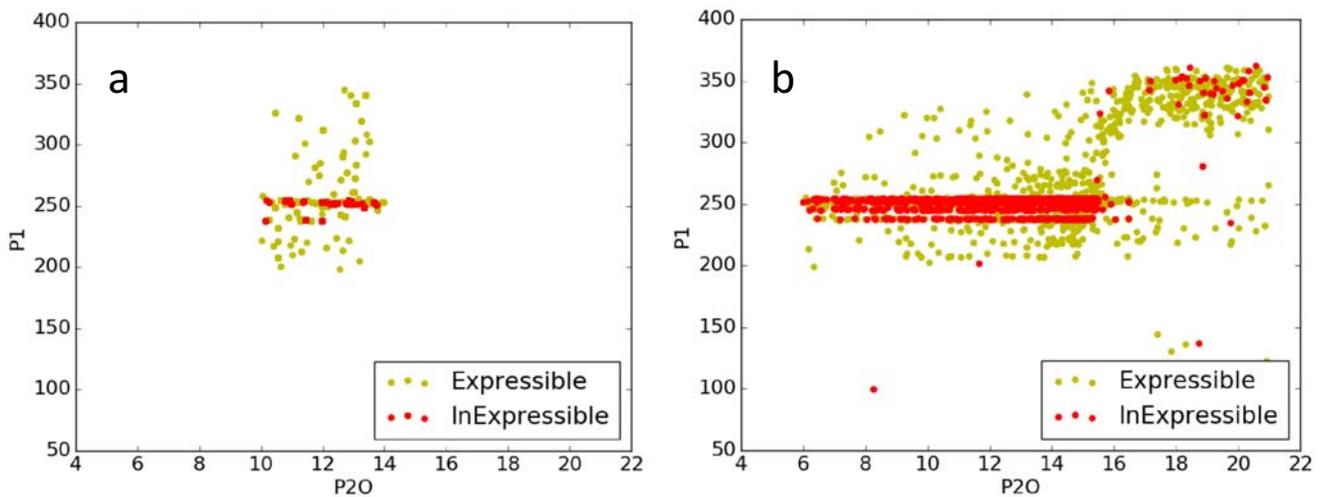


516

517 **Fig 8. Superimposed anthesis date results using NY6 and NY7.** Data illustrating that searches via database and
518 DE optimization over a much larger parameter space are equally unable to reproduce the observations for lines
519 shown as red dots.

520 Given these issues, a sensible follow-up question might be about what specific GSP estimates were
521 reported for the red points? Here we report answers only for P1.

522

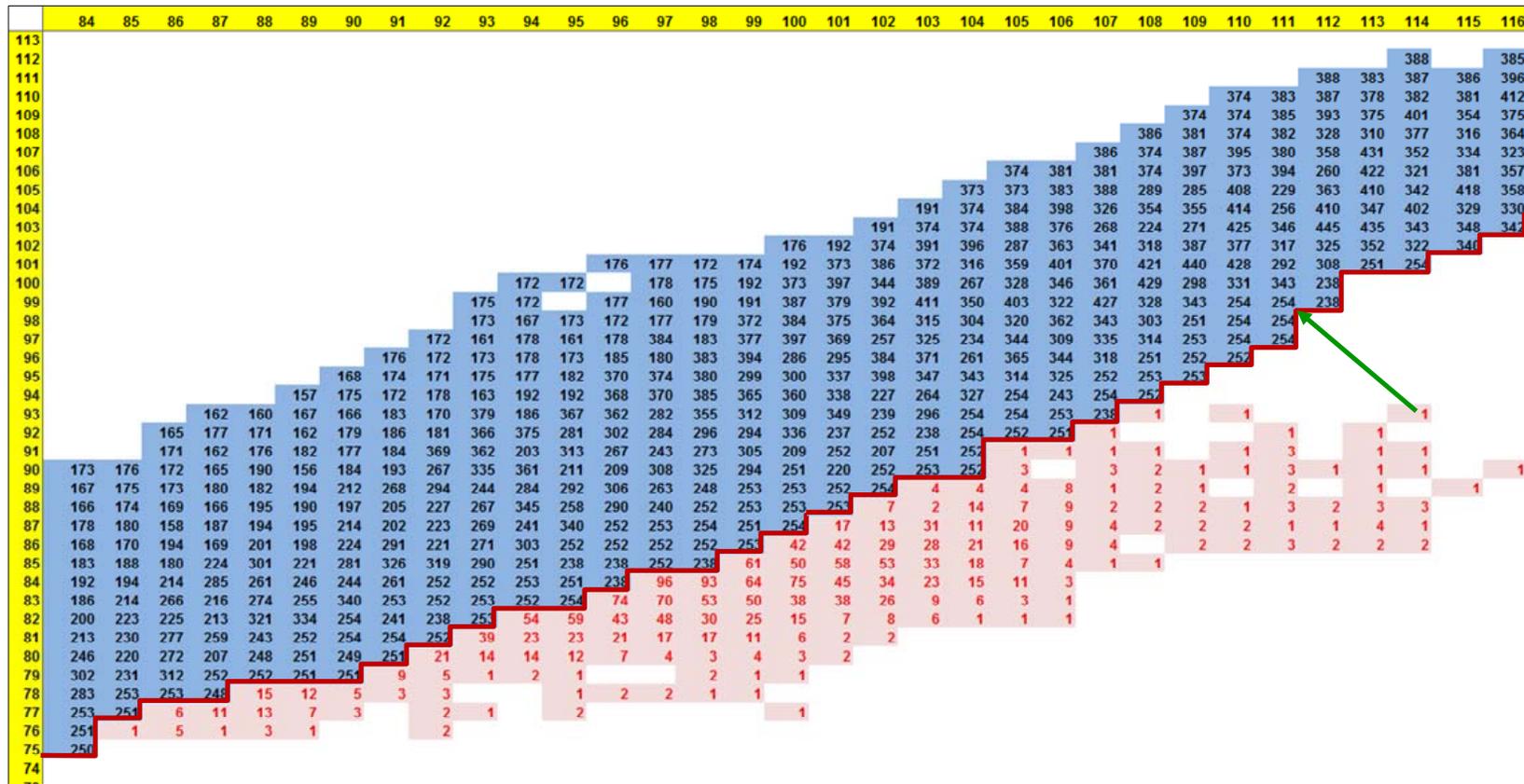


523

524 **Fig 9. Scatterplot of P1 vs. P2O estimates using data from NY6 and NY7 based on the database search (a) and**
525 **Differential Evolution (b).** Yellow and red dots are, respectively, observations characterized as expressible and
526 inexpressible by model predictions.

527 Fig 9 shows scatterplots of P1 and P2O estimates generated using data from NY6 and NY7 via the
528 database search and DE. The color coding is consistent with that in Fig 7a. The pronounced bands at ca. P1=250
529 in both panels are immediately striking – although the scale is small, a corresponding band is quite evident at
530 the same position in Fig 6. A tabulation reveals that, of all 4,731 lines represented in the Fig 9a, 3,227 (68.2%)
531 have estimates of P1 ranging from 245 to 260. Of these, 1,493 are expressible (yellow) and 1,734 (red) are not
532 expressible. Out of the total 4,731 points in the graph 2,189 (46.2%) are expressible and 2542 (53.8%) not. The

533 Fig 9b has similar proportions of expressible and inexpressible points (2327, 49.1%; and 2404, 50.9%;
534 respectively), reinforcing the similarity of results for parameter estimates from DE and database searches. The
535 differences are likely due to the ability of DE to explore the parameter space continuously whereas the
536 database search is restricted to the predefined discrete Sobol points. Still, one may wonder why so many P1
537 estimates are near the 250 degree-days? Fig 10 reveals the answer.



540 **Fig 10. P1 estimates from the database search (black) and the numbers of lines with inexpressible observations (red).** Figure arranged in a tableau
 541 organized as a phenotypic space plot corresponding to the center portion of Fig 8. The dark red line is the expressibility frontier and the green arrow
 542 shows the P1 value (254) from the GSP combination that minimizes the RMSE for one illustrative line. Horizontal and vertical yellow strips are the
 543 anthesis dates for NY6 and NY7

544 The numbers in black are the “first-best-found” P1 estimated values that generate the corresponding
545 row × column anthesis date combinations. A comparison with the corresponding dot colors and sizes in Fig 5b
546 indicates that, on the frontier (red borders Figs 5a,b and 10) between expressible and inexpressible
547 observations, there was essentially no equifinality and, concomitantly, narrow ranges of P1 values. Fig 10
548 shows that the P1 values along the frontier were all quite close to 250. For lines with observations falling
549 outside the frontier, the RMSE was minimized by assigning GSP values associated with the closest achievable
550 dates, i.e. those directly on the frontier. Therefore, all the lines counted by the red numbers were assigned P1
551 values that are very close to 250 and have essentially no equifinality. The green arrow in Fig 10 illustrates this
552 phenomenon for one line. The nearest P1 estimate is 254 and the length of the arrow (ca. 5.8 days) is
553 proportional to that line’s RMSE. Specifically, in this case the length is $1/\sqrt{2}$ times the RMSE because there are
554 $n = 2$ site-years.

555 Recall that the upper limit placed on P1 was 450 (and 600 in the DE search), therefore this outcome is
556 likely not an artifact of constraints in the GSP search space but, rather, a result of poor model expressivity, that
557 is the model inability to predict anthesis date pairs beyond those on the frontier. This mechanism accounts for
558 the P1 band at 250 in Fig 9a. Furthermore, as previously presented, more data cannot improve the prediction
559 of inexpressible lines, the banding in Fig 6 is not surprising.

560 **P20 gap**

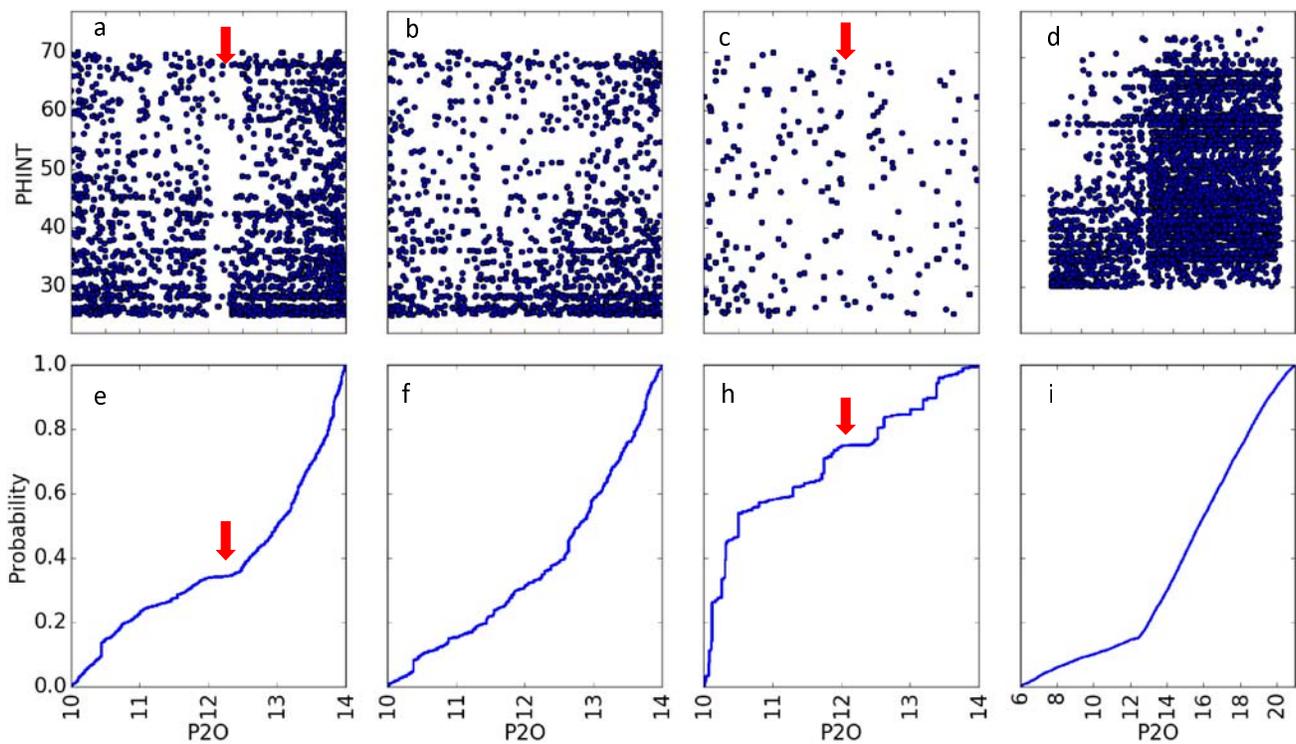
561 We now investigate the vertical gap in scatterplots involving P20 estimates (Fig 6), which documents
562 the intricacy of the interactions that can occur between model mechanisms, parameter ranges searched,
563 optimization algorithms used, and environments included. Exploratory re-tabulations of the Sobol-based
564 parameter database revealed that the P20 gap was clearly present in the three site-years having shorter day
565 lengths (FL6, FL7, and PR6) but absent in fits obtained by only including the remaining eight site-years with
566 longer days (Fig 11). Fig 12 shows that a substantial number of observations for short-day site-years are outside

567 the predicted phenotype ranges expressible by the model under either database or DE optimization. As
568 described in section “CERES-Maize model”, the model operated by calculating the number of leaves initiated by
569 the end of Stage 2 and predicts anthesis only after leaves are fully emerged. For any line, leaf number was a
570 constant across all site-years, namely $P1/(2 \times PHINT) + 5$. The variation of anthesis dates across plantings was
571 such that there were few, if any, combinations of P1 and PHINT that were compatible with the data from all
572 site-years. Therefore, the optimizer relied more heavily on the P2 and P2O parameters.

573 Specifically, the optimizer settled on very small P2O estimates, much smaller than the short southern
574 photoperiods. Instead, the optimizer relied on P2 estimates to generate anthesis date predictions that were
575 delayed to the greatest extent possible by lengthening Stage 2. Recall that P2O values above the day length
576 make Stage 2 only four days long, which is not enough time for temperature differences to accumulate the
577 needed variation. The abundance of low P2O estimates thus created the gap observed in scatterplots of P2O
578 with other GSPs (Fig 11a). In contrast, the photoperiods in the remaining longer-day site-years exceeded the
579 maximum allowed P2O values in the P2O database search during (and long after) the juvenile period.
580 Therefore, there was no empty band in the scatter plot (Fig 11b) because the optimizer was able to exploit
581 delays for any value of P2O.

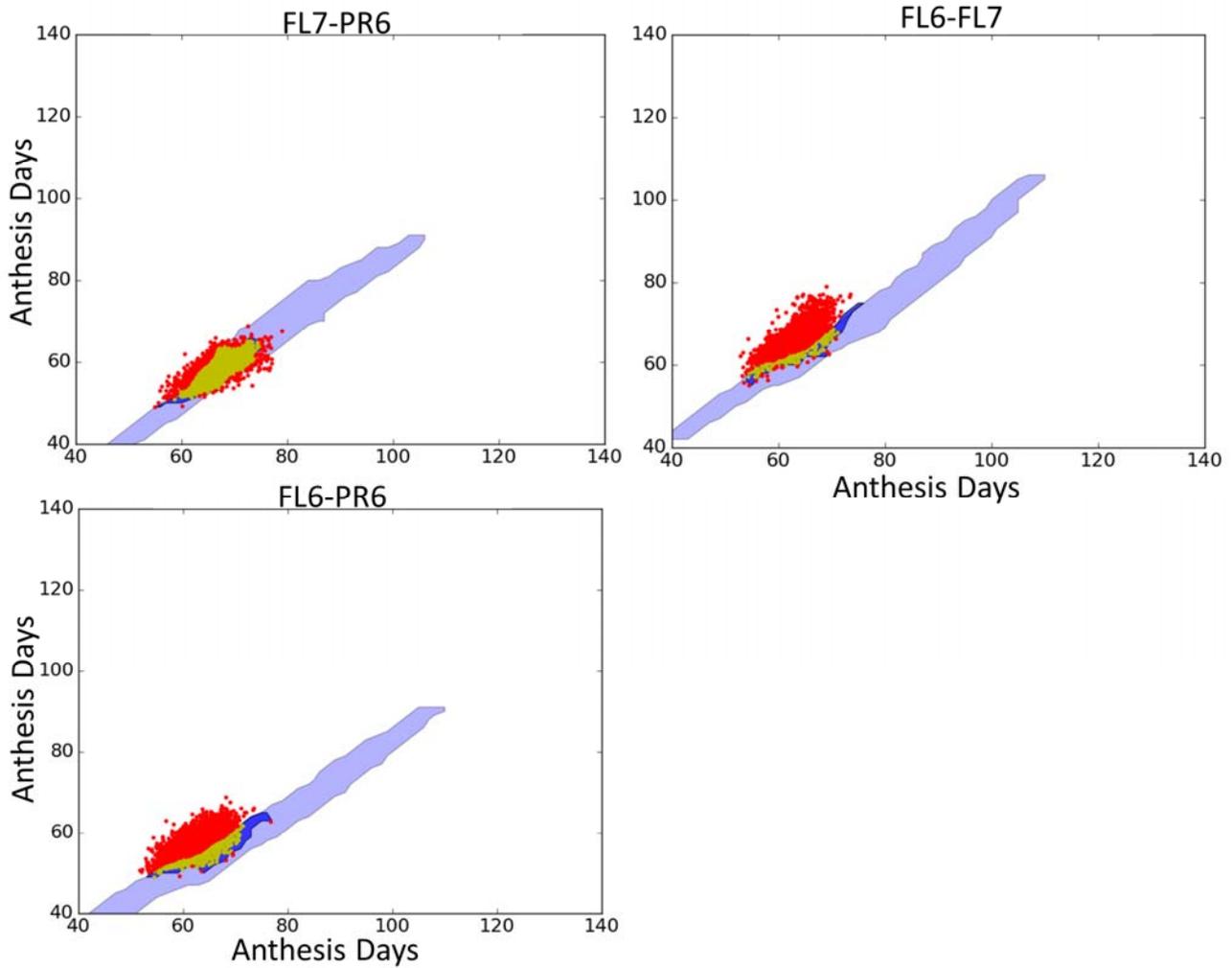
582 With the broader range of parameter values available to the DE runs and the increased flexibility
583 available between P1 and PHINT, other options became available. In particular, in many cases DE found GSP
584 combinations wherein P2O exceeded the southern day lengths so photoperiod had no influence on anthesis
585 date and no gap artifact was generated (Figs 11d,i). P1 and PHINT thus became the major explanatory
586 parameters. This is shown in Fig 13, whereby for each line, the parameter differences are plotted against the
587 RMSE differences that result from changing the estimation methods from database to DE optimization. The DE
588 estimate of P2O were larger in 4,507 out of 5,240 lines (87%; Fig 13d), almost always by enough to put it above
589 the local day lengths. In tandem, P1 values fell in 3,559 lines (Fig 13a), whereas PHINT rose in 4,102 lines (Fig
590 13c).

591 Note, however, that for *any* (P1, PHINT) combination, *any* P2O that exceeds the local day length will
592 give the same RMSE – a clear source of equifinality. Thus, the changes in P2O will not, in all likelihood, lead to
593 values that can be more closely related to genetics. Moreover, because of the limits on model expressivity,
594 none of the DE solutions gave significantly better fits than the database estimates. This is why virtually all
595 points in Fig 13 had DE RMSE's within 0.5 days (horizontal axes) of the database-based parameter estimates.
596 This, too, is an illustration of equifinality because the two optimizers were finding different GSP estimates
597 although the RMSE were of similar magnitude.



598

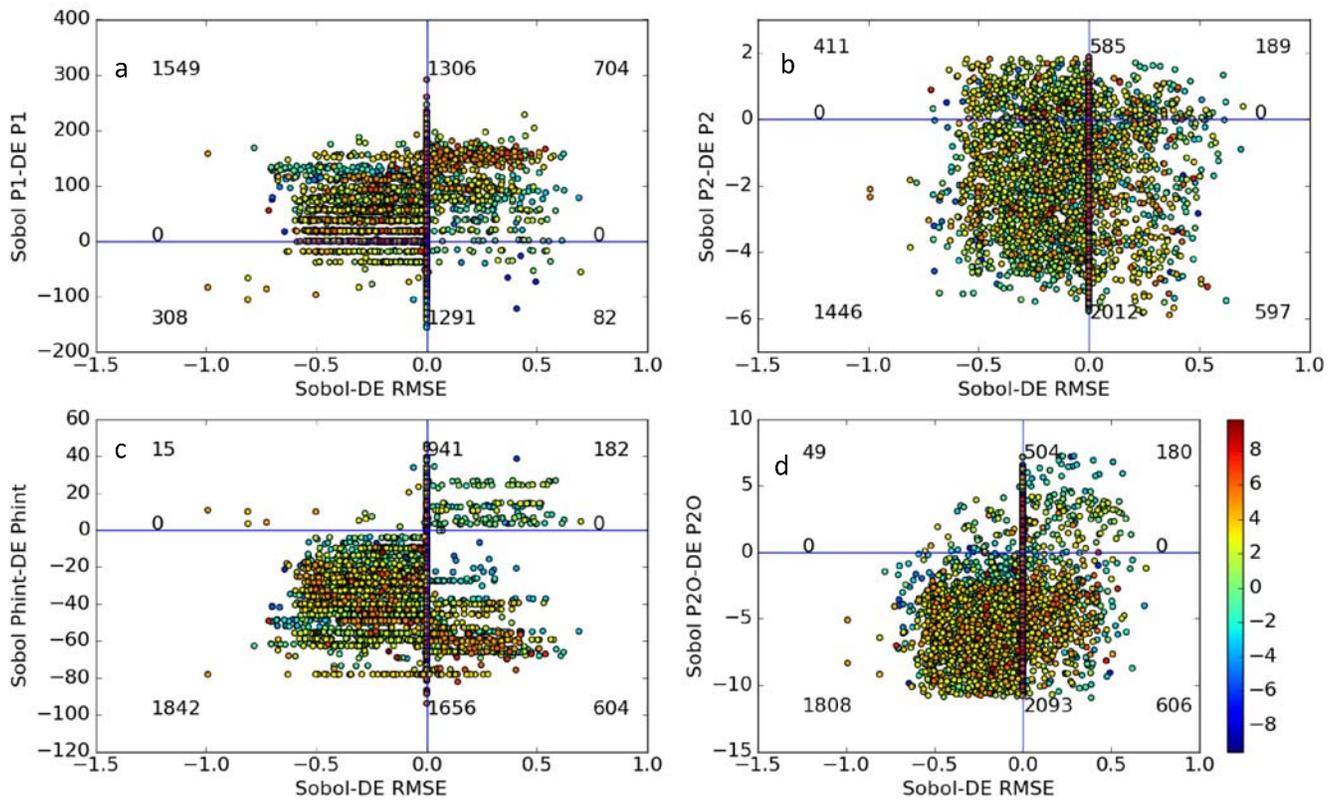
599 **Fig 11. P2O and PHINT scatter plots (top row) and P2O cumulative density functions (bottom row).** (a & e) all
600 11 site-years, (b & f) longer day site-years, (c & g) shorter day site-years based on the database approach, and
601 (d & i) shorter day site-years using the DE approach. All horizontal axes in both rows have the same scale.



602

603 **Fig 12. Phenotype space plots of observed and predicted values based on the three site-years with shorter**
604 **days.** Note the large number of points in the FL6-PR6 and FL6-FL7 plots that lie above the dark blue prediction
605 region based on DE.

606



607

608 **Fig 13. The differences in parameter estimates from database search vs. DE (vertical axes).** Figure plotted
 609 against the corresponding difference in RMSE for 5240 lines in FL6, FL7, and/or PR6. The color encodes the
 610 sum of residual (observed minus mean) across site-years for each line.

611 Tests for stability of GSP estimates

612 Table 5a shows the effect of including or excluding the effect of different subsets of site-years on the
 613 modeling of estimates (Equation 1) for each GSP for the base set. For all GSP parameters, AIC and BIC values
 614 were considerably smaller for models that included the random effect of site-year subsets, β_e , therefore
 615 suggesting non-negligible variability across site-year subsets on the GSP estimates. The table illustrates the size
 616 of the site-year set effects as follows. For scaling purposes, we provide the estimated intercept, $\hat{\mu}_p$, which also
 617 serves also as an estimated GSP grand mean across all lines and site-year subsets. The Index of Variability

618 (expressed as a percent) is the standard deviation of the β_e effect normalized by the grand mean. The
619 percentage of the total GSP variance ($\sigma_e^2 + \sigma_l^2 + \sigma_r^2$) attributable to site-year subsets is also shown. Both of
620 these descriptors indicate substantial variability between site-year sets, with indexes of variability ranging from
621 5.9% for P20 to 33.6% for P2 and over 20% of the total variance related to site-year sets for all GSP's.

622 The Chi square values from the likelihood ratio test and the associated p -values are presented in the
623 last two columns of Table 5a. The extreme p -values demonstrate that the GSP values depend on the set of site-
624 years used to estimate them. Therefore, the GSP's are not, in fact, genotype specific despite the goodness-of-fit
625 displayed in Fig 3. This result is completely understandable given the range of artifacts due to equifinality and
626 model expressivity issues identified above.

627 Table 5b shows the results when only estimates having ties are tested (left) vs. an analysis that includes
628 all estimates (right). The former corresponds to estimates for lines whose observations fall inside the
629 expressivity frontier and the latter includes the estimates for all lines. It is clear that the grand means, index of
630 variability, and percentages of GSP variance are highly similar between all three groupings in Table 5. Also, all p -
631 values are extremely significant and increase with the amount of data used.

632

633

634 **Table 5a. Estimated log likelihood, fit statistics, selected summary measures, and a likelihood ratio test for**
 635 **competing statistical models fitted on GSP estimates with and without the random effect of site-year subset,**
 636 **based on GSP estimates for the base group (N=60,834).**

GSP	Log likelihood w/o (top) and w/ (bot) a site-year set effect ^a	AIC w/o (top) and w/ (bot) a site-year set effect ^b	BIC w/o (top) and w/ (bot) a site-year set effect ^b	GSP Grand Mean $\hat{\mu}_p$	Index of Variability ^c $\sigma_e / \hat{\mu}_p$	Variance pcts. for site-year sets ^c $\sigma_e^2 / \sigma_{tot}^2$	Chi-square test statistic	Chi-square p -value ^d (df =0.5)
P1	-338046 -322689	676098 645386	676125 645422	264.625	12.30	34.38	30714	10^{-13334}
P2	-46154 -25237	92313 50482	92340 50518	1.037	33.55	33.92	41833	10^{-18163}
P2O	-105304 -95357	210614 190723	210642 190759	12.2440	5.88	27.83	19894	10^{-8635}
PHINT	-254875 -246903	509756 493815	509783 493851	44.167	15.44	22.62	15943	10^{-6919}

637 ^a Larger is better ^b Smaller is better ^c Chernoff upper bound on Chi-squared cum. dist. function.

638

640 **Table 5b. Summary measures and likelihood ratio p -values for competing statistical models fitted on GSP**
 641 **estimates with and without the random effect of site-year subset from data only having ties (left) and all data**
 642 **(right).**

GSP	GSP Grand Mean $\hat{\mu}_p$	Index of Variability ^c $\sigma_e/\hat{\mu}_p$	Variance pcts. for site-year sets ^c $\sigma_e^2/\sigma_{tot}^2$	Chi- square p -value ^d (df =0.5)	GSP Grand Mean $\hat{\mu}_p$	Index of Variability ^c $\sigma_e/\hat{\mu}_p$	Variance pcts. for site-year sets ^c $\sigma_e^2/\sigma_{tot}^2$	Chi- square p -value ^d (df =0.5)
With Ties (N=114,314)					With all Data (N=177,870)			
P1	273.5	11.37	29.77	10^{-23283}	270	11.48	29.94	10^{-34955}
P2	0.9137	36.33	35.23	10^{-34723}	0.9593	35.5	33.8	10^{-52518}
P2O	12.49	4.43	19.70	10^{-11883}	12.42	4.88	21.27	10^{-19806}
PHINT	43.57	18.65	26.31	10^{-17348}	43.94	17.3	24.35	10^{-23740}

643 ^a Larger is better ^b Smaller is better ^c Chernoff upper bound on Chi-squared cum. dist. function.

644 Discussion

645 Since their inception, ecophysiological models have been evaluated in terms of predictive ability, which
 646 are superb in many circumstances [60]. The model parameters were considered to be *inputs* whose genesis was
 647 secondary as long as the model outputs proved useful. However, as often happens in science, perceived needs,
 648 desiderata, and requirements escalate as technologies evolves. In particular, we are now demanding that the
 649 model inputs themselves be the accurate outputs of processes at the genetic level that can be modeled by

650 genomic prediction. It is not surprising, therefore, that modeling technologies (ranging from data collection to
651 estimation) that were adequate for past applications now require improvement.

652 From a fundamental but traditional perspective, there are several issues of perennial concern in crop
653 modeling. The first is model functional structure including both its degree of expressivity and its behavior under
654 optimization. For example, estimation procedures like DE, that primarily yield point estimates, are limited in
655 their ability to assess equifinality. At best, one can query the flatness of the goodness-of-fit function in the
656 neighborhood of the estimate, but this does not tell anything about the ubiquity of equifinality across the
657 parameter space. Nor do these procedures allow one to detect observations that fall outside of the model's
658 scope of expressivity unless the discrepancies are quite large. Doing so requires methods like the Sobol'
659 database scheme used here that can make broader assessments in both parameter and phenotype space. It
660 may well be that the rarity with which database methods have been used has led to an underappreciation as to
661 the prevalence of these adverse situations.

662 When expressivity issues are identified, results like those above are not likely to be solved merely by
663 acquiring more data of the same type. In such situations, better models will often be needed and modern genetic
664 studies can help. A great many plant component subsystems are currently under study at the molecular level.
665 Indeed, some of these (e.g., [61]) are even being combined into multi-scale organ and whole plant models. Even
666 without modeling directly at the genetic level one can use the derived insights to make informed choices
667 between alternative representations of individual ecophysiological processes. Tardieu (2003) refers to such
668 representations as "meta-mechanisms". It would seem plausible that building models from component parts of
669 increased biological realism should increase the ability to reproduce field variation – at the very least, it is hard
670 to see how it can hurt. As a concrete example, the B73 parent is photoperiod insensitive. In CERES-Maize,
671 however, the only way to express this is by setting P20 in excess of the observed photoperiods, with the
672 consequences we have seen.

673 This is not to say, however, that both more and better data are not needed. Indeed, data quality issues
674 can impact both expressivity and GSP stability. For example, while the date seed that are physically sown in a
675 field is usually known and not subject to error, researchers often report a subjective notion of “effective sowing
676 date” based on their interpretation of whether low soil moisture delayed germination. If errors in sowing date
677 push an anthesis observation across the expressivity frontier, erroneous GSP estimates will result. Such errors
678 can also arise if different personnel are involved across locations or growing seasons, especially for visually
679 evaluated phenotypes like most phenological traits. Providing the emergence date can provide a partial check
680 for these problems and also for errors in simulating time from sowing to emergence. Unfortunately,
681 emergence dates were not reported for the maize NAM dataset.

682 Another traditional modeling concern has always been the relationship between the observed
683 environmental data and the immediate environmental conditions actually experienced by individual plants.
684 Weather data can suffer from multiple sources of bias and error [63]. For example, stations that are not located
685 within or directly adjacent to experiments may have bias due to local variation in weather conditions.
686 Additionally, although of limited concern for anthesis dates, the quality of soil and management data. In this
687 study any systematic differences in protocols for collection of weather data between the sites as aggravated by
688 small sample effects, might have contributed to some degree to the significance levels in Table 5. It would
689 certainly be desirable to have a method by which this potential effect might be quantitatively assessed. Such a
690 method could be instrumental in designing experimental procedures for reducing the problem. One potential
691 example might be to eschew external measurements of some environmental variables (e.g., air temperature)
692 and use sensors onboard UAV’s or other automated vehicles to measure plant temperatures or other critical
693 features directly at high temporal and spatial frequencies.

694 More involved data types and structures are also needed to resolve issues of equifinality when they
695 arise. Equifinality is fundamentally a problem of discernment. In simple terms, given an equation $c = a + b$, if
696 one only has data on c , then estimates of a and b are doomed to be equifinal. If one desires otherwise, one

697 must find a way to measure either a or b . Current technological efforts to develop high throughput
698 phenotyping approaches might be quite helpful in this regard. For example, assuming that
699 $TOLN=P1/(PHINT \times 2)+5$ is the correct way to model the number of leaves at anthesis, data on total leaf number
700 would help constrain the parameter estimates. This leads toward a range of constrained and/or multiobjective
701 estimation procedures on which there has been significant amounts of research [64,65]. Maximum entropy
702 methods offer another opportunity wherein one identifies a probability distribution of values that is
703 constrained by but mathematically no more informed than is justified by a set of potentially diverse data types
704 [66]. Another alternative might be Bayesian methods with multivariate likelihood functions that combine
705 several observational variables [67].

706 Another approach to reduce equifinality would be to use simpler models. The fewer the number of
707 processes and GSP's in a model, the smaller the opportunity for hard-to-spot tradeoffs to exist wherein
708 adjustments to one parameter can be offset by tweaking another one. Of course, the tradeoff may be less
709 expressivity leading to other problems. However, Welch et al. (2005) presented 12 dichotomies comparing
710 gene network modeling and quantitative genetics approaches, where aspects of the former might also apply to
711 ecophysiological modeling. They opined that an optimal modeling approach should entail a synthesis of both.
712 The key features to be contributed from the network (i.e., ecophysiological) side would be (1) the ability to
713 handle time-varying dynamics, (2) a far more parsimonious approach to expressing biological and biology \times
714 environmental interactions, and (3) a more mechanistic explanation of how traits originate. It is at least
715 conceivable that some way station of moderate complexity exists between statistical genetics and full crop
716 models that can achieve this.

717 At whatever level of complexity proves appropriate, one cannot accurately estimate the parameters
718 controlling model components without collecting data on settings wherein the relevant processes operate
719 differentially. This is clear from the P2O gap phenomenon, which was apparent when only short day data was
720 used and absent under long days. Both settings distorted the results, in one case compressing estimates into a

721 restricted range, leaving a gap, and, in the other, allowing them to spread out. Furthermore, this interacted
722 with the range of values allowed, which caused shifts between (P1, PHINT) and (P2, P2O) as to which
723 parameters appeared to be “explanatory”. The debilitating influence of such behavior on linking parameter
724 values to genes is terribly obvious.

725 However, it also should not escape notice that the gap was evident even in a mixture of environments,
726 suggesting that good experimental design entails more than just making sure that a suitable range of
727 environments is included. There is some notion of balance that needs to be established and applied globally to
728 data selection. In this context, it is worth noting that despite the fact that thousands of lines were planted in
729 each location, there were only 539 lines where data were reported from all 11 trials. However, given the
730 expense of such large-scale trials and the multiple purposes each one will serve, “balance” cannot mean
731 “orthogonality” where all lines are planted at all sites. Of course, an established benefit of ecophysiological
732 models is to serve as guides to help prioritize experimentation over time. It seems likely that as their
733 integration with statistical genetic models expands, they might also be able to assist in the rational planning and
734 resource allocation for large, multi-site trials.

735 Another approach entirely would be to seek to move beyond a two-step “estimate and then map”
736 paradigm. Conventional mapping methods essentially isolate genetic markers whose pattern of assignment to
737 lines mirrors the pattern of phenotype values of interest. A general linear model is assumed to mediate
738 between marker states and realized phenotypes. There is no conceptual reason why that general linear model
739 might not be replaceable by a crop model. In effect, one could conceive a hierarchical model in which a first-
740 level model is specified on the data and higher order submodels are specified on the parameters that
741 characterize the behavior of observed data, much like proposed by Bello et al. (2010).

742 One could conceptually implement this hierarchy in the context of crops by fitting phenotypes with an
743 ECM whose GSP's are then specified as functions of genetic markers at another level of the hierarchical model.

744 Indeed, this is what the current paradigm attempts, except that the two-step estimation process curtails
745 smooth borrowing of information across hierarchical levels of the model that could potentially help resolve the
746 equifinality problem.

747 We acknowledge that one-step hierarchical model approach might not solve the sort of expressivity
748 problems described in the thought experiment and documented in our results (both in section “Model
749 expressivity”). Yet, it would enable the genetic structure of the population to inform the GSP estimation
750 process. The potential utility of this hierarchical modeling approach is currently under study in one of our labs.
751 The approach would also enable more efficient use of data. Currently, the two-step approach requires data
752 from multiple environments [39] for each line in order to estimate the GSP’s before mapping can proceed.
753 However, consider a line that was culled very early in the selection process, perhaps even after a single round.
754 Because the parameters estimated in putative one-step hierarchical modeling schemes would include marker
755 effects, even just one planting becomes a usable observation if the line is genotyped. This is a sufficiently
756 inexpensive operation now that some programs (e.g. CIMMYT; [69] are doing so routinely for the offspring of all
757 crosses.

758 A one-step hierarchical modeling approach might also make it possible to utilize data taken on lines
759 after they enter the market place. Analogously to high throughput phenotyping in breeding programs,
760 precision agricultural management is also investing in sensor- and model-based approaches to improve
761 productivity [70,71] while collecting a wealth of multivariate data. Usually, of course, hybrids are released into
762 areas where they show low G×E interactions. For example, a line with a particular P2O is not likely to be
763 released across a sufficient range of latitudes to have great differences in day length. This would make it
764 difficult to directly estimate P2O for the line using the methods described in this paper.

765 However, in a one-step hierarchical model approach, one would only be looking for markers that
766 influenced P2O. In this case, data from many lines and geographical areas could be used together. This would

767 also make such data usable for the sorts of hypothesis testing about genes discovered by other means, thus
768 facilitating genetically-informed ecophysiological modeling. For such approaches to be workable, however,
769 there are many policy issues to be resolved including information property rights and fair economic returns to
770 data, not to mention the need to greatly harden cybersecurity protections [72]. However, if this can be done
771 then issues of environmental coverage would likely be ameliorated due to the extent of the data that would
772 become available.

773 **Conclusions**

774 The original and seemingly simple goal of this study was to first fit the anthesis date component of the
775 CERES-Maize model to data from over 5000 genotyped lines and then genetically map the resulting GSP values.
776 However, we were unexpectedly detoured when we found that despite the high predictive quality of the values
777 obtained, there were numerous artifacts that emerged in the estimation process, thereby making our
778 immediate goal unachievable. We find it interesting that the problems we encountered would likely be
779 invisible, though present, in smaller data sets and, unless addressed by suitable research, these problems bode
780 ill for understanding any genetic underpinnings of ecophysiological models. This is worrisome given the recent
781 escalating attention that has been given to this method of melding ecophysiological and statistical genetic
782 models as a way of accelerating the crop improvement process so as to help meet global food and fiber needs
783 by 2050.

784 The constraining issues fall into two categories. The first arises in situations where the model is unable
785 to express the observed data for some line even by a relatively few days. In this circumstance, the line is
786 assigned the GSP associated with the nearest point on model's expression frontier – values which can, however,
787 change only slowly along that boundary. The result is that many and in some cases a large majority of lines are
788 assigned the same GSP values independent of their actual genetics.

789 The second symptom arises when the model can reproduce the data. In these instances, there can be
790 many combinations of GSP values that predict equally well. When such equifinality exists, there is no principled
791 way to assign the line a genetically relevant value. In short, when the model can express the data there is no
792 unique combination of GSP values and, when unique combinations do exist they are often values being given to
793 many lines because of a deficiency in model expressivity.

794 This finding is rather remarkable because in both breeding efforts and, indeed, genetic studies as a
795 whole, anthesis date is considered, if not a simple trait, at least one that has proved much easier to elucidate
796 than many others. In addition, it is generally, much more readily predicted by classical phenology models for
797 reasons that, themselves, have become generally understood [37]. This cannot but make one wonder, what
798 pitfalls might lie in wait for efforts to probe other, more involved traits.

799 Therefore, the next question to be asked by follow-on research is how prevalent are these phenomena.
800 The best way to do that would seem to be to use Sobol database search methods. This is because, unlike
801 optimizers that find single “best estimates”, the database approach will reveal the both the extent of the
802 expressible phenotype regions as well as a direct measure of the extent of any equifinality.

803 However, despite the ability to reuse results databases for many searches, undertaking such a program
804 in any broadly based fashion will be highly demanding computationally. For this reason, strong consideration
805 should be given to disaggregating comprehensive models into separate modules that can be studied
806 independently at much lower computational cost. (This is what we did for the limited DE run, although Python
807 certainly is not a high performance language.) A better long-term strategy would be to program future models
808 in a manner that supports single-module testing at the source code level. Doing so will facilitate the whole-
809 model verifications needed to ensure that fragmentation into modules for testing and improvement by
810 different labs does not compromise integration at the level of the scientific community.

811 As module testing and innovation progress, it will be of strategic value to ground improvements in
812 advancing genetic understanding at the molecular level. While this might seem daunting to those versed in
813 purely physiological approaches, it need not be so. One of the most venerable concepts in all of the life
814 sciences is that of the biological hierarchy that is, a series of many functional levels extending from molecules to
815 the biosphere. One of the perspectives emerging from molecular science is that that hierarchy might, be
816 operationally much flatter than commonly believed. That is, simple changes at lower levels can easily create
817 tangible responses multiple levels higher. To the extent that this is true, it greatly reduces the complexity of
818 bridging across those levels. This is the philosophy behind the meta-mechanism approach mentioned earlier
819 [62,73].

820 That approach has a proven ability to account for environmental interactions with sufficient skill to
821 eliminate observed G×E interactions from GSP's in the data sets used (Reymond et al., 2003). However, as
822 shown by the *p*-values in Table 5, the very large data set used herein conveyed an extraordinary power to
823 detect site-year dependencies in GSP estimation. Indeed, so powerful as to make one wonder if an insignificant
824 result is scientifically achievable by any even remotely feasible research effort? A better number to use for
825 practical evaluations might be the index of variability in Table 5. This would give a clear index of the size of the
826 effect as a percentage of the parameter values. Also, means exist for comparing such indices to see if
827 reductions in their values (i.e. by an improved model with lowered site-year set dependency) are statistically
828 significant (Vangel, 1996).

829 A final message from our research is that one cannot fix problems that one does not know exist.
830 Community interest in the fitting-and-mapping paradigm has been high as shown by the heavy citation rates for
831 the seminal papers in this area. For example, as of September, 2016, the Hammer et al. (2006) paper had been
832 cited 257 times and those publications, *themselves*, had been cited by 6,370 others (Source: Google Scholar).
833 There is also no doubt as to the importance of the ability to predict the behaviors of novel genotypes in novel
834 environments while crosses are still in the planning stage. Indeed, this is precisely the genotype-to-phenotype

835 problem, which has been declared by the National Research Council to be a top-priority goal for applied biology
836 (NRC, 2008). So these impediments need to be overcome. However, with methods now in hand to detect
837 adverse model behaviors under estimation, research that is probing ever more deeply into the control
838 mechanisms of plant growth and development, and concrete tests to document model improvements, there is
839 no reason to believe that we cannot do so.

840 **Acknowledgements**

841 The plan to use the maize NAM data to first developed through discussions with iPlant
842 (www.iplantcollaborative.org) on novel applications of cyberinfrastructure in plant science. The authors
843 acknowledge the Texas Advanced Computing Center (TACC; <http://www.tacc.utexas.edu>) at The University of
844 Texas at Austin and Beocat, Kansas State University for providing high performance computing resources that
845 have contributed to the research results reported within this paper. Support for this effort was also supplied by
846 the Department of Agronomy at Kansas State University.

847 **Author Contributions**

848 Conceptualization: JW KT SW. Methodology: AL SW KT JW. Analyzed the data: AL SW. Manuscript
849 preparation: AL SW JW KT

850

851

852

853

854 **References**

- 855 1. Stone T. Sustainability and the needs of 2050 agriculture: Developed and developing world perspectives.
856 2011. Report No.: 23.
- 857 2. Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, et al. Food security: the challenge of
858 feeding 9 billion people. *science*. 2010;327: 812–818.
- 859 3. Cooper M, Technow F, Messina C, Gho C, Totir LR. Use of Crop Growth Models with Whole-Genome
860 Prediction: Application to a Maize Multienvironment Trial. *Crop Sci*. 2016; Available:
861 <https://dl.sciencesocieties.org/publications/cs/articles/0/0/cropsci2015.08.0512>
- 862 4. Hammer G, Cooper M, Tardieu F, Welch S, Walsh B, van Eeuwijk F, et al. Models for navigating biological
863 complexity in breeding improved crop plants. *Trends Plant Sci*. 2006;11: 587–593.
- 864 5. Welch SM, Dong Z, Roe JL, Das S. Flowering time control: gene network modelling and the link to
865 quantitative genetics. *Crop Pasture Sci*. 2005;56: 919–936.
- 866 6. White JW, Hoogenboom G. Simulating effects of genes for physiological traits in a process-oriented crop
867 model. *Agron J*. 1996;88: 416–422.
- 868 7. Yin X, Stam P, Kropff MJ, Schapendonk AH. Crop modeling, QTL mapping, and their complementary role in
869 plant breeding. *Agron J*. 2003;95: 90–98.
- 870 8. Wit CT de. Photosynthesis of leaf canopies. Wageningen: Centre for Agricultural Publications and
871 Documentation; 1965.
- 872 9. Meuwissen TH, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker
873 maps. *Genetics*. 2001;157: 1819–1829.
- 874 10. Reymond M, Muller B, Leonardi A, Charcosset A, Tardieu F. Combining quantitative trait Loci analysis and
875 an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to
876 temperature and water deficit. *Plant Physiol*. 2003;131: 664–675. doi:10.1104/pp.013839
- 877 11. Jones JW, Hoogenboom G, Porter CH, Boote KJ, Batchelor WD, Hunt L, et al. The DSSAT cropping system
878 model. *Eur J Agron*. 2003;18: 235–265.
- 879 12. White JW, Hoogenboom G. Crop response to climate: ecophysiological models. *Climate change and food*
880 *security*. Springer; 2010. pp. 59–83.

- 881 13. Chenu K, Chapman SC, Tardieu F, McLean G, Welcker C, Hammer GL. Simulating the Yield Impacts of
882 Organ-Level Quantitative Trait Loci Associated With Drought Response in Maize: A “Gene-to-Phenotype”
883 Modeling Approach. *Genetics*. 2009;183: 1507–1523. doi:10.1534/genetics.109.105429
- 884 14. Hammer GL, Woodruff DR, Robinson JB. Effects of climatic variability and possible climatic change on
885 reliability of wheat cropping—a modelling approach. *Agric For Meteorol*. 1987;41: 123–142.
- 886 15. Hoogenboom G, Jones JW, Wilkens PW, Porter CH, Hunt LA, Singh U, et al. Decision Support System for
887 Agrotechnology Transfer (DSSAT) version 4.5 (<http://dssat.net>). Prosser, Washington; 2015.
- 888 16. Keating BA, Carberry PS, Hammer GL, Probert ME, Robertson MJ, Holzworth D, et al. An overview of
889 APSIM, a model designed for farming systems simulation. *Eur J Agron*. 2003;18: 267–288.
- 890 17. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, et al. Genetic Properties of the Maize
891 Nested Association Mapping Population. *Science*. 2009;325: 737–740. doi:10.1126/science.1174320
- 892 18. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, et al. The genetic architecture of
893 maize flowering time. *Science*. 2009;325: 714–718.
- 894 19. Hung H-Y, Shannon LM, Tian F, Bradbury PJ, Chen C, Flint-Garcia SA, et al. ZmCCT and the genetic basis of
895 day-length adaptation underlying the postdomestication spread of maize. *Proc Natl Acad Sci*. 2012;109:
896 E1913–E1921. doi:10.1073/pnas.1203189109
- 897 20. Wallach D, Goffinet B, Bergez J-E, Debaeke P, Leenhardt D, Aubertot J-N. Parameter Estimation for Crop
898 Models. *Agron J*. 2001;93: 757. doi:10.2134/agronj2001.934757x
- 899 21. Grimm SS, Jones JW, Boote KJ, Hesketh JD. Parameter estimation for predicting flowering date of soybean
900 cultivars. *Crop Sci*. 1993;33: 137–144.
- 901 22. Mavromatis T, Boote K, Jones J, Wilkerson G, Hoogenboom G. Repeatability of model genetic coefficients
902 derived from soybean performance trials across different states. *Crop Sci*. 2002;42: 76–89.
- 903 23. Thorp KR, DeJonge KC, Kaleita AL, Batchelor WD, Paz JO. Methodology for the use of DSSAT models for
904 precision agriculture decision support. *Comput Electron Agric*. 2008;64: 276–285.
- 905 24. Hunt L, White J, Hoogenboom G. Agronomic data: advances in documentation and protocols for exchange
906 and use. *Agric Syst*. 2001;70: 477–492.
- 907 25. Román-Paoli E, Welch SM, Vanderlip RL. Comparing genetic coefficient estimation methods using the
908 CERES-Maize model. *Agric Syst*. 2000;65: 29–41.
- 909 26. Koduru P, Welch SM, Das S. A particle swarm optimization approach for estimating parameter confidence
910 regions. Proceedings of the 9th annual conference on Genetic and evolutionary computation. ACM; 2007.
911 pp. 70–77. Available: <http://dl.acm.org/citation.cfm?id=1276969>

- 912 27. He J, Jones JW, Graham WD, Dukes MD. Influence of likelihood function choice for estimating crop model
913 parameters using the generalized likelihood uncertainty estimation method. *Agric Syst.* 2010;103: 256–
914 264. doi:10.1016/j.agsy.2010.01.006
- 915 28. Welch SM, Zhang, J, Sun N, Mak TY. Efficient estimation of genetic coefficients of crop models. *The Third*
916 *International Symposium on System Approaches for Agricultural Development.* 2000.
- 917 29. Irmak A, Jones JW, Mavromatis T, Welch SM, Boote KJ, Wilkerson GG. Evaluating methods for simulating
918 soybean cultivar responses using cross validation. *Agron J.* 2000;92: 1140–1149.
- 919 30. Luo Y, Weng E, Wu X, Gao C, Zhou X, Zhang L. Parameter identifiability, constraint, and equifinality in data
920 assimilation with ecosystem models. *Ecol Appl.* 2009;19: 571–574.
- 921 31. Medlyn BE, Robinson AP, Clement R, McMurtrie RE. On the validation of models of forest CO₂ exchange
922 using eddy covariance data: some perils and pitfalls. *Tree Physiol.* 2005;25: 839–857.
- 923 32. Jones CA, Richie JT, Kiniry JR, Godwin DC. Subroutine structure. CERES-Maize Simul Model Maize Growth
924 Dev CA Jones JR Kiniry Contrib PT Dyke AI. 1986; Available: [http://agris.fao.org/agris-](http://agris.fao.org/agris-search/search.do?recordID=US201301749745)
925 [search/search.do?recordID=US201301749745](http://agris.fao.org/agris-search/search.do?recordID=US201301749745)
- 926 33. Kiniry JR, Bonhomme R. Predicting maize phenology. *Predict Crop Phenol.* 1991;11: 5–131.
- 927 34. Major DJ, Kiniry JR. Predicting daylength effects on phenological processes. *Predict Crop Phenol.* 1991;
928 15–28.
- 929 35. Bratzel F, Turck F. Molecular memories in the regulation of seasonal flowering: from competence to
930 cessation. *Genome Biol.* 2015;16: 1.
- 931 36. Valentim FL, van Mourik S, Posé D, Kim MC, Schmid M, van Ham RC, et al. A quantitative and dynamic
932 model of the Arabidopsis flowering time gene regulatory network. *PLoS One.* 2015;10: e0116973.
- 933 37. Wilczek AM, Roe JL, Knapp MC, Cooper MD, Lopez-Gallego C, Martin LJ, et al. Effects of genetic
934 perturbation on seasonal life history plasticity. *Science.* 2009;323: 930–934. doi:10.1126/science.1165826
- 935 38. Dong Z, Danilevskaya O, Abadie T, Messina C, Coles N, Cooper M. A gene regulatory network model for
936 floral transition of the shoot apex in maize and its dynamic modeling. *PLoS One.* 2012;7: e43450.
- 937 39. Welch SM, Wilkerson G, Whiting K, Sun N, Vagts T, Buol G, et al. Estimating soybean model genetic
938 coefficients from private–sector variety performance trial data. *Trans ASAE.* 2002;45: 1163.
- 939 40. Lee M, Sharopova N, Beavis WD, Grant D, Katt M, Blair D, et al. Expanding the genetic map of maize with
940 the intermated B73 \times Mo17 (IBM) population. *Plant Mol Biol.* 2002;48: 453–461.
- 941 41. Flint-Garcia SA, Thuillet A-C, Yu J, Pressoir G, Romero SM, Mitchell SE, et al. Maize association population:
942 a high-resolution platform for quantitative trait locus dissection. *Plant J.* 2005;44: 1054–1064.

- 943 42. Quiring SM, Legates DR. Application of CERES-Maize for within-season prediction of rainfed corn yields in
944 Delaware, USA. *Agric For Meteorol.* 2008;148: 964–975.
- 945 43. Gill PE, Murray W, Wright MH. *Practical optimization.* Academic Press; 1981.
- 946 44. Press WH, Teukolsky SA, Vetterling WT, Flannery BP. *Numerical recipes in FORTRAN: the art of scientific*
947 *computing* [Internet]. Cambridge University Press Cambridge; 1992. Available:
948 <http://library.wur.nl/WebQuery/groenekennis/571574>
- 949 45. Saltelli A, Annoni P, Azzini I, Campolongo F, Ratto M, Tarantola S. Variance based sensitivity analysis of
950 model output. Design and estimator for the total sensitivity index. *Comput Phys Commun.* 2010;181: 259–
951 270.
- 952 46. Burhenne S, Jacob D, Henze GP. Sampling based on Sobol’ sequences for Monte Carlo techniques applied
953 to building simulations. *Proc Int Conf Build Simulat.* 2011. pp. 1816–1823. Available:
954 https://www.ibpsa.org/proceedings/BS2011/P_1590.pdf
- 955 47. Matsumoto M, Nishimura T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-
956 random number generator. *ACM Trans Model Comput Simul TOMACS.* 1998;8: 3–30.
- 957 48. Laurila H, Mäkelä P, Kleemola J, Peltonen J. A comparative ideotype, yield component and cultivation
958 value analysis for spring wheat adaptation in Finland. *Agric Food Sci.* 2012;21: 384–408.
- 959 49. Semenov MA, Stratonovitch P. Designing high-yielding wheat ideotypes for a changing climate. *Food*
960 *Energy Secur.* 2013;2: 185–196.
- 961 50. Andrés F, Coupland G. The genetic basis of flowering responses to seasonal cues. *Nat Rev Genet.* 2012;13:
962 627–639.
- 963 51. Mascheretti I, Turner K, Brivio RS, Hand A, Colasanti J, Rossi V. Florigen-encoding genes of day-neutral and
964 photoperiod-sensitive maize are regulated by different chromatin modifications at the floral transition.
965 *Plant Physiol.* 2015;168: 1351–1363.
- 966 52. Welch SM, Roe JL, Das S, Dong Z, He R, Kirkham MB. Merging genomic control networks and soil-plant-
967 atmosphere-continuum models. *Agric Syst.* 2005;86: 243–274.
- 968 53. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *ArXiv Prepr*
969 *ArXiv14065823.* 2014; Available: <http://arxiv.org/abs/1406.5823>
- 970 54. Akaike H. Maximum likelihood identification of Gaussian autoregressive moving average models.
971 *Biometrika.* 1973;60: 255–265.
- 972 55. Schwarz G. Estimating the Dimension of a Model. *Ann Stat.* 1978;6: 461–464.
973 doi:10.1214/aos/1176344136

- 974 56. Harrison SR, Tamaschke HU. Applied statistical analysis. Prentice-Hall of Australia; 1984.
- 975 57. Das S, Suganthan PN. Differential evolution: a survey of the state-of-the-art. IEEE Trans Evol Comput.
976 2011;15: 4–31.
- 977 58. Nelder JA, Mead R. A simplex method for function minimization. Comput J. 1965;7: 308–313.
- 978 59. Kennedy J. Particle swarm optimization. Encyclopedia of machine learning. Springer; 2011. pp. 760–766.
979 Available: http://link.springer.com/10.1007/978-0-387-30164-8_630
- 980 60. Batchelor WD, Basso B, Paz JO. Examples of strategies to analyze spatial and temporal yield variability
981 using crop models. Eur J Agron. 2002;18: 141–158.
- 982 61. Chew YH, Wenden B, Flis A, Mengin V, Taylor J, Davey CL, et al. Multiscale digital Arabidopsis predicts
983 individual organ and whole-organism growth. Proc Natl Acad Sci. 2014;111: E4127–E4136.
- 984 62. Tardieu F. Virtual plants: modelling as a tool for the genomics of tolerance to water deficit. Trends Plant
985 Sci. 2003;8: 9–14.
- 986 63. Fall S, Watts A, Nielsen-Gammon J, Jones E, Niyogi D, Christy JR, et al. Analysis of the impacts of station
987 exposure on the US Historical Climatology Network temperatures and temperature trends. J Geophys Res
988 Atmospheres. 2011;116. Available: <http://onlinelibrary.wiley.com/doi/10.1029/2010JD015146/full>
- 989 64. Rabotyagov S, Campbell T, Valcu A, Gassman P, Jha M, Schilling K, et al. Spatial multiobjective optimization
990 of agricultural conservation practices using a SWAT model and an evolutionary algorithm. J Vis Exp JoVE.
991 2012; e4009. doi:10.3791/4009
- 992 65. Tatsumi K. Effects of automatic multi-objective optimization of crop models on corn yield reproducibility
993 in the U.S.A. Ecol Model. 2016;322: 124–137. doi:10.1016/j.ecolmodel.2015.11.006
- 994 66. Hess CP, Liang Z-P, Lauterbur PC. Maximum cross-entropy generalized series reconstruction. 5th IEEE
995 EMBS International Summer School on Biomedical Imaging, 2002. 2002. p. 8 pp.-.
996 doi:10.1109/SSBI.2002.1233979
- 997 67. Franks SW, Beven KJ, Gash JH. Multi-objective conditioning of a simple SVAT model. Hydrol Earth Syst Sci
998 Discuss. 1999;3: 477–488.
- 999 68. Bello NM, Steibel JP, Tempelman RJ. Hierarchical Bayesian modeling of random and residual variance–
1000 covariance matrices in bivariate mixed effects models. Biom J. 2010;52: 297–313.
- 1001 69. Battenfield SD, Guzmán C, Gaynor RC, Singh RP, Peña RJ, Dreisigacker S, et al. Genomic selection for
1002 processing and end-use quality traits in the CIMMYT spring bread wheat breeding program. Plant
1003 Genome. 2016;9. Available:
1004 <https://dl.sciencesocieties.org/publications/tpg/abstracts/9/2/plantgenome2016.01.0005>

- 1005 70. Thompson LJ, Ferguson RB, Kitchen N, Frazen DW, Mamo M, Yang H, et al. Model and Sensor-Based
1006 Recommendation Approaches for In-Season Nitrogen Management in Corn. *Agron J.* 2015;107: 2020–
1007 2030.
- 1008 71. Thorp KR, Gore MA, Andrade-Sanchez P, Carmo-Silva AE, Welch SM, White JW, et al. Proximal
1009 hyperspectral sensing and data analysis approaches for field-based plant phenomics. *Comput Electron*
1010 *Agric.* 2015;118: 225–236.
- 1011 72. FBI. Smart Farming May Increase Cyber Targeting Against US Food and Agriculture Sector [Internet].
1012 United States Federal Bureau of Investigation. Private Industry Notification; 2016 Mar. Report No.:
1013 160331–1. Available: [http://docshare.tips/fbi-cyber-bulletin-smart-farming-may-increase-cyber-targeting-](http://docshare.tips/fbi-cyber-bulletin-smart-farming-may-increase-cyber-targeting-against-us-food-and-agriculture-sectorpdf_57596669b6d87f09938b4634.html)
1014 [against-us-food-and-agriculture-sectorpdf_57596669b6d87f09938b4634.html](http://docshare.tips/fbi-cyber-bulletin-smart-farming-may-increase-cyber-targeting-against-us-food-and-agriculture-sectorpdf_57596669b6d87f09938b4634.html)
- 1015 73. Tardieu F, Tuberosa R. Dissection and modelling of abiotic stress tolerance in plants. *Curr Opin Plant Biol.*
1016 2010;13: 206–212. doi:10.1016/j.pbi.2009.12.012
- 1017