**Simulation-Based Characterization of Phenology in Maize NAM Populations**

Jeffrey W. White, USDA ARS, Arid Land Agricultural Research Center

11 February 2010

*INCOMPLETE DRAFT of Internal report for the iPlant G2P project. Please do not circulate.*

**Executive Summary**

Simulation modeling provides a mechanism for characterizing plant processes at a more mechanistic level than is possible through direct observation of simple traits such as time of anthesis or plant height. Furthermore, by integrating effects of environment and crop management, a simulation model should serve as a filter to reduce effects of environment. Many models represent differences among lines or cultivars through parameters that are believed to have strong genetic control. In essence, the new phenotypic traits are the model parameters, and the broad hypothesis is that the parameters will show greater heritabilities than unprocessed traits such as time to anthesis or plant height. (Note that use of simulation modeling also has the benefit of reducing the size of the dataset for subsequent analyses.) In mapping studies, the expectation is that analysis of model parameters should provide more information than directly observed traits, reflected in detection of loci with larger effects than for conventional traits. Nonetheless, simulation-assisted mapping studies are rare. One reason is that the process of preparing data for simulation and then conducting the parameter estimation is moderately complex and is made more difficult by the potentially large numbers of lines and environments that must be simulated.

This report exercise outlines the work flow for estimating two model parameters affecting time to anthesis in approximately 5000 lines of the maize NAM populations evaluated in 11 environments (locations x years). The CSM-CERES-Maize model was used as incorporated in the DSSAT version 4.5 package. DSSAT contains various tools to assist data preparation, modeling and model analysis, but it was designed for experiments that typically involve less than 100 lines and has a limit of 999 treatments (e.g., lines by environments). The workflow is summarized in Table 1 and Fig. 1.

The first challenge encountered was that provenance data such as exact locations of trials, planting dates, soil conditions, irrigation dates and amounts, and daily weather were not readily available. The minimum data were assembled to permit running the experiments assuming no water or nutrient deficits. An initial single cultivar model control file was created using the DSSAT tool X-Build, which specified conditions for all 11 nurseries. That file was then used as a template for all lines to be simulated. DSSAT required individual files of observed days to anthesis, which were created with SAS.

CSM-CERES-Maize assumes that genetic differences in time to anthesis are controlled primarily by the duration of the juvenile phase (P1), photoperiod sensitivity (P2), and the phyllochron interval (PHINT). Phyllochron interval should be estimated based on main stem leaf number data, which were not available, so only P1 and P2 were estimated. Parameters was estimated using GenCalc2, which uses a simple grid search but has functions for semi-automated updating of the file containing cultivar parameters for the 5000 lines. Each optimization took approximately 70 seconds, so parameters were estimated over four days using three PCs. One PC crashed due to hardware problems. For several populations, it appeared that poor selection of initial parameter values had not allowed the optimization to converge, so parameters were re-estimated for those populations. GenCalc2 lacks a batch interface, so the program was controlled by a Python script that created appropriate responses to requests for input.

Initial inspection of the parameters and root mean squared errors of prediction (RMSE) indicated that the parameter estimations were reasonably successful. RMSE values were typically 2 to 4 days, which is good considering the limitations of the weather data, potential for unaccounted effects of water deficits, failure to estimate variation in PHINT, and other sources of variability. In most populations, P1 and P2 showed low correlations, suggesting that they characterized biologically different processes of development.

The biggest challenge in this process was dealing with the large numbers of lines and observations. The overall work flow appeared straightforward, although flexibility is needed throughout the process. Notably, this includes dealing with exceptions (e.g., large amounts of missing data) and inspecting intermediate results. The difficulties in assembling provenance data are a concern, especially considering that more complex traits require much better characterizations of initial conditions and management.

In terms of underlying biology, three issues are worth noting. The first is that much of the variability in the anthesis data is due to differences among environments. Simulations of a single, generic cultivar explained about 80% of the variability (Table 2, percent sums of squares). Thus, we are dealing with data where genetic differences have to be detected against a large background of environmental variability. This emphasizes the importance of provenance data and tools for accounting for effects of environment and management (E and M, so hence GEM to P). A second issue is that the large number of missing values suggests that some materials were very late flowering, so anthesis date was not observable. Simply treating the observations as missing ignores valuable information and would underestimate the level of photoperiod sensitivity. A third issue is that the phenology sub-model in CSM-CERES-Maize seems poorly suited for further work because it relies on the phyllochron interval as a surrogate for rate of reproductive development. A more mechanistic model might perform better.

Ed Buckler’s group kindly preformed a QTL analysis on P1 and P2. Only P1 showed QTL effects, and these were considered relatively small. Thus, although CSM-CERES-Maize explains a large portion of the observed variation in phenology, the fundamental hypothesis that the model coefficients would show strong QTL effects was not confirmed. More research is needed, considering:

* The CSM-CERES-Maize model as used has problems. This might include the assumption that the critical short daylength (CSD) is 12.5 h.
* The strategy for parameter estimation algorithm needs revision, including my decision to vary P1 but hold PHINT constant.
* The bias from missing values may be greater than one might hope. The population with male parent Ki3 was only represented by 129 lines in my dataset. Ki3 is elsewhere described as a "tropical yellow flint line from Suwan 1", so that might be one of the more photoperiod sensitive populations.
* The data flow was somehow bungled (e.g., lines were mismatched with data). The regression analyses would argue that the dataset is behaving as one would expect. Also, most of the data re-arranging was done with SAS, which reduces chances for major bungling.

**Introduction**

This project was undertaken to provide a concrete test case for the iPlant Genotype-to-Phenotype project, which seeks to provide the cyber-infrastructure required for quantitative prediction of phenotypes from genetic information. The goal was to identify issues related to simulating the large numbers of treatments (genotypes x environment) that might be useful in facilitating the genetic analysis of populations through QTL mapping.

One of the most powerful approaches available for associating genotypes with phenotypes is through use of mapping populations to detect QTL. For field-measured traits, the datasets typically include phenotypic data from multi-environment trials involving very large numbers of entries. One challenge is how to deal with interactions of genotypes with environment (G x E). Even for simple traits such as time of anthesis, basic physiology leads one to expect that genotypes will respond differently under different photoperiod, temperature and water regimes. Arguably, rather than analyzing raw phenotypic data, one should focus on underlying processes and traits. Ecophysiological models attempt to dynamically represent basic processes of plant growth and development, potentially providing a mechanism for characterizing the processes. Where differences in values of model parameters are required to account for differences among lines or cultivars, the assumption would be that this variation reflects important allelic variation for one or more loci. This expectation has been partially corroborated for very limited sets of germplasm in common bean, soybean, wheat, barley, maize, rice and sorghum, including at least three QTL studies. Besides re-expressing the phenotypic data as more basic, highly heritable traits, simulation-based analyses reduce the overall size of the dataset required for QTL analysis.

The reference set of phenotypic data were for the 27 populations x 200 lines evaluated in the project “Molecular and Functional Diversity of the Maize Genome” (Buckler et al., 2009; Gore et al., 2009; McMullen et al., 2009). Phenotypic data were provided through Jim Holland and his student, Hsiaoyi Hung, who is analyzing the phenology data for his PhD dissertation. Data on time to anthesis (“tasseling”) and silking were available for 11 environments, including three winter-season trials that were especially useful for estimating photoperiod sensitivity.

The reference modeling system was the Cropping Systems Model (CSM; Jones et al., 2003) as implemented in the Decision Support System for Agrotechnology Transfer version 4.5 (DSSAT; Hoogenboom et al., 2010), which contains tools to facilitate data preparation, model calibration, simulation *per se*, and viewing of outputs. DSSAT was developed mainly for conventional field experiments, typically having less than 50 treatments, so it was anticipated that the strategy for processing the data would have to be matched to the large number of treatments (> 5000 lines in 11 environments).

**The Work Flow**

The basic steps required are to prepare the data required to run the model for the set of lines and environments, estimate the model parameters of interest for the lines, and then analyze the parameters using QTL (Table 1).

**Provenance data**

 Crop simulation models need sufficient information to “reproduce” a given experiment. Ideally, these data includes date of planting, daily weather conditions (maximum and minimum temperature, solar radiation, rainfall), initial soil conditions, a description of the soil profile (or profiles if there is important variability across the field), the planting arrangement, and dates and amounts of irrigation water and nutrients applied. For the NAM populations, no provenance data were initially available in journal articles or from their web site. Hsiaoyi Hung provided the data on hand, including planting dates, partial weather data, names of locations, and partial information on the row spacing, plot size and populations used.

 Although the weather data were provided in a single MS Excel workbook, the sheets (one per location-year) varied in formatting and units. The sheets could have been processed using the DSSAT4.5 import tool, WeatherMan, but it appears that the tool can only process single sheets. Thus, rather than having to process 11 files (sheets in the workbook), a SAS program was written that read the workbook, made necessary data conversions, checked for the most obvious types of data errors, and exported the data in the required format. Additional required weather data were obtained from on-line sources such as the Florida Automated Weather Network and the NASA/POWER dataset. Approximate latitudes of the sites were obtained using Google Earth. Ideally, one should report the location of the experiment and of the corresponding weather station.

**Treatments and phenotypic data**

 Each of 26 populations contained 200 recombinant lines, while the 27th population consisted of a wide sample of germplasm. The materials were grown at six locations and for two seasons except at Ponce, PR, where a single nursery was grown. The phenotypic data were provided in CSV format, which was easily read, linked to locations and seasons, and then processed for modeling.

 The only issue of note in these data was that the file lacked an explanation of missing values. If values were missing because the plants failed to flower due to photoperiod sensitivity or intrinsic lateness that is potentially useful information, and those values should have been coded to distinguish between plots where data were not taken due to field problems (e.g., poor stand, herbicide damage or diseases) or errors in observations (e.g., plot skipped or raw data lost).

**Simulations**

 The normal approach within DSSAT4.5 is to create one control file per experiment to specify the treatments, field(s), soil conditions and crop management. The control file allows a maximum of 999 treatment levels, and it seemed likely that a large file would slow iterative parameter estimation, since 11 control files would have to be accessed for each iteration of the estimation process. The decision was made to create a template for the control files that contained a single maize line, with the eleven environments as treatments. Different lines would be simulated by replacing only the text that specifies the line being simulated. The initial template control file was created using the X-Build tool of DSSAT. Subsequent processing used SAS to extract the list of lines and Python to create the control files for each line and then launch the model for a single line.

 Although the model can simulate effects of water and nitrogen status on phenology, the deficiencies in the provenance data precluded simulating water or nitrogen dynamics. The corresponding model processes were excluded from the simulations by setting flags in the template model control file.

 Because of the large number of simulations that would be run additional changes were made to minimize run time. Most output files were turned off, and the parameter for duration of grain filling was set to a low value in order to cause very early maturation, thus halting the simulation of a single cultivar x environment. Although not measured, run times were considerably less than 1 second.

**Preparation of cross validation data**

 Model parameters are estimated by iteratively modifying the target parameters and comparing resulting simulated values with the observed data (phenotypes). Thus, the observed data for days to anthesis for all lines and environments had to be organized in a format suitable for the parameter estimation tool. The ATCreate tool in DSSAT can prepare such files, but it is meant for single experiments and lacks a batch interface. A SAS program was used to read the file of original data and create approximately 5000 files containing the observed data and respective treatment numbers required by DSSAT.

**Parameter estimation**

 CSM-CERES-Maize assumes that genetic differences in time to anthesis are controlled primarily by the duration of the juvenile phase (P1), photoperiod sensitivity (P2), and the phyllochron interval (PHINT). Phyllochron interval should be estimated based on main stem leaf number data, which were not available, the only P1 and P2 were estimated. Parameters was estimated using GenCalc2, which uses a simple grid search but has functions for semi-automated updating of the file containing cultivar parameters for the 5000 lines. GenCalc2 lacks a batch interface, so the program was controlled by a Python script that created appropriate responses to requests for input.

Each optimization took approximately 70 seconds, so parameters were estimated over four days using three PCs. (One older PC crashed due to hardware problems.) For several populations, it appeared that poor selection of initial parameter values prevented the optimization from converging, so parameters were re-estimated for those populations. In retrospect, these problems likely were due to missing values for lines that were late flowering.

**Preliminary analyses prior to QTL**

 Following any major simulation experiment, the outputs should be inspected for basic consistency. A logical first concern is how well the observed phenotypes compare with the simulated phenotypes (cross validation). Table 2 shows analyses of variance for five regressions of observed phenotypes vs. simulated values and associated explanatory variables such as population and environment. The first regression is the standard cross-validation regression (observed vs. simulated) and shows that the simulations account for approximately 91% of the total variation among lines. This comparison is also graphed in Fig. 2.

 A key but seldom examined question is whether simulating differences among lines is better than assuming just one “generic” cultivar. The second regression of Table 2 first tests the effect of a generic cultivar, whose values of P1 and P2 were the respective means for all lines. The sums of squares are estimated sequentially, so the effect for different lines represents a residual effect once the generic effect is considered. About 80% of the variation is accounted for by the generic cultivar, suggesting that most of the variability in the datasets reflects simple, essentially trivial variation among environments that does not require consideration of genetics. Of the remaining 20% of variation, the simulations assuming genetic differences among lines (i.e., allowing variation among the parameters P1 and P2) account for another 11.5%.

 The next two ANOVAs test how much of the remaining variation is somehow related unexplained effects of environment or population. Environment explained another 5.5%. This could reflect at least four factors:

1. Failure of the simulation model to account correctly for certain processes or genetic effects. One well-documented deficiency is that effect of warmer temperatures can increase photoperiod sensitivity.
2. Bias in weather data due to location of the station relative to the experiment, incorrect calibrations, etc.
3. Bias due to failure to consider effects of water or nitrogen deficits.
4. Bias among observers at different locations as to when plants were judged to have reached anthesis.

 These regression analyses are incomplete. One might consider effects of lines within populations or years within locations. There are valid arguments for using mixed modeling instead of ordinary least squares. Nonetheless, they suggest that the overall parameter estimation process was successful.

 An important genetic expectation is that if P1 and P2 correctly represent independently controlled processes and the major loci are not closely linked, then they should be uncorrelated. Correlations within populations were generally low (Table 3), but the consistent negative association suggests that was compensation between the two parameters – over estimation of one parameter was compensated for by a lower value of the other one.

**Discussion of the Work Flow *per se***

In retrospect, the basic work flow is straightforward. The required data are assembled, checked, formatted, used for calibration, and the resulting parameters exported for QTL analysis.

There is a lot of discussion in modeling communities about optimal strategies for parameter estimation, but for the maize NAM dataset, where there was a relatively limited set of observed data and parameters to estimate, the main constraint likely is not the optimization algorithm per se but the overall processing strategy for large numbers of lines or individuals. For this test case, more effort should have gone to initial data checking and testing of estimation strategies, which is a somewhat different issue than the algorithm. The initial estimates of P1 and P2 for each population might have been estimated by sampling a subset of lines first. More speculatively, a better approach for such large numbers of lines might be to create a matrix of potential values of P1 and P2, simulate expected days to anthesis for each combination of coefficients and locations, and select the parameter combination for each line that minimizes the RMSE. This could be used as the only form of parameter estimation, or it could be used to provide initial values, which would then be passed a more sophisticated estimation tool.

**Discussion of the Science**

In the CSM-CERES-Maize model, genetic differences in phenology are controlled by three parameters:

* P1 – Thermal time from seedling emergence to the end of the juvenile phase (expressed in degree days above a base temperature of 8°C) during which the plant is not responsive to changes in photoperiod.
* P2 -- Extent to which development (expressed as days) is delayed for each hour increase in photoperiod above the longest photoperiod at which development proceeds at a maximum rate (which is considered to be 12.5 hours).
* PHINT -- Phyllochron interval; the interval in thermal time (degree days) successive leaf tip appearances.

PHINT is used in the model essentially as the intrinsic rate of development, but since it should be estimated from leaf numbers, it was not estimated. Thus, only P1 and P2 were estimated. It is quite possible that the fitted values of P1 show excessive variability. Optional strategies would be to estimate PHINT with P1 and P2, to hold P1 constant and fit P2 and PHINT, or to hold P2 constant and fit P1 and PHINT. These are being explored with a subset of the lines.

 While the example presented is for phenology, the basic approach is readily extensible to other traits. Crop models such as CSM-CERES-Maize can estimate a large number of other parameters that have strong genetic influences but also vary with environment. These include grain number, grain size and grain protein content. CSM-CERES-Maize has relatively simplistic physiology, but more detailed models are available. The basic approach is applicable to traits measured on much shorter time scales that relate to processes such as photosynthesis and transpiration. With a leaf-level sub-model coupled to a canopy model, one might simulate processes on a time scale of minutes and characterize parameters such as stomatal response to soil water deficits or to humidity. Observed data might be estimated from more complex modeling using remotely sensed data for fluorescence, leaf nitrogen levels, canopy architecture and canopy temperature.

Ed Buckler’s group kindly preformed a QTL analysis on P1 and P2. Only P1 showed QTL effects, and these were considered unexpectedly small. Thus, although CSM-CERES-Maize explains a large portion of the observed variation in phenology, the fundamental hypothesis that the model coefficients would show strong QTL effects was not confirmed. It was especially alarming that the locus i1063, which has a large effect on photoperiod response, had no effect on P2. More research is needed, considering whether:

* The bias from missing values is greater than one might hope. Population 13, with male parent Ki3, is only represented by 129 lines in the dataset of estimated P1 and P2. Ki3 is elsewhere described as a "tropical yellow flint line from Suwan 1", so that might be one of the more photoperiod sensitive populations.
* The CSM-CERES-Maize model as used has problems. This might include the assumed critical shirt daylength of 12.5 h and my decision to vary P1 but hold PHINT constant. Evidence from a controlled environment study (Kiniry et al., 1983) suggests that the CSD varies among cultivars and can have values from below 10 h to over 12.5 h, but one could also question whether a CSD estimated from the constant temperature regime is quantitatively applicable to field responses.
* The parameter estimation algorithm needs revision. The grid search in GenCalc2 seemed to select a limited range of variation in P2. This may relate back to the assumed CSD.
* The data flow was somehow bungled (e.g., lines were mismatched with data). The regression analyses would seem to argue that the dataset is behaving as one would expect. Also, much of the dataset re-arranging was done with SAS, which provides strong controls on merging datasets, reduces such errors.

An important lesson may be that the model should have been tested more thoroughly before embarking on a relatively intensive period of computing and analysis. One is reminded that “fools rush in where angels fear to tread” or more secularly, “haste makes waste.”

**References**

Buckler ES, et al., 2009. The genetic architecture of maize flowering time. Science 325: 714-718.

Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, Ware DH, Buckler ES, 2009 A first-generation haplotype map of maize. Science 326: 1115-1117.

Hoogenboom et al., 2010. Decision Support System for Agrotechnology Transfer. Ver. Version 4.5. University of Hawaii, Honolulu, HI. On CD-ROM.

Jones, J.W., G. Hoogenboom, C.H. Porter, K.J. Boote, W.D. Batchelor, L.A. Hunt, P.W. Wilkens, U. Singh, A.J. Gijsman, and J.T. Ritchie. 2003. The DSSAT Cropping System Model. European Journal of Agronomy 18:235-265.

Kiniry, J.R., J.T. Ritchie, R.L. Musser, E.P. Flint, and W.C. Iwig. 1983. The Photoperiod Sensitive Interval in Maize. Agron J 75:687-690.

McMullen M, et al., 2009. Genetic properties of the maize Nested Association Mapping population. Science 325: 737-740.

**Table 1. Summary of overall workflow**

|  |  |  |
| --- | --- | --- |
| **Step in workflow** | **Processing** | **DSSAT tool** |
| Organize provenance data: |  |  |
| * Planting date
 | Entered directly into X-Build  | X-Build |
| * Daily weather
 | Processed using SAS from Excel workbook | WeatherMan |
| * Soil profiles
 | (pending) | S-Build |
| * Irrigation
 | (pending) | X-Build |
| * Fertilizers
 | (pending) | X-Build |
| * Planting arrangement
 | (two locations are pending) | X-Build |
| Convert provenance data to model-compatible formats | Used X-Build to create the template file and SAS to extract a complete list of populations and lines. Then used Python to create files on the fly for simulations. | X-Build |
| Convert phenotypic data to model-compatible formats | Used SAS to create files because AT-Create is for individual files. | AT Create |
| Prepare file for model parameters estimates of each line | Used SAS to create file | (no software in DSSAT) |
| Initial testing of model sensitivity and parameter estimation strategies | [Not done] | DSSAT S.A. &GenCalc2 |
| Parameter estimation | Used GenCalc2 but controlled with Python script that takes control of Windows interface. | GenCalc2 |
| Preliminary evaluation of results | Used Python script to run model and extract relevant data. The used SAS for regressions, plotting, etc. | G-Build |
| Estimation of loci effects | (pending) | (no software in DSSAT) |

Table 2. Analyses of variance for linear regressions with observed days to anthesis as the dependent variable. Sources of variation are introduced sequentially. The dataset excludes lines with less than six values of observed anthesis because calibrations for these lines showed large RMSE. Due to the very large dataset, all effects are significant at P << 0.001. The most meaningful indicator is the percentage of sums of squares (SS).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Source | DF | SS | Percent SS | MS | F |
| Sim. for different lines | 1 | 5055992.05 | 90.72 | 5055992.05 | 481557.97 |
| Residual | 49233 | 516909.01 | 9.28 | 10.50 |  |
|  |  |  |  |  |  |
| Sim. of generic line | 1 | 4416655.23 | 79.25 | 4416655.23 | 420674.91 |
| Sim. for different lines | 1 | 639360.31 | 11.47 | 639360.31 | 60897.404 |
| Residual | 49232 | 516885.53 | 9.27 | 10.50 |  |
|  |  |  |  |  |  |
| Sim. of generic line | 1 | 4416655.23 | 79.25 | 4416655.23 | 1232333.9 |
| Sim. for different lines | 1 | 639360.31 | 11.47 | 639360.31 | 178394.14 |
| Environment | 9 | 311175.12 | 5.58 | 34575.01 | 9647.1109 |
| Sim. by environment | 10 | 29332.19 | 0.53 | 2933.22 | 818.42607 |
| Residual | 49213 | 176378.21 | 3.16 | 3.58 |  |
|  |  |  |  |  |  |
| Sim. of generic line | 1 | 4416655.23 | 79.25 | 4416655.23 | 430322.83 |
| Sim. for different lines | 1 | 639360.31 | 11.47 | 639360.31 | 62294.049 |
| Population | 26 | 4329.30 | 0.08 | 166.51 | 16.223522 |
| Sim. by population | 26 | 7793.08 | 0.14 | 299.73 | 29.203631 |
| Residual | 49180 | 504763.14 | 9.06 | 10.26 |  |
|  |  |  |  |  |  |
| Sim. of generic line | 1 | 4416655.23 | 79.25 | 4416655.23 | 1362266 |
| Sim. for different lines | 1 | 639360.31 | 11.47 | 639360.31 | 197203.26 |
| Population | 26 | 4329.30 | 0.08 | 166.51 | 51.358539 |
| Environment | 9 | 314524.55 | 5.64 | 34947.17 | 10779.049 |
| Pop. by environment | 250 | 39338.72 | 0.71 | 157.35 | 48.534282 |
| Residual | 48947 | 158692.96 | 2.85 | 3.24 |  |
|  |  |  |  |  |  |
| Sim. of generic line | 1 | 4416655.23 | 79.25 | 4416655.23 | 1362266 |
| Population | 26 | 307623.29 | 5.52 | 11831.67 | 3649.3397 |
| Environment | 9 | 512199.07 | 9.19 | 56911.01 | 17553.539 |
| Pop. by environment | 250 | 44362.44 | 0.80 | 177.45 | 54.732311 |
| Sim. for different lines | 1 | 133368.07 | 2.39 | 133368.07 | 41135.833 |
| Residual | 48947 | 158692.96 | 2.85 | 3.24 |  |

Table 3. Mean values of P1 and P2, number of lines included, and correlation between P1 and P2 for each population. Critical value for correlation with N = 190 and P < .05 is 0.137.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Population** | **P1** | **(std)** | **P2** | **(std)** | **N** | **Correlation** | **Pedigree** | **Male parent** |
| 1 | 224.1 | 14.4 | 0.869 | 0.063 | 195 | -0.438 | (B73×B97)S5 | B97 |
| 2 | 256.8 | 14.2 | 0.857 | 0.040 | 198 | -0.329 | (B73×CML103)S5 | CML103 |
| 3 | 268.7 | 17.7 | 0.825 | 0.053 | 196 | -0.182 | (B73×CML228)S5 | CML228 |
| 4 | 275.5 | 17.6 | 0.817 | 0.052 | 198 | -0.157 | (B73×CML247)S5 | CML247 |
| 5 | 267.1 | 15.3 | 0.858 | 0.049 | 189 | -0.164 | (B73×CML277)S5 | CML277 |
| 6 | 251.1 | 12.5 | 0.843 | 0.036 | 186 | -0.131 | (B73×CML322)S5 | CML322 |
| 7 | 255.9 | 12.2 | 0.827 | 0.067 | 194 | -0.089 | (B73×CML333)S5 | CML333 |
| 8 | 283.0 | 15.6 | 0.845 | 0.059 | 198 | -0.525 | (B73×CML52)S5 | CML52 |
| 9 | 264.5 | 11.8 | 0.842 | 0.058 | 192 | -0.177 | (B73×CML69)S5 | CML69 |
| 10 | 231.2 | 11.0 | 0.833 | 0.086 | 195 | 0.213 | (B73×Hp301)S5 | Hp301 |
| 11 | 218.6 | 11.0 | 0.685 | 0.121 | 196 | 0.052 | (B73×Il14H)S5 | Il14H |
| 12 | 263.1 | 12.8 | 0.843 | 0.051 | 197 | -0.159 | (B73×Ki11)S5 | Ki11 |
| 13 | 243.0 | 13.7 | 0.831 | 0.062 | 129 | -0.102 | (B73×Ki3)S5 | Ki3 |
| 14 | 246.4 | 11.4 | 0.856 | 0.040 | 197 | -0.050 | (B73×Ky21)S5 | Ky21 |
| 15 | 257.6 | 14.8 | 0.829 | 0.053 | 190 | -0.252 | (B73×M162W)S5 | M162W |
| 16 | 256.9 | 13.0 | 0.826 | 0.066 | 195 | -0.245 | (B73×M37W)S5 | M37W |
| 17 | 219.5 | 15.0 | 0.777 | 0.148 | 200 | -0.367 | ?? |  |
| 18 | 262.2 | 17.1 | 0.846 | 0.023 | 196 | -0.066 | (B73×Mo18W)S5 | Mo18W |
| 19 | 214.3 | 9.0 | 0.739 | 0.142 | 199 | -0.570 | (B73×MS71)S5 | MS71 |
| 20 | 252.0 | 13.2 | 0.823 | 0.056 | 191 | -0.225 | (B73×NC350)S5 | NC350 |
| 21 | 240.1 | 11.5 | 0.845 | 0.035 | 190 | -0.435 | (B73×NC358)S5 | NC358 |
| 22 | 231.8 | 9.4 | 0.800 | 0.113 | 195 | -0.003 | (B73×Oh43)S5 | Oh43 |
| 23 | 244.8 | 11.1 | 0.850 | 0.061 | 191 | -0.045 | (B73×Oh7B)S5 | Oh7B |
| 24 | 210.4 | 12.6 | 0.687 | 0.126 | 189 | -0.212 | (B73×P39)S5 | P39 |
| 25 | 260.3 | 14.1 | 0.853 | 0.039 | 175 | -0.185 | (B73×Tx303)S5 | Tx303 |
| 26 | 257.7 | 13.5 | 0.858 | 0.036 | 195 | -0.107 | (B73×Tzi8)S5 | Tzi8 |
| 27 | 257.5 | 33.2 | 0.843 | 0.085 | 281 | 0.186 | [Diverse lines] |  |

Figure 1. Outline of work flow used to estimate model coefficients for the 5000 NAM lines.



Figure 2. Comparison of observed vs. simulated days to anthesis for all populations. Note that the total number of observations is > 50,000, so many points overlap. R2 = 0.91(prob. < 0.00001). RMSE = 3.2 days. Observed = -3.92 + 1.05 \* Simulated.

**Appendix A. Detailed listing of issues encountered while processing the data. (Probably not of interest for most readers.)**

**Provenance data**

Crop management data:

1. Planting dates provided by H. Hung.
2. Row spacing and populations (approximate in some cases) provided by H. Hung for four of six sites.
3. Plot sizes are small and may have bias results due to border effects.
4. No irrigation data were available.
5. No fertilizer data were available.
6. No additional information on crop condition, residues present at planting, etc. were available.

Weather data:

1. Different formats – increased labor for preparing files and chances of errors not being detected.
2. Insufficient days of data – for Aurora, additional days are needed, especially for late season. For several sites, the data start at planting, which doesn’t permit running a fallow before hand to estimate initial soil moisture (in the absence of measured data).
3. Lack of solar radiation data, data in incorrect units, outliers, etc. – I used data from NASA/POWER to fill but even these have missing values.
4. No information on location of weather stations in relation to actual experiments.

Utility of DSSAT for provenance data:

1. Model control file (“file-X”): X-Build was used to create a single template containing

**Initial Simulations**

1. 2 char + 4 digit cultivar id

**Comments on Science in the CSM-CERES-Maize Model**

1. Genetic differences in phenology are controlled by three parameters:
	1. P1 – Thermal time from seedling emergence to the end of the juvenile phase (expressed in degree days above a base temperature of 8°C) during which the plant is not responsive to changes in photoperiod.
	2. P2 -- Extent to which development (expressed as days) is delayed for each hour increase in photoperiod above the longest photoperiod at which development proceeds at a maximum rate (which is considered to be 12.5 hours).
	3. PHINT -- Phylochron interval; the interval in thermal time (degree days) successive leaf tip appearances.

PHINT is used in the model essentially as the intrinsic rate of development, but since it should be estimated from leaf numbers, it was not estimated. Thus, only P1 and P2 were estimated. It is quite possible that the fitted values of P1 show excessive variability. Optional strategies would be to estimate PHINT with P1 and P2, hold P1 constant and fit P2 and PHINT, hold P2 constant and fit P1 and PHINT. This is where a cluster for fitting coefficients would be useful.

1. Phenology in maize is affected by factors other than temperature and daylength. Photoperiod sensitivity