Development of a Generalized Additive Model (GAM) for Soybean Maturity Prediction in African Environments.

SUPPLEMENTARY MATERIALS

Analysis of residuals and influential observations.

Cook's coefficients are distance measures calculated as the residual obtained by fitting a model with all the observations relative to a model with an observation removed at the time. This difference is normalized by the inverse of an error or cost metric (e.g. RMSE). A higher value for the coefficient indicates a higher likelihood for an observation being too much influential in the overall estimation of model parameters. Distance metrics were adjusted for genotype-grouping level effects.

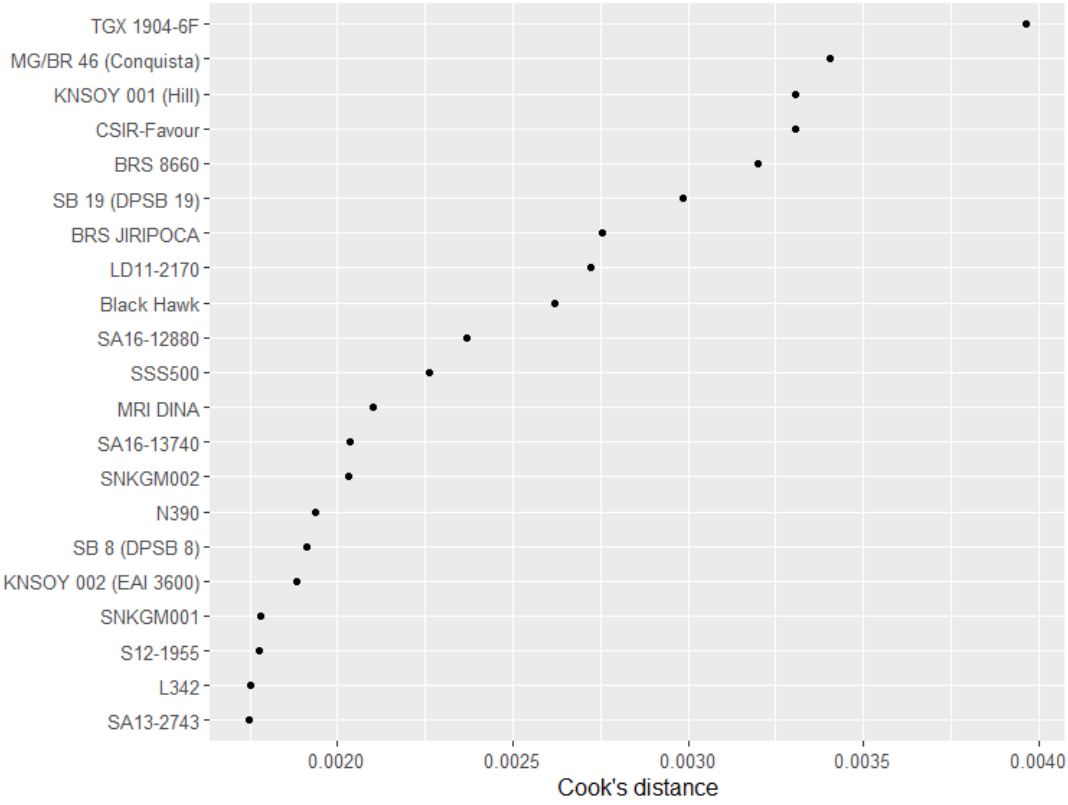

Figure S1. Ranking of influential genotypes sorted by Cook's coefficient values

Influential analysis based on Cook's distance should not necessarily lead to automatic deletion of an observation. Here, we complemented the influence analysis with visual inspection of residuals. We assumed an acceptance threshold region bounded by 3-times the standard error of the residuals from the random effects model. Overall, 28 outliers (GxLxS) departed from this assumption (1.5 % of all BLUPS). Furthermore, only three out of 21 genotypes with a large distance coefficient were part of an environment with a large residual. (L342-Chilanga-2019, N390-Chilanga-2019, and SP 8 DPSB – Thika 2016/2017). The random effects model was re-estimated with these three observations removed.

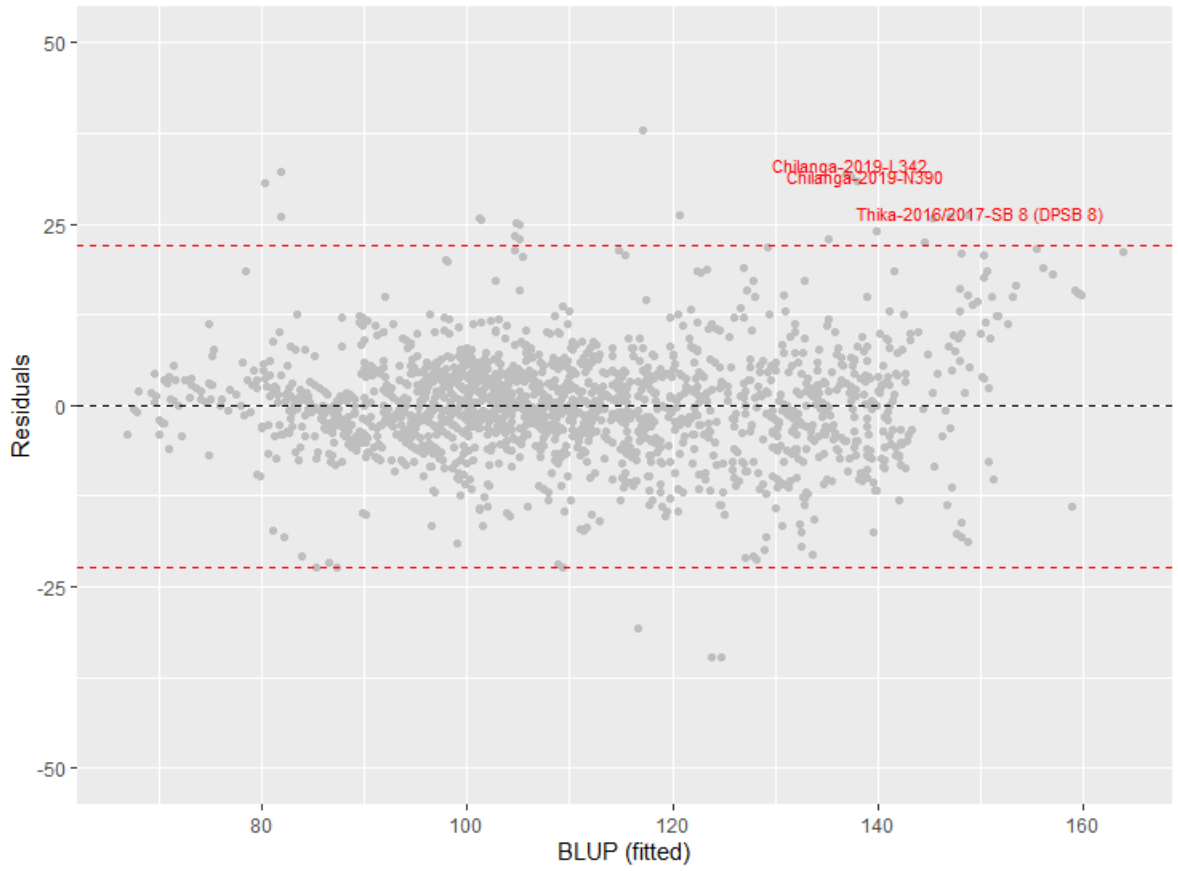

Figure S2. Outliers corresponding to observations displaying both a large Cook's coefficient and a large residual.

Modeling: Training/testing sets based on planting date availability:

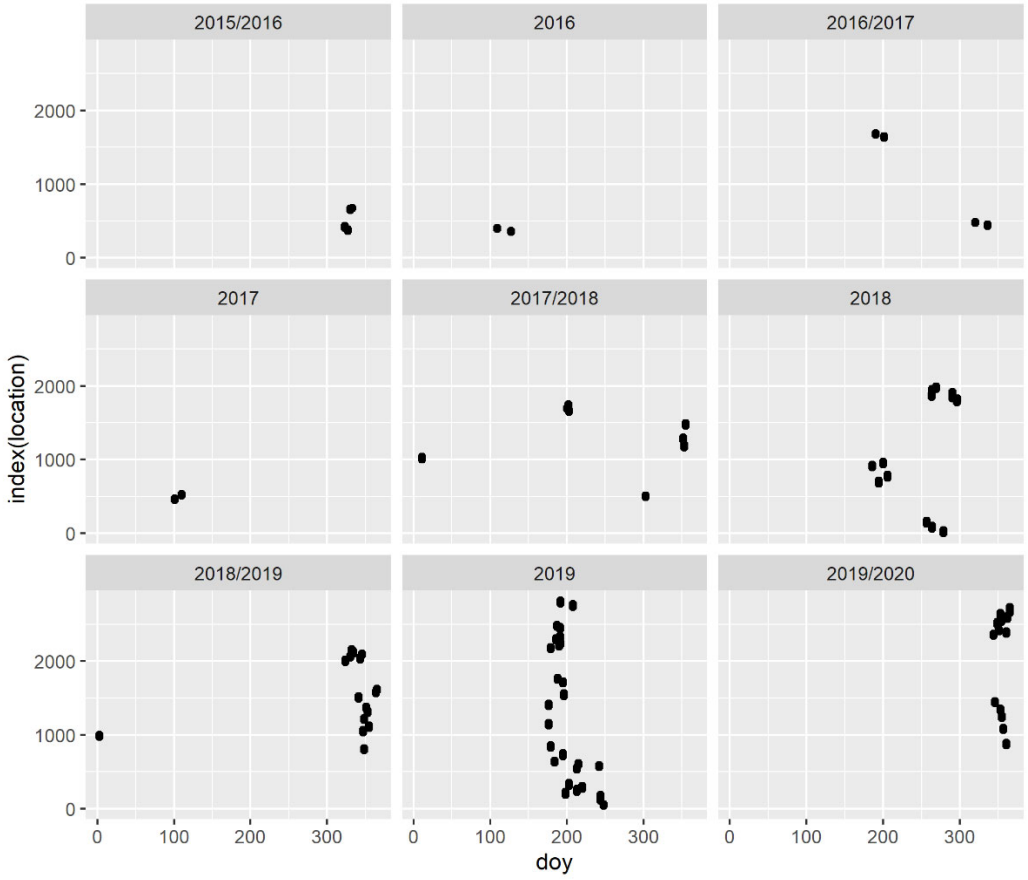

Figure S3. Planting date variation in the SIL-PAT Network (2015-2020). Numbers in the y-axis are identification labels for the trialing fields in the database. The x-axis shows planting dates for a given field converted to Julian-days (day of year, doy). Crops planted between-years (~ day of year > 300) are indicative of a winter season. Summer crops instead are planted around the mid-months within a year (~ day of year 200). To account for this variation and use the most information available during the training phase, the training and testing datasets included at least a winter and summer soybean crop for the 2018/2019, 2019, and 2019/2020 seasons. Data from previous years were held-out for model validation.

Modeling: Feature Selection

Beginning with one predictor at a time, stepwise forward regression was performed at (n+1) steps. No further improvement was noted after the seventh iteration, which included the features TMINMM, TMEANM, TMEANDIFF, DLMEANDIFF, ALT, lat, and long.

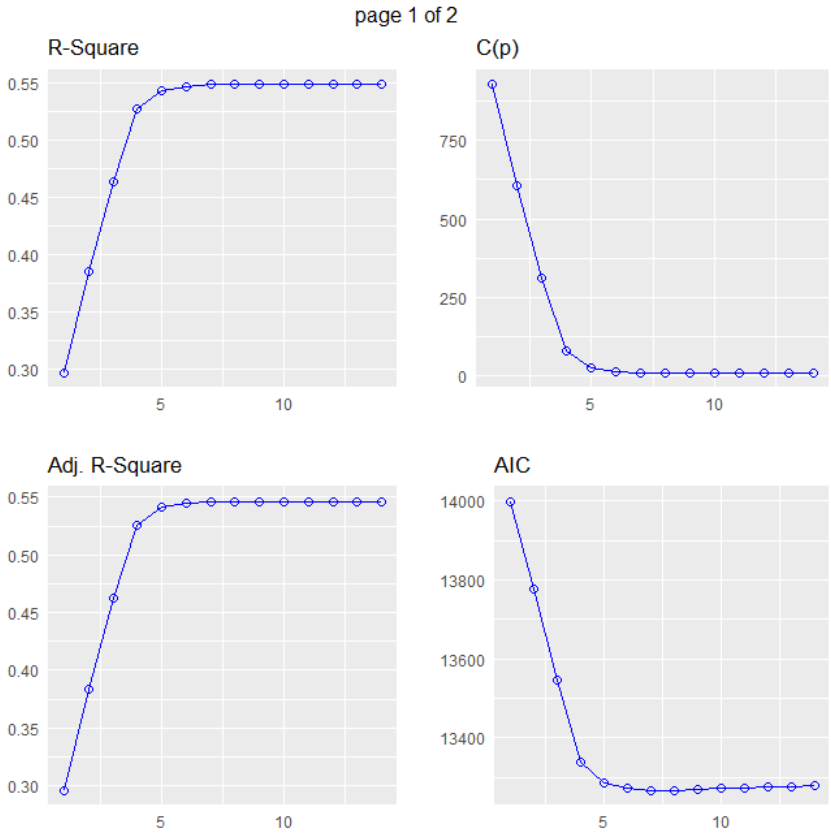

Figure S4. Stepwise forward selection for the best feature-set to predict maturity times using seasonal variables. Model agreement improved no more after the seventh iteration. The best candidate features were TMINMM, TMEANM, TMEANDIFF, DLMEANDIFF, ALT, lat, and long.

To remove collinearity, models with one temperature-based predictor at the time along with DLMEANDIFF and geolocation (lat, long) were evaluated independently. Also, a model with ALT was considered. The best feature subset was:

TMINMM +DLMEANDIFF+ latitude + longitude (AIC= 13, 817, R^2 = 0.36)

This final feature-set was used next to enhance the prediction accuracy of soybean TTM via GAM modeling.

Table S1. Evaluation and testing of GAM models used to predict Soybean maturity timing. Linear regression models (lm) included for comparison.

Model	5-fold validation				All-train	
	RMSE (days)		R^2-adjusted			
	Training	Testing	Training	Testing	AIC	BIC
gam s(TMINM, 3)	13.55	13.67	0.46	0.46	13542	13564
gam s(TMINM, 6)	13.17	13.31	0.49	0.48	16609	16646
s(TMINM, 3) + s(lat, long, 4)	13.17	13.35	0.48	0.48	16617	16657
lm(TMINM + long)	--	16.24	--	0.19	20057	20080
gam s(TMINM,3)+s(long,4)	13.14	13.32	0.49	0.48	16631	16670
gam s(TMINM, 3) + long	13.38	13.49	0.47	0.47	16669	16697
gam s(TMINM, 6) + s(long, 6)	11.95	12.30	0.58	0.56	16236	16304
gam s(TMINM, 6) + s(long, 4)	12.38	12.65	0.54	0.53	16370	16426
gam s(TMINM, 6) + long	12.54	12.73	0.53	0.53	16415	16460
lm(TMINM+long+lat)	--	16.25	--	0.18	20059	20087
gam s(TMINM,6)+s(lat,long,4)	12.43	12.67	0.54	0.53	16391	16497
gam s(TMINM,6)+s(lat,long,6)	12.41	12.67	0.54	0.53	16392	16457
gam s(TMINM,6)+s(long,4)+s(lat,4)	12.20	12.51	0.56	0.54	16322	16395
gam s(TMINM,6)+s(long,4)+s(lat,6)	11.80	12.08	0.59	0.57	16037	16122
gam S(TMINM,6)+s(long,6)+s(lat,4)	11.72	12.04	0.59	0.58	16191	16272
gam s(TMINM,6)+s(long,6)+s(lat,6)	11.65	12.01	0.60	0.58	15947	16043
gam s(TMINM,6)+s(long,6)+lat	11.95	12.23	0.57	0.56	16193	16266
lm(TMINM+long+lat+DLMEANDIFF)	14.90	15.48	0.34	0.32	13817	13850
gam s(TMINM,4)+s(long,4)+s(lat,4)+ DLMEANDIFF	11.15	11.40	0.63	0.62	15944	16010
gam s(TMINM,3)+s(long,6)+s(lat,6)+ DLMEANDIFF	10.76	11.10	0.65	0.64	15803	15587

The indicator s is the basis function with k-break points or knots. e.g., s(TMINM, 3) is the response of maturity to TMINM with 3 knots

Table S1. (Continuation) Evaluation and testing of GAM models used to predict Soybean maturity timing. Linear regression (lm) included also for comparison.

Model	5-fold validation				All-train	
	RMSE (days)		R^2-adjusted			
	Training	Testing	Training	Testing	AIC	BIC
lm(TMINM)	15.43	15.79	0.29	0.30	14000	14016
lm(TMINM+ DLMEANDIFF)	14.96	15.48	0.32	0.33	13924	13945
lm(TMINM + long)	--	16.24	--	0.19	20057	20080
gam s(TMINM, 3) + long	13.38	13.49	0.47	0.47	16669	16697
gam S(TMIN, 3) + DLMEANDIFF	12.50	12.70	0.54	0.53	13280	13307
gam S(TMIN, 3) + s(DLMEANDIFF,3)	10.78	11.02	0.66	0.65	12793	12825
gam s(TMINM, 6) + s(long, 6)	11.95	12.30	0.58	0.56	16236	16304
gam s(TMINM, 6) + s(long, 4)	12.38	12.65	0.54	0.53	16370	16426
lm(TMINM+long+lat)	--	16.25	--	0.18	20059	20087
gam s(TMINM,6)+s(lat,long,6)	12.41	12.67	0.54	0.53	16392	16457
gam s(TMINM,6)+s(long,4)+s(lat,6)	11.80	12.08	0.59	0.57	16037	16122
gam S(TMINM,6)+s(long,6)+s(lat,4)	11.72	12.04	0.59	0.58	16191	16272
gam s(TMINM,3)+s(long,6)+s(lat,6)	11.65	12.01	0.60	0.58	15947	16043
lm(TMINM+long+lat+DLMEANDIFF)	14.90	15.48	0.34	0.32	13817	13850
gam S(TMINM,3) + DLMEANDIFF	12.50	12.70	0.54	0.53	13280	13307
gam S(TMINM,3) + s(DLMEANDIFF,3)	10.78	11.02	0.66	0.65	12793	12825
gam s(TMINM,3)+s(long,6)+s(lat,6)+DLMEANDIFF	10.76	11.10	0.65	0.64	15803	15587
gam S(TMINM,3) + S(DLMEANDIFF,3)+lat + long	10.05	10.3	0.70	0.69	12576	12620
gam S(TMIN,3)+s(DLMEANDIFF,3)+s(lat,long,4)	9.99	10.35	0.70	0.69	12564	12613

The indicator s is the basis function with k-break points or knots. e.g., s(TMINM, 3) is the response of maturity to TMINM with 3 knots

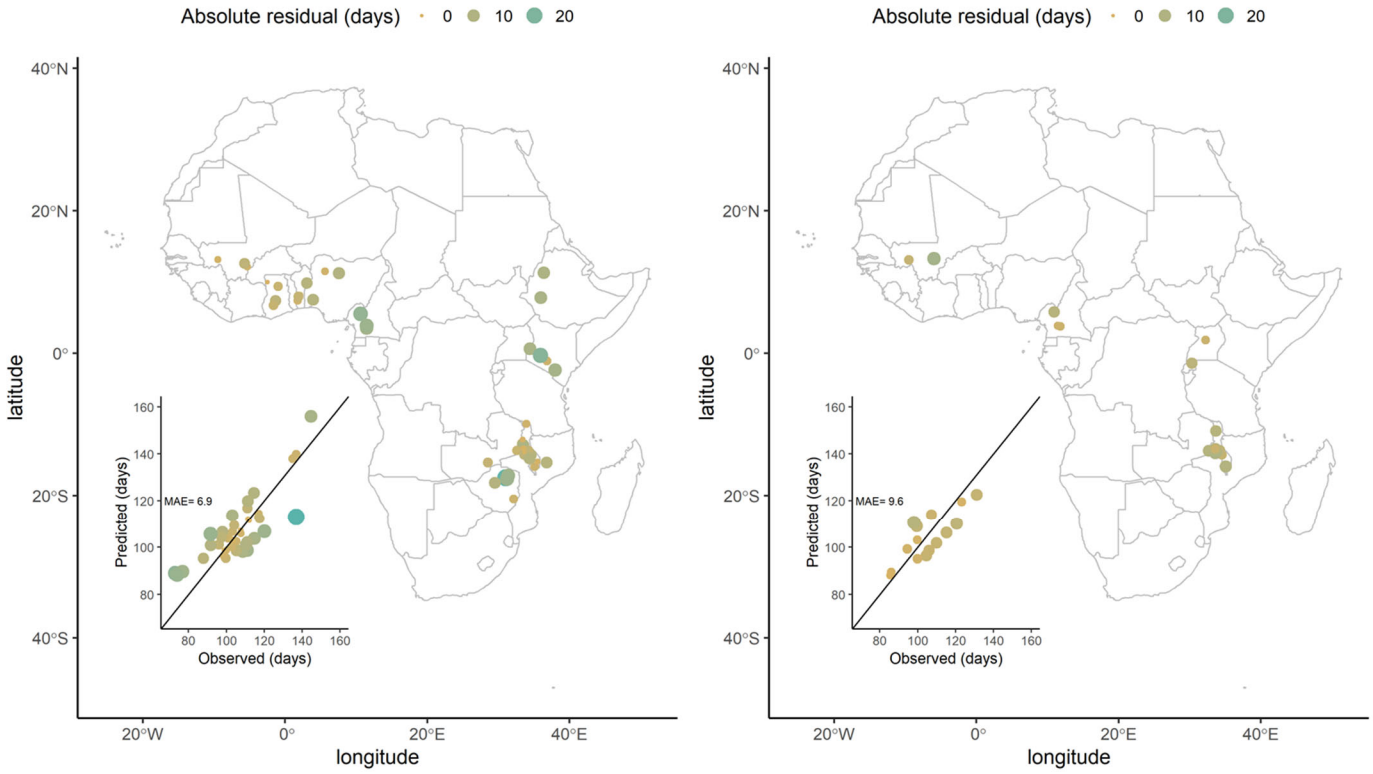

Figure S5. Geographical distribution of Soybean Time to Maturity (TTM) predictions from the GAM Model. Each dot represents the mean absolute error for all the cultivars in a given environment. Plots at the bottom left corner in each plot display the agreement between observed and predicted TTM (Dots closer to the 1:1 diagonal are environments with a better match). Left and Right panels are for the training/testing and validation phases.

Time to maturity Cultivar x Environment Interactions

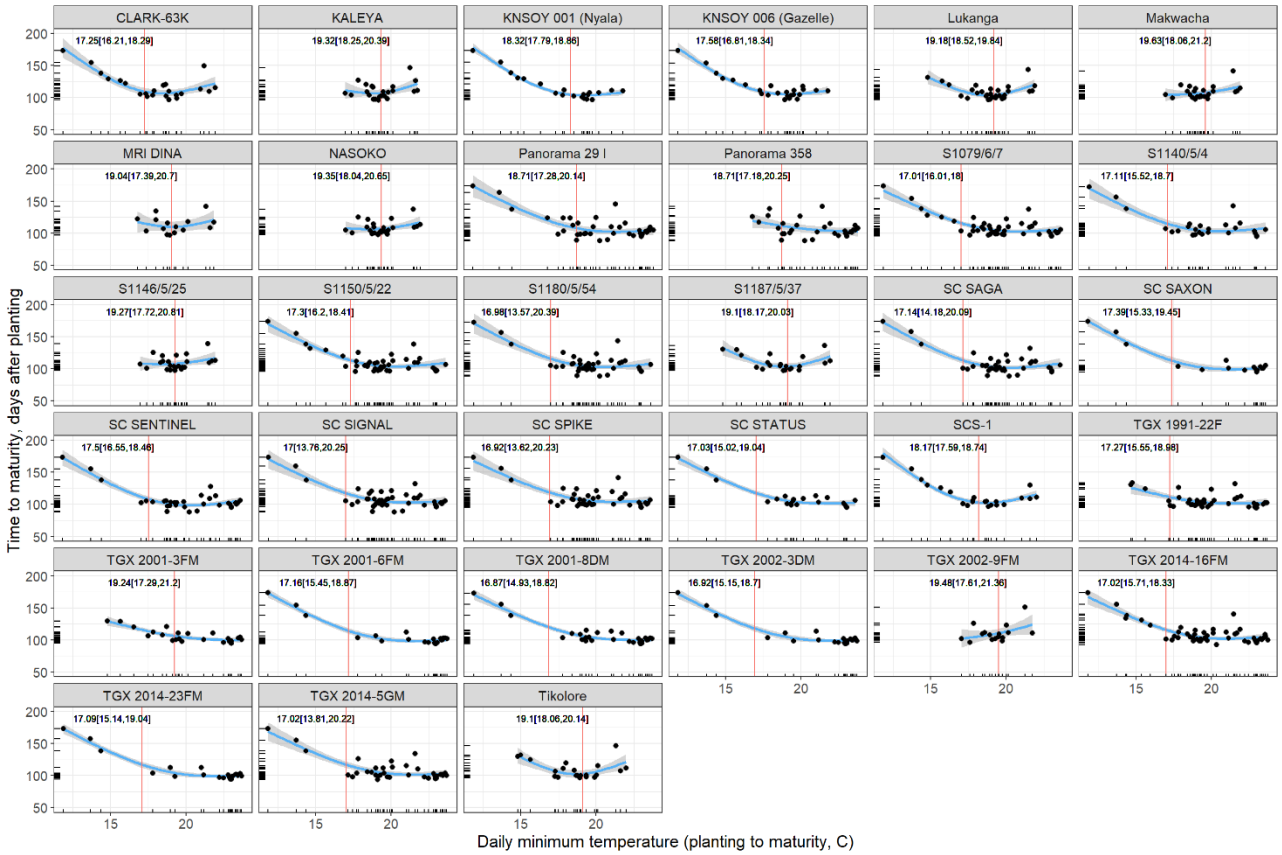

Figure S6. GAM smoothed response of soybean maturity time to minimum temperature. Each dot represents a testing site where the cultivar was tested. Red parallel lines represent a change in the direction of the response, approximated using segmented regression. 33 cultivars showed a significant relationship.

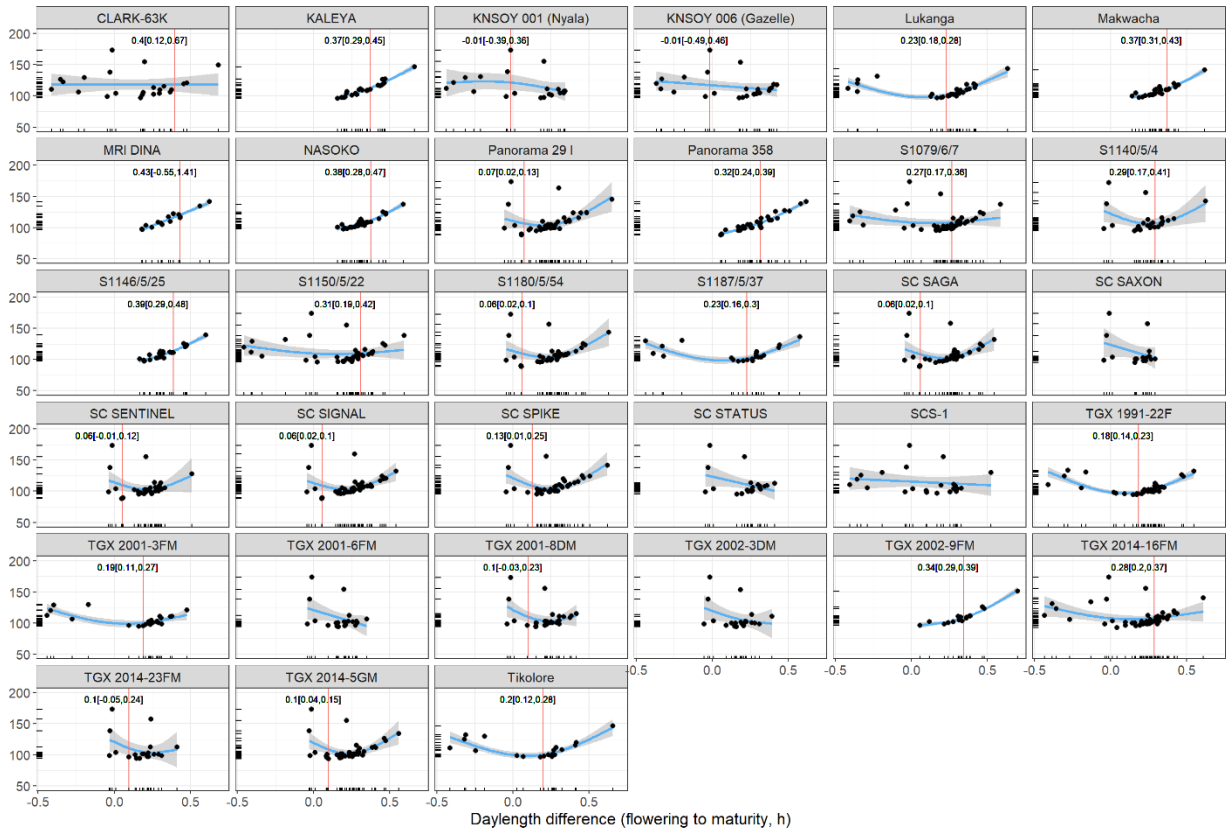

Figure S7. GAM smoothed response of soybean maturity time to post-flowering daylength. Each dot represents a testing site where the cultivar was tested. Red parallel lines represent a change in the direction of the response, approximated using segmented regression. 5 out of 33 cultivars failed to display a significant change in response.