

MODELLING CASH CROP GROWTH IN TENNESSEE IN RESPONSE TO CLIMATE

Spencer Weston

Submitted in partial fulfillment of the requirements for Chancellor's
Honors and Bachelor of Arts with Honors in Geography at the
University of Tennessee – Knoxville

Advised by Dr. Nicholas Nagle, Associate Professor of Geography

May 8, 2017

Contents

Abstract.....	2
Introduction	3
Data and Data Preparation	5
Linear and Generalized Additive Methods	6
Linear and Generalized Additive Results	8
Generalized Linear Mixed Model Methods	10
Generalized Linear Mixed Model Results	14
Discussion.....	19
Appendices.....	22
Appendix A-Introduction	22
Appendix B- Data and Data Preparation.....	23
Appendix C- Linear and Generalized Additive Methods	24
Appendix D- Linear and Generalized Additive Results.....	25
Appendix E- Generalized Linear Mixed Models Results	27
Appendix F- Evaluating Chain Convergence	28
Log Posterior Histograms.....	28
Coefficient Traceplot Evaluation.....	30
Posterior Coefficient Density Plots	32
Appendix G- Posterior Prediction Checking.....	34
References	38

Abstract

Driven by climate change, the climate of the 21st century may be vastly different with changes in El Niño, the jet stream, thermohaline circulation, and other climate systems causing shifts in long term climates across the globe. Given the vulnerability of soybean, corn, and cotton to such factors and their predominance in Tennessee's agriculture, evaluating climate's historical influence on these cash crops' yields may provide actionable information for current and future farms and farmers. This study applies linear regression, generalized additive models (GAM), and generalized linear mixed models (GLMM) to determine crop yield responses to evaporation, precipitation, and maximum, minimum, and average temperature in Tennessee from 1947-2015. The linear models do not meet the residual independence assumption of regression due to a temporal relationship between yield and year. Next, the GAMs incorporate a smoothing factor for year but fail to remove temporal autocorrelation. Finally using Bayesian statistical inference with Markov Chain Monte Carlo(MCMC) sampling, the GLMMs incorporate an autoregressive correlation structure of the residuals. The GLMMs are the preferred and final models though software restrictions prevent their full optimization. Model validation and evaluation of chain convergence, posterior distributions, and posterior predictive checking are presented in the appendix.

Introduction

Agriculture defines significant portions of Tennessee's landscape, culture, and economy. Approximately half of Tennessee's land area is farmland with over 200,000 jobs in the sector. The state earns \$20 billion annually from agriculture with cotton, soybeans, and corn producing \$186, \$147, and 135\$ million in revenue respectively in 2001 (Farm Facts, 2016). Given anticipated climate changes, ascertaining climate's effect on crop yields may aid decision making for crop choice, land use, and agricultural investment.

According to Climate Action Tracker (2014), an independent scientific organization tracking global efforts to minimize climate change, current policies could lead to 2.6°- 4.9° C warming by 2100 under current policies and 2.3°- 3.5° C if all current pledges to reduce emissions are met. Currently, Tennessee's average summer temperatures are in the mid-80°'s F during the summer months. Temperature shifts of the degree predicted by Climate Action Tracker could prove problematic as some counties would reach near 95°F average temperatures in the summer months. Soybean yield would likely decrease in this scenario as their yield begins to diminish above 85°F (Tacarindua, Shiraiwa, Homma, Kumagai, and Sameshima, 2013). While corn typically is heat resistant up to 95°F, when four consecutive days of temperature between 93°F and 98°F or above are accumulated, yield will decrease by 1% with additional 2% and 4% reductions for the fifth and sixth consecutive days ("High Temperature Effects", 2012). As average temperatures rise, yield reduction resulting from consecutive hot days could be expected to rise as well. Furthermore, increased night time temperatures can cause excessive night time respiration reducing available sugars, causing earlier maturity, and reducing final dry weight ("High Temperature Effects", 2012). While higher temperatures have potential to reduce yields, they would also expand the growing season. Currently, most of the state has a greater than 50%

chance of experiencing a freeze lethal to plants by the middle of October as shown in Appendix A.1. While corn and cotton are mostly harvested by the end of September, only 56% of Tennessee's soybeans are harvested by the end of October (McClure and Cannon, 2016). Additionally, longer growing seasons could increase the chances of success for late planted crops or replanted stands. In this study, evaporation serves as a proxy for transpiration, or the evaporation of water from plants. Corn, cotton, and soybeans are all vulnerable to water stress generated by increased transpiration. In contrast, precipitation facilitates growth, an effect which is especially strong during the seed growth phase in corn, cotton, and soybeans.

This study will apply linear regression, GAMs, and GLMMs to crop yields in bushels/acre (pounds/acre for cotton) as a response to the climate variables from 1947-2015. Models iterate from linear regression to GLMMs to find a model which fits the assumptions of regression. In spring and fall months, increases in temperature are expected to increase yield while decreasing yield in the summer. Evaporation in all months is expected to decrease yield because of increased water stress. Finally, precipitation is expected to increase yield in all months.

Data and Data Preparation

Corn, cotton, and soybean production in bushels/acre (pounds/acre for cotton) by county and year are downloaded from the United States Department of Agriculture (USDA, 2017). Climate data was downloaded from the National Community for Atmospheric Research's (NCAR) Climate Change Scenarios GIS download portal (NCAR, 2016).

The climate data is in a 1° latitude by 1.25° longitude grid. The downloaded output is in the form of a point shapefile representing the centroid of each polygon with an approximate spatial resolution of 125 km² (NCAR, 2016). The climate variables can only be downloaded individually, so the county climate combinations are finally joined together resulting in a final data frame of the climate variables. To join the climate data to Tennessee counties, a generic county level shapefile for Tennessee was imported into R from ArcMap (ESRI). The crop data is then attached to the county data by FIPS codes. Next, a Voronoi triangle transformation is applied to the climate point grid. Voronoi transformation applied to an equally spaced grid results in a square grid. The grid is projected to match the Tennessee shapefile projection. The shapefile and the Voronoi grid are joined with a spatial overlay by each county's centroids presence in the Voronoi grid as shown in Appendix B.1. Appendix B.2 displays a sample result of the overlay showing the average temperature across Tennessee for October 2010.

Linear and Generalized Additive Methods

First, biological frameworks for selecting which months to include in the analysis are developed for each crop. The month in which each crop is typically planted is chosen. Damage to young plants soon after seeding due to weather would be reflected in final yields, and planting month weather could be predictive of the length of the growing season for a specific year. Under these criteria, June is chosen for soybeans while April is chosen for corn and cotton.

Next, a mid-summer month is chosen to reflect the weather around pivotal growth stages for each plant. For corn, the R1 growth phase, during and shortly after fertilization, is when the plant is most susceptible to heat or water stress (Nielsen, 2016). Soybean and cotton yields follow a similar process (Tacarindua et al, 2013). This phase occurs in August for soybeans and July for corn and cotton. Finally, harvest date months are included with the desire of capturing late season freezes. Histogram, scatterplot, and bivariate regression analysis of the data indicates all variables except evaporation are approximately Gaussian. Evaporation has little variability across the state in any given month, so it is unimodal with a few counties either above or below the mode. As expected, precipitation, average temperature, maximum temperature, and minimum temperature are collinear. The resulting linear models are presented in equation 1.

The residuals versus fitted plots for each model were significantly non-linear. Yield by acre increased significantly for each plant from the beginning to the end of the study period, so the non-linear results appear to result from temporal effects. Appendix C.1 shows the relationship between time and soybean yield. Bivariate regression indicates yield and year are strongly correlated, and the scatter plot reflects that conclusion. Similar trends are reflected in corn and cotton plots.

Equation 1- Initial Linear Models for Soybean, Corn, and Cotton

Soybean	$\text{Yield} \sim \beta_1 \text{TMINJune} + \beta_2 \text{TMINAug} + \beta_3 \text{TMINOct} + \beta_4 \text{TMAXJune} + \beta_5 \text{TMAXAug} +$ $\beta_6 \text{TMAXOct} + \beta_7 \text{TAAJune} + \beta_8 \text{TAAug} + \beta_9 \text{TAAOct} + \beta_{10} \text{EVPJune} + \beta_{11} \text{EVPAug} +$ $\beta_{12} \text{EVPOct} + \beta_{13} \text{PPTJune} + \beta_{14} \text{PPTAug} + \beta_{15} \text{PPTOct} + \varepsilon$
Corn	$\text{Yield} \sim \beta_1 \text{TMINApril} + \beta_2 \text{TMINJuly} + \beta_3 \text{TMINSep} + \beta_4 \text{TMAXApril} + \beta_5 \text{TMAXJuly} +$ $\beta_6 \text{TMAXSep} + \beta_7 \text{TAApril} + \beta_8 \text{TAAJuly} + \beta_9 \text{TAASep} + \beta_{10} \text{EVPApril} + \beta_{11} \text{EVPJuly} +$ $\beta_{12} \text{EVPSep} + \beta_{13} \text{PPTApril} + \beta_{14} \text{PPTJuly} + \beta_{15} \text{PPTSep} + \varepsilon$
Cotton	$\text{Yield} \sim \beta_1 \text{TMINApril} + \beta_2 \text{TMINJuly} + \beta_3 \text{TMINSep} + \beta_4 \text{TMAXApril} + \beta_5 \text{TMAXJuly} +$ $\beta_6 \text{TMAXSep} + \beta_7 \text{TAApril} + \beta_8 \text{TAAJuly} + \beta_9 \text{TAASep} + \beta_{10} \text{EVPApril} + \beta_{11} \text{EVPJuly} +$ $\beta_{12} \text{EVPSep} + \beta_{13} \text{PPTApril} + \beta_{14} \text{PPTJuly} + \beta_{15} \text{PPTSep} + \varepsilon$

To compensate for the temporal effects, generalized additive models are created for each crop. These models are equivalent to the linear models, but they include a smoothing factor for year. The family is Gaussian with an identity link function for each model. In each model, multiple variables become nonsignificant when the smooth on year is included. Because of the multicollinearity among the variables, nonsignificant variables are removed. The K value, or number of knots, for the smoothing factor defaults to 9 by the `mgcv::gam` function in R. By increasing K, autocorrelation and GCV are reduced, and the deviance explained increases. However, the risk of overfitting directly increases with K. Weighing the improved results versus the possibility of overfitting, 20 knots are incorporated into each final model though its selection is ultimately arbitrary.

Linear and Generalized Additive Results

The initial linear model results are all significant with p-values approaching 0. Adjusted R-squared values are moderately low ranging from 0.27 to 0.41. However, there are significant issues with non-linearity as the residuals versus fitted plot in Appendix D.1 visualizes. As previously discussed, the non-linearity is likely a result of not counting for the increase in yield over time. The heteroskedasticity is prevalent in each crop model. Therefore, little evidence remains that simple multivariate linear regression is a suitable model for the data.

Rejecting the linear model, time is incorporated as a 20 knot smooth parameter in the generalized additive models. Equation 2 contains the final equation for the cotton GAM and Table 1 contains coefficient results and p-values. With the added smoothing factor, each model explains above 70% of the variability in crop yield whereas the linear models were all below 42%. The smoothing factor is significant in each model, and each model drops nonsignificant variables when the smoothing factor is incorporated. Evaluative plots, shown in Appendix D.2, are generally acceptable. However, there is some clustering of the residuals. Some of this is likely due to the stratified evaporation variables, but it is unclear what other factors may be causing the spread.

Equation 2- Cotton GAM Model

$$\text{Yield} \sim \beta_1 TMINJuly + \beta_2 TMINSep + \beta_3 TMAXJuly + \beta_4 TAJuly + \beta_5 TASEp + \beta_6 EVPApril + \beta_7 EVPSep + \beta_8 S(\text{Year}, k = 20) + \varepsilon$$

Table 1- Cotton GAM Results

Deviance explained- 74.9%, GCV- 10030, N=1866 – s(year): F=140.9, p=0.000***

Variable	Coefficient	Standard Error	P-value
TMINJuly	-169.78	32.54	0.000***
TMINSep	59.84	11.37	0.000***
TMAXJuly	-149.48	28.37	0.000***
TAJuly	334.70	60.28	0.000***
TASep	-51.78	13.00	0.000***
EVPApril	247.95	82.70	0.003**
EVPSep	-432.24	72.42	0.000***
PPTSep	-1.83	0.26	0.000***

TMIN = Minimum Temperature, TMAX = Maximum Temperature, TA = Average Temperature, EVPApril = Evaporation, PPT = Precipitation

However, the GAMs still fail to remove temporal autocorrelation of the residuals. The ACF and PACF plots in Appendix D.3 show the temporal autocorrelation for cotton. The corn and soybean models have similar ACF and PACF plots. Because using time as a smooth independent variable fails to remove autocorrelation, it is concluded that the residuals are not independent due to their structure across time. To incorporate the error structure, the model needs to include fixed and random effects which is outside the scope of generalized additive modeling. Therefore, the GAMs cannot meet the residual independence assumption of regression and are rejected.

Generalized Linear Mixed Model Methods

Moving from generalized additive modeling to a model which incorporates a residual dependence structure necessitates several key changes. The model is initially reset to the simple linear model removing time as an independent factor. The linear model is converted to a GLM which dictates a distribution of the response variable, specifies a systemic component resulting from the explanatory variables, and contains a link between the response variable and the systemic component (Zuur, Ieno, Walker, Saveliev, and Smith, 2009). Next, Temporal autocorrelation structure is added based on the time series of yield in each county creating a GLMM. Finally, Markov Chain Monte Carlo(MCMC) is used to fit the model due to increased simplicity in incorporating autocorrelation (Zuur et al., 2009).

The initial GLMM for cotton is shown in equation 3. The model assumes a Gaussian distribution of Y_i with mean μ_i and variance σ^2 . The identity matrix is the Gaussian link function meaning the linear predictor, η , plus error, ε_i , equals μ_i which is the expected value of Y_i . The error term is autoregressive with a lag of 1. The relationship between ε_t and ε_{t-k} is signified by ρ raised to lag k leading to a quick deterioration in ε_t and ε_{t-k} 's relationship as k increases. By introducing the autoregressive component, the model incorporates the relationship of the residuals across time theoretically removing the temporal autocorrelation. Ideally, the model would explore implementing a Poisson or gamma distribution, but the package used for applying MCMC does not allow these families when incorporating temporal autocorrelation at the time of writing.

Equation 3- Base GLMM Equation and Priors

$$Y_i \sim N(\mu_i, \sigma^2)$$

$$E(Y_i) = \mu_i = \eta(X_{i1}, \dots, X_{i15}) + \varepsilon_i \quad \text{and} \quad \text{var}(Y_i) = \sigma^2$$

$$\text{cor}(\varepsilon_t, \varepsilon_{t-k}) = \rho^k$$

$$\begin{aligned} \eta = & \alpha + \beta_1 TMINApril + \beta_2 TMINJuly + \beta_3 TMINSep + \beta_4 TAApril + \\ & \beta_5 TAJuly + \beta_6 TASep + \beta_7 TMAXApril + \beta_8 TMAXJuly + \beta_9 TMAXSep + \\ & \beta_{10} PPTApril + \beta_{11} PPTJuly + \beta_{12} PPTSep + \beta_{13} EVPApril + \beta_{14} EVPJuly + \\ & \beta_{15} EVPSep \end{aligned}$$

(The soybean model has June, August, and October in the place of April, July, and September respectively)

Climate Coefficient and Autoregression Priors = Improper Flat Prior Over the Reals

$$\text{Prior Sigma}_{crop} = \text{Student's } T(DF_{crop}, 0, \sigma_{crop}^2)$$

To perform the Bayesian analysis, the *brms* package is applied (Bürkner, 2017). The package functions as an R interface to Stan using common R syntax. Stan is statistical software written in C++ which allows Bayesian operations. In contrast to many other options, Stan uses Hamiltonian Monte Carlo (Duane, Kennedy, Pendleton, and Roweth, 1987) and the No-U-Turn Sampler(NUTS) (Hoffman and Gelman, 2014) which require fewer samples for chain convergence though at the cost of increased computation time per sample.

The default prior in brms is an improper flat prior over the reals (Bürkner, 2017). This default prior was used for all model components except σ which was given a Student's t-distribution with a mean of 0 across all crops and unique degrees of freedom and variance by crop. The priors are obtained with the `get_prior()` function. The improper flat prior over the reals is extremely diffuse indicating to the model that there's little prior knowledge of the distribution of the coefficients. Therefore, the posterior distribution of the coefficients will be predominantly influenced by the data (Zuur et al., 2009).

The analysis incorporates 4 chains with 2000 iterations each. The first 1000 samples of each chain are discarded as warmup leaving 4000 total post-warmup samples. As mentioned, attributes specific to the NUTS sampler lead to quick convergence and small chains compared to other sampling techniques. After running each model, the chains are assessed for convergence. Methods and analysis for chain convergence are presented in Appendix F. Coefficient Rhats, another measure of convergence, are also examined. Rhat values greater than 1.1 indicate the chains have not converged (Bürkner, 2017).

Instead of r^2 , deviance, p-values, and GCV as returned by the generalized additive models, the Bayesian model outputs information on the likelihood distribution of each coefficient with an estimated mean, error, and 95% credibility intervals. For model evaluation, the Watanabe-Akaike information criterion (WAIC) and leave-one-out cross-validation (LOO) are used. Initial model variables are selected on the biological basis discussed in the linear and generalized additive methods section (i.e. months of planting, peak water stress sensitivity, and harvesting). Coefficients whose credibility intervals contain zero are removed as dictated by WAIC and LOO. Coefficients with zero near the center of the distribution are removed before edge cases. Removing edge cases is often penalized by WAIC indicating their removal does not

improve parsimony. Insignificant coefficients whose removal is penalized by WAIC are kept in the model. Finally, model fit is evaluated with graphical posterior predictive checking which compares observed data Y with simulated predictions from the model Y^{Rep} . Posterior Predictive checking is shown in Appendix G.

Generalized Linear Mixed Model Results

Following the model selection methodology described above, the final linear components of the GLMMs are presented in Table 2. All models have identity link functions and normal distributions of Y_i as shown in Equation 3. Table 3 shows the model selection for corn with equivalent tables available for cotton and soybeans in Appendix E. In the corn model selection, only three variables out of fifteen are removed because of WAIC and LOO score penalization when more nonsignificant variables are removed. The soybean model required adding and subtracting several variables to maximize WAIC and LOO. Most of the nonsignificant cotton variables led to lower WAIC scores when removed. July Precipitation's 95% credibility intervals maintain zero, but with zero at the far end of the posterior distribution, removing the variable would imply more confidence in the credibility intervals than is perhaps warranted. In fact, most precipitation variables have coefficients near zero and correspondingly small standard deviations, but their removal does not seem to improve parsimony. Tables 4-6 present the final results for each model.

Referencing Tables 4-6, Rhat values for all coefficients across all models are approximately 1 indicating chain convergence. To visually confirm this conclusion, coefficient trace and density plots, presented in Appendix F, are analyzed for convergence and stationarity. Log-posterior histograms are also analyzed in Appendix F. Summarizing the models' convergence, Rhat and evaluative graphs uniformly indicate Markov Chain convergence across the models confirming the selection of 1000 samples for both burn-in and posterior sampling is sufficiently large.

Table 2- Final Linear Predictor Equations

Cotton	$\eta = \alpha + \beta_1 TMIN_{April} + \beta_2 TMIN_{July} + \beta_3 TMIN_{Sep} + \beta_4 TAA_{April} + \beta_5 TA_{July} + \beta_6 TMAX_{April} + \beta_7 TMAX_{July} + \beta_8 PPT_{July} + \beta_9 PPT_{Sep} + \varepsilon_i$
Soybean	$\eta = \alpha + \beta_1 TMIN_{Aug} + \beta_2 TMIN_{Oct} + \beta_3 TA_{June} + \beta_4 TAA_{Aug} + \beta_5 TAO_{Oct} + \beta_6 TMAX_{June} + \beta_7 TMAX_{Oct} + \beta_8 PPT_{Aug} + \beta_9 PPT_{Oct} + \beta_{10} EVP_{June} + \beta_{11} EVP_{Aug} + \beta_{12} EVP_{Oct} + \varepsilon_i$
Corn	$\eta = \alpha + \beta_1 TMIN_{April} + \beta_2 TMIN_{July} + \beta_3 TMIN_{Sep} + \beta_4 TAs_{ep} + \beta_5 TMAX_{April} + \beta_6 TMAX_{July} + \beta_7 TMAX_{Sep} + \beta_8 PPT_{July} + \beta_9 PPT_{Sep} + \beta_{10} EVP_{April} + \beta_{11} EVP_{July} + \beta_{12} EVP_{Sep} + \varepsilon_i$

Table 3- Model Selection for Corn, Selecting Model 3

Model Number	Model Changes	WAIC	LOO
1	Full Model	46511.05	46511.06
2	Remove TA April, July	46507.39	46507.38
3	Remove PPTApril	46506.64	46506.64
4	Remove TAs _{ep}	46508.03	46508.04

Table 4- Final Cotton Results Table

	Rhat	Mean	SD	2.50%	50%	97.50%
Intercept	1	-23972.4	1874.5	-27741.1	-23962.5	-20259.1
TMINApril	1	111.1	30.2	52	110.7	170.7
TMINJuly	1	-103.5	37.1	-174.7	-103.9	-28.8
TMINSep	1	27.1	3.6	20	27.1	34
TMAXApril	1	118.3	29	60.1	118	174.7
TMAXJuly	1	-111.8	31.4	-171.8	-112.4	-48.4
TAApril	1	-234.2	59	-350.4	-233.2	-118.4
TAJuly	1	275.3	69.3	133.1	276.3	407.9
PPTJuly	1	0.3	0.2	0	0.3	0.7
PPTSep	1	-1	0.2	-1.4	-1	-0.6
AR(1)	1	0.7	0	0.6	0.7	0.7
Sigma	1	125.3	2.1	121.4	125.3	129.4
Log-posterior	1	-9943.8	2.4	-9949.3	-9943.4	-9940

Table 5- Final Soybean Results Table

	Rhat	Mean	SD	2.50%	50%	97.50%
Intercept	1	-746.9	58	-862.7	-747.5	-632.4
TMINAug	1	-2.2	0.4	-3.1	-2.2	-1.3
TMINOct	1	4.4	1	2.4	4.4	6.4
TMAXJune	1	-0.8	0.4	-1.6	-0.8	0
TMAXOct	1	7.1	0.9	5.3	7.1	8.8
TAJune	1	3.7	0.4	2.8	3.7	4.6
TAAug	1	1.5	0.4	0.6	1.5	2.3
TAOct	1	-11.1	1.9	-14.7	-11.1	-7.3
EVPJune	1	-23.2	3.6	-30.1	-23.2	-16.3
EVPAug	1	-9.2	2.3	-13.8	-9.2	-4.7
EVPOct	1	-11.8	8	-27.5	-11.9	3.8
PPTAug	1	0.1	0	0.1	0.1	0.1
PPTOct	1	0	0	0	0	0.1
AR(1)	1	0.6	0	0.6	0.6	0.7
Sigma	1	6.3	0.1	6.2	6.3	6.5
Log-posterior	1	-9211.4	2.7	-9217.7	-9211	-9207.1

Table 6- Final Corn Results Table

	Rhat	Mean	SD	2.50%	50%	97.50%
Intercept	1	-2585.8	215.8	-3013.2	-2581.7	-2162
TMINApril	1	-7	0.5	-8	-7	-5.9
TMINJuly	1	4.9	0.8	3.4	4.9	6.5
TMINSep	1	-8.2	3.4	-15.1	-8.2	-1.5
TMAXApril	1	9.2	0.8	7.6	9.2	10.7
TMAXJuly	1	2.9	0.5	1.9	2.9	3.9
TMAXSep	1	-9	3	-14.7	-9	-3.3
TASep	1	16.1	6.3	3.8	16.1	28.3
EVPApril	1	-50	9.4	-68.6	-50.1	-31
EVPApril	1	-30	7.6	-44.9	-30.1	-14.9
EVPSep	1	17.1	8.6	0.5	17	34.3
PPTJuly	1	-0.1	0	-0.1	-0.1	0
PPTSep	1	0	0	0	0	0.1
AR(1)	1	0.7	0	0.7	0.7	0.8
Sigma	1	21.5	0.2	21.1	21.5	21.9
Log-posterior	1	-18481.3	2.7	-18487.5	-18481	-18477

The corn and cotton models have a residual autocorrelation, ρ , of approximately 0.7 while the soybean has a ρ of approximately 0.6. Using a ρ of 0.7, the one lag autoregressive model dictates ρ of approximately 0.7 between ε_t and ε_{t-1} . Residuals with a lag of 2 have a ρ of approximately .49, residuals with a lag of 3 have a ρ of approximately .24 with ρ quickly decreasing as lag increases. Unsolved complications with the brms package prevent acquisition of residual autocorrelation in the GLMMs. Nevertheless, the autocorrelation component very likely improves the model. As Zuur et al. (2009) discuss, creating the perfect autocorrelation model is less important than ensuring an autocorrelation structure is included if the model dictates its inclusion.

The results, in Tables 4-6, present the means of each coefficient posterior distribution with 95% credibility intervals, standard deviations, and quantiles. The coefficient represents the mean likelihood of the coefficient given the data. The credibility intervals indicate the range in which 95% of the coefficient's posterior distribution resides. As written in Glickman and Van Dyk (2007), "The posterior distribution is a complete summary of what we know about the parameters, both from the data and – as quantified via the prior distribution – from other sources of information." Whereas frequentist methods give a p-value for hypothesis testing, Bayesian methods give a probability of the parameters in the form of the posterior distributions.

Finally, posterior predictive checking, detailed in Appendix G, proves problematic. Because the observed distributions of crop yield are not completely Gaussian, use of the Gaussian family means posterior predictive checks create distributions which pick up the mean and variance of the observed distribution but do not reflect skewness or non-Gaussian perturbations in the distribution. Use of a different family, most likely Poisson, would allow better fit models, but it cannot be implemented with autocorrelation in brms. The issue is most prevalent in the corn and soybean models, but the cotton model posterior prediction check is similarly flawed to a lesser degree.

Discussion

The initial linear and generalized additive models used to evaluate the relationship between crop yield and climatic variables were rejected due to autocorrelation of the residuals. The dependent residuals could mean these models' p-values and errors were spuriously low. The GLMM incorporates a residual dependence structure to ameliorate this problem. Despite this correction likely improving or eliminating autocorrelation, complications prevent verification of autocorrelation removal. The Markov Chain Monte Carlo specific components such as Rhat and coefficient trace and density plots indicate convergence and sufficient posterior sampling. Posterior predictive checking shows the Gaussian family and identity link function fail to completely capture the nuances of the observed yield distributions despite fitting the mean and variance. Nevertheless, the GLMMs represent significant improvements over the linear and generalized additive models and are clearly the preferred model for interpretation after incorporating the aforementioned caveats.

Across all crops, higher temperatures in spring and fall months were expected to increase yield while decreasing yield in summer months. Evaporation was expected to uniformly decrease yield due to water stress while precipitation was predicted to increase yield. Though the models' weaknesses prevent definitive conclusions, the GLMM results may still be interpreted in terms of the hypotheses to obtain exploratory evidence of their validity.

Temperature's correlation with yield does not uniformly support or deny the hypotheses for any crop. If we trust the coefficients, then temperature's effects on growth seem to be working in biological or climatic systems outside the scope of the temperature hypotheses. Three out of the four evaporation variables in the final models were negatively correlated with yield.

September's evaporation positively correlates with corn yield with 95% credibility intervals spanning from 0.5 to 34.3. Given the model's caveats, the true mean correlation of September evaporation could indeed be zero or negative considering the posterior distribution's proximity to zero. In totality, more evidence supports the hypothesis of an inverse relationship between evaporation and yield than a positive relationship though definitively confirming the hypothesis would be overzealous given the data. Precipitation's small effect across all models does not align with expectations considering plant hydration's importance to growth. If most Tennessee farms are irrigated, precipitations' effect on yield may be mitigated by the fact that shortages in precipitation are overcome with irrigation. Therefore, the hypothesis that precipitation uniformly increases yield is neither confirmed nor denied though the small effect size may indicate alternative or mitigating factors, such as irrigation, may be unaccounted for.

While the models improved with each iteration, there are clear remaining issues that need to be fixed to fully optimize the model. At the very least, the GLMMs incorporation of autocorrelation directly attempts to fix the known issues with the previous models and represents a definitive improvement, but limitations precluded obtaining autocorrelation measures in the GLMM. To verify autocorrelation's removal and that an autocorrelation lag of 1 is appropriate, obtaining residual autocorrelation is a clear next step for further research. Moreover, the Gaussian distribution of yield is inadequate, and implementation of a Poisson or another family would likely improve the posterior predictive checks. At this time, implementing another family would require a different R package or perhaps different software. Finally, Tennessee's counties could be grouped into similar physiographic regions. From the Mississippi flood plain in West Tennessee to the Appalachian mountains in the East, the ample intrastate variation does not support the models' assumption of uniform climatic effects across the state. By grouping

counties by physiography, a model including group effects by physiographic region would better control for this variation by allowing climatic influences to vary by region.

Moving towards a more general analysis of the research process, creating a more thorough biological framework including farming methodology is needed. Irrigation, which could be mitigating precipitation's effect, is an example of an unincorporated farming practice which may be influencing the results. If the effects of temperature presented by the model are true, then biological factors may be at play which are superior or alternative to the hypothesis that warmer weather promotes increased yields in spring and fall but decreases it in summer. Further, additional variables are available from download from the NCAR GIS portal such as skin temperature which may better reflect the temperature plants experience at their surface. Replacing evaporation with evapotranspiration would remove a layer of abstraction in the model and soil type may be an important factor that went untouched in these models. Nevertheless, the model does prove that linear and general additive modeling are likely incorrect or incomplete methods for predicting climate's effect on crop yields. If it is confirmed that the inclusion of temporal autocorrelation improves the GLMMs, then the research also implies that mixed effects need to be incorporated into yield. The next step for further research is clearly implementing a different family in the GLMMs in order to create models capable of incorporating the nuances in actual crop yields across the state.

Appendices

Appendices A-E contains images, plots, and graphs referenced in their respective sections. Secondly, Appendix F contains information on the Bayesian GLMMs' model evaluation. Log posterior histograms, coefficient trace plots, and coefficient density plots are presented to evaluate chain convergence. In Appendix G, posterior predictive checks are performed by comparing posterior sample distributions with the observed data in both distribution and scatter plots.

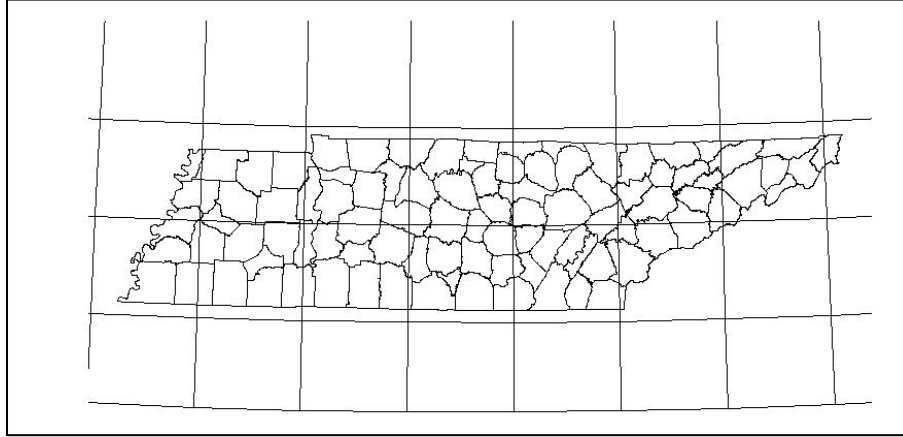
Appendix A-Introduction

A.1- Probability of having not experienced a frost by the given date and location

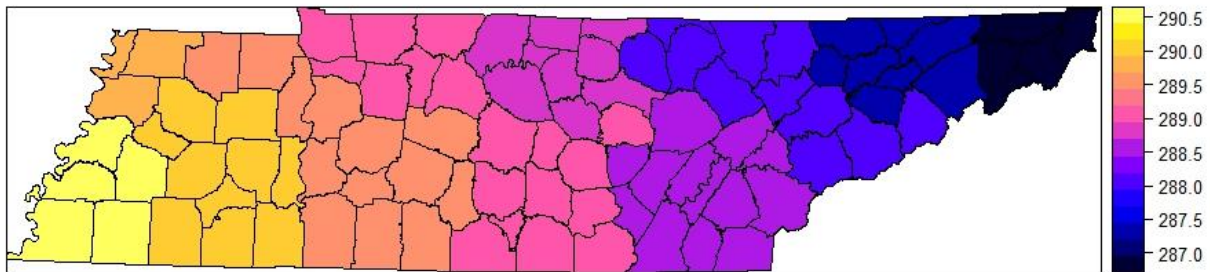
	Probability		
Location	50%	25%	10%
Bolivar	11-Oct	31-Oct	18-Nov
Brownsville	8-Oct	5-Nov	30-Nov
Covington	26-Oct	12-Nov	26-Nov
Jackson	24-Oct	31-Oct	7-Nov
Martin	15-Oct	7-Nov	28-Nov
Memphis	6-Nov	13-Nov	19-Nov
Milan	17-Oct	7-Nov	26-Nov
	13-Oct	4-Nov	25-Nov
Newbern	13-Oct	6-Nov	28-Nov
Samburg	15-Oct	6-Nov	25-Nov
Savannah	20-Oct	10-Nov	29-Nov

Appendix B- Data and Data Preparation

B.1- Visualization of Voronoi transformation of climate data and Tennessee counties overlay

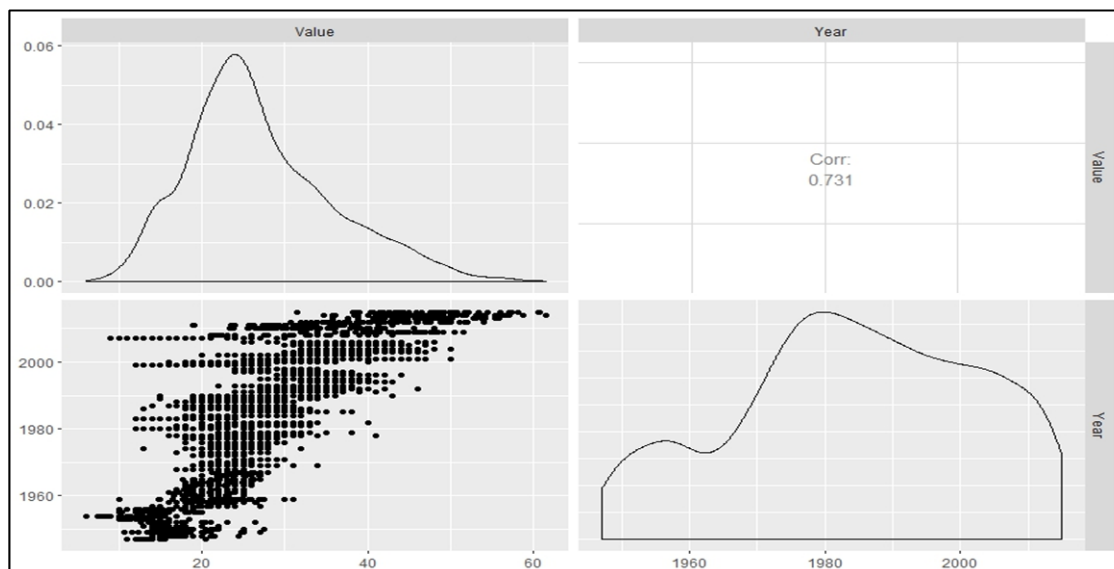


B.2- Sample map of climate data joined to Tennessee counties showing average temperatures(K) in October 2010.



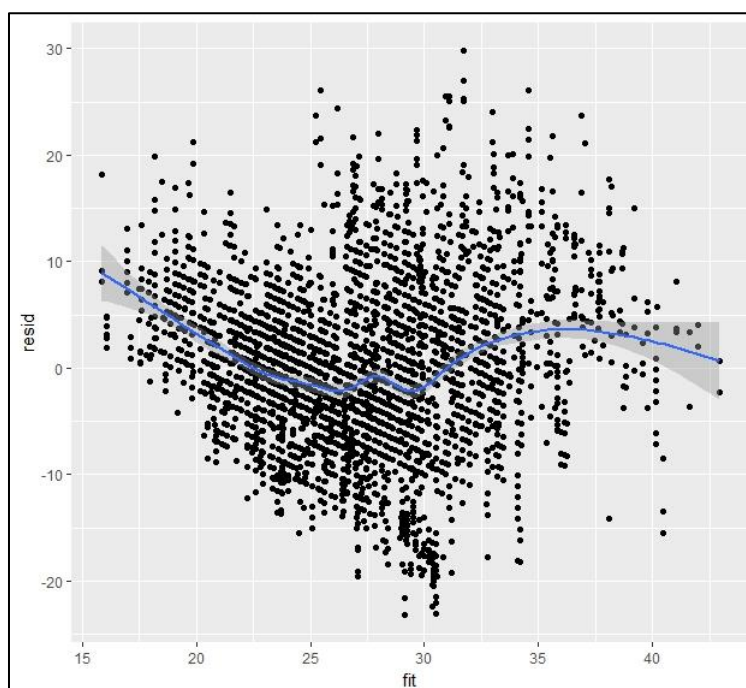
Appendix C- Linear and Generalized Additive Methods

C.1- Pair plot of soybean yield, Value, and Year

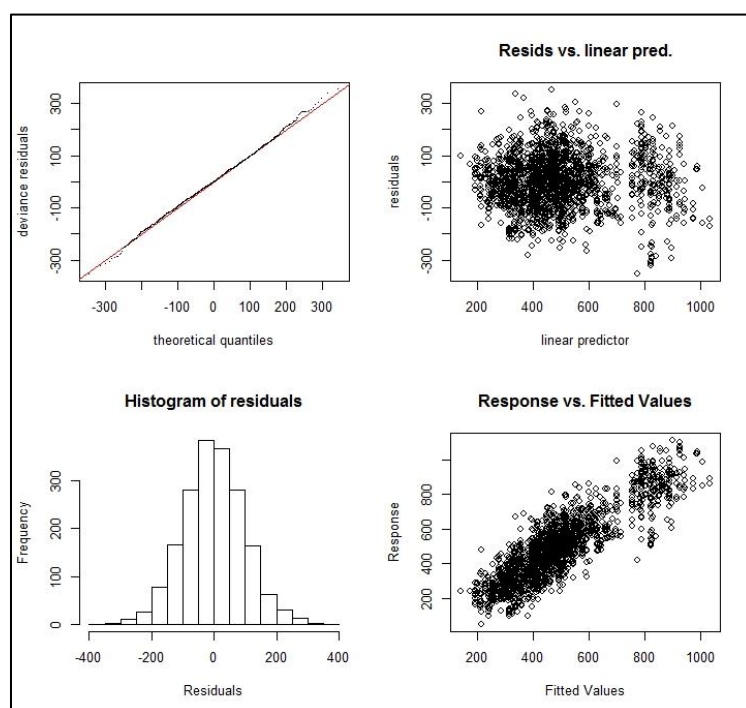


Appendix D- Linear and Generalized Additive Results

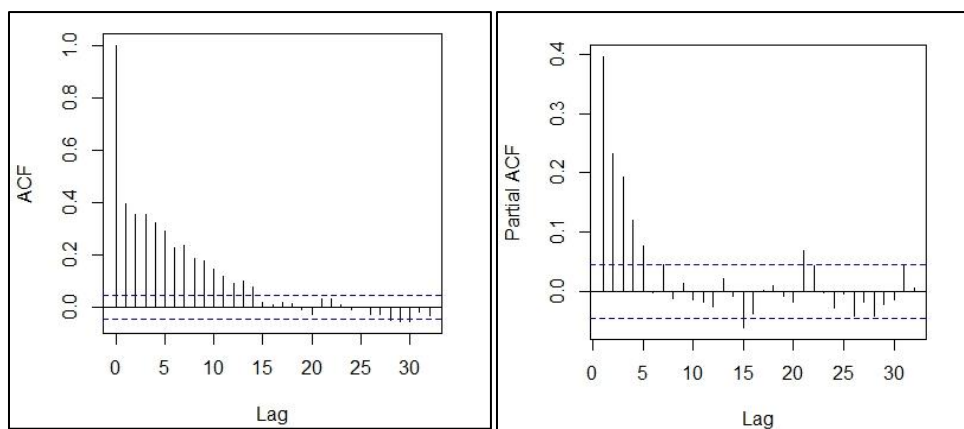
D.1- Soybean linear model residuals versus fitted values plot showing heteroskedasticity



D.2- Cotton GAM evaluative plots



D.3- Cotton ACF and PACF plots, respectively, showing temporal autocorrelation



Appendix E- Generalized Linear Mixed Models Results

E.1- Model Selection by WAIC and LOO for cotton, selecting model 4

Model Number	Model Changes	WAIC	LOO
1	Full Model	23341.27	23341.28
2	Remove all EVP variables	23339.54	23339.55
3	Remove PPTApril, TMAXSep	23338.07	23338.07
4	Remove TASEp	23336.23	23336.24

E.2- Model Selection by WAIC and LOO for soybean, selecting model 6

Model Number	Model Changes	WAIC	LOO
1	Full Model	25667.48	25667.48
2	Remove PPT June	25666.12	25666.12
3	Remove TA and TMAX Aug	25673.35	25673.35
4	ADD TA and TMAX Aug; Remove EVP Oct	25666.94	25666.94
5	ADD EVP Oct; Remove TMIN June	25665.78	25665.78
6	REMOVE TMAX Aug	25664.40	25664.40

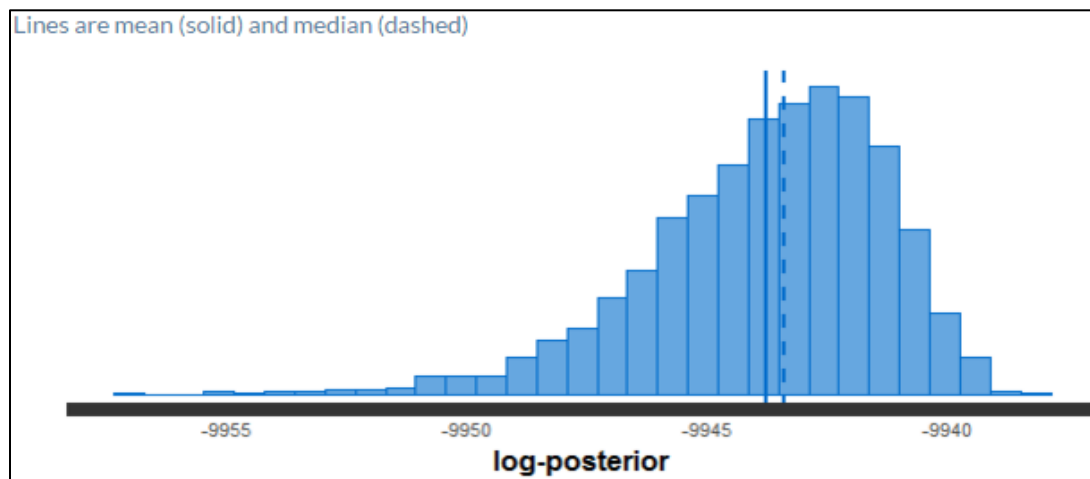
Appendix F- Evaluating Chain Convergence

Log Posterior Histograms

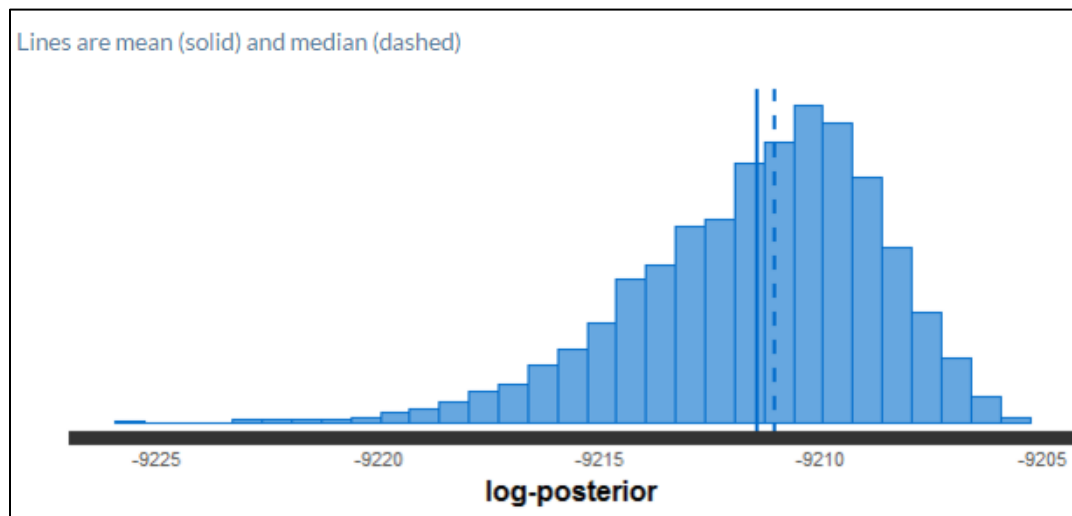
Log posterior histograms are used to assess Markov Chain convergence. If the histograms have multiple modes, then the Markov Chains have not converged on the posterior mean.

Essentially, if there were more than one mode, it would indicate the chains are still evaluating multiple points as plausible posterior means. Across all models, the log posterior has centered on a central point indicating convergence.

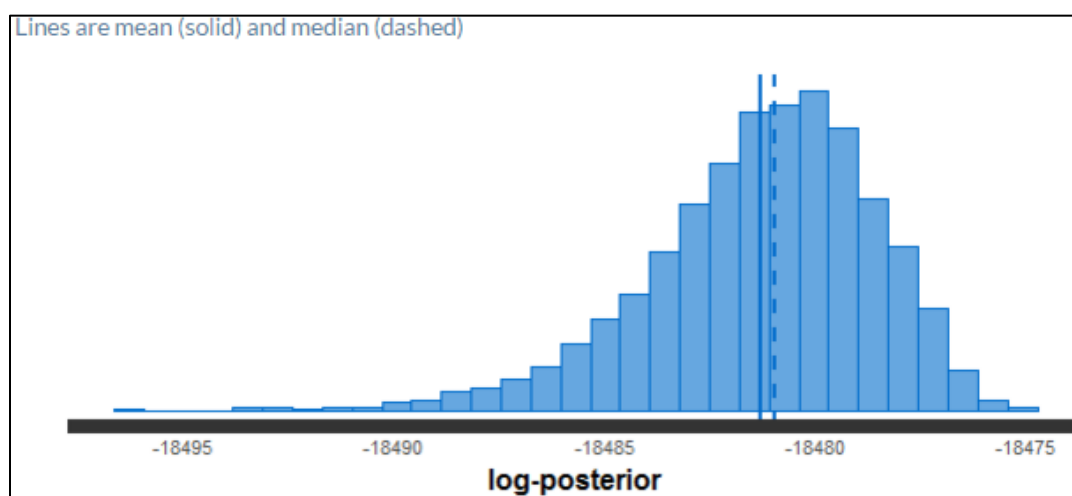
F.1- Cotton Log-posterior Histogram



F.2- Soybean Log-posterior Histogram



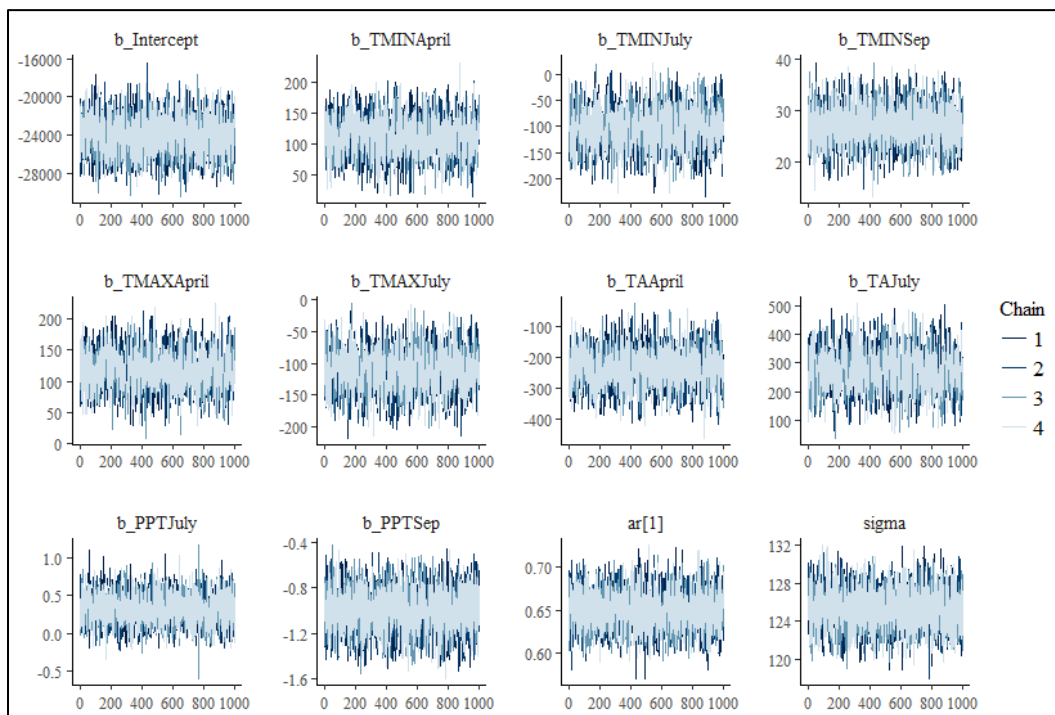
F.3- Corn Log-posterior Histogram



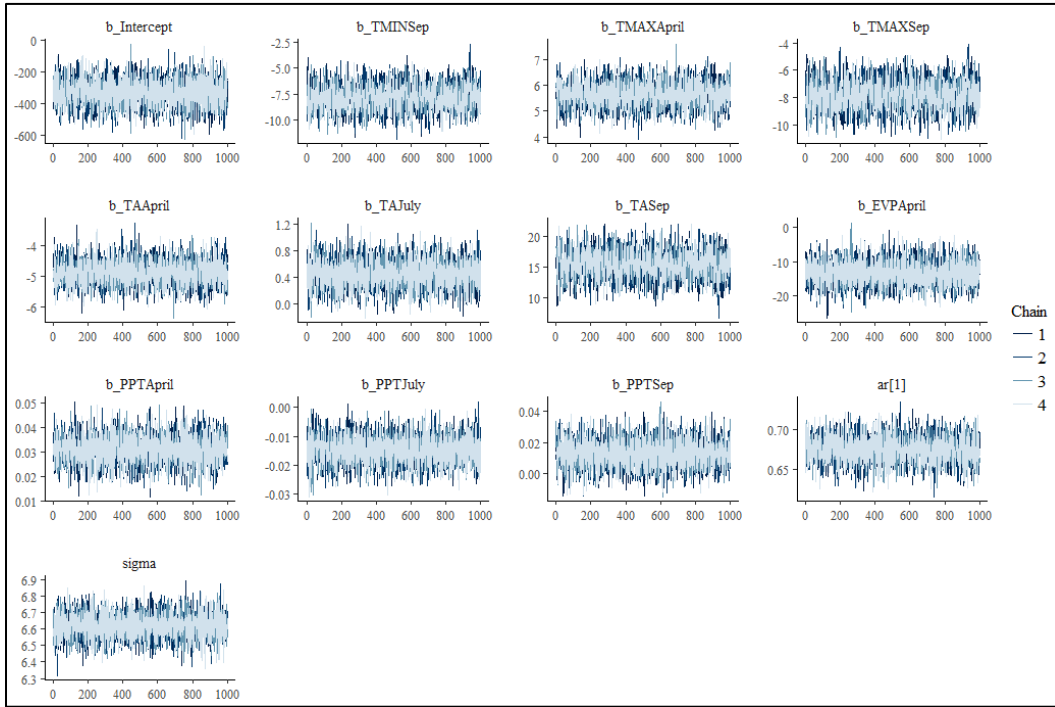
Coefficient Traceplot Evaluation

Coefficient traceplots are another visual method of confirming convergence. The four chains for each model are plotted by variable including the autoregressive components and sigma. If the chains do not converge on a single point, show trends of increasing or decreasing, or are not stationary, then they are concluded to have not converged. Across all models and variables, the chains appear to have converged indicating the selection of 1000 burn-in samples and 1000 posterior samples is sufficient.

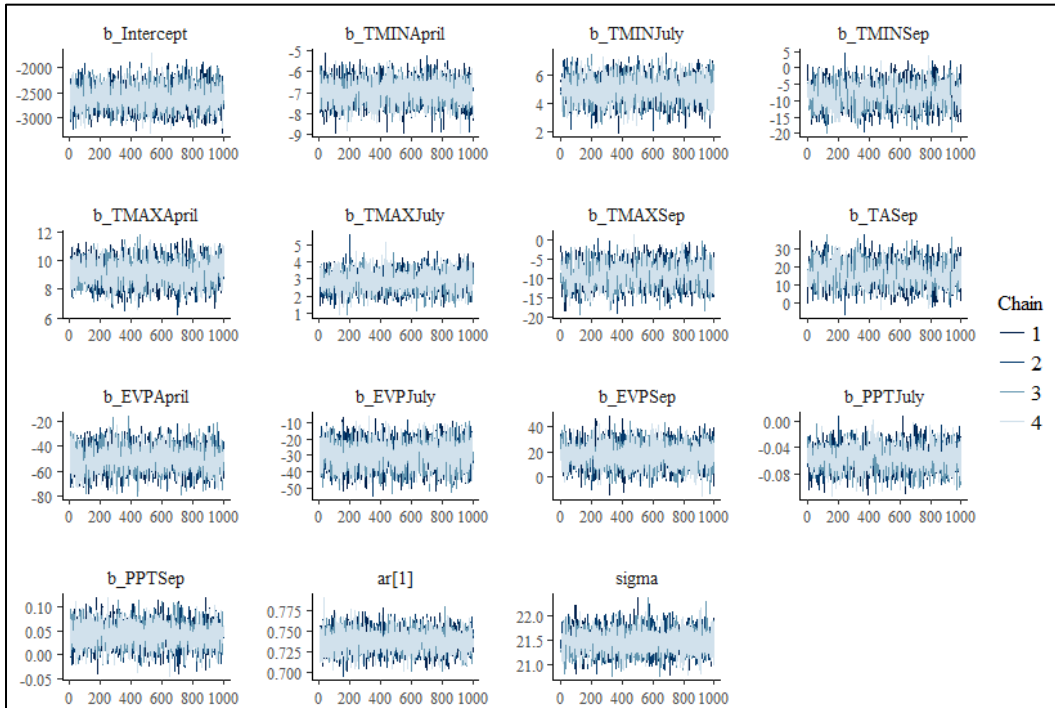
F.4- Cotton Coefficient Traceplot



F.5- Soybean Coefficient Traceplot



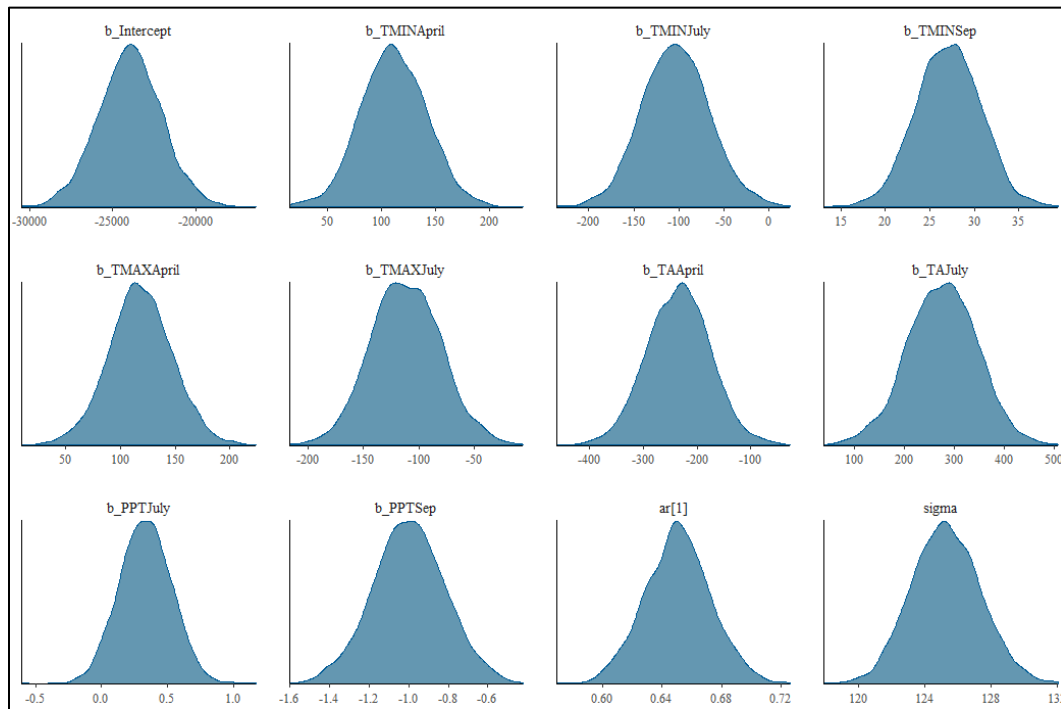
F.6- Corn Coefficient Traceplot



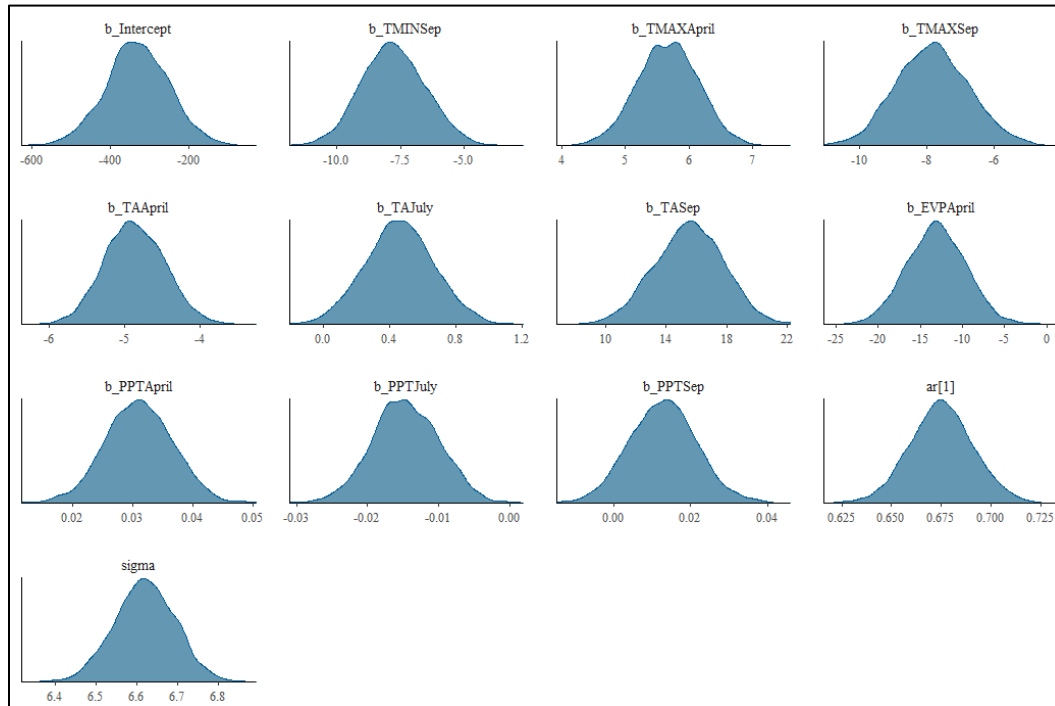
Posterior Coefficient Density Plots

The posterior coefficient density plots represent the posterior distribution of the models' components. Like the log posterior histograms, multiple modes represent the Markov Chains still searching for the posterior mean while a unimodal distribution indicates convergence. Across all components in all models, the distributions are unimodal indicating convergence and sufficient posterior sampling.

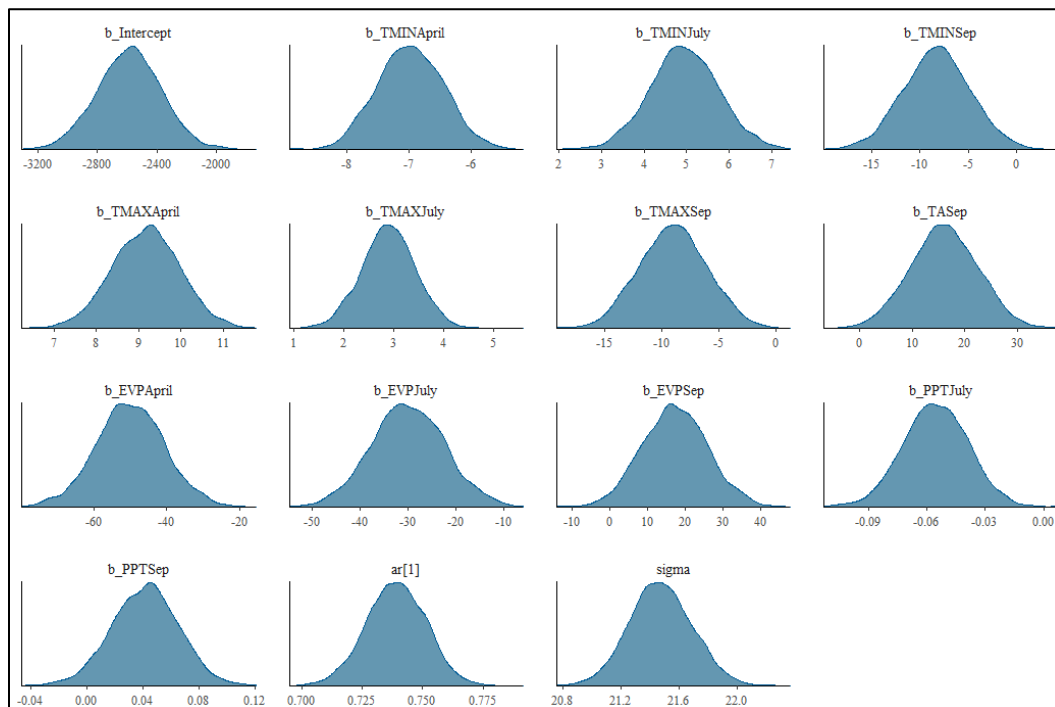
F.7- Cotton Coefficient Density Plot



F.8- Soybean Coefficient Density Plot



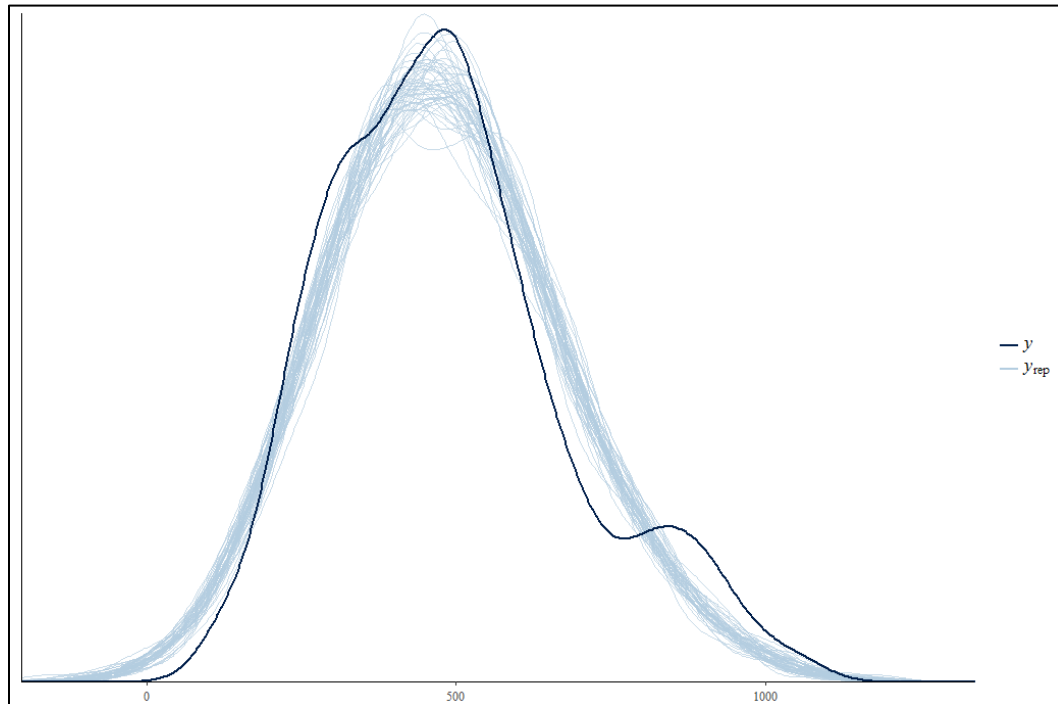
F.9- Corn Coefficient Density Plot



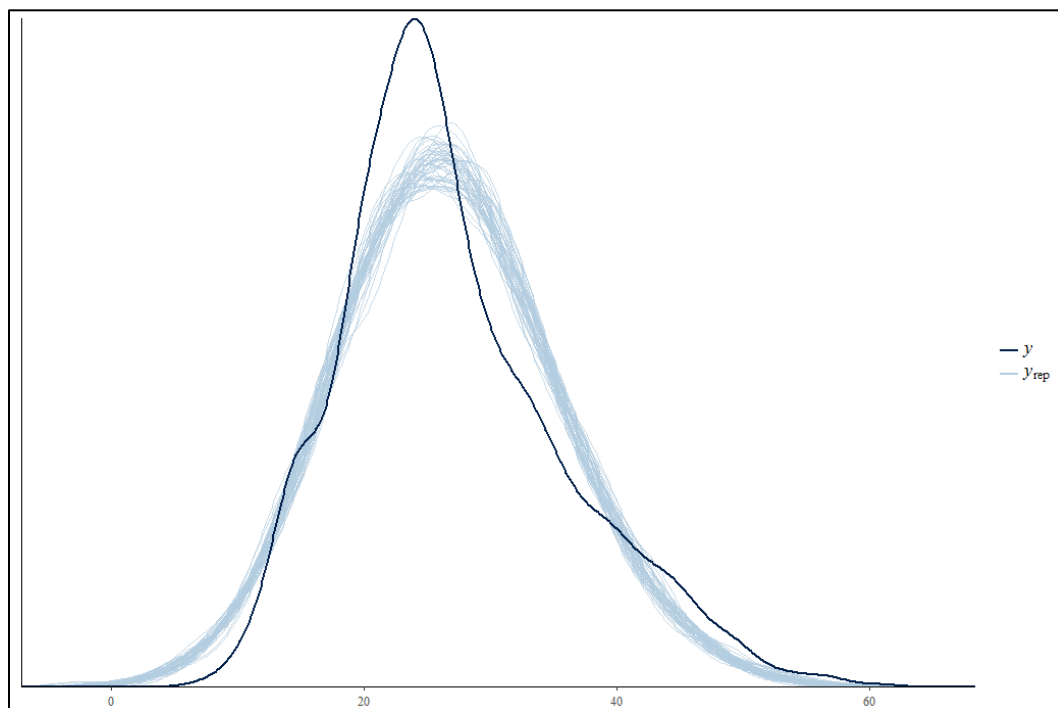
Appendix G- Posterior Prediction Checking

Model fit is evaluated with distribution and scatter plots comparing the observed distribution Y with samples from the posterior distribution Y^{rep} . Using the Gaussian family and identity link function creates Gaussian posterior distributions. For the cotton model, the posterior distribution approximately fits the observed observation, but the soybean and corn models show discrepancies. The soybean model has a steeper observed distribution than the Gaussian distribution of Y^{rep} indicating underestimation of the frequency of soybean yields around 25 bushels/acre. The corn model has a right skew causing Y^{rep} to underestimate around 45 bushels/acre and overestimate around 75 bushels an acre. Further, the Gaussian distribution is not capable of picking up fluctuations such as the uptick in Corn yield around 90 bushels/acre. The scatterplots indicate a strong correlation between Y and Y^{rep} albeit with slight overestimation at the lowest yields and underestimation at the highest yields across the models. These flaws are the strongest repudiation against the model in its current format. Ideally, other distributions, such as a Poisson, would be used to improve the posterior predictive checks, but limitations in the brms package do not allow Poisson distributions while also incorporating autocorrelation.

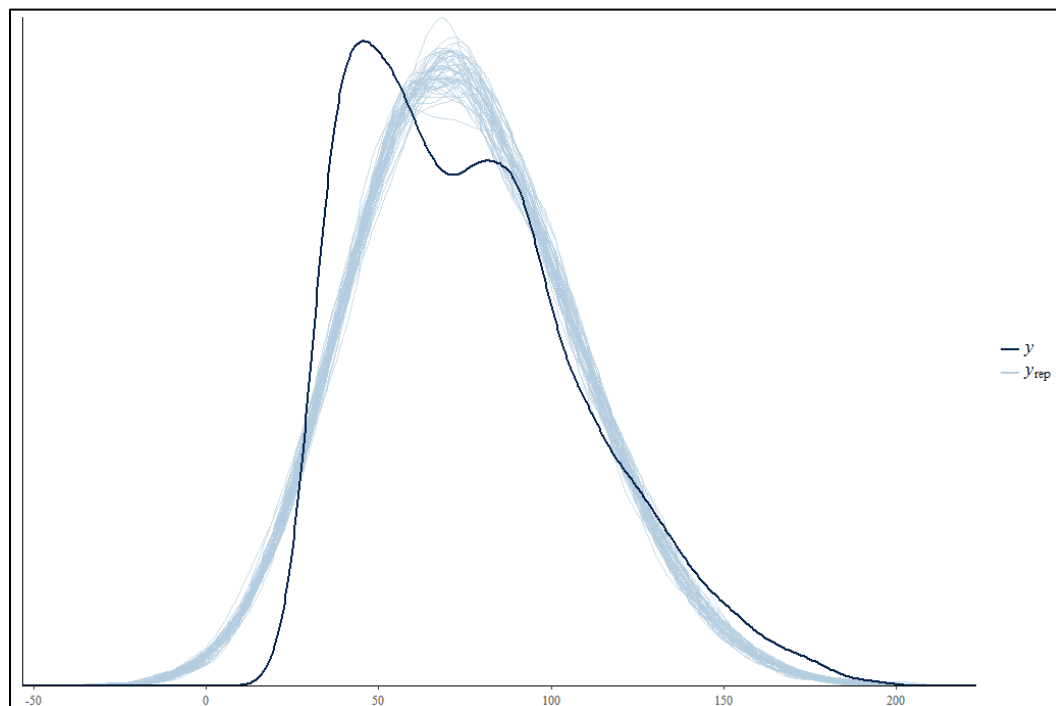
G.1-Cotton Posterior Prediction Test with Distributions (50 posterior samples):



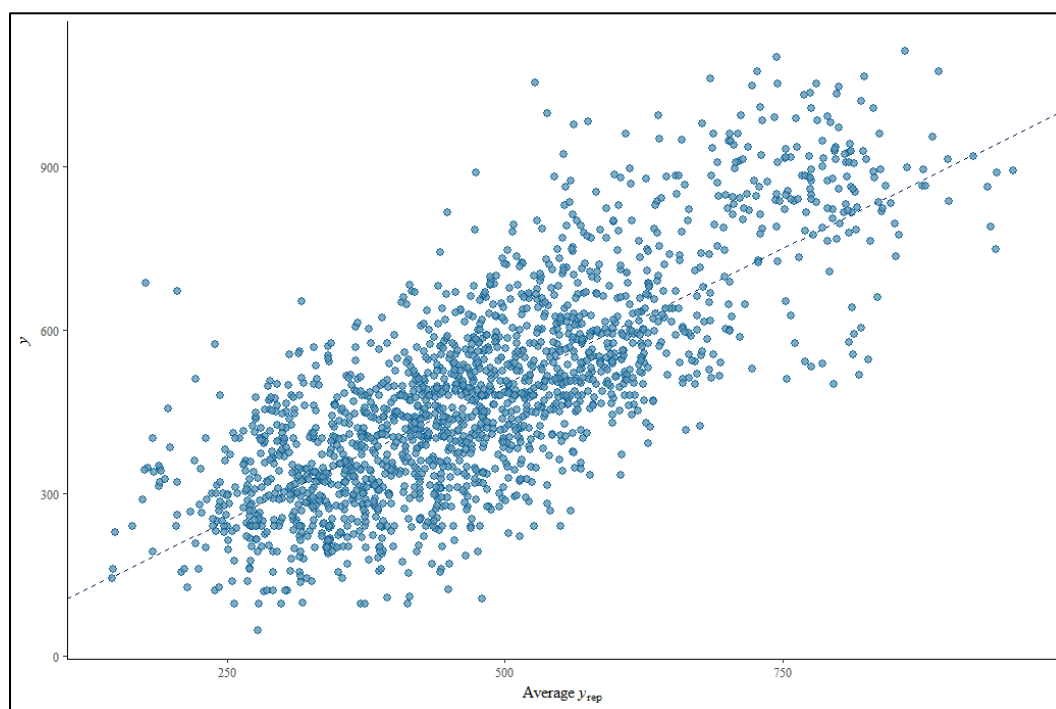
G.2- Soybean Posterior Prediction Test with Distributions (50 posterior samples)



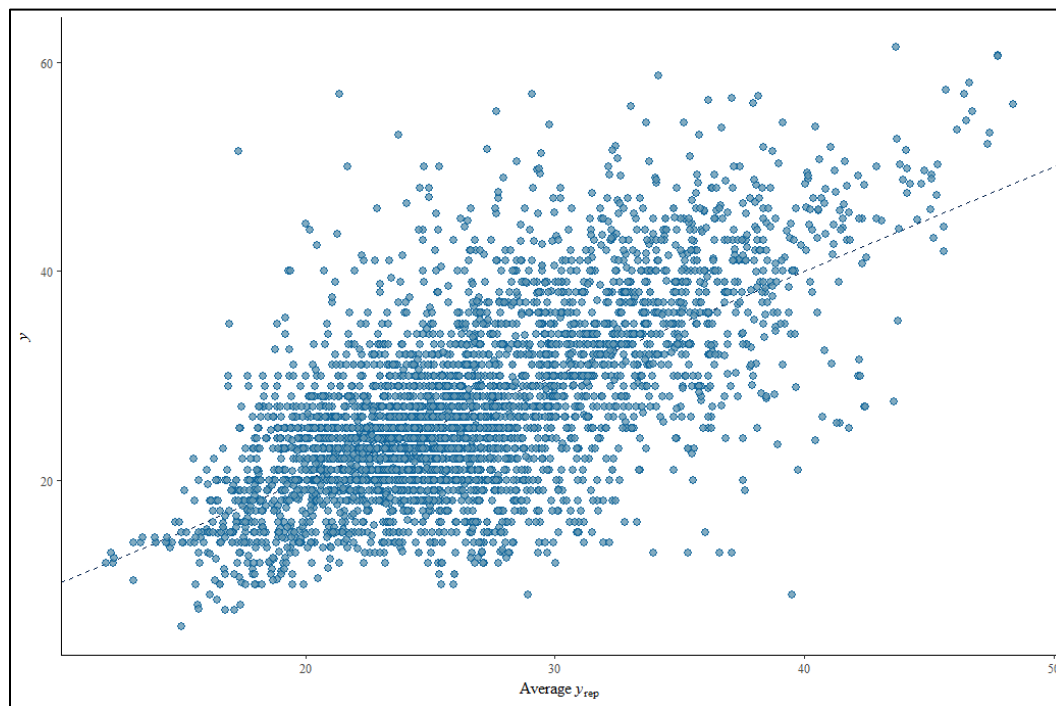
G.3- Corn Posterior Prediction Test with Distributions (50 posterior samples)



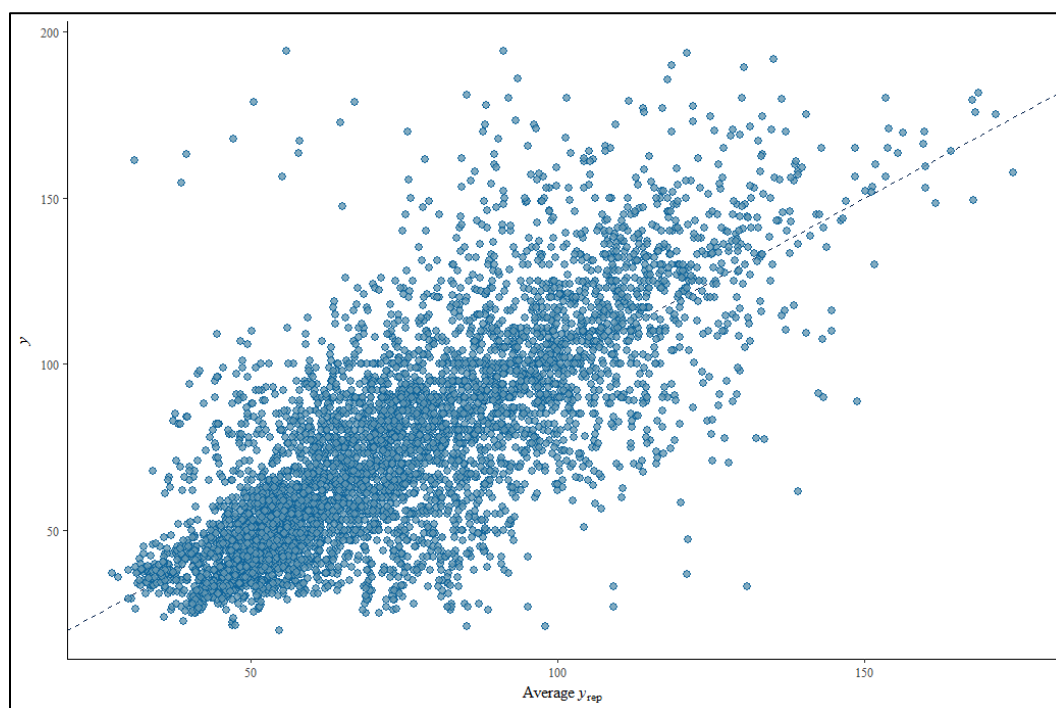
G.4- Cotton Posterior Predictive Check with Scatterplot



G.5- Soybean Posterior Predictive Check with Scatterplot



G.6- Corn Posterior Predictive Check with Scatterplot



References

- Auguie, Baptiste. “gridExtra: Miscellaneous Functions for ‘Grid’ Graphics.”. R Package Version 2.2.1 29 Feb. 2016. <https://CRAN.R-project.org/package=gridExtra>.
- Baddeley, Adrian, Ege Rubak, and Rolf Turner. “Spatial Point Patterns: Methodology and Applications with R.” London: Chapman and Hall/CRC Press, 2015. R Package Version 1.50-0. 23 Mar. 2017. <http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/>.
- Bivand, Roger and Nicholas Lewin-Koh. “maptools: Tools for Reading and Handling Spatial Objects.” R Package Version 0.8-39. 25 Mar. 2017. <https://CRAN.R-project.org/package=maptools>.
- Bivand, Roger, Tim Keitt, and Barry Rowlingson. “rgdal: Bindings for the Geospatial Data Abstraction Library.” R Package Version 1.1-10. 25 April, 2017. <https://CRAN.R-project.org/package=rgdal>.
- Bürkner, Paul-Christian. “brms: An R Package for Bayesian Multilevel Models using Stan.” R Package Version 1.6.1. *Journal of Statistical Software*, 17 Apr. 2017.
- "Climate Action Tracker." *Climate Action Tracker*. Climate Action Tracker, 2016. Accessed on 08 Dec. 2016.
- Craig, Chism. "Managing Late-Planted Cotton." UT Extension. University of Tennessee Institute of Agriculture. Web. Accessed on 04 Dec. 2016. <http://www.utcrops.com/cotton/VarietyTestingData/W074.pdf>.

Duane S, Kennedy AD, Pendleton BJ, Roweth D. “Hybrid Monte Carlo.” 1987. *Physics Letters B*, 195(2), 216–222. Accessed on 02 May, 2017.

Environmental Systems Research Institute (ESRI). *ArcGIS Release 10.4.1*. Redlands, CA, 2015. Accessed on 01 Nov. 2016.

"High Temperature Effects on Corn, Soybeans." *Corn Content from Corn and Soybean Digest*. Ohio State University, 24 July 2012. Web. Accessed on 06 Dec. 2016.

Hoffman MD, Gelman A. “The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo.” *The Journal of Machine Learning Research*, 2014. 15(1), 1593– 1623. Accessed on 02 May, 2017.

Glickman, Mark E., and David A. Van Dyk. "Basic Bayesian Methods." *Topics in Biostatistics*. Vol. 404. Humana Press, 2007. 319-38. Glicko.net. Accessed on 04 May 2017.

McClure, Angela T., and Cannon, K. *2016 Tennessee Corn Quick Facts*. 2016. Accessed on December 4, 2016.

National Center for Atmospheric Research. *Climate Change Scenarios*, 2016. Web. Accessed on 08 Dec. 2016.

Nielsen, R. L. "Silk Development and Emergence in Corn." *Silk Development and Emergence in Corn - Corny News Network (Purdue University)*. Agronomy Department Purdue University, July 2016. Web. Accessed on 06 Dec. 2016.

R Core Team. “R: A language and environment for statistical computing.” *R Foundation for Statistical Computing*. Vienna, 2016. <https://www.R-project.org/>.

- Ritchie, Glen L., Graig W. Bednarz, Phillip H. Jost, and Steve M. Brown. "Cotton Growth and Development." *The University of Georgia Cooperative Extension*, June 2007. Web. Accessed on 08 Dec. 2016.
<http://cotton.tamu.edu/General%20Production/Georgia%20Cotton%20Growth%20and%20Development%20B1252-1.pdf>.
- Schloerke, Barret, Jason Crowley, Di Cook, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, and Joseph Larmarange. "GGally: Extension to 'ggplot2'." R Package Version 1.3.0. 13 Nov, 2016. <https://CRAN.R-project.org/package=GGally>.
- Tacarindua, Custodio R.P., Tatsuhiko Shiraiwa, Koki Homma, Etsushi Kumagai, and Ryoji Sameshima. "The Effects of Increased Temperature on Crop Growth and Yield of Soybean Grown in a Temperature Gradient Chamber." *Field Crops Research* 154 (2013): 74-81. Elsevier. Accessed on 06 Dec. 2016.
- Turner, Rolf. "deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation." R Package Version 0.1-12. 22 Apr. 2017. <https://CRAN.R-project.org/package=deldir>.
- Stan Development Team. "StanHeaders: Headers for the {R} interface to {Stan}." R Package Version 2.15.0-1. 19 Apr. 2017. <http://mc-stan.org/>.
- United States Department of Agriculture. *National Agricultural Statistics Service*, 2017. Web. Accessed on 29 Apr. 2017. <https://www.nass.usda.gov/>.
- Wickham, Hadley, and Romain Francois. "dplyr: A Grammar of Data Manipulation." R Package Version 0.5.0. 24 June 2016. <https://CRAN.R-project.org/package=dplyr>.

Wickham, Hadley. “reshape2: Flexibly Reshape Data: A Reboot of the Reshape Package.”

Journal of Statistical Software, 21(12). 1-20. R Package Version 1.4.2 22 Oct. 2016.

<http://www.jstatsoft.org/v21/i12/>.

Wickham, Hadley. “ggplot2: Elegant Graphics for Data Analysis.” Springer-Verlag, 2009. R

Package Version 2.2.1. <http://ggplot2.org>.

Wickham, Hadley. “tidyr: Easily Tidy Data with ‘spread()’ and ‘gather()’ Functions.”. R

Package Version 0.6.0. 10 Jan. 2017. <https://CRAN.R-project.org/package=tidyr>.

Wood, Simon. “Generalized Additive Models: An Introduction with R.” *Chapman and*

Hall/CRC. R Package Version 1.8-17. 02 Feb, 2017. [https://cran.r-](https://cran.r-project.org/web/packages/mgcv/index.html)

[project.org/web/packages/mgcv/index.html](https://cran.r-project.org/web/packages/mgcv/index.html).

Zuur, Alain F., Elena N. Ieno, Neil J. Walker, Anatoly A. Saveliev, and Graham M. Smith.

Mixed effects models and extensions in ecology with R. New York: Springer, 2011.

Springer Link. 2009. Accessed on 01 Oct. 2016.