Does Spatial Correlation Matter in Econometric Models of Crop Yield Response and Weather?

Seong Do Yun
Graduate Research Assistant, Purdue University | yun16@purdue.edu

Benjamin M. Gramig
Associate Professor, Purdue University | bgramig@purdue.edu

Michael S. Delgado
Assistant Professor, Purdue University | delgado2@purdue.edu

Raymond J.G.M. Florax
Professor, Purdue University and VU University Amsterdam | rflorax@purdue.edu

Selected Paper prepared for presentation at the Agricultural & Applied Economics Association and Western Agricultural Economics Association Annual Meeting, San Francisco, CA, July 26-28, 2015.

Does Spatial Correlation Matter in Econometric Models of Crop Yield Response and Weather?*

Seong Do Yun
Department of Agricultural Economics, Purdue University
yun16@purdue.edu

Benjamin M. Gramig
Department of Agricultural Economics, Purdue University
bgramig@purdue.edu

Michael S. Delgado
Department of Agricultural Economics, Purdue University
degado2@purdue.edu

Raymond J.G.M. Florax
Department of Agricultural Economics, Purdue University
Department of Spatial Economics, VU University Amsterdam, The Netherlands
rflorax@purdue.edu

Abstract

Due to the rapidly growing availability and accessibility of spatially gridded weather data products, significant effort has been devoted to handling weather and climate variables properly in econometric models. It is, however, noteworthy that relatively less econometric attention is paid to how spatial correlation in weather variables and econometric models can be specified and performed. To fill this gap, this study scrutinizes the main source spatial correlation in econometric models of weather and climate variables, and implements in-sample and out-of-sample prediction analyses with spatial panel model specifications of crop yield response function. First, this paper theoretically and empirically demonstrates that the aggregation bias is a main source of spatial correlation rather than omitted weather variables. With soil variables, we specify six competing specifications of crop yield response function with pooled, fixed effects and random effects with spatially robust standard errors. From the results of prediction performances, we demonstrate that the choice of predictor (prediction models) can be motivated from the purpose of models rather than a better prediction performance. In addition, we empirically argue that the omitted socio-economic variables are not a serious econometric concern in crop yield response function of this study.

Keywords: Spatial Correlation, Panel Estimation Approach, Crop Yield Response Function, Weather Variables, Climate Change
JEL Codes: C33, C53, Q51, Q54

*. This paper is prepared for a presentation at the 2015 AAEA & WAEA Joint Annual Meeting, San Francisco, July 26 - 28, 2015.

1. Introduction

As a result of the growing availability and accessibility of spatially gridded weather data, significant effort has been devoted to the proper handling of weather and climate variables in econometric models. A review of the recent climate-economy literature describes well-developed models and estimation methods that provide reasonable solutions to econometric issues, such as nonlinearity, identification of causality, estimation of a damage function, and model specification differences between weather and climate (Dell et al., 2014). Yet, there still remain a number of unaddressed issues. Auffhammer et al. (2013) point out five major econometric pitfalls associated with using observed weather data and climate model output in economic analyses: the choice of weather data set, averaging station-level data across space, correlation between weather variables, endogenous weather data coverage, and spatial correlation. While the first four pitfalls can be solved by proper data management and have been addressed in the previous literature (Auffhammer et al., 2013; Dell et al., 2014), spatial correlation has received relatively less attention in applied econometric studies. This study fills this gap by scrutinizing spatial correlation in econometric models of crop yield response and analyzing the relative performance of alternative models starting from Schlenker and Roberts (2009) as an example of the (Deschênes and Greenstone, 2007) panel estimation approach.

Various socio-economic sectors and phenomena—e.g., agriculture, forestry and land use, population and human settlement, energy supply and demand, or industry—can be affected (or expected to be affected) by climate change and weather extremes (IPCC, 2014). Among them, agriculture has been the focus of much of the recent research on climate impacts because of the given strong relation between the physical environment and agricultural output—temperature and precipitation are direct inputs in the biological processes of plant growth (Dell et al., 2014). An agronomic crop yield response function is one of the most frequently adopted models in econometric analyses and it has a well-developed economic and

econometric story about how to include weather and climate variables. By briefly summarizing the current methodological debates between Deschênes and Greenstone (2007, 2012) and Fisher et al. (2012), this study compares various spatial econometric specifications that model spatial correlation by extending the nonlinear crop yield response function suggested by Schlenker and Roberts (2009). Because the crop yield response function itself is often adopted as a base model in the fields of climate change, food security, nutrition, and development economics, the methods and results in this study can be directly applied to studies of these and other topics. Considering that model specifications and interpretations explained in this study are not specific to crop yields, we argue that the econometric approaches in this study can be applied, without loss of generality, to a broad array of topics, that involve the relation between scio-economic outcomes and weather variables.

In the previous empirical economic studies of weather fluctuation and climate change, two fields of econometrics have attempted to take spatial correlation into account. Main-stream econometrics[1] (Auffhammer et al., 2013; Deschênes and Greenstone, 2007; Schlenker and Roberts, 2009), have adopted and broadly applied a nonparametric approach to estimate the variance-covariance (VC) matrix suggested by Conley (1999, 2008) (henceforth, Conley's method). In the spatial econometrics literature, the presence of spatial correlation in regression models is the central focus and the forms of spatial processes are explicitly specified (Anselin, 2001; Anselin et al., 2004; Baylis et al., 2011). However, both branches of econometrics have paid less attention to the question about how spatial correlation between weather variables and in econometric models can be presented, specified and performed. Given the increased use of spatially gridded weather data products—and geo-referenced data more generally—and computationally expensive methods, answers to this question can provide more appropriate model specification strategies. This study adopts spatial econometric techniques to model a

1. We borrow the notion of classification between the mainstream econometrics and spatial econometrics from spatial econometricians' view (Anselin, 2010; Gibbons and Overman, 2012).

crop yield response function that accounts for spatial correlation and heterogeneity using the best practices identified in the prior literature (Auffhammer et al., 2013; Dell et al., 2014).

The aim of this study is to compare prediction performance between non-spatial and spatial panel estimation approaches that take spatial correlation into account. To specify alternative models, we first delve into the data generating process that gives rise to spatial correlation and motivates the need for this research. This study compares the Moran's I measure of spatial autocorrelation between weather variables and their aggregation over grid cells, counties and states, and argues that aggregation over geographic units introduces bias and can be one of the main causes of spatial correlation in regression disturbance terms. Additionally, this study discusses the economic and biophysical reasoning behind spatial correlation in crop yield response based upon two spatial econometric motivations—omitted variables and spatial heterogeneity–described in LeSage and Pace (2009). In the performance comparisons, this study focuses on better prediction capability as results of temperature and precipitation impacts rather than better coefficient estimates. This is because the true data generating process that relates crop yield to spatially varying explanatory variables remains unknown and generating the best prediction is the main purpose of many climate change related studies. This also helps to address specific solutions for the controversial debate on identification and specification in spatial econometrics models pointed out by Gibbons and Overman (2012), McMillen (2012), and Pinkse and Slade (2010).

The data used in the performance comparisons are county level corn yields, temperature, total precipitation from 1981 to 2013, and soil characteristics. The spatially gridded weather and soil data are aggregated up to county levels. Due to the intensively managed nature of irrigated crops that mask the impact of precipitation on yield, our study counties are limited to those east of the 100th Meridian line as in Schlenker et al. (2006) and Schlenker and Roberts (2009). The prediction capabilities of candidate models and specifications are compared using the root-mean squared prediction error (RMSE) by performing in-sample

and out-of-sample prediction analysis.

The rest of the paper is divided into five parts. The second section provides a brief background on panel estimation approaches and the crop yield response function. The third section presents the motivation to account for spatial correlation in the specification of crop yield response models. The data are described in detail in the fourth section. The results of the performance comparison analysis are discussed in the fifth section. The paper concludes with a summary and discussion.

2. The Econometrics of Weather

Although weather is closely related to climate, recognizing the differences between these two is very important when developing an econometric model and identification strategy. Weather is the condition of the atmosphere over a short period of time, whereas climate is the behavior of the atmosphere over a relatively long period of time (Auffhammer et al., 2013). For instance, daily measured temperature is a weather variable. On the other hand, 30-year averaged temperature is a climate variable, referred to a climate "normal" by climatologists. Due to the conceptual differences, econometric setups using weather and climate measures of the same variables can result in different results (Dell et al., 2014). For the prediction capability comparison analysis, this study utilizes a panel estimation approach (Deschênes and Greenstone, 2007), which can analyze climate change impacts from the estimated sensitivity of economic outcomes to weather extremes and fluctuation. Thus, we focus on weather shocks rather than climate change in our exposition to avoid conceptual or econometric confusion. This section concludes by presenting the panel estimation approach (Deschênes and Greenstone, 2007; Schlenker and Roberts, 2009) that is the baseline model we compare to alternative models that account for spatial correlation.

2.1 General Econometric Concept

Based on Dell et al. (2014), this section briefly introduces the general econometric approaches to understand the impact of weather on the socio-economic output. An unknown functional relation of econometric models can be written as Equation (1) (Dell et al., 2014):

$$\mathbf{y} = f(\mathbf{C}, \mathbf{X}), \tag{1}$$

which links weather variables ($\mathbf{C}$) and other non-weather exogenous variables ($\mathbf{X}$) on socio-economic output ($\mathbf{y}$). For instance, $\mathbf{y}$ is crop yield, $\mathbf{C}$ is temperature and precipitation. $\mathbf{X}$ can include any characteristics that are correlated with $\mathbf{C}$ and also affect the outcome of interests, possibly by conditioning the weather response, e.g., fertilizer usage, elevation or slope of the land. A typical linear form of regression Equation (1) can be estimated using cross-sectional data:

$$y_i = \mathbf{C}_i\boldsymbol{\beta} + \mathbf{X}_i\boldsymbol{\gamma} + \varepsilon_i, \tag{2}$$

where i is an index of individual observations and ε_i is the disturbance term. The functional form can be more flexible and a nonlinear form of $\mathbf{C}$ is generally modeled (Dell et al., 2014; Schlenker and Roberts, 2009). The main issues of estimating ε_i are concerned with endogeneity caused by reverse causality, omitted variables, and over-controlling (Dell et al., 2014). The biased estimators that result from this endogeneity often distort the net effect of weather. The error process is typically modeled using robust standard errors. Particularly, if observations i are geographical units (e.g., counties, countries, or subnational regions), spatial correlation is embedded in the variance-covariance matrix by clustering at a larger spatial resolution, or assuming a spatial error process (Anselin, 2006) or distance decay structure (Conley, 1999). When applying spatial fixed effects on the data generated by a spatial dependence structure (spatial lag or spatial error), the estimation could be spurious due to the removal of spatial correlation by the spatial fixed terms (Anselin and Arribas-Bel,

2012). When interpreting the results of Equation (2), this can lead to a misinterpretation of climate impacts. Even though cross-sectional models are assumed to be the long-run equilibrium, the climate changes in the long-run make the socio-economic mechanism variant. The general equilibrium interpretation based upon a consistent mechanism in Equation (2), therefore, may not be valid.

Equation (2) can be extended to standard panel models to investigate the effects of weather shocks as:

$$y_{it} = \mathbf{C}_{it}\boldsymbol{\beta} + \mathbf{Z}_{it}\boldsymbol{\gamma} + \mu_i + \theta_t + \varepsilon_{it}, \tag{3}$$

where t indexes time, μ_i is a spatial fixed effects term, θ_t is a time fixed effects term, and $\mathbf{Z}_{it}$ contains non-weather time-varying observables. The panel estimation approach of Equation (3) is often adopted to investigate the effects of weather shocks. As stated above, weather events are different from climate. It is plausible that weather variables in $\mathbf{C}_{it}$ vary randomly over t as random draws from the distribution in a given spatial area, i.e., weather draws from the climate distribution (Dell et al., 2014). From this intrinsic property of random replication, most importantly, Dell et al. (2014) note that this "weather-shock" panel estimation approach in their article has strong identification properties through the two fixed effects terms. The fixed effects of μ_i for the spatial area absorb fixed spatial characteristics, which can be observed or unobserved including many possible omitted variables. The time fixed effects of θ_t can reflect any common time trend like technology that help ensure that the relationships of interest are identified from idiosyncratic local shocks.[2] Since Equation (3) is mainly defined as an explicit reduced form equation, it is relatively less plagued by the causal inference problem than methods that include weather variables as instruments.

In empirical applications of Equation (3), researchers often encounter a number of methodological decisions to implement the panel estimation approach (Dell et al., 2014).

2. Dell et al. (2014) point out that, in empirical studies, time fixed effects may enter separately by subgroups of the spatial area to allow for differential trends in sub-samples of the data. As an example of this, Deschênes and Greenstone (2007) include state by year fixed effects term for their county-level analysis.

Even though including $\boldsymbol{Z}_{it}$ is helpful to capture additional residual variations, the over-controlling problem stated in the discussion of the cross-section model of Equation (2) can be problematic if the endogeneity of $\boldsymbol{Z}_{it}$ caused by $\mathbf{C}_{it}$ plays a role. If the time lag of the dependent variable, y_{it-1}[3], is assumed to be a part of the true data generating process, the inclusion of the time lag with a short panel may bias coefficient estimates. Since the exclusion of the lag variable may lead to omitted variable bias, enough length of panel data is required to be adopted in most cases. The functional form of $\mathbf{C}_{it}$ is generally modeled as a flexible form rather than the linear from in Equation (3). The frequently adopted method is using a level value of $\mathbf{C}_{it}$ (e.g., exposure to growing degree days over the growing season) and giving them as several regressors (Deschênes and Greenstone, 2007; Schlenker and Roberts, 2009).

Despite its identification benefits, panel estimation of Equation (3) has difficulty connecting short-run weather fluctuation to long-run climate change. The estimation and interpretation of standard panel models of Equation (3) explain the variation in differences between observations and the grand mean. Under this logic, Seo (2010) argues that the panel estimation with fixed effects describes weather fluctuation rather than climate change that requires the concept of grand mean variation. For this reason, the application of estimates of Equation (3) using weather data to climate should start from the careful logic about how to match the short-term results to the mid-term or long-term implication. The crop yield response function adopted in this study provides a very well-developed example of this conceptual matching.

2.2 Crop Yield Response Function as a Panel Estimation

This study adopts agronomic crop yield response function to perform comparison analyses. To derive the final model specification of the baseline model suggested by Schlenker and Roberts (2009) in this study, we briefly summarize the methodological discussions from the previous literature.

3. It is noteworthy that this paper investigate inclusion of spatial lag (spatial dependence) in the later sections.

We can find the early econometric investigation in the impacts of climate change on agricultural production from the literature of production approaches. The production approach specifies a relationship between climate and agricultural output, and uses this estimate to simulate the impacts of climate change (Adams, 1989; Adams et al., 1995). Among many different types of climate change impacts, some studies start adopting an agronomic process of crop yield in economic production function (Dixon et al., 1994). The recent production approaches give more flexible forms of production function or stochastic term of inefficiency (Zhengfei et al., 2006). The early work of production approaches highly depends on cross-section data whereas the recent work is based on panel estimation. The most important contribution of production approaches in climate impact studies is involving weather (or climate) variables as inputs of production function considering agronomic knowledge. For instance, crop yield can be represented as yield production function in the form of Equation (1):

$$\mathbf{y} = f(\mathbf{H}, \mathbf{P}, \mathbf{S}, \mathbf{X}), \tag{4}$$

where $\mathbf{H}$ is temperature, $\mathbf{P}$ is total precipitation, $\mathbf{S}$ is soil property, and $\mathbf{X}$ is other non-weather exogenous variables[4]. The production function, $f(\cdot)$ is often given as a production function like Cobb-Douglass and high-order translog production function. In other words, we can adopt nonparametric forms in estimation of the production function. Both approaches can be referred to general discussion of Stochastic Frontier Analysis (SFA) and Data Environment Analysis (DEA).

The major econometric concern of adopting Equation (4) on weather impact analysis is the downward estimates due to the disregards of compensatory responses to change in weather made by profit-maximizing farmers. For example, farmers may alter their input bundles in fertilizer, mix of crops, or changes of land use due to changes of climate (Deschênes

4. The weather (or climate) vector of the general form of Equation (1) become more specified as $\mathbf{C} = (\mathbf{H}, \mathbf{P}, \mathbf{S})$. The weather input factors in crop yield response function comes from agronomic studies, for example, Williams et al. (2008)

and Greenstone, 2007). This myopic assumption of farmer's behavior is named "dumb-farmer scenario." (Mendelsohn et al., 1994)

To come up with the limited farmer's compensatory behaviors in the production function approach, Mendelsohn et al. (1994) develop the Ricardian approach, a type of hedonic approach from the crop mix change scenario. By keeping the same weather variables in Equation (4), we can present the Ricardian approach of panel data with the notation of Equation (3) as:

$$y_{it} = \sum_{j=1}^{J} \boldsymbol{\beta}_j g_j(\overline{\mathbf{C}}_{it}) + \mathbf{Z}_{it}\boldsymbol{\gamma} + \varepsilon_{it}, \tag{5}$$

where y_{it} is farmland value (or crop revenues) rather than crop yields, $\overline{\mathbf{C}}$ is a series of climate variables[5] and $g(\cdot)$ is a functional form of climate variables. For example, $g(\cdot)$ can be a combination of monthly average temperature, their squared terms, and total precipitation (Mendelsohn et al., 1994). The main philosophy adopted in Mendelsohn et al. (1994) is that the farmland value (land rent) is equal to the net yield of the highest and best use of the land under the competitive markets. Therefore, the rent of farmland can take into account the direct impacts of climate on yields of different crops as well as the indirect substitution of different inputs, introduction of different activities and other potential adaptations to different climates.

Consistent estimation of the vector $\boldsymbol{\beta}$ requires $E[g_j(\overline{\mathbf{C}}_{it})\varepsilon_{it}|\mathbf{Z}_{it}] = 0$ for each climate variable j. This assumption will be invalid if there are unmeasured permanent and transitory factors that co-variate with the climate variables (Deschênes and Greenstone, 2007). Schlenker et al. (2005) show that the irrigation factors are critical in the Ricardian approach while Schlenker et al. (2006) adopt spatial weights matrix in the disturbance process. Furthermore, Deschênes and Greenstone (2007) apply Conley's method to take into account spatial correlation, which

5. The notations of Equation (5) is adopted and modified from Deschênes and Greenstone (2007). It is noteworthy that $\overline{\mathbf{C}}$ represents climate variables (mostly, the averaged values) whereas $\mathbf{C}$ indicates weather variables (mostly, a realization of weather) in this study.

can violate the exogeneity stated above.

Deschênes and Greenstone (2007) point out that the Ricardian approach can suffer from the general specification and identification problem. They argue that it has been recognized that unmeasured characteristics (e.g., soil quality and the option value to convert to a new use) are important determinants of output and land values in agricultural settings. Consequently, the Ricardian approach may confound climate with other factors, and the sign and magnitude of the omitted variable bias is unknown. Instead of these confounding complexities, Deschênes and Greenstone (2007) exploit the random year-to-year variation in temperature and precipitation to estimate whether agricultural profits are higher or lower in years when it was warmer and wetter. The panel estimation approach suggested by Deschênes and Greenstone (2007) can be written as:

$$y_{it} = \sum_{j=1}^{J} \boldsymbol{\beta}_j g_j(\mathbf{C}_{it}) + \mathbf{Z}_{it}\boldsymbol{\gamma} + \mu_i + \theta_t + \varepsilon_{it}, \tag{6}$$

where y_{it} is agricultural profits and $\mathbf{C}_{it}$ is a realization of weather. The replacement of agricultural profits of y_{it} in Equation (4) is due to the fact that land values capitalize long-run characteristics of sites and, conditional on spatial fixed effects (μ_i), annual realizations of weather should not affect land values. In addition, it is impossible to estimate the effect of the long-run climate averages in a model with spatial fixed effects, because there is no temporal variation in climate variables of $\overline{\mathbf{C}}_{it}$. Therefore, the climate variables in Equation (5) are replaced with weather variables of $\mathbf{C}_{it}$. The inclusion of a full set of spatial fixed effects of μ_i allows absorbing all unobserved space-specific time-invariant determinants of the dependent variable. The existence of time-indicator of θ_t is to control for time differences in the dependent variable that are common regardless of location. Deschênes and Greenstone (2007, 2012) argue that inclusion of the state by year fixed effects (θ_{rt}) is the proper specification of θ_t for their county-level agricultural revenue analysis. The orthogonality condition is now given as

$E[g_j(\mathbf{C}_{it})\varepsilon_{it}|\mathbf{Z}_{it}, \mu_i, \theta_t] = 0.$

The major contribution of the panel estimation approach is that Deschênes and Greenstone (2007) provide a theoretical and empirical framework of weather variables to incorporate long-term behaviors of climate change. However, their model is only valid under the assumption that farmers cannot undertake the full range of adaptation in response to a single year's weather realization. This assumption is supported by the fact that farmers are unlikely to switch crops upon a year's weather realization but adjust the mixture of inputs. If the degree of climate change is small, which is expected in climate change studies, the panel estimation approach provides the long-run hedonic equilibrium. The value of this panel estimation approach is that it provides an alternative of production approach by simply replacing y_{it} as crop yields. The theoretical and empirical rigidity are not harmed from this replacement.

By adopting a county-level empirical study, Deschênes and Greenstone (2007) demonstrate that there is no statistically significant relationship between weather and U.S. agricultural profits, corn yields, or soybean yields. They also argue that if short-run fluctuation have no impact, then in the long-run when adaptation is possible, climate change will plausibly have little impact or could even be beneficial (Dell et al., 2014). This conclusion leads the renowned controversial debate between Deschênes and Greenstone (2007, 2012) and Fisher et al. (2012). Fisher et al. (2012) point out to the data errors in Deschênes and Greenstone (2007) and demonstrate the indeed negative impacts of weather fluctuation from the corrected data. They also argue that the state by year fixed effects (θ_{rt}) absorb almost all variations in weather and thus, it is inappropriate in the panel estimation process. Besides, they argue that the hedonic approach is still useful in the analysis and the recent nonparametric contribution can resolve the problems pointed by Deschênes and Greenstone (2007). In the following reply to Fisher et al. (2012), Deschênes and Greenstone (2012) defend that their conclusions are not harmed by the data errors or other critics by Fisher et al. (2012). In the aspect of spatial correlation, an interesting point in the debate is that they are all agree that spatial

correlation is an important factor to be considered and they all adopts Conley's method to their results.

As a solution to the debate, Schlenker and Roberts (2009) suggest a very substantial model crop yield response function that is the baseline model of this study as:

$$y_{it} = \int_{\underline{h}}^{\overline{h}} g(h)\phi_{it}dh + \mathbf{z}_{it}\gamma + \mu_i + \varepsilon_{it}, \tag{7}$$

where the first term of integral represents a nonlinear form of temperature (the detailed in Equation (9)), and $\mathbf{z}_{it}$ includes total precipitation (and its squared) and time (and its squared). The soil property is given in the spatial fixed effects terms of μ_i. Equation (7) includes the important three agronomic determinants of crop yields—temperature, precipitation, and soil—described in Equation (4). It is also based upon all substantial factors stated in the methodological summary above. Since the Equation (7) follows the panel estimation approach with the dependent variable of crop yields, it is an alternative to the production approach. Schlenker and Roberts (2009) consider spatial correlation by adopting Conley's method.

The major methodological contribution of Schlenker and Roberts (2009) is that Equation (7) resolves nonlinearity, correlation, and endogeneity of weather variables. The nonlinearity of temperature is given as a flexible functional form and the empirical demonstration of this nonlinearity is evaluated as the most crucial impact of this literature. By considering the conclusions of Roberts et al. (2012) concerning nonlinearity of precipitation and correlation between temperature and precipitation[6], Equation (7) takes into account the correlations between weather variables. Finally, the fixed effects terms and time trends allow Equation (7) to control potential issues from the omitted variables. In addition to the methodological

6. Roberts et al. (2012) empirically show that the quadratic form is statistically enough to reflect the nonlinearity of precipitation. They also demonstrate that the correlation between temperature and precipitation is not statistically significant and therefore, the inclusion of interaction terms of these two is not necessary to be essential. However, they argue that there can be serious bias from the omission of other weather factors like evaporation rate.

benefits of Equation (7), this study attempts to fill the gap of relatively less attention-paid to but important factor, i.e., spatial correlation.

3. Spatial Correlation and Panel Estimation Approach

Weather or climate variables are inherently correlated across space (Auffhammer et al., 2013) and it is well known from the spatial econometrics literature that the presence of spatial correlation in regression models can lead to serious statistical problems (Anselin, 1988, 2006). In addition, the potentially omitted weather variables can be taken into account by including spatial correlation in the regression disturbance terms through Conley's method or spatial econometrics techniques (Auffhammer et al., 2013; Deschênes and Greenstone, 2007; Schlenker and Roberts, 2009). This study, however, argues that inherent spatial correlation stated in Auffhammer et al. (2013) is not particularly in relation to econometrics. Furthermore, spatial correlation in the disturbance terms mainly comes from aggregation bias rather than the omitted weather variables. We scrutinize spatial correlation in different aggregation levels and connect an aggregation bias to the spatial specification in econometric models of the panel estimation approach. And then, we extend the panel estimation approach by Schlenker and Roberts (2009) to various spatial econometric specifications for the comparison analyses of prediction capabilities.

3.1 Weather Data: Spatial Correlation and Aggregation Bias

Due to the recent increased availability of access to a number of different types of weather (or climate) data, it is not difficult to find research adopting weather variables from many difference sources[7]. Among the various types of data products based on the output of global climate models (GCMs) (often called atmosphere-ocean GCMs, AOGCMs) or regional climate

7. Auffhammer et al. (2013) and Dell et al. (2014) describe many useful data sources for weather and climate data products.

models (RCMs), spatially gridded weather and climate data products have become popular recently. The PRISM, the Climate Research Unit (CRU) at the University of East Anglia, and data by University of Delaware (UDEL) are the three representative examples. We particularly focus on the PRISM data due to its popularity in socio-economic literature. The general logic pointed out here, however, is valid for any type of non-gridded or gridded weather data products as well.

Before discussing spatial correlation in weather variables, it is prerequisite to understand the data generating processes of the gridded (or areal) weather variables. This is because spatial correlation is already incorporated into the values of this type of weather variables. In addition, we can explain that spatial correlation issues in econometrics setup can be related to the aggregated up or averaged variables to the larger geographical boundary than their own provided geographical level, which are frequently used in economics.

One of the common mistakes from economists and non-climate specialists is that weather variables can be measured in an area as a gridded weather data provides. Unfortunately, however, this is not what weather variables stand for. Most of the weather variables including temperatures and precipitations are continuous over space and time[8] and they are measured at a certain point not at an area *per se*. Even though gridded or aggregated weather variables are computationally efficient and tactically convenient, they make unclear management issues of spatial correlation in econometric models. For example, spatial correlation in temperature means geostatistical correlation often described with the measures based on variogram rather than areal measure like Moran's I or G-statistic.

To briefly understand the meaning of spatial correlation in a data generating process of weather variable, our interest is to find a relationship between a county-level corn yield

8. In geo-statistics, this type of data is defined as geostatistical data. The spatially collected data with a certain areal boundary like county-level corn yield is called areal data. The event happens at a certain location with uncertainty such as earthquakes is classified as point process. Depending upon these data types, modeling strategy are different. For the further details, refer Cressie (1993).

in Indiana and the averaged temperature for April in the year of 2014. Figure 1 describes spatial units of data collection and their symbolic generalization.

– FIGURE 1 about here –

In the left panel of Figure 1, the averaged temperature of April comes from the grid cell data provided by the PRISM[9]. As descried above, temperature is continuous and they are measured at a certain geographic locations. The ground stations dispersed irregularly over the US measure temperature at first. However, there are not enough ground stations[10] to cover all regions as seen in '$*$' of Figure 1. To cover the missing regions, a weather data provider adopts interpolation, extrapolation or other statistical methods by adding additional information like satellite measured data. Considering all important factors affecting weather, the PRISM adopts a climate-elevation regression to produce weather a data of the 30 arcsec ($\sim$800m) sized grids for contiguous US (Daly et al., 2008). The size of 30 arcsec is the USDA-NRCS standard to describe an agricultural climate data set. To generate a tractable size of grids with 4 Km $\times$ 4 Km, which shown in Figure 1, the inverse distance weights are applied to an observation and its neighbor cells, which is the same method commonly applied in the spatial econometric literature to construct a spatial weights matrix. It is noteworthy that constructing grid cell level weather variables is the process of a transformation from a point-wise system to an area-wise system. Since this is a domain transformation between infinite to finite, it is not viable to deliver all the information of point units to area units. The PRISM data with 4 Km $\times$ 4 Km grid size, therefore, already includes spatial correlation in its generation process and the size itself is the largest grid-size to reflect the actual spatial variations in the weather variables for the contiguous U.S.

A grid cell of the PRISM is smaller than a county and each county is consisted of different

9. For the further details on data, refer to the PRISM webpages: http://www.prism.oregonstate.edu/.
10. For simplicity, we only represent the weather stations operated by the National Oceanic and Atmospheric Administration (NOAA). In the PRISM, much more weather stations operated by other institutions are reflected to their process. For the details, refer Daly et al. (2008)

numbers of grid cells. Since the available corn yields data from the National Agricultural Statistics Services (NASS) are county-level, we need to aggregated up the grid-cell to the county-level. The right panel of Figure 1 generalizes this process of data collecting units. We can assume that there are regions $A = A_1 \cup A_2 \cup \cdots \cup A_n$ and $A_i \cap A_j = \phi$ for $\forall i \neq j$. For A_i, this region is consisted of several sub-regions such as $A_i = a_{i1} \cup a_{i2} \cup \cdots \cup a_{in_i}$ and $a_{ir} \cap a_{is} = \phi$ for $\forall r \neq s$[11]. In our example, A is Indiana State, A_i is a county of Indiana, and a_{ij} is a PRISM grid cell. If the temperature of a grid cell a_{ij} given by PRISM is H_{ij} and A_i has n_i number of the PRISM grids, we can have $\overline{H}_i$ as a county temperature by applying an area weighted average for n_i number of grids. As stated above, we cannot keep all spatial variation in the PRISM grid cells due to the information loss of transformation. If we suppose the disappeared spatial variation, a grid-cell temperature can be written as:

$$H_{ij} = \overline{H}_i + \nu_{ij} = \sum_{j=1}^{n_i} w_{ij} H_{ij} + \nu_{ij}, \tag{8}$$

where ν_{ij} is a grid cell-level temperature variation from the a county mean temperature $\overline{H}_i$ and w_{ij} is an areal weights to a grid cell a_{ij}. From Equation (8), we have the two propositions in spatial correlation.

Proposition 1: The larger aggregation losses the more spatial variation.
As $n_i \to \infty$, the absolute sum of ν_{ij} diverges. i.e., $\lim_{n_i \to \infty} \sum_{j=1}^{n_i} |\nu_{ij}|$ does not define.

Proposition 2: The larger aggregation has the less spatial correlation by Proposition 1.
If k is a larger geographic aggregation level than i, then $\overline{H}_k$ is in $[\min(H_{ij}), \max(H_{ij})]$. Therefore, the spatial correlation with a larger level aggregation level (ρ_k) is less than its lower

11. In this example, we assumes that a county is consisted of several exclusive gridcells. It is, however, not a problem having the cases that some grids cells are cut by two or more counties. When this happens, researchers often apply areal weighting or omission of those cells.

level aggregation (ρ_i). i.e., $\rho_k \leq \rho_i$.

To empirically support the above two propositions, we calculate Moran'I of the grid-cell, county-levle, and state-level for yearly average temperature, growing season degree days (GDD) and total precipitation (Mar. to Aug.) from the PRISM data. All maps in Figure 2 are adopted 2013 data and the Moran's I are based upon yearly changes.

– FIGURE 2 about here –

– TABLE 1 about here –

In case of temperature and total precipitation, it is clear that the grid cell-level variables are highly spatially correlated due to their generating process. The yearly temperature of Figure 2 and Table 1 shows that county-level spatial correlations are similar. When using GDD for Mar. to Aug., the correlation itself is smaller than temperature. However, county-level Moran's I shows the similar magnitude of spatial correlation. Therefore, we can say that there has relatively less loss of spatial correlation in county level for temperature and GDD. In case of total precipitation of Table 1, county-level Moran's I are notably reduced. This is because precipitations are more spatially heterogeneous and topography and other factors affect precipitations a lot. In all three variables, state-level spatial correlations are shrunken a lot. Therefore, state-level is not a fine aggregation level to consider spatial correlation in all three variables.

From two propositions and the results of data analysis, we have two important implications on model specification with spatial correlation of weather. First, we need to use a proper level of aggregation on weather variables. Two propositions say that the area weighted weather variables do not have enough level spatial variation information. Therefore, too large scale aggregation of weather variable cannot be used in the models of spatial correlation. Second, the main source of spatial correlation in the disturbance terms is the aggregation bias rather than the omitted weather variables. In our example regression relation, the dependent variable

is crop yields that the sum of total yields for a county whereas the temperature variable is the area weighted value. Therefore, the spatial variation on the grid-cell levels within a county is reflected to the dependent variable while those are eliminated in county-level averaged temperature. Since the eliminated spatial variation plays a role to explain the variation of crop yields, the disturbance term has to take into account spatial correlation unexplained by county-level weather variables.

From the two model implications above, this study argue that spatial correlation in weather variables does not necessarily mean grid-cell level spatial correlation stated in Figure 2[12] of Auffhammer et al. (2013). Spatial correlation can exist in any geographically aggregated level of weather variables. In addition, the disappeared spatial variations due to aggregation play a main cause of spatial correlation in the regression disturbance terms. Auffhammer et al. (2013) mention that the spatial dependence of the regressors will not be a problem if the model correctly accounts for all weather variables. This study, however, argues that the spatial correlation in disturbance terms can be a problem even though researcher can include all omitted weather variables that cannot generally happen. Besides, the omitted variables in panel structure are possibly not a right reasoning to include spatial dependence structure. Even though LeSage and Pace (2009) motivate the omitted variables is one of reason to include spatially lagged variables, this is not rigid argument in the panel structure. Since spatial fixed effects terms are presented in Equation (7), for example, these fixed effects terms will take the role of omitted weather variables. If we motivate the omitted weather variables to use spatial correlation structure in the disturbance terms of the fixed effects panel model, it possibly double-counts the omitted variables and the estimates are likely to be confounded. The motivation of omitted variables will be discussed again in the next section.

12. At p.189 in Auffhammer et al. (2013), they draw the map of spatial correlation calculated based upon the values of eight surrounding neighbors' cells. This study believes that spatial correlation requires to be shown as spatially weights version of correlation like Moran's I or Geary's C rather than the Pearson type correlation.

We have an additional proposition needed to be considered in spatial correlation in the disturbance terms.

Proposition 3: The center of an area in an aggregated weather variable is unknown. Within A_i, the given weather variable is a constant over spatial boundary of A_i. Therefore, the center of weather variables in a finer level geography $\sum_{j=1}^{n_i} w_{ij} A_{ij}$ does not necessary to be matched with the centroid of A_i.

The most frequently adopted center of weather variable within the geographical boundary of an area is the centroid. Many approaches to generate spatial dependence structure are based upon the distance between two centroids, which are assumed as a known center of weather variable. Any point inside of A_i, however, is possible to be a candidate of representative point of weather. The centroid of A_i in the right panel of Figure 1 is the blue dot (●). If the true center of H_i is the blue triangle (▲), then the additional errors will be added into the estimating model. A centroid is the physical center of mass over homogeneous areal unit, but there is no evident reason that a centroid is a good representative center for deriving distances of weather variables. And obviously, there is no proper answer to why the Euclidean distance is a proper measure. Due to this incorrect measurement of distance, Conley (1999) studies two different cases with the suggested model for an exact measure and for an inexact measure of distance. The Conley's method, therefore, can be an appropriate way to resolve the indicated issues. Kelejian and Prucha (2007) propose a non-parametric spatial heteroscedasticity and autocorrelation consistent (henceforth, SHAC) estimator of the VC matrix and this is a generalized version of the Conley's Method. Both methods can provide a better alternative than the fixed structure of spatial dependence in Anselin (1988, 2006).

3.2 Spatial Specification of Crop Yield Response Function

In the prediction performance comparison analyses, this study replicates the crop yield response function used in Schlenker and Roberts (2009) with different time periods[13]. The county-level corn yield from 1981 to 2013 is adopted as an example and Equation (7) suggested by Schlenker and Roberts (2009) is named FE for notational simplicity. In an explicit form of Equation (7) can be written as:

$$\text{FE:} \qquad y_{it} = \int_{\underline{h}}^{\overline{h}} g(h)\phi_{it}dh + \gamma_1 P_{it} + \gamma_2 P_{it}^2 + \gamma_3 t + \gamma_4 t^2 + \mu_i + \varepsilon_{it}, \tag{9}$$

where P_{it} is total precipitation and t is time trend variable. In the nonlinear form of temperature, $\phi_{it}(h)$ is the time distribution of heat over the growing season in county i and year t. Following by Schlenker and Roberts (2009), we use the growing season to months March through August for corn yields. Observed temperatures during this time period range between the lower bound $\underline{h}$ and the upper bound $\overline{h}$. A time-invariant county fixed effects μ_i is to control heterogeneity, such as soil type and quality. The time trend t is included to take into account advances of technology and other time dependent trends.

As stated in the previous section, we discuss the omitted variable motivation stated in Auffhammer et al. (2013) briefly. Auffhammer et al. (2013) point out that the omitted weather variables are the main source of spatial correlation in the regression disturbance terms. This study, however, demonstrate that the main source of spatial correlation in the disturbance terms are aggregation bias that force to vanish spatial variation. Further, the omitted variable motivation by Auffhammer et al. (2013) is possible to be invalid argument from the econometrics and spatial econometrics context. Of course, the most preferred way

13. We adopt the period of 1981 to 2013 whereas Schlenker and Roberts (2009) use the period of 1950-2005. When Schlenker and Roberts (2009) published, the daily PRISM data was not available yet and they, therefore, interpolate ground station-level weather data by themselves. The PRISM recently releases the daily data for 1981-2014 (but the data of 2014 is provisional) and Roberts recommends to use this data for a replication of their study in their G-FEED blog: http://www.g-feed.com/.

is including the omitted variables as control variables, which is generally unavailable. The next general attempts to resolve the omitted variable bias can be instrumental variables (IVs) or control function approach. It is, however, really difficult to find an appropriate IV. In the weather data generating process, almost of geographical characteristics are adopted. For example, the PRISM uses elevation, longitude and latitude, topography, distance from the coast, and many other weather related factors. If we, therefore, use these variables as an IV or control factors, these can be doubly counted in the regression estimates. In addition to this, it is noteworthy that panel estimation approach (Deschênes and Greenstone, 2007) adopted in the baseline model is based upon the philosophy of absence of temporal correlation due to the intrinsic property of weather fluctuation, which is randomly and exogenously given. For this reason, the past weather data cannot be a proper IV or control factor as well.

The next approach to resolve the omitted weather variables issue can be adding fixed effects terms and this is a particularly proper way in panel model. If this is the case in the crop yield response function of Equation (7), then we will encounter difficulty to discern time invariant soil factors from the omitted weather variables. It is possible to leave the spatial fixed effects taking these confounding factors if the fixed effects terms are not dominating all the regression variation. As Auffhammer et al. (2013) argued, if there is spatial correlation in the omitted weather variables, then this is a motivation to have random effects rather than the fixed effects (LeSage and Pace, 2009). Since we have to keep soil factors in the control variable set, soil factors have to be present in the random effects panel set up as:

$$\text{RE:} \qquad y_{it} = \int_{\underline{h}}^{\overline{h}} g(h)\phi_{it}dh + \gamma_1 P_{it} + \gamma_2 P_{it}^2 + \gamma_3 t + \gamma_4 t^2 + \boldsymbol{\gamma}\mathbf{S}_i + \varepsilon_{it}, \tag{10}$$

where $\mathbf{S}$ is a vector of soil properties at county i. In empiricial analysis, we adopt four soil factors—water holding capacity (whc), soil erosivity of K-factor, organic matters in top soil, and soil pH. Equation (10) is an alternative specification motivated from the omitted weather

variables mentioned in Auffhammer et al. (2013).

One important assumption of the panel estimation is possibility of input mix even though change of crop is not allowed. Under this assumption, the possibility of spatial correlation in input mix can be questioned. Separating from the omitted weather variable motivation, this can be named the omitted socio-economic variables. Unlike the omitted weather variables, the omitted socio-economic variables are the factor of the dependent variable because the given panel estimation approach assumes the crop yield response function as the optimized value function. This motivation, therefore, provides the necessity of spatially lagged dependent variable that is believed as a fine proxy of omitted spatially correlated variables (Anselin, 2006; LeSage and Pace, 2009). Our alternative spatial panel specification from the omitted socio-economic variables can be represented as:

$$\text{FE SLAG:} \qquad y_{it} = \rho \mathbf{w}_i \mathbf{y}_t + \int_{\underline{h}}^{\overline{h}} g(h)\phi_{it} dh + \gamma_1 P_{it} + \gamma_2 P_{it}^2 + \gamma_3 t + \gamma_4 t^2 + \mu_i + \varepsilon_{it}, \qquad (11)$$

where $\mathbf{w}_i$ is ith row of spatial weights matrix.

In addition to the specification above, we can derive another possibility of spatial correlation from the geophysical processes. As we discussed weather data generating process, weather variables are inherently spatially correlated. Even though Auffhammer et al. (2013) state that the spatial dependence of the regressors will not be a problem if the model correctly accounts for all weather variables, this cannot be agreed with spatial econometricians (Anselin, 1988, 2006). One interesting point on this motivation is that the literature adopting crop yield response function with spatial econometric approaches are not taken into account this motivation (Anselin et al., 2004; Baylis et al., 2011). Particularly, Baylis et al. (2011) perform the panel estimation approaches by using farmland values as the spatial panel version of Deschênes and Greenstone (2007). They, however, overview the possibility of spatial econometric extension of the panel estimation approach and don't provide any detailed specification

motivation. For this reason, this study attempts to involve geophysical motivation as a testable simulation. The equation of geophysical motivation can be written as:

$$\text{SLX:} \qquad y_{it} = \int_{\underline{h}}^{\overline{h}} g(h)\phi_{it}dh + \gamma_1 P_{it} + \gamma_2 P_{it}^2 + \gamma_3 t + \gamma_4 t^2 + \mathbf{w}_{it}\mathbf{C}_{jt}\boldsymbol{\theta} + \mu_i + \varepsilon_{it}, \tag{12}$$

where $\mathbf{C}_{jt}$ is a vector of heat and precipitation variables. In empirical specification, we adopt all heat variables (x_{it}) and total precipitation (P) but its squared. The spatially lagged explanatory (SLX) weather variables can be combined the above specification as well.

In addition to four specifications above, we perform two additional specifications. By following the textbook type panel regressions, the pooled regression model can be played as a comparable model in Equation (13).

$$\text{Pooled:} \qquad y_{it} = \int_{\underline{h}}^{\overline{h}} g(h)\phi_{it}dh + \gamma_1 P_{it} + \gamma_2 P_{it}^2 + \gamma_3 t + \gamma_4 t^2 + \varepsilon_{it}, \tag{13}$$

The spatially robust standard errors are usually adopted with Conley's method and SHAC. Therefore, we additionally calculate Conley-type standard errors for FE and SLX. For FE SLAG model, SHAC is added for the counterpart. Since these additions are based on the residuals derived from the $N^{1/2}$ consistent estimator, we estimate FE, FE SLAG, and SLX with Generalized Method of Moments (GMM) estimation. In the case of RE model, standard errors are not separately calculated from the estimators. By considering the assumptions of no-serial (or temporal) correlation, we have the spatial random effects model suggested by Kapoor et al. (2007) as:

$$\text{KKP-RE:} \qquad y_{it} = \int_{\underline{h}}^{\overline{h}} g(h)\phi_{it}dh + \gamma_1 P_{it} + \gamma_2 P_{it}^2 + \gamma_3 t + \gamma_4 t^2 + \boldsymbol{\gamma}\mathbf{S}_i + u_{it}. \tag{14}$$

With matrix notation, $\mathbf{u}_N = \rho(I_T \otimes W_N)\mathbf{u}_N + \boldsymbol{\varepsilon}_N$ where N is the size of spatial observations, T is the size of time period, I is identity matrix, W is spatial weights matrix, and $\otimes$ is the

Kronecker product.

One may question the identification issues in spatial econometrics pointed by Gibbons and Overman (2012), McMillen (2012), and Pinkse and Slade (2010). We emphasize the fact that this study is not trying to estimate the best coefficient (given as λ or ρ in the above) of spatially dependent variables. Since the crop yield response function play a base model in many fields to have prediction caused by climate change, this study implement prediction capability comparison as the main purpose. With this specific purpose, this study follows the usage of spatial econometric models as a predictor in Dormann et al. (2005). The interested reader may find more abundant spatial panel specifications from Elhorst (2014) and Millo (2014).

3.3 Additional Specifications

To implement empirical estimations for the regression models above, we need to have additional specification in the functional from of heat and the spatial weights matrix of FE SLAG, KKP-RE, and spatial kernel density function in spatially robust standard errors. Schlenker and Roberts (2009) adopts three different specification on the heat integrals—step function, m-th order Chebychev polynomials, and piecewise liner–for the approximation of $g(h)$. We apply the m-th order Chebychev polynomials to have a smooth representation of temperature impacts as:

$$\begin{aligned}\int_{\underline{h}}^{\overline{h}} g(h)\phi_{it}dh &= \sum_{k=1}^{m}\delta_k \sum_{h=-1}^{39} T_k(h+0.5)\left[\Phi_{it}(h+1)-\Phi_{it}(h)\right] \\ &= \sum_{k=1}^{m}\delta_k x_{it,k},\end{aligned}$$

where $T_k(\cdot)$ is an m-th order Chebychev polynomial. Adopting the results of Schlenker and Roberts (2009), we choose eighth order Chebychev polynomial and thus, $k = 8$ in our

specification.

To keep the consistency of spatial correlation arguments in the previous section, we adopt the first order queen spatial weights matrix and apply the row standardization scheme that supports the consistent estimator of irregular spatial process (Kelejian and Prucha, 1999). In Conley's method and SHAC estimator, we need to specify the spatial kernel density function. Conley (2008) states that a uniform kernel density can be an operationally convenient choice and, therefore, we adopt it. In the SHAC variance-covariance matrix estimator, we apply the Parzen kernel density that is used in Kelejian and Prucha (2007). In both kernel density functions, the bandwidth is given as six-nearest distances to satisfy $N^{1/2}$ consistency condition descried in Kelejian and Prucha (2007). It is noteworthy that six is very close to the average number of spatial links (5.6130) in the spatial weights matrix adopted in FE SLAG and RE of the following empirical analysis.

For all data management and estimation process, the most recent version of R (3.2.0) at the point analysis is adopted[14] with *plm* packages (Croissant and Millo, 2008) and *splm* packages (Millo and Piras, 2012).

4. Data and Estimation Results

To implement prediction performance comparison among the models above, we adopt corn yields from the National Agricultural Statistics Services (NASS) by the United States Department of Agriculture (USDA), weather variables from the PRISM data, and soil data from the gridded Soil Survey Geographic (gSSURGO) database. The constructed balanced panel data is 1,964 counties for 33 years (i.e., $N = 1,964 \times T = 33$). Area weighted average is applied to all weather and soil variables from the gridded PRISM (4 Km $\times$ 4 Km resolution) and gSSURGO (10 m $\times$ 10 m resolution).

14. In data management steps, parallel computing is adopted on the Intel(R) Core(TM) i5-2450M CPU of 2.50GHz with 8 GB RAM and 64 bits Windows system.

4.1 Temporal and Geographical Range

This study adopts the daily temperature and precipitation data from the PRISM. The fully available daily weather data period in the PRISM is 1981 to 2014. Since the current version of the year 2014 data is provisional[15], this study takes the study period as 1981 to 2013 (33 years). The geographical boundary of this study is the east counties of the 100th Meridian to avoid irrigation issue, which forms an endogeneity in the crop yield response function. This is the same geographical boundary adopted in Schlenker and Roberts (2009). Since the county and state geographical boundaries have been changed for the past 33 years, we adopt the most detailed 500k (1:500,000) county boundaries map of 2013 from the Census Bureau[16]. All variables are matched with this map. The Figure 3 shows the geographical boundary and frequencies of yield data from 1981 to 2013.

– FIGURE 3 about here –

The histogram of Figure 3 (a) describes county-level yearly corn harvesting frequencies of the east counties of the 100th Meridian line. Among 2,510 east counties, 2,393 counties have produced corn at least once. Within 2,393 counties, 429 counties have not produced corn for some years while 1,964 counties have done for full 33 years. The map of Figure 3 (b) represents geographical boundaries of study area. The non-colored counties are the west counties of the 100th Meridian line or non-corn growing counties. Among 2,393 colored counties, the skyblue-colored areas are 1,964 counties of 33 years corn growing history while the pink regions are excluded 429 counties due to its shorter corn growing history.

In the empirical studies of this study, we construct $N = 1,964 \times T = 33$ of balanced panel data by using only 1,964 counties of 33 years corn growing data. While estimation methods of non-spatial panel regression models are broadly applicable in both balanced and unbalanced data, spatial panel models with unbalanced data are currently under suggesting

15. The last accessed to the PRISM FTP is January 10^{th}, 2015.
16. https://www.census.gov/geo/maps-data/data/cbf/cbf_counties.html

and developing stages (Pfaffermayr, 2009; Wang and fei Lee, 2013). By excluding the pink colored counties, 14.64 % of available observations are not used. Selection issues can be questioned from this exclusion. We argue, however, that the selection bias is not serious in this study due to two reasons. Schlenker and Roberts (2009) empirically demonstrate robustness of their estimation results by comparing the east counties only to the all counties of 48 contiguous states. In addition, the study area in Figure 3 covers most major corn belt regions and the exclusion itself does not create any island county that causes non-link spatial process in spatial economic models.

4.2 Variables

The six comparable models in this study need variables of corn yield, temperature, precipitation, and soil. Table 2 presents the explanation of variables adopted in this study and descriptive statistics.

– TABLE 2 about here –

The past county-level corn yields (bu/ac) are extracted from the NASS. Based on the map of 2013, we adjust some counties aggregated or disaggregated counties by using the bridge table provided by the NASS[17]. The counties disappeared in 2013 are excluded. In regression analysis, we transform corn yields as the logarithmic form. Therefore, the corn yield represented in regression equations is $y_{it} = log(\text{corn yield}_{it} + 1)$.

In crop yield response function, many different temperature measures can be applied. The most simplest measure year or monthly average temperature and their interactions, for example, Lobell and Burke (2010) and Mendelsohn et al. (1994). Deschênes and Greenstone (2007) point out that the cumulative heat exposure is agronomic ally proper measure of the role of temperature in any plant growth. Many recent careful studies now use GDD as the heat measure on crop yield response function (Deschênes and Greenstone, 2007; Roberts

17. http://www.nass.usda.gov/Data_and_Statistics/County_Data_Files/Frequently_Asked_Questions/county_list.txt

et al., 2012; Schlenker et al., 2006). By following Schlenker and Roberts (2009), this study estimates the GDD by using a sinusoidal curve between the daily minimum and maximum temperature in PRISM data[18]. We construct the GDD of Mar. to Aug. in each 1° C degree temperature interval between -5 °C and +50 °C over whole 4Km by 4Km PRISM grid cells. At each degree over all PRISM grid cells in a county, we finally derive county-level GDD with the area-weighted average. The agricultural area in each cell is obtained from Schlenker's web-link[19]. In Table 2, the descriptive statistics of optimal GDD for corn growing between +5 °C and +35 °C is presented. In regression analysis, we lump all time a corn plant is exposed to a temperature below 0 °C into one category that indicates freezing level as Schlenker and Roberts (2009) do. Similarly, the harmful temperature above 39 °C is lumped into one category. The total precipitation in mm between March to August within each year is adopted as *ppt* variable as shown in Table 2. This 1° C degree heat measure covers all the optimal or harmful temperature ranges stated in agronomy literature. Figure 4 describes distribution of yield, temperature and total precipitation.

– FIGURE 4 about here –

It is noteworthy that the GDD distribution in Figure 4 has a thicker for higher temperature range than the GDD distribution in Schlenker and Roberts (2009). This is because our study period includes more recent climate change impacts (the warmer temperature trends and 2012 droughts) for 2006 to 2013, which are not included in Schlenker and Roberts (2009). It is, however, both distributions are very similar.

In this study, we very carefully select soil variables for RE and KKP-RE models. Many previous economic-climate literature adopt soil structure (proportion of sand, clay, and silt) and other soil properties such as drainage or erosivity as a separate variable. For example, Schlenker et al. (2006) and Baylis et al. (2011) adopt the percentage of clay and the

18. The detailed methods can be referred from the University of California Statewide Integrated Pest Management (UCIPM) program: http://www.ipm.ucdavis.edu/WEATHER/ddconcepts.html
19. http://www.wolfram-schlenker.com/dailyData.html

premeability in a regression model. It is, however, noteworthy that soil structure is a major determinant of other soil properties. The more clay means higher premeability, the more sandy soil shows higher drainage and erosivity factor. In many cases, therefore, inclusion of soil composition and other soil properties in a regression model can be double counting of a soil property. Wolkowski (2005) classifies ten important soil factors on crop production system[20]. Among those, we select four factors—organic matter (om), erosion (K-factor), drainage (water holing capacity, whc), and soil pH—without including soil composition in empirical analysis. From 10m by 10m gSSURGO data, we first calculate each depth averaged soil factors across soil horizon. Then, we apply the area weighted average on all soil variables within each county. The values in this study are depth and area weighted averages of each county. Since county-level soils are assumed to be invariant over times, the number of observations in soil variables of Table 2 is 1,964 counties. To verify the rigidity of our soil calculation, we plot four soil values over study area as shown in Figure 5.

– FIGURE 5 about here –

From the US Geological Survey (USGS), we confirm that all four soil variables show the similar spatial distribution with the USGS maps.

4.3 Estimation Results

With variables defined in Table 2, we estimate six comparable models—Pooled, FE, FE SLAG, RE, and KKP-RE—in the previous section with proper estimation methods. Table 3 shows the model estimation results.

– TABLE 3 about here –

In Table 3, we present standard errors in parenthesis and spatially (robust) standard errors in brackets. Due to two standard errors with one estimate in some models, we mark asterisks

20. This is an extension article from University of Wisconsin: http://www.soils.wisc.edu/extension/area/horizons/2005/SoilQualityCropProduction.pdf. The ten important factors are 1. Organic matter, 2. Crop appearance, 3. Earthworms, 4. Erosion, 5. Tillage ease, 6. Drainage, 7. Soil structure, 8. Soil pH, 9. Soil test P and K, and 10. Yield.

of p-values on the standard errors instead of on the estimates. As Deschênes and Greenstone (2007, 2012) and Fisher et al. (2012) discussed, spatially robust standard errors with Conley's method and SHAC estimator are larger than the standard errors in parenthesis and thus, the p-values with spatially robust standard errors are improved.

As shown in Table 3, overall individual significance levels are good enough for all models. As expected, FE and FE SLAG models show the same directions of sign for all estimates as RE and KKP-RE do. It is noteworthy that the terms reflecting spatial correlation are statistically significant. The spatial lag coefficient, λ in FE SLAG is positive and statistically significant with 1 % significance level. All spatially lagged variables of temperature and precipitation variables in SLX model are statistically significant. The ρ coefficient in KKP-RE model derived from the first stage of optimization of the moment condition is 0.6406 and the magnitude itself is very similar to the estimated λ.

Across all models, precipitation and time trend variables shows the same and expected direction of signs. In contrast, the estimates of Chebyshev polynomial coefficients change relatively largely across models. This means that the nonlinearity in each model are much different from each other due to the different assumptions of spatial correlation. It is, however, noteworthy that the signs of the highest Chebyshev polynomial coefficients are all negative and this supports the inverse U-shaped global nonlinear temperature trends in Schlenker and Roberts (2009). All soil variables in RE and KKP-RE are statistically significant except organic matters (om) and have the same positive signs.

As discussed earlier, the true data generating process of crop yield response function is unknown. The six comparable models are competing models in having better estimation and prediction. Due to its high importance of crop yield response function in prediction in many related studies, this study aims to implement prediction performance comparison analyses. The estimation results in Table 3 are directly used for in-sample prediction analysis in the next section.

5. Performance Comparison Analysis

This study implements two types of prediction analysis. The first is in-sample prediction performances from the results of Table 3. The second is out-of-sample prediction by simulating 1,000 year-to-tear sampling replications.

5.1 In-Sample Prediction Performance

By adopting the estimates in Table 3, we calculate the mean root squared errors (RMSE) between observed corn yields (y_{it}) and predicted corn yields ($\hat{y}_{it}$). Figure 6 presents the averaged RMSE over 33 years in each county.

– FIGURE 6 about here –

Across all six models, the third quantile (75 %) of RMSE is less than 0.8 except FE SLAG. To make all results are comparable with the same scale, we assign the counties having the RMSE greater than one as one in Figure 6. In in-sample prediction, FE SLAG model shows the poorest in-sample prediction performance among all six models and its county-level average RMSEs are greater or equal to one over all study area. Contrast to the results of FE SLAG, the other five models show relatively less RMSEs across all counties. From the RMSEs in five models, we can indicate obvious spatial patterns in RMSE. The west counties in study area show less accuracy in in-sample prediction. In addition, relatively larger RMSEs are recorded along with the Appalachian Mountains. As discussed in the earlier sections on spatial correlation, this result can be explained by two geographical factors. First, we intentionally select the east counties of the 100th Meridian line to avoid irrigation issues. The west counties in the study area, however, are not fully free from the irrigation efforts. The other reason is topographical complexity of the counties along with Appalachian Mountains. The temperature and precipitation in bumpy areas shows more dynamical changes than those in flat regions. Due to our county-level aggregation, this alteration becomes smoothed and makes less accurate predictions.

The first column of Table 4 represents the average of in-sample RMSE over whole 64,812 observations.

– TABLE 4 about here –

SLX and FE shows the best prediction performance in in-sample prediction. The averaged RMSE of RE is only 0.004 smaller than the first two, and its value of KKP-RE is approximately 0.03 smaller than the averaged RMSE of RE. Due to these tiny differences, it is hard to confirm the goodness of prediction performances among the first four models. Considering the fact that the Pooled model has no spatial correlation consideration, the averaged RMSE of FE SLAG is notably worse than other models. At least in in-sample prediction performance, the spatial correlation come from changes of input mix seems not to be a strong motivation.

5.2 Out-of-Sample Prediction Performances

The in-sample prediction can play a role to explain the goodness of fit measure. To compare direct prediction performances, the more general approach is (pseudo) out-of-sample prediction. Among 1,000 times of sampling replication, we randomly select 27 of the 33 years in our full sample. Relative performance is measured according to the accuracy of each model's prediction for the omitted 6 years of the sample (approximately 18 %). By the rule of thumb, 15 % to 20 % of out-sample selection is generally performed. Due to considerable spatial correlation across study areas, we sample years of year-to-year random weather fluctuation instead of observations (Schlenker and Roberts, 2009).

The second column of Table 4 represents the averaged RMSE of out-of-sample prediction. The order of models in table is arranged by the superiority of out-of-sample prediction performances, i.e., the smaller RMSE to the larger. From the 1,000 replications, SLX and FE show the best prediction among six models. The following models are KKP-RE and RE. Again, the differences between the first two and the second two are marginally small as about 0.03. From the results, we can conclude that FE SLAG is a relatively poor predictor of crop

yield response function.

The rest of five columns present pair-wise Welch t-test against the null hypothesis of equal RMSE under unequal variances. The smaller statistic means two models are performing the same prediction according to the produced RMSE. The larger means tow models are not performing the same predictions. The Welch test results indicate that SLX and FE produce the same RMSE but the other models do not. Particularly, FE SLX is expected to have the similar performances with FE. Yet, it turns out that this expectation is statistically not true. The RMSE from KKP-RE is statistically congruent to the RMSE of RE. The other models, however, do not comparable to RE or KKP-RE.

With the motivations of six models, the results of in-sample and out-of-sample performances provide two important intuitions for researchers and practitioners interested in prediction of crop yield response function corresponding to climate change impacts. First, the inclusion of spatial correlation into crop yield response function can be motivated by the management or adaptation purposes rather than the improvement of prediction performances. As discussed, the four models—SLX, FE, RE and KKP-RE—are not statistically discernible in prediction sense. Due to the different motivation in four models, each model can be selected to fit the purpose of prediction. SLX model is better to reflect direct geo(bio)physical process. The impact of heat islands or freezing poles can be an appropriate case of SLX. FE models can reflect direct spatial heterogeneity in fixed effects terms. In economics, this spatial heterogeneity sometimes play a more important role rather than spatial dependence of weather variables. RE and KKP-RE models have soil variables as regressors. Since many climate impacts on crop yields and its adaptation processes are related to carbon emission, N-fertilizer usages, and irrigation matters, RE or KKP-RE model can be a proper model strategy to these cases.

The second modeling intuition is that the omitted soico-economic variables (FE SLAG) are possibly not an important factor to be considered. Since the model of Deschênes and

Greenstone (2007) allows the changes input mixes, the omission of important socio-economic variables in the model sometimes becomes a serious modeling issue. From the results of in-sample and out-of-sample prediction performances, at least, we demonstrate that the inclusion of spatially lagged corn yields is not helpful to increase prediction performance. In addition to this, there is a possibility that **WY** is not a good instrument or proxy of the omitted socio-economic variables unlike spatial econometricians' expectation. It is noteworthy that this is the exactly opposite conclusion from the panel estimation approaches on land values in Baylis et al. (2011). As an alternative of production approach, the crop yield response function of panel estimation does not support the role of spatially lagged dependent variable in prediction performance aspect.

6. Conclusion

Due to the rapidly growing availability and accessibility of spatially gridded weather data, significant effort has been devoted to handling weather and climate variables properly in econometric models. Among many potential pitfalls, this paper aims to analyze the role of spatial correlation in economics of weather and climate. For the model specification purposes, this study first demonstrates empirical necessities of embedding spatial correlation in econometric models by scrutinizing spatial correlations with the geographic aggregation levels and discussing potential economic and geo(bio)physical reasoning. By adopting the crop yield response function in Schlenker and Roberts (2009), this study classifies six panel models to be compared—Pooled, FE, FE SLAG, RE, and KKP-RE. With the corn yields during 1981 to 2013, this paper empirically implements prediction comparison analysis through in-sample and out-of-sample prediction performance. We remark six conclusion and contributions of this study as follows.

First, the major source of spatial correlation in weather and climate variable is aggregation bias rather than the omitted variable bias. With the three propositions and empirical analysis

of temperature, GDD and precipitation, we argue that the aggregation of weather variables is not properly managed in the regression setup. To take into account this aggregation bias, we need to consider specification strategies on how to reflect this spatial correlation.

Second, spatial correlation caused by aggregation bias and omitted variables can be involved as different model specification strategy. If the aggregation bias is the major issue, then these spatial correlations are lumped into the disturbance terms of an econometric model. By adopting spatially robust standard errors like Conley's method or SHAC, this can be properly managed. The direct inclusion of spatially lagged weather variables can be another specification strategy. If the omitted socio-economic variable plays an important role, then spatially lagged dependent variable can be an instrument or proxy variable of it. In the case of spatial correlation caused by the omitted weather variables, then the random effects model with soil variables is an alternative specification rather than fixed effects model. Based on these specifications, we classify six competing models of crop yield response function.

Third, soil composition variables can create double counting problem with other soil properties in crop yield response function. With careful review on agronomic-soil literature, this paper argues that many soil properties are derived from the soil composition of sand, clay and silt. In crop yield response function, we include four soil properties—water holding capacity, erosivity K-factor, organic matters, and soil pH—but soil structure in random effect models. We believe that this soil specification is valid for other models including soil characteristics.

Fourth, the spatially robust standard errors by Conley's method and SHAC estimators can provide a better confidence intervals of prediction. From the estimation results of six models in Table 3, we can find the general patterns of larger standard errors in Conley's method or SHAC than non-spatial standard errors. This, therefore, makes a better statistical test result in individual estimates and a narrower confidence interval of predictor.

Fifth, the motivation to choose prediction models can be a better economic motivation

rather than a better prediction. As shown in in-sample and out-of-sample prediction performances, the four models of SLX, FE, KKP-RE and RE produce the statistically indiscernible prediction performances. This means, therefore, the selection prediction model can be originated from the benefits of each model setup. SLX model can be better in regional level analysis, FE is possibly good in spatial heterogeneity analysis, and RE/KKP-RE can provide benefits of adaptation processes corresponding to climate change.

Lastly, there is a possibility that the bias from omitted socio-economic variable in crop yield response function is not serious as expected. The prediction performance of FE SLAG model is the worst in both in-sample and out-of-sample prediction. If the spatially lagged dependent variable is a good proxy as stated in spatial econometrics, then the changes of input mix in Deschênes and Greenstone (2007) is well reflected in the corn yields itself and therefore, there is little spatial correlations are generated from this omitted variables. At least, in empirical crop yield response function of this study, it turns out that the omitted socio-economic variables are not serious source of spatial correlation.

References

Adams, R. M.: 1989, Clobal climate change and agriculture: An economic perspective, *American Journal of Agricultural Economics* **71**(5), 1272–1279.

Adams, R. M., Fleming, R. A., Chang, C.-C., a. McCarl, B. and Rosenzweig, C.: 1995, A reassessment of the economic effects of global climate change on u.s. agriculture, *Climatic Change* **30**(2), 147–167.

Anselin, L.: 1988, *Spatial Econometircs: Methods and Models*, Dordrecht: Kluwer Academic Publishers.

Anselin, L.: 2001, Spatial effects in econometric practice in environmental and resource economics, *American Journal of Agricultural Economics* **83**(3), 705–710.

Anselin, L.: 2006, Spatial econometrics, *in* T. C. Mills and K. Patterson (eds), *Palgrave Handbook of Econometrics: Volume 1, Econometrics Theory*, Basingstoke: Palgrave Macmillan.

Anselin, L.: 2010, Thirty years of spatial econometrics, *Papers in Regional Science* **89**(1), 3–25.

Anselin, L. and Arribas-Bel, D.: 2012, Spatial fixed effects and spatial dependence in a single cross-section, *Papers in Regional Science* **92**(1), 3–18.

Anselin, L., Bongiovanni, R. and Lowenberg-DeBoer, J.: 2004, A spatial econometric approach to the economics of site-specific nitrogen management in corn production, *American Journal of Agricultural Economics* **86**(3), 675–687.

Auffhammer, M., Hsiang, S. M., Schlenker, W. and Sobel, A.: 2013, Using weather data and climate model output in economic analyses of climate change, *Review of Environmental Economics and Policy* **7**(2), 181–198.

Baylis, K., Paulson, N. D. and Piras, G.: 2011, Spatial approaches to panel data in agricultural economics: A climate change application, *Journal of Agricultural and Applied Economics* **43**(3), 325–338.

Conley, T. G.: 1999, GMM estimation with cross sectional dependence, *Journal of Econometrics* **92**(1), 1–45.

Conley, T. G.: 2008, Spatial econometrics, *in* S. N. Durlauf and L. E. Blume (eds), *The New Palgrave Dictionary of Economics*, Basingstoke: Palgrave Macmillan, pp. 741–747.

Cressie, N. A. C.: 1993, *Statistics for Spatial Data*, New York: Wiley.

Croissant, Y. and Millo, G.: 2008, Panel data econometrics in R: The plm package, *Jornal of Statistical Software* **27**(2), 1–43.

Daly, C., Halbleib, M., Smith, J. I., Gibson, W. P., Doggett, M. K., Taylor, G. H., Curtis, J. and Pasteris, P. P.: 2008, Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous united states, *International Journal of Climatology* **28**(15), 2031–2064.

Dell, M., Jones, B. F. and Olken, B. A.: 2014, What do we learn from the weather? the new climate-economy literature, *Journal of Economic Literature* **52**(3), 740–798.

Deschênes, O. and Greenstone, M.: 2007, The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather, *The American Economic Review* **91**(1), 354–385.

Deschênes, O. and Greenstone, M.: 2012, The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather: Reply, *The American Economic Review* **102**(7), 3761–3773.

Dixon, B. L., Hollinger, S. E., Garcia, P. and Tirupattur, V.: 1994, Estimating corn yield response models to predict impacts of climate change, *Journal of Agricultural and Resource Economics* **19**(1), 58–68.

Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M. and Wilson, R.: 2005, Methods to account for spatial autocorrelation in the analysis of species distributional data: A review, *Ecography* **30**(5), 609–628.

Elhorst, J. P.: 2014, *Spatial Econometrics: From Cross-sectional Data to Spatial Panels*, SpringerBriefs in Regional Science, Dordrecht: Springer.

Fisher, A. C., Hanemann, W. M., Roberts, M. J. and Schlenker, W.: 2012, The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather: Comment, *The American Economic Review* **102**(7), 3749–3760.

Gibbons, S. and Overman, H. G.: 2012, Mostly pointless spatial econometrics?, *Journal of Regional Science* **52**(2), 172–191.

IPCC: 2014, Summary for policymakers, *in* O. Edenhofer, R. Pichs-Madruga, E. F. Y. Sokona, S. Kadner, K. Seyboth, A. Adler, I. Baum, S. Brunner, P. Eickemeier, B. Kriemann, J. Savolainen, S. Schlomer, C. von Stechow, T. Zwickel and J. Min (eds), *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

Kapoor, M., Kelejian, H. H. and Prucha, I. R.: 2007, Panel data models with spatially correlated error components, *Journal of Econometrics* **140**(1), 97–130.

Kelejian, H. H. and Prucha, I. R.: 1999, A generalized momonets estimator for the autogressive parameter in a spatial model, *International Economic Review* **40**(2), 509–533.

Kelejian, H. H. and Prucha, I. R.: 2007, Hac estimation in a spatial framework, *Journal of Econometrics* **140**(1), 131–154.

LeSage, J. and Pace, R. K.: 2009, *Introduction to Spatial Econometrics*, New York: CRC Press.

Lobell, D. B. and Burke, M. B.: 2010, On the use of statistical models to predict crop yield responses to climate change, *Agricultural and Forest Meteorology* **150**(11), 1443–1452.

McMillen, D. P.: 2012, Perspectives on spatial econometrics: Linear smoothing with structured models, *Journal of Regional Science* **52**(2), 192–209.

Mendelsohn, R., Nordhaus, W. D. and Shaw, D.: 1994, The impact of global warming on agriculture: A ricardian analysis, *The American Economic Review* **84**(4), 753–771.

Millo, G.: 2014, Maximum likelihood estimation of spatially and serially correlated panels with random effects, *Computational Statistics and Data Analysis* **71**, 914–933.

Millo, G. and Piras, G.: 2012, splm: Spatial panel data models in R, *Jornal of Statistical Software* **47**(1), 1–38.

Pfaffermayr, M.: 2009, Maximum likelihood estimation of a general unbalanced spatial random effects model: a monte carlo study, *Spatial Economic Analysis* **4**(4), 467–483.

Pinkse, J. and Slade, M. E.: 2010, The future of spatial econometrics, *Journal of Regional Science* **50**(1), 103–117.

Roberts, M. J., Schlenker, W. and Eyer, J.: 2012, Agronomic weather measures in econometric models of crop yield with implications for climate change, *American Journal of Agricultural Economics* **95**(2), 236–243.

Schlenker, W., Hanemann, W. M. and Fisher, A. C.: 2005, Will u.s. agriculture really benefit from global warming? accounting for irrigation in the hedonic approach, *The American Economic Review* **95**(1), 395–406.

Schlenker, W., Hanemann, W. M. and Fisher, A. C.: 2006, The impact of global warming on u.s. agriculture: An econometric analysis of optimal growing conditions, *The Review of Economics and Statistics* **88**(1), 113–125.

Schlenker, W. and Roberts, M. J.: 2009, Nonlinear temperature effects indicate severe damages to u.s. crop yields under climate change, *Proceedings of the National Academy of Sciences of the United States of America (PNAS)* **106**(37), 15594–15598.

Seo, S. N.: 2010, An essay on the impact of climate change on us agriculture: Weather fluctuations, climatic shifts, and adaptation strategies, *Climate Change* **121**(2), 115–124.

Wang, W. and fei Lee, L.: 2013, Estimation of spatial panel data models with randomly missing data in the dependent variable, *Regional Science and Urban Economics* **43**(3), 521–538.

Williams, C. L., Liebmana, M., Edwardsb, J. W., Jamesc, D. E., Singerc, J. W., Arritta, R. and Herzmanna, D.: 2008, Patterns of regional yield stability in association with regional environmental characteristics, *Crop Science* **48**, 1545–1559.

Zhengfei, G., Lansink, A. O., van Ittersum, M. and Wossink, A.: 2006, Integrating agronomic principles into production function specification: A dichotomy of growth inputs and facilitating inputs, *American Journal of Agricultural Economics* **88**(1), 203–214.

Table 1: Moran's I: Yearly Average Temperature, Growing Season Degree Days (GDD) (Mar. to Aug.), and Total Precipitation (Mar. to Aug.)

	Avg. Temperature			GDD			Total Precipitation		
year	Grid	County	State	Grid	County	State	Grid	County	State
1981	0.9876	0.9778	0.7233	0.7189	0.6511	0.0075	0.9484	0.6532	0.1729
1982	0.9889	0.9827	0.7581	0.7599	0.7974	0.1763	0.9463	0.6922	0.2831
1983	0.9875	0.9761	0.7192	0.7477	0.7683	0.0719	0.9460	0.6684	0.3582
1984	0.9890	0.9802	0.7337	0.7563	0.7828	0.0753	0.9516	0.7750	0.1996
1985	0.9894	0.9830	0.7517	0.7337	0.7666	0.1085	0.9547	0.6934	0.2869
1986	0.9884	0.9808	0.7530	0.7066	0.6063	0.0522	0.9459	0.6009	0.2094
1987	0.9876	0.9744	0.7334	0.6859	0.6544	-0.0206	0.9432	0.6150	0.1180
1988	0.9877	0.9765	0.7164	0.7082	0.7657	0.0470	0.9417	0.6741	0.1626
1989	0.9883	0.9809	0.7317	0.7476	0.7391	0.1756	0.9630	0.7628	0.1744
1990	0.9885	0.9796	0.7468	0.7142	0.7027	0.1161	0.9581	0.6430	0.2065
1991	0.9885	0.9792	0.7350	0.7570	0.7812	0.0375	0.9513	0.7935	0.3496
1992	0.9877	0.9783	0.7313	0.7405	0.7628	0.2801	0.9510	0.7523	0.1544
1993	0.9892	0.9814	0.7440	0.7311	0.6010	0.0767	0.9549	0.6393	0.1791
1994	0.9879	0.9797	0.7381	0.7298	0.7822	0.2854	0.9658	0.7822	0.2767
1995	0.9883	0.9794	0.7345	0.7514	0.7615	0.0717	0.9468	0.6389	0.2762
1996	0.9890	0.9818	0.7394	0.7411	0.7500	0.2704	0.9568	0.7670	0.1901
1997	0.9879	0.9784	0.7215	0.7512	0.7892	0.2390	0.9546	0.7059	0.1820
1998	0.9886	0.9786	0.7530	0.7083	0.7383	0.2720	0.9543	0.7934	0.3074
1999	0.9886	0.9791	0.7367	0.7245	0.6975	0.0656	0.9473	0.6385	0.2648
2000	0.9885	0.9803	0.7420	0.7118	0.7229	0.1504	0.9556	0.7112	0.1720
2001	0.9883	0.9781	0.7314	0.7300	0.7971	0.2792	0.9582	0.7004	0.3177
2002	0.9886	0.9793	0.7295	0.7552	0.7997	0.2004	0.9641	0.7220	0.3470
2003	0.9877	0.9798	0.7366	0.7438	0.8327	0.1438	0.9617	0.8447	0.2785
2004	0.9882	0.9803	0.7376	0.7404	0.7437	0.2167	0.9602	0.7449	0.1518
2005	0.9886	0.9797	0.7386	0.7126	0.7467	0.0953	0.9494	0.7667	0.2904
2006	0.9887	0.9788	0.7258	0.6961	0.6999	0.0755	0.9481	0.6927	0.2615
2007	0.9884	0.9808	0.7485	0.7047	0.7270	0.1039	0.9630	0.5954	0.0821
2008	0.9892	0.9823	0.7328	0.7479	0.7859	0.1687	0.9674	0.7171	0.1904
2009	0.9892	0.9827	0.7362	0.7417	0.7651	0.3062	0.9626	0.7618	0.2920
2010	0.9891	0.9787	0.7381	0.7123	0.6820	0.1040	0.9483	0.6394	0.1842
2011	0.9897	0.9829	0.7452	0.7439	0.7358	0.3715	0.9606	0.7551	0.2183
2012	0.9887	0.9788	0.7309	0.7072	0.7607	0.3090	0.9521	0.6802	0.2998
2013	0.9886	0.9819	0.7212	0.7318	0.8298	0.2703	0.9643	0.7930	0.2556

Table 2: Descriptive Statistics

Variable	Explanation	# of Obs.	Mean	Median	S.D.	Min.	Max.
yield	Corn Yields during 1981-2013 (Bushel/ac)	64,812	104.52	103.00	35.31	0.00	234.20
Optimal GDD	Optimal Corn Growing Degree (5 °C to 35 °C) Days (GDD)	64,812	167.67	169.98	10.76	121.91	183.99
ppt	Total Precipitation (mm)	64,812	595.34	583.82	158.45	60.62	1463.40
whc	Area weighted Water Holding Capacity (cm/cm)	1,946	23.37	23.15	6.11	4.06	41.05
kfactor	Soil Erosivity K-factor	1,964	0.29	0.28	0.09	0.04	0.54
om	Organic Matters in 2 mm Top Soil (%)	1,964	2.61	1.41	4.06	0.29	53.71
soil pH	Soil pH	1,964	6.07	5.94	0.93	4.53	8.19

Table 3: Estimation Results

	Pooled	FE	FE SLAG	SLX	RE	KKP-RE
	Eq (13)	Eq (9)	Eq (11)	Eq (12)	Eq (10)	Eq (14)
$x_{it,1}$	−8.8880	−2.3365	−0.7016	0.1379	−3.4155	−3.2652
	(0.0928)***	(0.1049)***	(0.1214)***	(0.5678)	(0.0967)***	[0.1719]***
		[0.1313]***	[0.1316]***	[0.1919]		
$x_{it,2}$	−7.3300	−1.5055	−0.4492	−0.6053	−3.6886	−3.9155
	(0.0400)***	(0.0862)***	(0.0888)***	(0.4922)	(0.0565)***	[0.0981]***
		[0.0504]***	[0.0600]***	[0.1740]***		
$x_{it,3}$	−6.5000	−0.8503	−0.2409	0.3119	−1.9659	−1.7950
	(0.0992)***	(0.0992)***	(0.0845)***	(0.4423)	(0.0942)***	[0.1665]***
		[0.1319]***	[0.1336]*	[0.1462]**		
$x_{it,4}$	−7.0840	0.1395	0.0745	1.0313	−1.8551	−1.9326
	(0.0643)***	(0.0916)	(0.0718)	(0.4020)**	(0.0721)***	[0.1270]***
		[0.0847]*	[0.0895]	[0.1138]***		
$x_{it,5}$	−3.4450	0.0407	0.0324	0.5264	−0.6758	−0.5089
	(0.1082)***	(0.0942)	(0.0738)	(0.3631)	(0.0925)***	[0.1589]***
		[0.1329]	[0.1363]	[0.0539]***		
$x_{it,6}$	−6.3250	0.0681	0.0511	0.6078	−1.4149	−1.3172
	(0.0904)***	(0.0958)	(0.0750)	(0.3481)*	(0.0868)***	[0.1488]***
		[0.1207]	[0.1198]	[0.0312]***		
$x_{it,7}$	−1.8550	−0.9734	−0.3068	0.4853	−1.2050	−0.8748
	(0.0932)***	(0.0754)***	(0.0693)***	(0.2787)*	(0.0755)***	[0.1289]***
		[0.1131]***	[0.1154]***	[0.1948]**		
$x_{it,8}$	−4.7480	−1.0387	−0.3379	0.3425	−1.8827	−1.4327
	(0.0878)***	(0.0792)***	(0.0729)***	(0.2777)	(0.0760)***	[0.1294]***
		[0.1094]***	[0.1155]***	[0.0319]***		
P_{it}	0.0018	0.0011	0.0004	0.0015	0.0016	0.0016
	(0.0001)***	(0.0001)***	(0.0001)***	(0.0001)***	(0.0001)***	[0.0001]***
		[0.0001]***	[0.0001]***	[0.0001]***		
P_{it}^2	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	(0.0000)***	(0.0000)***	(0.0000)***	(0.0000)***	(0.0000)***	[0.0000]***
		[0.0000]***	[0.0000]***	[0.0000]***		
t	0.0137	0.0065	0.0022	0.0064	0.0069	0.0081
	(0.0007)***	(0.0005)***	(0.0005)***	(0.0005)***	(0.0005)***	[0.0012]***
		[0.0008]***	[0.0009]**	[0.0008]***		
t^2	0.0001	0.0002	0.0001	0.0002	0.0002	0.0002
	(0.0000)***	(0.0000)***	(0.0000)***	(0.0000)***	(0.0000)***	[0.0000]***
		[0.0000]***	[0.0000]***	[0.0000]***		
whc					0.0125	0.0089
					(0.0010)***	[0.0011]***

Note: * p-value < 10%, ** p-value < 5%, *** p-value < 1%. Standard errors are in parenthesis. [·] are spatial standard errors with Conley (1999, 2008), Kelejian and Prucha (2007), or Kapoor et al. (2007) .

Table 3 (Continued)

	Pooled	FE	FE SLAG	SLX	RE	KKP-RE
	Eq (13)	Eq (9)	Eq (11)	Eq (12)	Eq (10)	Eq (14)
kfactor					0.4146	0.2690
					(0.0710)***	[0.0814]***
om					0.0000	0.0000
					(0.0000)*	[0.0000]
soil pH					0.2072	0.2038
					(0.0042)***	[0.0067]***
$wx_{it,1}$				−2.6786		
				(0.5888)***		
				[0.2403]***		
$wx_{it,2}$				−1.0424		
				(0.5089) **		
				[0.1796]***		
$wx_{it,3}$				−1.3228		
				(0.4669)***		
				[0.1839]***		
$wx_{it,4}$				−1.0355		
				(0.4243) **		
				[0.1629]***		
$wx_{it,5}$				−0.6314		
				(0.3897)		
				[0.0916]***		
$wx_{it,6}$				−0.6529		
				(0.3752)*		
				[0.1244]***		
$wx_{it,7}$				−1.6120		
				(0.3004)***		
				[0.1791]***		
$wx_{it,8}$				−1.5719		
				(0.3006)***		
				[0.1284]***		
wP_{it}				−0.0004		
				(0.0001)***		
				[0.0000]***		
λ (ρ)			0.6845			0.6406
			(0.0374)***			
			[0.0045]***			
County FE	NO	YES	YES	YES	NO	NO

Note: * p-value < 10%, ** p-value < 5%, *** p-value < 1%. Standard errors are in parenthesis. [·] are spatial standard errors with Conley (1999, 2008), Kelejian and Prucha (2007), or Kapoor et al. (2007) .

Table 4: Model Comparison Test for Prediction Accuracy

		Out-of-Sample					
	In-Sample		Welch Test for Equal Forecasting Accuracy				
Model	RMSE	RMSE	FE	KKP-RE	RE	Pooled	FE SLAG
SLX	0.0655	0.1402	0.0032	23.3950	24.8429	33.8952	88.0270
FE	0.0657	0.1402		23.3965	24.8449	33.8973	88.0270
KKP-RE	0.1022	0.1861			0.4656	13.0959	84.1451
RE	0.0702	0.1871				13.0243	84.1470
Pooled	0.1218	0.2182					81.1990
FE SLAG	0.8547	1.2352					

Note: The first column reports the average root mean squared in-sample prediction error from Table 3. The second column represents the average root mean squared out-of-sample prediction error from 1,000 replications. Rows are sorted from the best forecast performance (lowest average RMSE) to worst. The last five columns present pair-wise Welch t-test against the null hypothesis of equal RMSE.

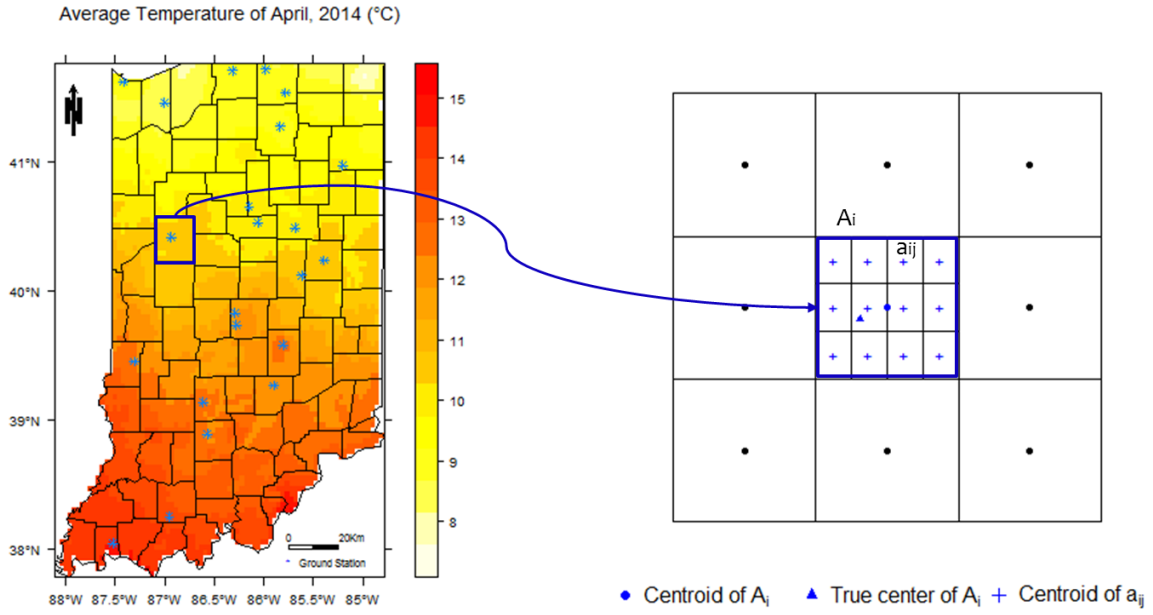
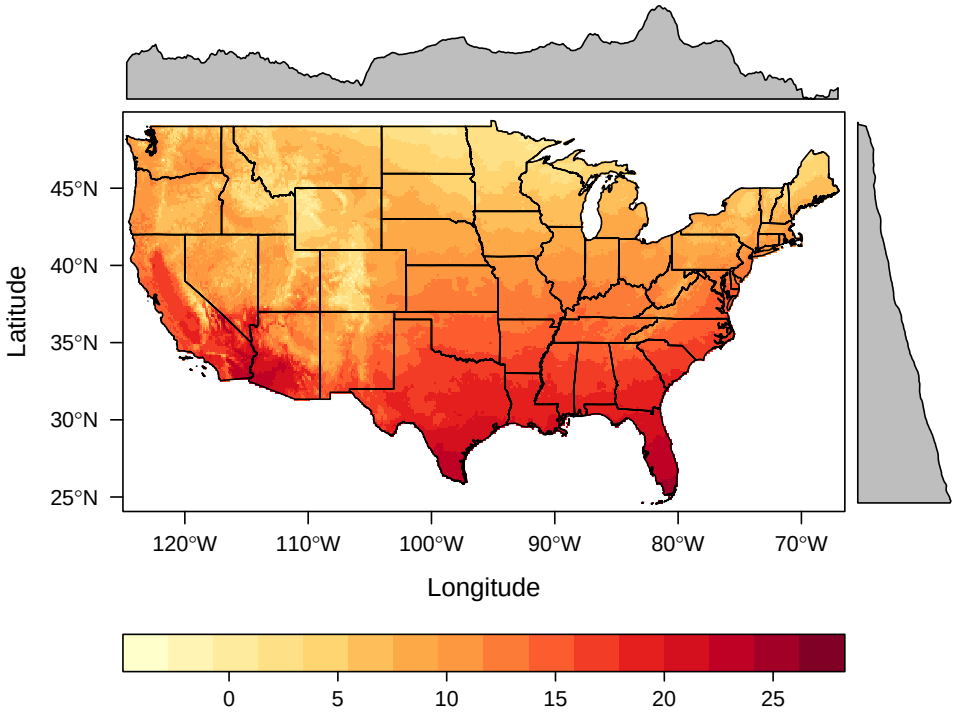
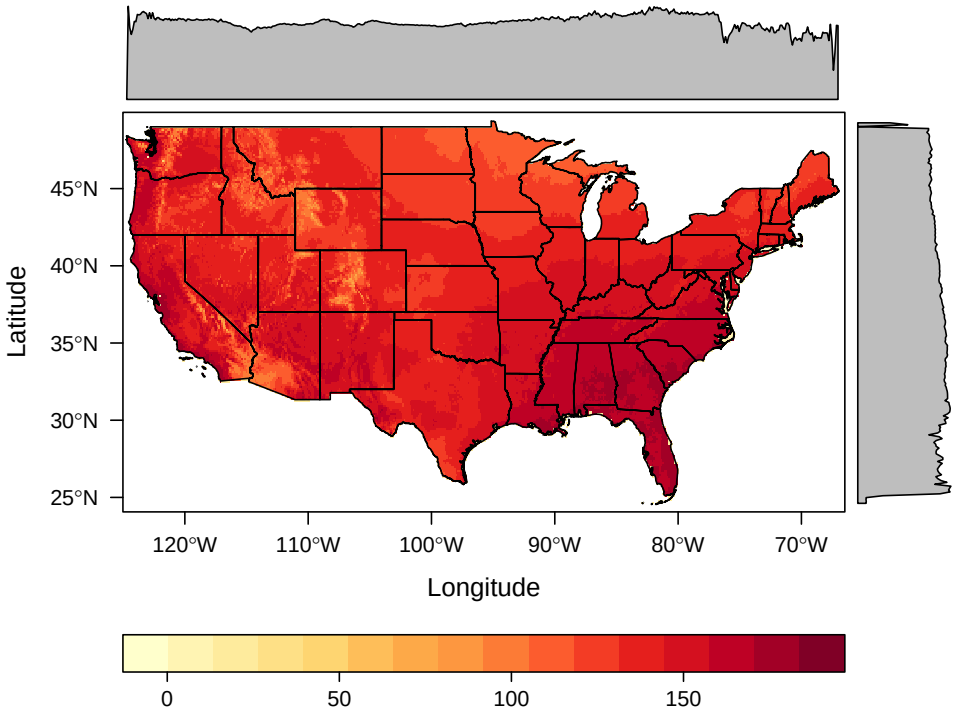


Figure 1: Spatial Units of Weather Variable: Average Temperature of Indiana State, April, 2014

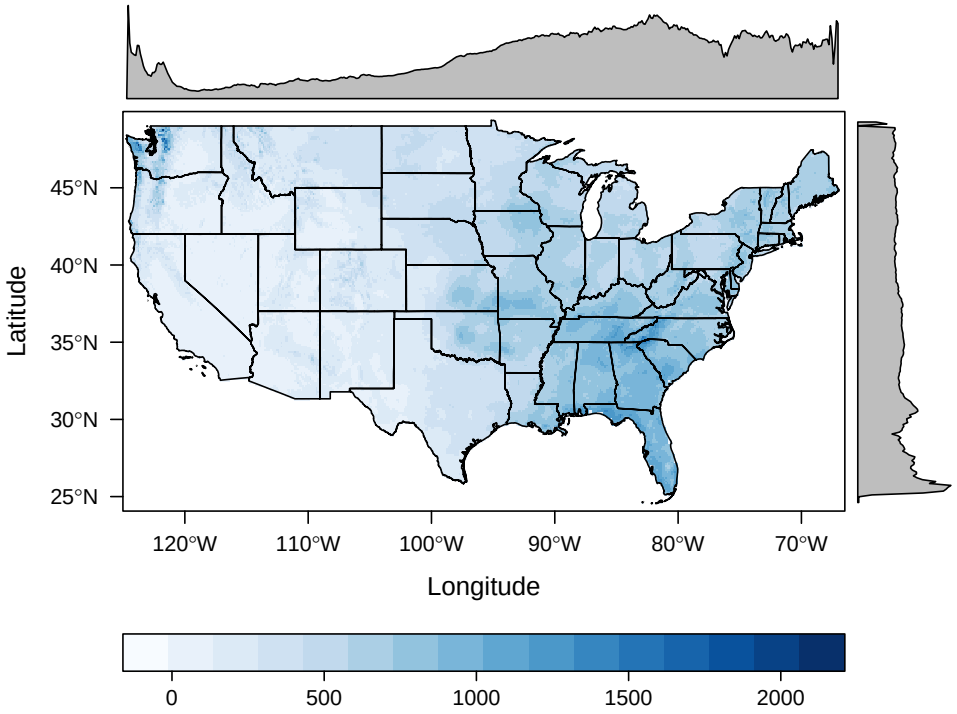


(a) Yearly Average Temperature of 2013 (Celsius)



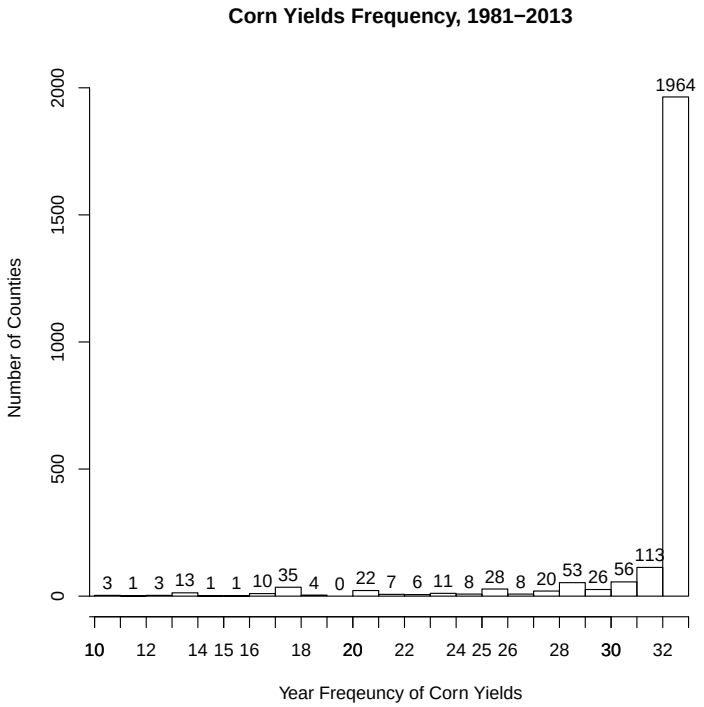
(b) Exposure During Growing Season of 2013 (Degree Days)

Figure 2 (Continued on the next page)

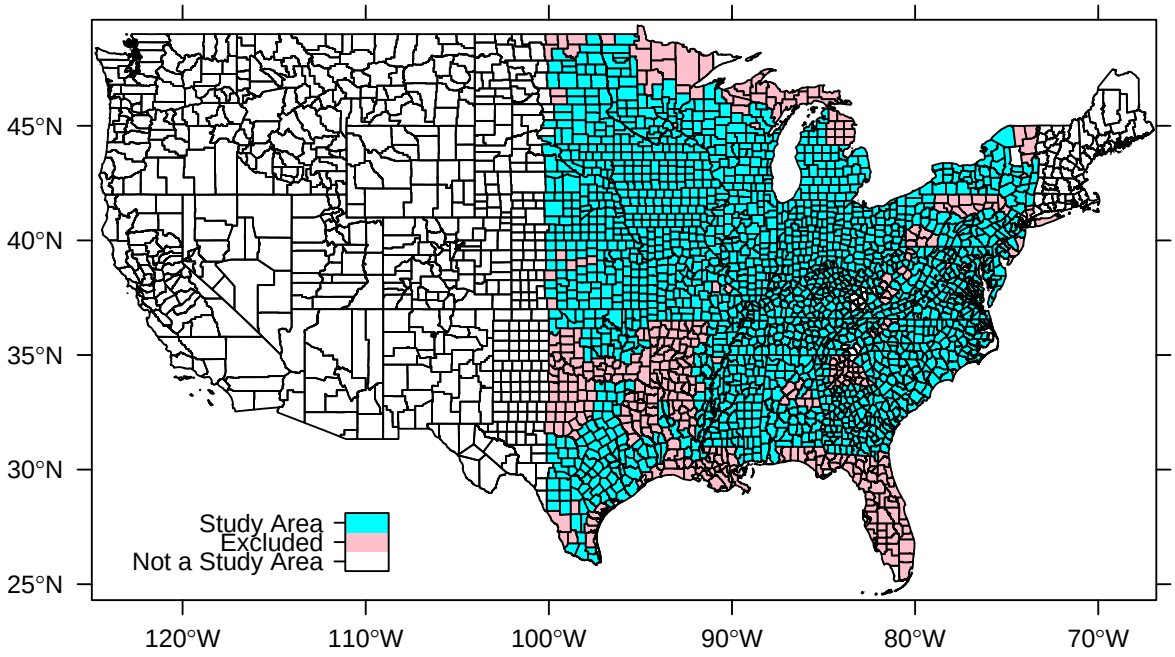


(c) Total Precipitation of 2013 (mm)

Figure 2: Spatial Correlation in 2013 Weather Variables



(a) Year Frequency of the East Counties of the 100th Meridian Line



(b) Study Area: 1,964 Counties with 33-year Corn Growing History for 1981-2013

Figure 3: Geographical Boundaries of Study Area

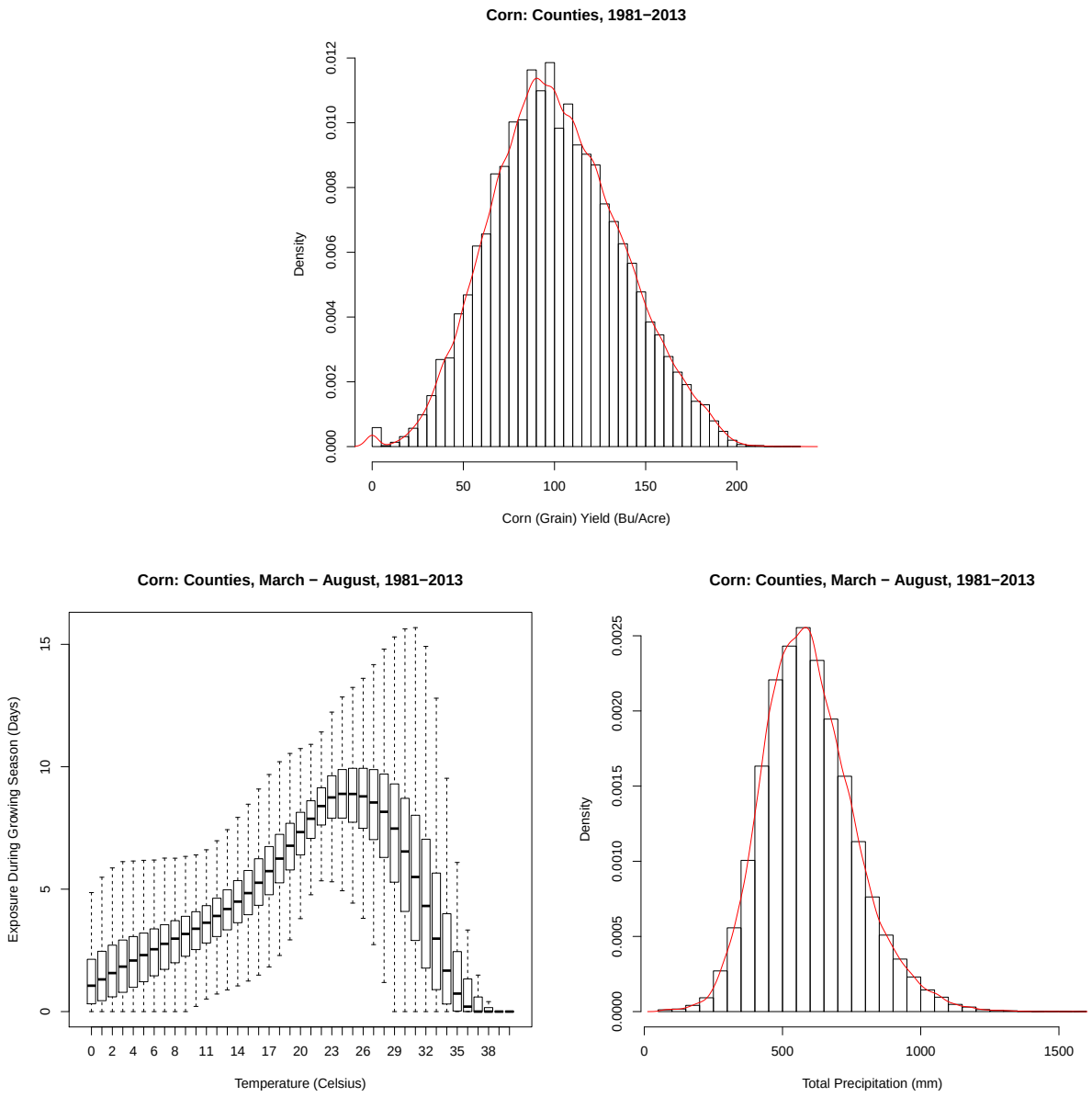


Figure 4: Data Description: Corn Yield, Exposure During Growing Season, and Total Precipitation

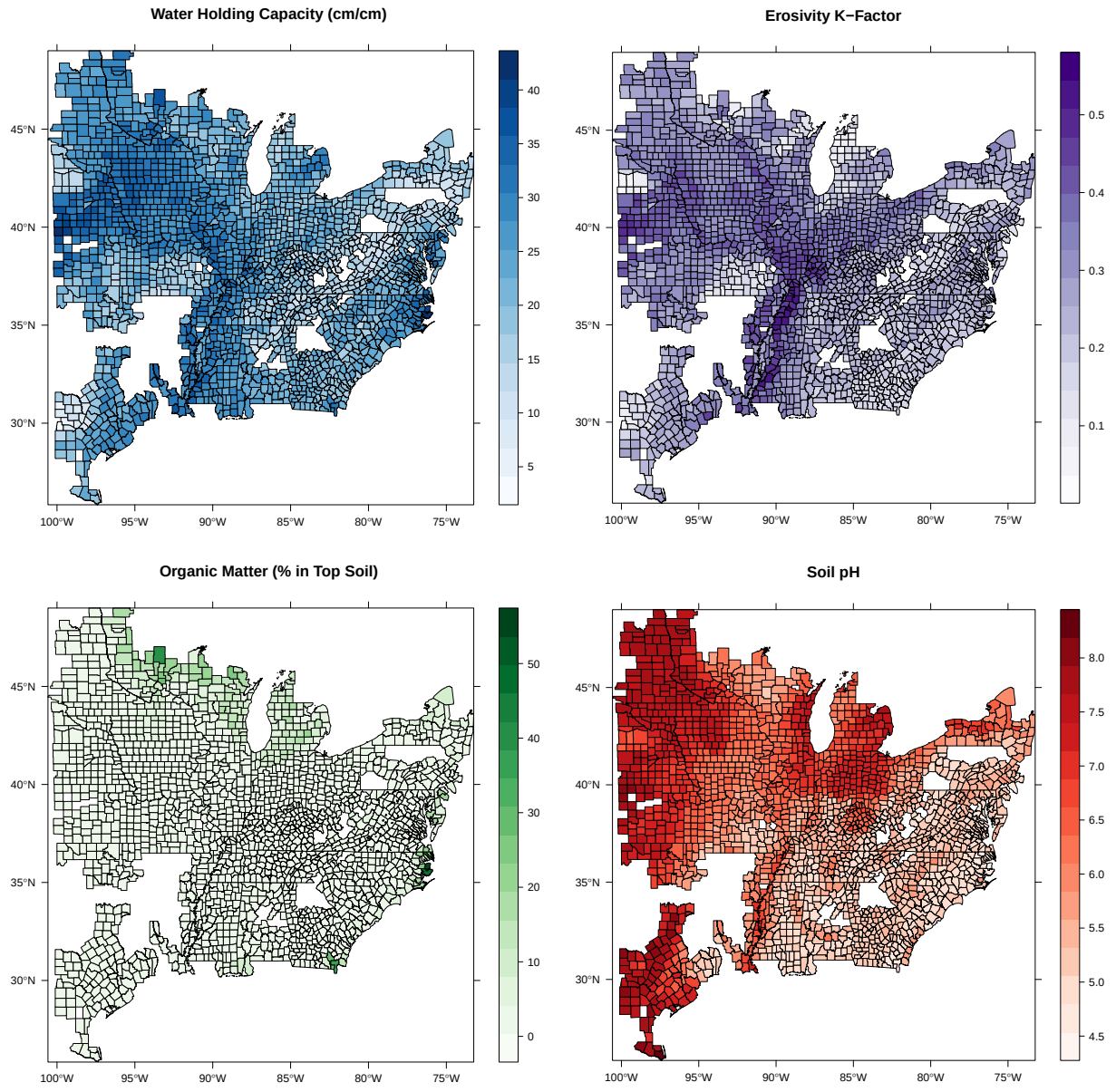


Figure 5: Soil Properties: Water Holding Capacity, Erosivity K-factor, Organic Matter, and Soil pH

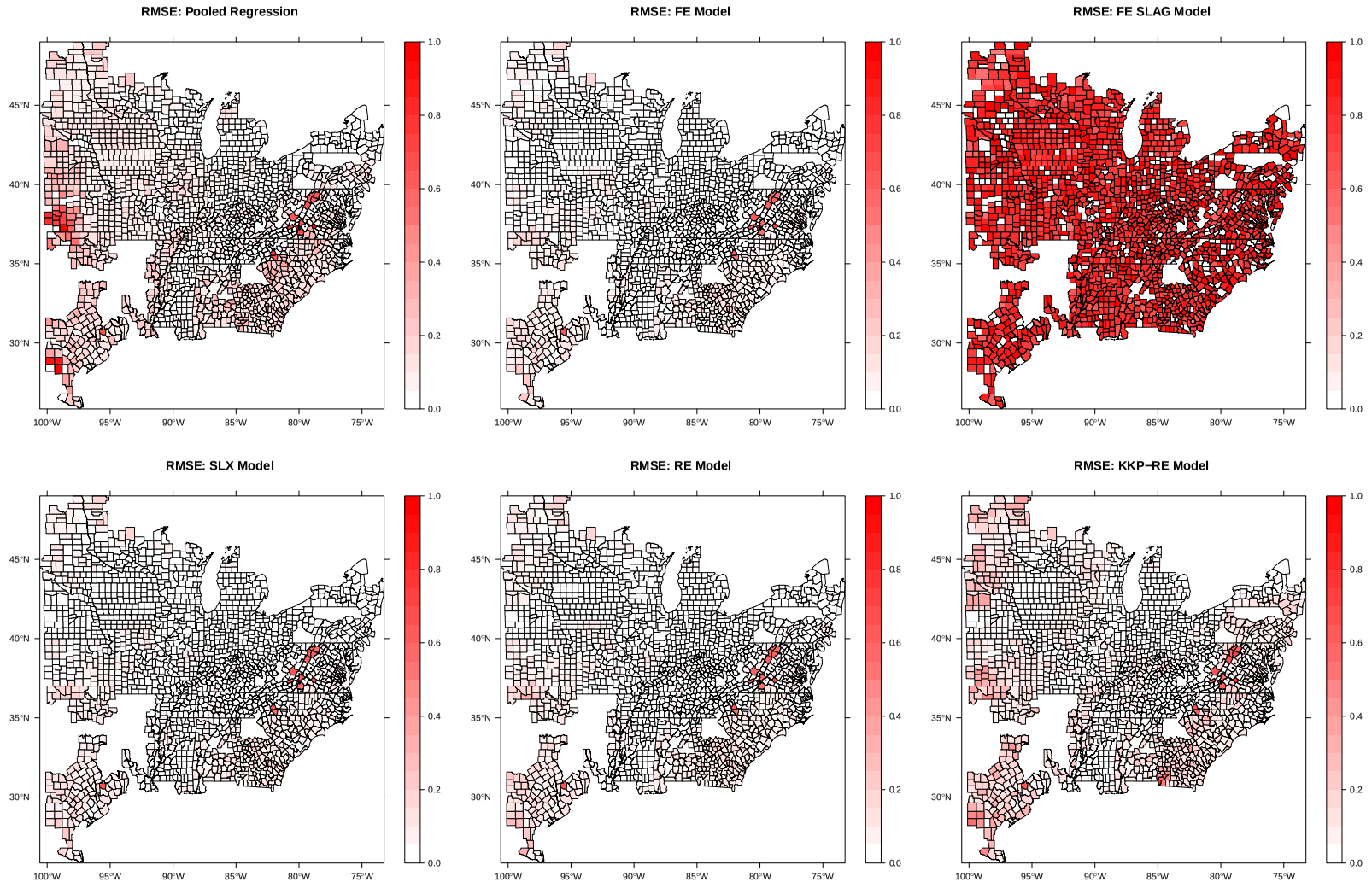


Figure 6: In-Sample-Prediction Accuracy: RMSE across 1,964 Counties
Note: The RMSE values greater than 1 are assigned as 1.